

LREC 2026

2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026

Workshop Proceedings

Editors
Mustafa Jarrar, Mo El-Haj, Amal Haddad Haddad, Serin Atiani, Shadi Abudalfa, Khalil Sima'an, Paul Rayson, and Camille Mansour

11 May 2026
https://sina.birzeit.edu/nakba-nlp/2026/

Proceedings of 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026

©ELRA Language Resources Association (ELRA), 2026
These proceedings are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 978-2-493814-57-9

Message from the General Chair

Welcome to the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), held as part of the Fifteenth Biennial Language Resources and Evaluation Conference (LREC 2026) in Palma, Mallorca, Spain. Nakba-NLP 2026 took place in a hybrid format on May 11, 2026.

The narratives of the ongoing Palestinian Nakba carry profound historical, cultural, and academic significance. Preserving these narratives—and critically examining the role of AI in documenting and shaping them—is essential to ensure their accessibility and relevance for present and future generations. Nakba testimonies exist in diverse formats, making their transformation into machine-understandable resources a key challenge and highlighting the need for accessible, well-annotated archives.

The workshop aims to explore how artificial intelligence, natural language processing, and corpus linguistics can support the documentation, analysis, and curation of Nakba narratives, while addressing ethical considerations in such contexts. It also seeks to develop resources that enable researchers and the public to generate and share insights across Arabic, English, and other languages.

This year's workshop featured two main tracks: Regular Papers and Shared Tasks. We received nine regular paper submissions, of which six were accepted. Five shared task proposals were submitted, four of which were accepted (NakbaVirality, NakbaArchiveClassifier, StanceNakba, and Arabic Manuscript Understanding). These tasks attracted teams worldwide to participate in developing systems related to Nakba narratives.

Overall, the final program comprised 47 papers: 6 regular papers, 4 shared task overview papers, and 37 shared task system papers. These contributions were co-authored by 148 researchers from 18 countries, reflecting the workshop's growing international reach. Each submission was reviewed by at least two members of the Program Committee.

This year's workshop featured a keynote by Paul Biggar, founder of the Tech for Palestine Initiative, presenting a global, community-driven effort to build digital tools and open datasets that document the ongoing Nakba and support research and preservation.

The panel discussion, *"NLP: Reconstructing Narratives about the Ongoing Nakba,"* was moderated by Khalil Sima'an and featured Kareem Darwish, Hanine Shehadeh, Jamila J. Ghaddar, and Rama Salahat. The discussion explored how NLP can contribute to reconstructing and analyzing narratives, highlighting both methodological challenges and opportunities.

We thank all authors for their submissions and the Program Committee members for their valuable reviews.

Workshop organizers: Mustafa Jarrar, Mo El-Haj, Amal Haddad Haddad, Serin Atiani, Shadi Abudalfa, Khalil Sima'an, Paul Rayson, Terry Regier and Camille Mansour.

Workshop page: `https://sina.birzeit.edu/nakba-nlp/2026/`

Organizing Committee

Nakba-NLP 2026:

Organising Committee & Workshop Chairs: Mustafa Jarrar (Hamad Bin Khalifa University, Qatar)
Mo El-Haj (VinUniversity, Vietnam; Lancaster University, UK)
Shadi Abudalfa (King Fahd University of Petroleum and Minerals, Saudi Arabia)
Serin Atiani (Princess Sumaya University, Jordan)
Amal Haddad Haddad (Universidad de Granada, Spain)
Camille Mansour (Institute for Palestine Studies, Lebanon)
Paul Rayson (Lancaster University, UK)
Khalil Sima'an (University of Amsterdam, The Netherlands)
Terry Regier (University of California, USA)

Programme Committee: Abdelkader El Mahdaouy, Mohamed VI Polytechnic University
Abdellah El Mekki, School of Computer Sciences, Mohammed VI Polytechnic University
Abdelrahim Qaddoumi, NYU
Abdulrahman Abdulsalam, University of Utah
Abed Alhakim Freihat, University of Trento
Abualsoud Hanani, Birzeit University
Adnan Yahya, Birzeit University
Ala Alazzeh, Birzeit University
Alaa Aljabari, Birzeit University
Ali AlKhathlan, Assistant Professor King Abdulaziz University
Almoataz B. Al-Said, Cairo University
Areej Jaber, Technical University Khadouri
Ashraf Elnagar, University of Sharjah
Ayah Soufan, Strathclyde University
Azzeddine Mazroui, University Mohammed First, Faculty of Sciences
Aziz Mohaisen, University of Central Florida
Badr AlKhamissi, EPFL
Baker Abdalhaq, Annajah National University
Basem Ezbidi, Birzeit University
Bassam Haddad, University of Petra
Bayan AbuShawar, Al Ain University
Bashar Tahayna, Birzeit University
Dana Abdulrahim, University of Bahrain
Dima Taji, Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
ElMoatez Billah Nagoudi, The University of British Columbia
Eyad Elyan, Robert Gordon University
Fadhl Eryani, University of Tübingen
Fadi Zaraket, American University of Beirut
Faisal Awartani, Insights for Research Polling and Training
Fatemah Husain, Kuwait University
Fatima Haouari, Qatar University
Fethi Bougares, LIUM- Le Mans Université
Fouzi Harrag, Ferhat Abbas University
Ghassan Mourad, Lebanese university
Ghada Almadbouh, Birzeit University
Go Inoue, Mohamed bin Zayed University of Artificial Intelligence

Habiba Dahmani, Mohamed Boudiaf University
Haithem Afli, ADAPT Centre, Munster Technological University
Hamada Nayel, Benha University
Ibrahim Abu Farha, University of Sheffield
Ibrahim Bounhas, Université de la Manouba Engineering and Technology
Imed Zitouni, Google
Imene Bensalem, ESCF de Constantine and MISC Lab, Constantine 2 University, Algeria
Injy Hamed, Institute for Natural Language Processing, University of Stuttgart
Kamel Gaanoun, National Institute of Statistics and Applied Economics
Kamel Smaili, LORIA
Kaoukab Chebaro, Columbia University
Karim Bouzoubaa, Mohammed V University in Rabat
Khaled Shaalan, The British University in Dubai
Khalid Choukri, ELRA/ELDA
Khalil Mrini, Bytedance
Khloud Al Jallad, SySSR
Kholoud Aldous, Northwestern University in Qatar
Labib Arafeh, AlQuds University
Lama Nachman, Intel Labs
Lamia Hadrich-Belguith, ANLP Research Group, MIRACL Lab, FSEGS, Sfax University
Majdi Sawalha, The University of Jordan
Manar Alkhatib, British University in Dubai
Maram Hasanain, Qatar Computing Research Institute
Mo El-Haj, Lancaster University
Mohamed Lichouri, Centre de Recherche Scientifique et Technique pour le Développement de la Langue Arabe (CRSTDLA)
Mohamed Yahya, NLP Researcher
Mohammad Abuoudeh, Al Hussein Bin Talal University
Mohammed Attia, Google Inc.
Mohammed Khalilia, Birzeit University
Mohammed Salah Al-Radhi, Budapest University of Technology and Economics, Department of Telecommunications and Media Informatics
Mohammad Azza, Princess Sumaya University for Technology
Mahmoud Al-Ayyoub, Jordan University of Science and Technology
Mona Baker, University of Oslo
Moustafa Al-Hajj, Lebanese University
Muhammad Abdul-Mageed, The University of British Columbia
Mohammad Ateeq, Xi'an Jiaotong-Liverpool University
Mourad Abbas, University of Science and Technology Houari Boumediene
Munir Fakher Eldin, Birzeit University
Munther Dahleh, MIT
Nada Ghneim, Damascus University - Information Technology Engineering Faculty
Nagham Hamad, Birzeit University
Nizar Habash, New York University
Omar Shehabi, Yale Law School
Omar Tesdell, Birzeit University
Omar Trigui, University of Sousse in Tunisia
Omar Qawasmeh, Princess Sumaya University for Technology
Osama Hamed, Palestine Technical University
Oussama Ben Khiroun, University of Manouba
Owen Rambow, Stony Brook University

Radi Jarrar, Birzeit University
Radwan Tahboub, Palestine Polytechnic University
Raia Abu Ahmad, DFKI
Rana Malhas, Qatar University
Razan Jaber, Stockholm University
Rebhi Baraka, Islamic University of Gaza
Rehab Duwairi, Jordan University of Science and Technology
Reem Suwaileh, Qatar University
Saad Ezzini, Lancaster University
Sabri Boughorbel, Qatar Computing Research Institute
Salima Harrat, ENS Bouzaréah, Algiers
Salima Mdhaffar, LIA - University of Avignon
Samhaa R. El-Beltagy, Newgiza University/Optomatica
Sara Nabhani, De Rijksuniversiteit Groningen
Samir Abdaljalil, Texas A&M University
Samir Tartir, Al Hussein Technical University
Sanaa Kaddoura, Zayed University - United Arab Emirates
Seif Mechti, ISSEPS
Serin Atiani, Princess Sumaya University for Technology
Shady Elbassuoni, American University of Beirut
Sultan Alrowili, University of Delaware
Susan Akram, Boston University
Tamer Elsayed, Qatar University
Thaher Gharabeh, Universidad de Granada, Spain
Wajdi Zaghouani, Northwestern University, Qatar
Walid Magdy, The University of Edinburgh
Wassim El-Hajj, American University of Beirut
Wissam Antoun, Inria
Yahya Mohamed Elhadj, Arab Center for Research and Policy Studies
Zenina Jallad, Birzeit University

Table of Contents

The NakbaEcho Dataset: From Oral Testimonies to a Transcribed Arabic History Corpus
Batool Najeh Balah, Mahmoud Fawzi, Houda Elmimouni and Walid Magdy 1

Mining the Pre-1948 Palestinian Press: Unsupervised Keyphrase Extraction and Temporal Discourse Analysis from Five Historical Arabic Newspapers
Basel Barakat and Nizam Barakat 23

Nakba Discourse 2025: A Bilingual Social Media Dataset for Collective Trauma Analysis
Wajdi Zaghouani, Mabrouka Bessghaier and Kais Attia 33

ChronoLearn: A GRAG LLM-Based System for Structuring and Exploring Historical Narratives
Mohammad O. ALADDASI, Shahd L. Abu Hijleh and Omar Qawasmeh 43

Credibility Assessment for Arabic News on the Gaza War: A Hybrid Neural-Symbolic Pipeline
Sanaa Abril, Sihame Mouanid, El habib Ben lahmar and Omar Zahour 50

Tarikhi: Arabic Temporal Information Extraction from Arabic Historical Documents
Qusay Abdo, Serin Atiani, Tariq Sraiji and Adnan Saeed 60

NAKBA NLP 2026: Shared Task on Arabic Handwritten Manuscript Understanding (Palestine Memory–Omar Al-Saleh Memoir)
Hadi Hamoud, Ahmad Ali Chamseddine, Bilal Shalash, Firas Ben Abid, Mustafa Jarrar, Chadi Abou Chakra, Bernard Ghanem and Fadi A. Zaraket 70

StanceNakba Shared Task: Actor and Topic-Aware Stance Detection in Public Discourse
Kholoud Khalil Aldous, Md. Rafiul Biswas, Mabrouka Bessghaier, Shimaa Amer Ibrahim, Kais Attia and Wajdi Zaghouani 80

The NakbaArchiveClassifier Shared Task on Nakba Image Classification
Alexei Abrahams, Shadi Abudalfa, Mustafa Jarrar and George Mikros 91

The NakbaVirality Shared Task on MultimodalVirality Prediction in High-Stakes Discourse
Saad Ezzini, Salima Lamsiyah, Shadi Abudalfa, Samir El-Amrany and Walid Alsafadi . . 98

Faisal_Adam at NakbaArchiveClassifier Shared Task: Archival Image Classification for Structural Destruction: A Robust Pipeline Using ResNet-50 and Test-Time Augmentation
Faisal Muhammad Adam and Salisu Aliyu 104

Doaa Sulaiman at AR-MS NakbaNLP 2026: Faithful Diplomatic Transcription of Arabic Manuscripts Using a Human-Centred Annotation Framework
Doaa Bahjat Sulaiman 108

KvochurHegel at StanceNakba: Robust Stance Detection with Regularized Natural Language Inference
Minh-Hoang Le 113

KvochurHegel at NakbaArchiveClassifier Shared Task: Nakba Image Classification via ConvNeXt-V2 and Label Smoothing
Minh-Hoang Le 118

HCMUS_TheFangs at NakbaArchiveClassifier Shared Task: Foundation Models and Advanced Training Strategies for Conflict Damage Classification
Duy Minh Dao Sy, Trung Kiet Huynh, Nguyen Chi Tran, Phu Quy Nguyen Lam and Phu Hoa Pham 121

Baflah-lamri at NAKBA-NLP 2026: Manual Ground Truth Enrichment
baflah abdelouahab and Lamri Mohamed 128

Free-Gaza at NakbaArchiveClassifier Shared Task: Towards Distinguishing the Destructive Effect of Nakba: NakbaImage Classification Using Artificial Intelligence Techniques
Nisreen I. R. Yassin and Enas A. Hakim Khalil 133

HCMUS_TheFangs at NakbaVirality Shared Task: The Audience is the Message: Escaping the Deep Learning Trap in Conflict-Domain Virality Prediction
Duy Minh Dao Sy 137

Xin1212 at NakbaVirality Shared Task: Frozen CLIP with Residual Adapter for Multimodal Virality Classification
Xinyan Zhang and Bingzhou Yang 144

A2NLP at StanceNakba Shared Task: Fine-Tuned AraBERT for Topic-Based Arabic Stance Detection
Alaa Nairat and Aysar Mahmoud Nairat 147

Ketaba-OCR at AR-MS NakbaNLP 2026: Efficient Adaptation of Vision-Language Models for Handwritten Recognition
Hassan Barmandah, Fatimah Emad Eldin, Khloud Al Jallad and Omer Nacar 160

No Overfit at NakbaArchiveClassifier Shared Task: A Swin Transformer-Based System for Destruction Image Classification
Mohamed Fathy Mohamed, Samar Mahmoud Abd El-Mageed and Ensaf Mohamed . . 171

Yafa at StanceNakba: Actor-Level Stance Detection Using Cross-Lingual Approach
Tasnim Zayet, Osama Hamed and Tasneem Duridi 177

The Resistant Word at StanceNakba Shared Task: A Topic-Aware Model for Cross-Topic Stance Detection
Sarah Yassine 182

"Hope" at NakbaArchiveClassifier Shared Task: Transfer Learning-Based CNN Models for Infrastructure Damage Detection
Lojien AlKhidir and HebaTalla Abdelhady 187

Al-Warraq at AR-MS NAKBA-NLP 2026: Adapting Vision-Language and Transformer Models for Automatic Manuscript OCR/HTR
Ahmad Edris Youssef, Aya Hafiz Faris, Alhasan Hamood, Zainab Kamil, Jana Alqasem and SARA Ali hamed" 191

Misraj AI at AR-MS NAKBA-NLP 2026: A State-of-the-Art VLM in Arabic Handwritten Text Recognition
Khalil Hennara, Muhammad Hreden, Zeina Aldallal, Sara Chrouf and Safwan AlModhayan 196

PushingBoundaries at NakbaVirality Shared Task: Recursive Prompt Improvement for Multimodal Virality Classification
Ashhadul Islam, Md Rafiul Biswas, Samir Brahim Belhaouari and Wajdi Zaghouani . . . 201

U4RASD at StanceNakba Shared Task: Data Augmentation and Auxiliary Objectives for Arabic Stance Detection
Nancy Hamdan, Aya Jouni, Aya Saïd and Fadi Zaraket . . . 206

AyahVerse at NakbaArchiveClassifier Shared Task: Architectural Trade-offs and Decision Calibration for Humanitarian Image Classification
Ibad-ur-Rehman Rashid and Akhtar Ali . . . 212

mlenthusiast at NakbaArchiveClassifier Shared Task: A Lightweight SVM-Gated Ensemble of EfficientNets for Image Classification
Md. Ajwad Hossain . . . 217

Pixel at NakbaArchiveClassifier Shared Task: ConvNeXt-Based Ensemble for Destruction Detection
Rahaf Jaber . . . 221

MennaAly at NakbaArchiveClassifier Shared Task: Transfer Learning with ResNet for Historical Image Classification
menna aly . . . 226

DLRG@ NakbaArchiveClassifier Shared Task: Deep Transfer Learning for Destruction Detection in Nakba Archive Images Using EfficientNet-B3
Ramesh R. Kannan and Ratnavel Rajalakshmi . . . 229

Oblevit at AR-MS NAKBA NLP 2026 Subtask 2: Hybrid CNN–BiLSTM–CTC Framework with Linguistic Refinement for Arabic Handwritten Manuscript Recognition
Reem Juhaysh, Abuelgasim Sami Abusonoun and Sara Ayad . . . 234

Viva_Palestine at StanceNakba Shared Task: Actor and Topic-Aware Stance Detection in Public Discourse
Wafaa El-Kassas, Enas Khalil and Enas El Houby . . . 239

KUET at StanceNakba Shared Task: StanceMoE: Mixture-of-Experts Architecture for Stance Detection
Abdullah Al Shafi, Md. Milon Islam, Sk. Imran Hossain and K. M. Azharul Hasan . . . 244

The Blackwell Collective at StanceNakba Shared Task: PAST-TIDE: Prototype-Anchored Statement Tuning with Topic-Invariant Normalization for Stance Detection
Md. Shakhoyat Rahman Shujon, MD Jahid Hasan Jim, Md. Milon Islam, Md Rezwanul Haque and Fakhri Karray . . . 252

AlSaifTeam at AR-MS NAKBA-NLP 2026: Building Expert-Quality Ground Truth for Arabic Handwritten Manuscripts
Joud Fahad AlSaif, Alhasan Mohammad Hamood and Jana Mohammad Alseed . . . 258

PalNLP at AR-MS Shared Task: Guidelines Paper for Arabic Manuscript Transcription
Mutaz Ayesh . . . 261

Digilians at NakbaVirality Shared Task: Bidirectional Cross-Attention for Multimodal Virality Prediction
Ahmed Eid Hassan, Noureldeen H. Mohamed, Abdelrhman M. Fawzy, Mohamed A. Abdelghany, Ahmed A. Hassan, Ahmed S. Qassim, Fady A. Abd El Sayed, Mohamed H. Mohamed, Rahma M. Mohamed, Arwa M. Abou-Attia, Mayar M. Mohamed, Shahd A. Sawla and Shahd O. Mahmoud 265

Not Gemma at AR-MS NakbaNLP 2026: Mubsir OCR: End-to-End Recognition of Arabic Handwritten Text
Ali Adel Ali, Mona Khaled Ali, Mohamed Emad Sayed and Ibrahim Naser Mostafa 269

Latent Narratives at AR-MS NakbaNLP 2026: Reducing Character Errors in Arabic Manuscript Transcription: A CER Oriented System
Sara Abdulmonem Al desouky 275

shroukgbr at StanceNakba Shared Task: Transformer-Based Ensemble Learning for Actor-Level Stance Detection in Palestinian–Israeli Social Media Discourse
Shrouk Anwar Gbr and Mohamed Ibrahim Ragab 280

NU_Hallucinators at NakbaArchiveClassifier Shared Task: A CLIP-Based Approach for Destruction Detection in Historical Image Archives
Salma Khaled Hegazy and Mohamed Ibrahim Ragab 285

ZAHIRA BOULANOUAR at NakbaArchiveClassifier Shared Task: Detecting Infrastructure Destruction in Gaza with a ConvNeXt Ensemble
ZAHIRA BOULANOUAR 290

EGCSS at StanceNakba Shared Task: Cross-Topic Arabic Stance Detection for Two Middle East Issues
Asmaa Qindeel, Toka Khaled, Batool Najeh Balah, Eman Elrefai and Mahmoud Fawzi 294

Workshop Program

14:00–14:05 *Opening Session: Welcome by Workshop Chairs*
Mustafa Jarrar

14:05–14:45 Keynote: The Tech for Palestine Initiative, Paul Biggar
Chair: Mustafa Jarrar

14:45–15:25 Session 1: Nakba Narratives and Resources
Chair: Serin Atiani

14:45–14:58 *The NakbaEcho Dataset: From Oral Testimonies to a Transcribed Arabic History Corpus*
Batool Najeh Balah, Mahmoud Fawzi, Houda Elmimouni and Walid Magdy

14:58–15:11 *Mining the Pre-1948 Palestinian Press: Unsupervised Keyphrase Extraction and Temporal Discourse Analysis from Five Historical Arabic Newspapers*
Basel Barakat and Nizam Barakat

15:11–15:25 *Nakba Discourse 2025: A Bilingual Social Media Dataset for Collective Trauma Analysis*
Wajdi Zaghouani, Mabrouka Bessghaier and Kais Attia

15:25–16:00 Session 2: Nakba Information Extraction and AI Modeling
Chair: Mo El-Haj

15:25–15:37 *ChronoLearn: A GRAG LLM-Based System for Structuring and Exploring Historical Narratives*
Mohammad O. ALADDASI, Shahd L. Abu Hijleh and Omar Qawasmeh

15:37–15:49 *Credibility Assessment for Arabic News on the Gaza War: A Hybrid Neural-Symbolic Pipeline*
Sanaa Abril, Sihame Mouanid, El habib Ben lahmar and Omar Zahour

15:49–16:00 *Tarikhi: Arabic Temporal Information Extraction from Arabic Historical Documents*
Qusay Abdo, Serin Atiani, Tariq Sraiji and Adnan Saeed

16:00–16:30 **Coffee Break**

16:30–17:00 **Session 3: Shared Tasks Papers**
Chair: Shadi Abudalfa

16:30–16:37 *NAKBA NLP 2026: Shared Task on Arabic Handwritten Manuscript Understanding (Palestine Memory–Omar Al-Saleh Memoir)*
Hadi Hamoud, Ahmad Ali Chamseddine, Bilal Shalash, Firas Ben Abid, Mustafa Jarrar, Chadi Abou Chakra, Bernard Ghanem and Fadi A. Zaraket

16:37–16:44 *StanceNakba Shared Task: Actor and Topic-Aware Stance Detection in Public Discourse*
Kholoud Khalil Aldous, Md. Rafiul Biswas, Mabrouka Bessghaier, Shimaa Amer Ibrahim, Kais Attia and Wajdi Zaghouani

16:44–16:51 *The NakbaArchiveClassifier Shared Task on Nakba Image Classification*
Alexei Abrahams, Shadi Abudalfa, Mustafa Jarrar and George Mikros

16:51–17:00 *The NakbaVirality Shared Task on MultimodalVirality Prediction in High-Stakes Discourse*
Saad Ezzini, Salima Lamsiyah, Shadi Abudalfa, Samir El-Amrany and Walid Alsafadi

17:00–17:50 **Panel Discussion: NLP, Reconstructing narratives about the ongoing Nakba**
Chair: Khalil Sima'an

17:50–18:00 *Closing Session*
Mustafa Jarrar

Faisal_Adam at NakbaArchiveClassifier Shared Task: Archival Image Classification for Structural Destruction: A Robust Pipeline Using ResNet-50 and Test-Time Augmentation
Faisal Muhammad Adam and Salisu Aliyu

Doaa Sulaiman at AR-MS NakbaNLP 2026: Faithful Diplomatic Transcription of Arabic Manuscripts Using a Human-Centred Annotation Framework
Doaa Bahjat Sulaiman

KvochurHegel at StanceNakba: Robust Stance Detection with Regularized Natural Language Inference
Minh-Hoang Le

KvochurHegel at NakbaArchiveClassifier Shared Task: Nakba Image Classification via ConvNeXt-V2 and Label Smoothing
Minh-Hoang Le

HCMUS_TheFangs at NakbaArchiveClassifier Shared Task: Foundation Models and Advanced Training Strategies for Conflict Damage Classification
Duy Minh Dao Sy, Trung Kiet Huynh, Nguyen Chi Tran, Phu Quy Nguyen Lam and Phu Hoa Pham

Baflah-lamri at NAKBA-NLP 2026: Manual Ground Truth Enrichment
baflah abdelouahab and Lamri Mohamed

Free-Gaza at NakbaArchiveClassifier Shared Task: Towards Distinguishing the Destructive Effect of Nakba: NakbaImage Classification Using Artificial Intelligence Techniques
Nisreen I. R. Yassin and Enas A. Hakim Khalil

HCMUS_TheFangs at NakbaVirality Shared Task: The Audience is the Message: Escaping the Deep Learning Trap in Conflict-Domain Virality Prediction
Duy Minh Dao Sy

Xin1212 at NakbaVirality Shared Task: Frozen CLIP with Residual Adapter for Multimodal Virality Classification
Xinyan Zhang and Bingzhou Yang

A2NLP at StanceNakba Shared Task: Fine-Tuned AraBERT for Topic-Based Arabic Stance Detection
Alaa Nairat and Aysar Mahmoud Nairat

Ketaba-OCR at AR-MS NakbaNLP 2026: Efficient Adaptation of Vision-Language Models for Handwritten Recognition
Hassan Barmandah, Fatimah Emad Eldin, Khloud Al Jallad and Omer Nacar

No Overfit at NakbaArchiveClassifier Shared Task: A Swin Transformer-Based System for Destruction Image Classification
Mohamed Fathy Mohamed, Samar Mahmoud Abd El-Mageed and Ensaf Mohamed

Yafa at StanceNakba: Actor-Level Stance Detection Using Cross-Lingual Approach
Tasnim Zayet, Osama Hamed and Tasneem Duridi

The Resistant Word at StanceNakba Shared Task: A Topic-Aware Model for Cross-Topic Stance Detection
Sarah Yassine

"Hope" at NakbaArchiveClassifier Shared Task: Transfer Learning-Based CNN Models for Infrastructure Damage Detection
Lojien AlKhidir and HebaTalla Abdelhady

Al-Warraq at AR-MS NAKBA-NLP 2026: Adapting Vision-Language and Transformer Models for Automatic Manuscript OCR/HTR
Ahmad Edris Youssef, Aya Hafiz Faris, Alhasan Hamood, Zainab Kamil, Jana Alqasem and SARA Ali hamed"

Misraj AI at AR-MS NAKBA-NLP 2026: A State-of-the-Art VLM in Arabic Handwritten Text Recognition
Khalil Hennara, Muhammad Hreden, Zeina Aldallal, Sara Chrouf and Safwan AlModhayan

PushingBoundaries at NakbaVirality Shared Task: Recursive Prompt Improvement for Multimodal Virality Classification
Ashhadul Islam, Md Rafiul Biswas, Samir Brahim Belhaouari and Wajdi Zaghouani

U4RASD at StanceNakba Shared Task: Data Augmentation and Auxiliary Objectives for Arabic Stance Detection
Nancy Hamdan, Aya Jouni, Aya Saïd and Fadi Zaraket

AyahVerse at NakbaArchiveClassifier Shared Task: Architectural Trade-offs and Decision Calibration for Humanitarian Image Classification
Ibad-ur-Rehman Rashid and Akhtar Ali

mlenthusiast at NakbaArchiveClassifier Shared Task: A Lightweight SVM-Gated Ensemble of EfficientNets for Image Classification
Md. Ajwad Hossain

Pixel at NakbaArchiveClassifier Shared Task: ConvNeXt-Based Ensemble for Destruction Detection
Rahaf Jaber

MennaAly at NakbaArchiveClassifier Shared Task: Transfer Learning with ResNet for Historical Image Classification
menna aly

DLRG@ NakbaArchiveClassifier Shared Task: Deep Transfer Learning for Destruction Detection in Nakba Archive Images Using EfficientNet-B3
Ramesh R. Kannan and Ratnavel Rajalakshmi

Oblevit at AR-MS NAKBA NLP 2026 Subtask 2: Hybrid CNN–BiLSTM–CTC Framework with Linguistic Refinement for Arabic Handwritten Manuscript Recognition
Reem Juhaysh, Abuelgasim Sami Abusonoun and Sara Ayad

Viva_Palestine at StanceNakba Shared Task: Actor and Topic-Aware Stance Detection in Public Discourse
Wafaa El-Kassas, Enas Khalil and Enas El Houby

KUET at StanceNakba Shared Task: StanceMoE: Mixture-of-Experts Architecture for Stance Detection
Abdullah Al Shafi, Md. Milon Islam, Sk. Imran Hossain and K. M. Azharul Hasan

The Blackwell Collective at StanceNakba Shared Task: PAST-TIDE: Prototype-Anchored Statement Tuning with Topic-Invariant Normalization for Stance Detection
Md. Shakhoyat Rahman Shujon, MD Jahid Hasan Jim, Md. Milon Islam, Md Rezwanul Haque and Fakhri Karray

AlSaifTeam at AR-MS NAKBA-NLP 2026: Building Expert-Quality Ground Truth for Arabic Handwritten Manuscripts
Joud Fahad AlSaif, Alhasan Mohammad Hamood and Jana Mohammad Alseed

PalNLP at AR-MS Shared Task: Guidelines Paper for Arabic Manuscript Transcription
Mutaz Ayesh

Digilians at NakbaVirality Shared Task: Bidirectional Cross-Attention for Multimodal Virality Prediction
Ahmed Eid Hassan, Noureldeen H. Mohamed, Abdelrhman M. Fawzy, Mohamed A. Abdelghany, Ahmed A. Hassan, Ahmed S. Qassim, Fady A. Abd El Sayed, Mohamed H. Mohamed, Rahma M. Mohamed, Arwa M. Abou-Attia, Mayar M. Mohamed, Shahd A. Sawla and Shahd O. Mahmoud

Not Gemma at AR-MS NakbaNLP 2026: Mubsir OCR: End-to-End Recognition of Arabic Handwritten Text
Ali Adel Ali, Mona Khaled Ali, Mohamed Emad Sayed and Ibrahim Naser Mostafa

Latent Narratives at AR-MS NakbaNLP 2026: Reducing Character Errors in Arabic Manuscript Transcription: A CER Oriented System
Sara Abdulmonem Al desouky

shroukgbr at StanceNakba Shared Task: Transformer-Based Ensemble Learning for Actor-Level Stance Detection in Palestinian–Israeli Social Media Discourse
Shrouk Anwar Gbr and Mohamed Ibrahim Ragab

NU_Hallucinators at NakbaArchiveClassifier Shared Task: A CLIP-Based Approach for Destruction Detection in Historical Image Archives
Salma Khaled Hegazy and Mohamed Ibrahim Ragab

ZAHIRA BOULANOUAR at NakbaArchiveClassifier Shared Task: Detecting Infrastructure Destruction in Gaza with a ConvNeXt Ensemble
ZAHIRA BOULANOUAR

EGCSS at StanceNakba Shared Task: Cross-Topic Arabic Stance Detection for Two Middle East Issues
Asmaa Qindeel, Toka Khaled, Batool Najeh Balah, Eman Elrefai and Mahmoud Fawzi

The NakbaEcho Dataset: From Oral Testimonies to a Transcribed Arabic History Corpus

Batool Balah[1], Mahmoud Fawzi[2], Houda Elmimouni[3], Walid Magdy[2]
[1]Maqsam, Jordan
[2]The University of Edinburgh, United Kingdom
[3]University of Manitoba, Canada
batoolnajeh@gmail.com, m.f.g.ibrahim@sms.ed.ac.uk
houda.elmimouni@umanitoba.ca, wmagdy@inf.ed.ac.uk

Abstract

We present *NakbaEcho*, a large-scale dataset derived from Palestinian testimonies about the 1948 Nakba. The resource is constructed from transcribing over 2,180 hours of recorded interviews gathered through the Palestine Remembered Oral History index and linked to multiple repositories, including the Palestinian Oral History Archive (POHA) and YouTube-hosted interviews. We harmonize interview-level metadata and generate timestamp-aligned transcripts from the original Arabic recordings using an automatic transcription pipeline configured for Palestinian Arabic. The dataset includes speaker-labeled segments and auxiliary annotations designed to support downstream research in Arabic speech processing, natural language processing, digital humanities, and oral-history analysis. We further provide an empirical evaluation of transcription quality and descriptive analyses of geographic coverage, narrator demographics, and linguistic patterns across the corpus. *NakbaEcho* contributes a structured and extensible computational resource for studying Palestinian oral testimony while expanding the availability of dialectal Arabic materials for speech, text, and social research, supporting reproducibility, benchmarking, and future cross-lingual and multimodal research directions.

Keywords: Oral history, Arabic, Levantine, Palestine, Nakba, automatic transcription, digital humanities

1. Introduction

Oral history has been vital to documenting, preserving, and transmitting Palestinian memory of the 1948 Nakba (Nur, 2008). Testimonies by survivors and their descendants across archives and initiatives document displacement, loss, return, and everyday life before and after 1948. These recordings serve as both historical evidence and linguistic records of spoken Palestinian Arabic across generations.

Despite the richness, Nakba's oral histories remain difficult to process at scale. Much of the material exists only as dispersed audio/video recordings, with limited transcription suitable for NLP or corpus-based analysis(Lamar et al., 2025). Existing computational work has therefore relied either on small translated subsets of archival interviews or on transcripts prepared for specific studies (Ashqar, 2025), rather than on a unified large-scale Arabic speech-and-text resource.

Nakba testimonies are typically produced in spoken Arabic, mostly in Palestinian dialect rather than Modern Standard Arabic (MSA), and they contain historical and geographical references to villages, districts, displacement routes, kinship relations, and political actors. These characteristics make the material particularly valuable for research on dialectal Arabic speech recognition, oral-history processing, narrative analysis, information extraction, and digital humanities. Nevertheless, they also make the data difficult to process using resources designed for edited text or for contemporary broadcast speech (Hamed and Zaidkilani, 2025).

In this paper, we introduce The *NakbaEcho* dataset, a structured textual dataset of interviews transcribed from the *Palestine Remembered*[1] community. The dataset provides useful metadata from the interviews that facilitate direct computational research, including places of origin, recording dates, and significant mentions. It assembles a structured transcription resource with timestamped segments, speaker labels, and auxiliary annotations that enable alignment between text and audio. We show some preliminary interesting patterns that motivate future research directions on the dataset.

The *NakbaEcho* dataset is provided for free to the research community[2]. The current repository release corresponds to the 708 transcripts analyzed in this paper, and additional transcribed hours will be added in future updates as processing continues.We hope it would facilitate producing future studies in the fields of NLP, social science, and political science to analyze the narrative on Nakba through the lens of the people who lived it.

[1]The Website of Palestine Remembered: https://www.palestineremembered.com/index.html
[2]https://github.com/batouln/NakbaEcho

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 1–22
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

2. Related Work

2.1. Historical Context

The Palestinian Nakba of 1948 refers to the mass displacement and dispossession of Palestinians during the war surrounding the end of the British Mandate and the creation of the state of Israel (United Nations Committee on the Exercise of the Inalienable Rights of the Palestinian People, 2025; Encyclopaedia Britannica, 2026; Khalidi, 1992). The term *Nakba*, meaning "catastrophe" in Arabic, refers both to the events of 1948 and to their enduring political, social, and cultural consequences for Palestinians in exile, under occupation, and within present-day Israel (United Nations Committee on the Exercise of the Inalienable Rights of the Palestinian People, 2025; Masalha, 2008; Sa'di and Abu-Lughod, 2007). The rupture of 1948 followed the final years of the British Mandate and the adoption of United Nations General Assembly Resolution 181, which proposed the partition of Palestine into Arab and Jewish states (United Nations General Assembly, 1947; Encyclopaedia Britannica, 2026). During the war, more than 700,000 Palestinians were displaced, and hundreds of villages and urban neighborhoods were destroyed, depopulated, or rendered inaccessible to their inhabitants (United Nations Relief and Works Agency for Palestine Refugees in the Near East, 2026; Khalidi, 1992). *All That Remains* documents the destruction and depopulation of more than 400 Palestinian villages, underscoring the scale of this transformation (Khalidi, 1992). The consequences of 1948 extended far beyond the initial displacement. Palestinian refugees and their descendants became dispersed across the West Bank, Gaza, neighboring Arab countries, and wider diasporas, while the question of return remained central to Palestinian political life and historical memory (Sayigh, 1994; Sa'di and Abu-Lughod, 2007). United Nations General Assembly Resolution 194, adopted in December 1948, established return and compensation as enduring reference points in discussions of Palestinian refugeehood (United Nations General Assembly, 1948; United Nations Relief and Works Agency for Palestine Refugees in the Near East, 2026).

The Nakba, for Palestinians, is not only a historical event but also a framework for understanding ongoing dispossession, exile, and the transmission of collective memory across generations (Masalha, 2008; Sa'di and Abu-Lughod, 2007; Kassem, 2011). Oral history has been central to preserving these experiences where official archives and dominant narratives have marginalized Palestinian perspectives (Sayigh, 1994; Masalha, 2008; Kassem, 2011). Testimony, family narrative, and village-based remembrance remain key to how the Nakba is narrated and sustained socially, including through accounts of life before displacement, expulsion, and meaning of home and return (Sa'di and Abu-Lughod, 2007; Sayigh, 1994; Kassem, 2011). This context situates Nakba's oral recordings as more than speech data:They are records of memory, place, loss, and survival, shaped by the long afterlife of 1948 and the social worlds in which testimony is produced, archived, and transmitted(Sa'di and Abu-Lughod, 2007; Sleiman and Chebaro, 2018; The Nakba Archive, 2002).

2.2. Nakba Oral History Archives

Nakba oral history has been preserved through archival initiatives focused on documentation and access rather than NLP-oriented data release (Nur, 2008). Two important collections are the Nakba Archive (The Nakba Archive, 2002) and the Palestinian Oral History Archive (POHA) (American University of Beirut). The Nakba Archive documents testimonies of Palestinians displaced in 1948 through audio/video interviews, but only a small subset has been transcribed and translated into English, limiting computational use (Lamar et al., 2025). Similarly, POHA preserves over one thousand hours of testimonies with rich metadata, but remains a media archive rather than a text corpus for speech and language research (Sleiman and Chebaro, 2018).

Thus, despite their importance, these archives highlight a persistent challenge: preservation infrastructures do not readily translate into reusable computational corpora. Researchers must often reconstruct transcripts, metadata, and analysis-ready files from heterogeneous sources.

2.3. Computational Work on Nakba

Recent work in Nakba NLP has begun converting parts of existing archives into analyzable resources (Jarrar et al., 2025). Lamar et al. (2025) present a geo-referenced dataset built from translated Nakba Archive interviews, showing how oral-history transcripts can support analysis of place, geography, and displacement. However, it is limited to about thirty English-translated interviews and focuses on manually annotated geographic references. Awad et al. (2025) analyze narrative cohesion in POHA oral histories using derived transcripts, demonstrating the value of large-scale collections for studying collective memory, shared origin, and gendered narratives. However, their contribution remains analytical and does not provide a publicly released, unified Arabic transcript corpus. Other resources draw on literary rather than oral sources. Hamed and Zaidkilani (2025) introduce a topic-classification corpus from eight written Nakba short stories. While useful, it differs from oral testimony in genre, language, and interactional structure, lacking features such as

spontaneous speech, interviewer–interviewee dynamics, and disfluency. Overall, existing work highlights the value of Nakba narratives but also their fragmentation: translated subsets, study-specific transcripts, and small literary corpora support specific tasks, while a reusable large-scale Arabic oral-history corpus remains lacking.

2.4. Arabic speech and dialect resources

The scarcity of large, well-structured Arabic speech resources has been widely noted, especially for dialectal and conversational varieties (Djanibekov et al., 2025). Recent datasets such as Casablanca (Talafha et al., 2024) have helped address this problem by providing dialectal Arabic speech with transcriptions and annotations for downstream tasks. These resources are important because they move beyond Modern Standard Arabic and highlight the need for datasets grounded in naturally occurring spoken language (Keleg et al., 2023). However, existing Arabic speech corpora generally come from conversational media, interviews, online videos, or benchmark-oriented collection settings rather than from historical oral testimony (Talafha et al., 2024). As a result, they do not capture the linguistic and documentary properties that make oral-history interviews distinctive: long-form remembrance, intergenerational narrative, historically specific vocabulary, place-rich discourse, and close coupling between archival metadata and speech content (Sa'di and Abu-Lughod, 2007; Sayigh, 1994; Sleiman and Chebaro, 2018). Nakba testimony therefore occupies a unique space at the intersection of dialectal Arabic speech technology, digital archives, and computational humanities (Jarrar et al., 2025; Lamar et al., 2025; Awad et al., 2025).

3. Data & Methods

3.1. Data Statistics

The *Palestine Remembered* Oral History index contains 1,341 interview entries, where a single entry may include more than one speaker (e.g., a narrator and their spouse). As shown in Table 1, recordings span the period from 1994 to the present, with durations ranging from 10 minutes to over 25 hours, yielding more than 3,000 hours of content.

Of these, **708 interviews** (approximately **2,180 hours**) were successfully transcribed to form the *NakbaEcho* dataset.

Geographic coverage The transcribed interviews cover all 14 historical Palestinian districts. Safed accounts for the largest share (213 interviews, 31.6%), followed by Haifa (124, 18.4%), Ramla (52, 7.7%), and Tiberias (47, 7.0%). Together, Safed and Haifa represent just over half (50.1%) of all transcribed interviews.

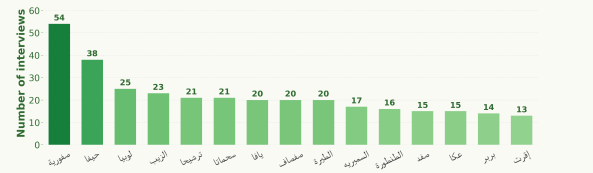


Figure 1: Count of interviews per village.

At the village level, the dataset covers 239 unique villages. Of these, 94 are represented by a single interview, reflecting broad geographic coverage alongside an uneven distribution of documentation across locations.

Narrator demographics The 708 transcribed interviews involve 719 individual narrators, as some recordings include multiple speakers. Gender was automatically detected using Claude Opus from narrator names and manually reviewed for gender-ambiguous Arabic names (e.g., وسام, شحادة, نعمة, جهاد), yielding 536 male narrators (74.5%) and 183 female (25.5%). To validate these labels, we compared them against speaker-level gender annotations independently produced by the Gemini transcription model, which infers narrator gender from acoustic features during diarization. Despite relying on entirely different signals—name morphology versus voice characteristics—the two methods produce near-identical distributions (74.6% male, 25.4% female) and agree on 98.6% of the 579 directly comparable cases (8 mismatches), supporting the reliability of both annotations.

Gender representation differs substantially by source. POHA recordings approach parity (44.1% female), whereas YouTube-sourced interviews are heavily male-dominated (8.7% female), reflecting the different collection contexts of each archive.

Birth years are available for 582 narrators and range from 1897 to 1962, with a median of 1929. The majority (488 of 582, 83.8%) were born between 1920 and 1939, indicating that most narrators were adolescents or young adults at the time of the 1948 Nakba.

3.2. Transcription Pipeline

Here we describe the transcription pipeline used to convert the recordings into text with rich metadata.

Source	Recording period	# interviews	Duration (mins)	Duration (hrs)
POHA	1994–2009	349	25,852	≈431
YouTube	2003–2026	323	102,878	≈1,715
Zochrot	2002–2026	36	2,050	≈34
Total	1994-2026	708	130,779	≈2,180

Table 1: Source-level statistics for the transcribed *NakbaEcho* dataset (708 interviews, ≈2,180 hours).

Start	End	Speaker	Transcript
00:00:22.500	00:00:29.800	SPK_01	السلام عليكم ورحمة الله وبركاته نرحب بكم في حلقة جديدة
00:01:45.000	00:01:49.000	SPK_01	بداية حابين تعرفنا باسمك الكامل الله يخليك
00:01:49.000	00:01:52.000	SPK_02	عبد المجيد محمد أبو سعيد

Figure 2: Examples of transcribed segments and the fields associated with each segment.

Timestamp	Speaker	Gender	Text	Emotion
[00:01:31 → 00:01:37]	SPEAKER_01	Male	عمي تسمح لنا نباشر بأسئلة القسم الأول؟	neutral
[00:12:13 → 00:12:17]	SPEAKER_00	Male	طيب. يا سيدي الله يديم عليك الصحة والعافية إن شاء الله	happy
[01:14:24 → 01:14:35]	SPEAKER_02	Male	سنة الغلاء تروح تجيب بواخر قمح وترميهن بحيفا بالبحر، شايف كيف؟	sad

Table 2: A snapshot of transcribed segments as they appear in the corpus, showing speaker labels, timestamps, and emotion annotations.

3.2.1. Segments

All audio files were automatically transcribed using the Gemini 3 Pro API[3]. The model was configured for dialectal Palestinian Arabic, with explicit constraints to preserve spoken language without normalization into Modern Standard Arabic (MSA).

To minimize hallucination, the transcription prompt enforced: 1) No content invention; unintelligible speech marked as `[inaudible]`; 2) Dialect preserved as spoken; 3) Proper nouns (including village names) retained without correction; and 4)Absolute timestamps required for each segment.

Audio files were processed in chunks to accommodate long interviews. Timestamps were recorded as absolute positions relative to the original audio file. Speaker diarization was performed automatically by the transcription model. Speaker identifiers are consistent within each processed audio chunk. However, speaker identity resolution across chunk boundaries was not performed. Therefore, speaker labels are locally consistent but not globally merged across entire interviews. As shown in Figure 2, each transcript was structured into segments, with each segment showing its `start` timestamp, `end` timestamp, `speaker`, and verbatim `text`. All timestamps follow a strict `HH:MM:SS.mmm` format to enable alignment with the original audio.

[3]Gemini 3 Pro: https://ai.google.dev/gemini-api/docs/gemini-3

3.2.2. Metadata

Speaker metadata: Table 2 shows the information generated for each speaker appearing within a processed chunk. The fields are divided into global ones like `gender` and local ones per segment like `emotion`, which can be *neutral*, *sad*, *angry*, *fear*, *happy*, or *tired*.

These attributes were model-inferred and were not manually validated. They should therefore be interpreted as automatic annotations rather than ground-truth demographic labels.

Literal Mention Flags: To support computational analysis of historical references, the presence of the following three terms was evaluated both on the chunk and the segment level: 1) Nakba, 2) Jews/Israel, and 3) British Mandate.

For each chunk, a boolean flag indicates whether the term is mentioned; if mentioned, a strictly literal summary is generated describing the immediate context of the reference, otherwise, the summary field is left empty. At the segment level, corresponding boolean flags indicate whether a given segment had the term mentioned or not. The annotation strategy follows a literal-first principle: summaries reflect only explicitly stated content without interpretive abstraction.

The terminology reflects the original language used by narrators, including historically situated references such as اليهود; these are preserved verbatim and should be interpreted within their historical

and discursive context rather than as standardized contemporary labels.

3.3. Transcription Quality Evaluation

To assess transcription quality, we conducted a manual evaluation on a stratified sample of 22 clips, each approximately 15 minutes in length (around **5.5 hours** in total). The sample reflects dataset diversity across dialect, recording conditions, generation, and gender.

The sample includes speakers from major Palestinian regions (Galilee, Haifa, Gaza, Yaffa, Ramla–Lyd), with varied recording conditions and generational coverage. **Recording conditions** were also balanced: **9 clips** originate from **POHA** recordings, while the remaining **13 clips** were sourced from YouTube, PalRemembered and Zochrot sources. The narrators span birth years between **1897 and 1937**, covering four generational cohorts (**pre-1915 through 1935–1944**), and represent individuals who directly experienced the events of the 1948 Nakba. The **gender distribution** in the evaluation set (**18 male** and **4 female** speakers) reflects that of the broader corpus.

Transcription quality was measured using Word Error Rate (WER), computed by comparing the automatic transcripts against manually corrected references. The system achieves an **overall micro-averaged WER of 7.00%**, with a **mean per-clip WER of 7.25%** and a **median WER of 4.34%**. Performance is consistent across the sample: **59% of clips fall below 5% WER**, and **86% fall below 10% WER**. While most clips exhibit low error rates, a small number of outliers contribute to a **standard deviation of 9.59%**, with WER ranging from **0.78% to 46.22%**.

Error analysis shows that **deletions (1,018)** constitute the largest proportion of errors, followed by **substitutions (833)** and **insertions (305)**. These results indicate that the transcription system achieves high accuracy overall, with most errors arising from missed tokens rather than incorrect or spurious insertions, which is consistent with the challenges of transcribing dialectal speech under variable acoustic conditions.

4. Ethical Considerations

This dataset is derived from oral testimonies documenting lived experiences of displacement and loss during and after the 1948 Nakba. These materials originate from archival initiatives aimed at preserving historical memory, and this work builds on those efforts by enabling computational access rather than replacing the original archival context.

All transcripts and annotations are generated automatically and may contain errors or simplifications. In particular, model-inferred attributes (e.g., emotion labels and speaker characteristics) are not manually validated and should not be interpreted as ground truth. Analyses based on these annotations should therefore be treated as exploratory.

The dataset preserves the original language used by narrators, including historically situated terminology. These expressions reflect the context of the testimonies and are included as documentary evidence rather than normative labels.

We encourage responsible use of this resource. The dataset is intended for research in language technology and digital humanities, and should be used with attention to historical context and the limitations of automated processing. In particular, references to social, political, or religious groups (e.g., اليهود), "Jews", reflect historically situated usage in the testimonies. These terms are retained to preserve the integrity of the source material and should not be interpreted as generalized or decontextualized labels.

5. NakbaEcho Dataset Analysis

5.1. NakbaEcho Dataset Summary Stats

Table 3 summarises the main statistics of the transcribed *NakbaEcho* dataset. The transcription process enables the extraction of additional descriptive signals from the testimonies, including the number of exchanges per interview, the frequency of named entities (NEs), speaker gender distribution, and inferred emotional attributes.

It is important to emphasize that this analysis is exploratory in nature. The reported statistics rely on automatically generated transcripts and downstream processing pipelines (e.g., NER, speaker attribution, and emotion detection), which are subject to errors and biases. Factors such as transcription inaccuracies, diarization limitations, and model imperfections may affect the reliability of these derived metrics. Therefore, the results should be interpreted as indicative trends rather than precise measurements.

5.2. Emotion, Lexicon, and Gender

Overall Emotion Distribution: Figure 3 presents the distribution of emotion labels assigned to transcript segments across the 708 interviews. The vast majority of segments (988,941; 94.1%) carry a NEUTRAL label, consistent with the long-form, conversational nature of oral-history testimony. Among the 62,219 non-neutral segments (5.9%), SAD constitutes the largest share (33,581; 54.0%), followed by HAPPY (18,839; 30.3%), ANGRY (5,049; 8.1%), and FEAR (3,202; 5.1%). The remaining categories—TIRED, PROUD, SERIOUS, CURIOUS, IN-

Source	Interviews	Exchanges	Tokens	NEs	Male%	Female%
POHA	349	126,506	2,203,952	109,413†	51.5	48.5
YouTube	323	911,144	10,445,044	626,260†	97.7	2.3
Zochrot	36	13,510	201,760	11,959†	54.8	45.2
Total	708	1,051,160	12,850,756	747,632†	91.6	8.4

Table 3: Statistics on the transcribed *NakbaEcho* dataset by source. Gender percentages reflect segment-level speaking time distribution.

QUISITIVE, and NOSTALGIC—together account for 2.3% of non-neutral segments.

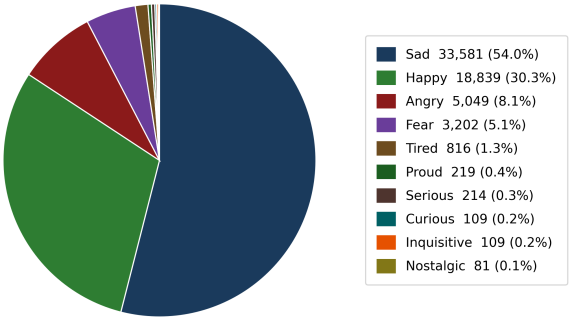


Figure 3: Distribution of non-neutral emotion labels across transcript segments (n = 708 interviews; 62,219 non-neutral segments, 5.9% of total). Neutral segments (94.1%) are excluded to highlight the relative distribution of affective categories.

The predominance of SAD segments reflects the thematic focus of the corpus: testimonies predominantly address displacement, loss, violence, and forced migration during the events of 1948. Because the dataset consists of first-person recollections of these experiences, segments describing expulsion, separation, destruction, and departure are expected to be frequent, resulting in a higher proportion of sadness-associated labels.

Lexical Patterns by Emotion: We examined how emotions are linguistically realized across the corpus. Segments labelled as SAD are dominated by displacement and loss vocabulary: طلعنا (we left), البلد (the town/village), الناس (the people), and اليهود appear prominently, frequently co-occurring in narrative passages describing departure from villages and the disruption of community life. طلعنا in particular functions as a recurring marker of forced departure across both SAD and FEAR segments.

HAPPY segments are characterized by vocabulary associated with social and communal life. High-frequency terms include كنا (we were), عنا (at our place / with us), and العريس (the groom), which occur in passages recounting weddings, family gatherings, and memories of pre-displacement daily life. The formulaic expression شاء (part of *in shā Allāh*, God willing) also appears prominently, reflecting the embedding of historical narration within everyday devotional language.

Segments labelled as ANGRY show higher prominence of collective conflict actors: اليهود, العرب (the Arabs), فلسطين (Palestine), and جيش (army) all rank among the top content words, reflecting descriptions of political confrontation and organised violence. FEAR segments similarly foreground اليهود and علينا (upon us / against us), suggesting narrative passages that describe perceived threat directed at the speaker and their community, alongside قالوا (they said) and قلت (I said), indicating reported speech as a vehicle for conveying danger and urgency.

The distribution of emotion labels and their associated lexicon indicates that emotional expression in the corpus is closely tied to narrative content. Conflict-related events and displacement episodes are predominantly associated with SAD, ANGRY, or FEAR labels, whereas segments recounting domestic life, religious observances, or social celebrations are more frequently associated with HAPPY. This pattern reflects the dual structure of many testimonies: recollections of everyday pre-1948 life alongside accounts of conflict and displacement. The analysis reports surface-level lexical co-occurrence within automatically labelled segments and does not interpret speaker intention beyond the textual evidence. Top content words stratified by emotion label are presented in Appendix E (Table 15).

Lexical Variation by Gender: To examine whether lexical salience differs across speaker demographics, we computed content-word frequency distributions separately for male and female narrators after stop-word removal (male: 8.24M tokens; female: 942K tokens). To test whether observed differences exceed chance variation, we applied log-likelihood ratio (G^2) tests and chi-square tests to all word types appearing at least five times across both subcorpora (110,954 types tested). Frequencies were normalised per 1,000 tokens for reporting purposes, while statistical tests were computed on



(a) Female Narrators



(b) Male Narrators

Figure 4: Word clouds of content terms by gender. Word size is proportional to frequency. Both groups share core narrative anchors (فلسطين, البلد, اليهود), while female speakers show greater prominence of kinship address forms and domestic vocabulary, and male speakers emphasise geographic and spatial terms. Differences are statistically supported via G^2 log-likelihood tests and Bonferroni-corrected chi-square tests (6,161 significant word types; see Table 17).

raw frequency counts. Of the tested types, 23,235 are significant at $p < 0.05$, and 6,161 remain significant after Bonferroni correction ($\alpha = 1.103 \times 10^{-7}$), indicating that the observed differences are statistically robust under the given corpus partition. The full significance table is provided in Appendix F (Table 17).

Both groups share core narrative anchors, including اليهود, البلد, and references to villages and land, indicating strong thematic overlap in how displacement and conflict are narrated. Statistically significant differences emerge in lexical prominence across groups. Male-favoured terms are dominated by spatial and geographic vocabulary: القرية (the village) appears at 2.46 per 1,000 tokens in male speech versus 0.20 in female speech ($G^2 = 3{,}104.8$); منطقة (area/region, $G^2 = 728.2$), متر (metre, $G^2 = 665.3$), and حوالي (approximately, $G^2 = 735.7$) follow a similar pattern, reflecting detailed descriptions of village layout, land boundaries, and geographic orientation. Female-favoured terms are predominantly kinship address forms and domestic vocabulary: يما (mother) appears at 1.08 per 1,000 tokens in female speech versus 0.04 in male speech ($G^2 = 3{,}170.7$); إنتي (you.FEM, $G^2 = 1{,}725.9$), بنت (girl/daughter, $G^2 = 1{,}676.2$), إمي (my mother, $G^2 = 1{,}552.3$), and جوزي (my husband, $G^2 = 1{,}354.5$) follow the same pattern, alongside movement verbs such as طلعنا (we left, $G^2 = 1{,}056.4$), reflecting a narrative register centred on family relationships and lived household experience.

These differences should not be interpreted as categorical distinctions but rather as statistically supported distributional tendencies within a shared narrative framework. We further note that the two gender subcorpora differ substantially in source composition: female segments are concentrated in POHA recordings (44.1% female), while male segments are dominated by YouTube-sourced interviews (97.7% male). As a result, the observed lexical contrasts may reflect not only speaker gender but also differences in recording context, interview structure, collection conditions, and dialectal variation. The reported results should therefore be interpreted as distributional differences between subcorpora rather than purely gender-driven effects.

Appendix F presents the top 15 content words per group (Table 16) and the top Bonferroni-significant words by gender (Table 17).

5.3. Named Entity Analysis

To characterize the informational density of the transcribed corpus, we applied named entity recognition (NER) using the CAMeL-Lab Arabic BERT model (Inoue et al., 2021), fine-tuned on the ANERcorp dataset, to all 708 interviews. The NER component extracts entity mentions and assigns each mention to one of four coarse-grained types: locations (LOC), persons (PERS), organizations (ORG), and a residual MISC category. We report mention counts (how often an entity is referenced) as an indicator of referential density, and unique entity counts (how many distinct entities appear) as an indicator of lexical richness. We note that BERT sub-word tokenisation occasionally produces fragment artefacts; the curated entity lists in Appendix A exclude such fragments.

Overall Entity Yield:

Type	Count	%
LOC	309,655	41.4
PERS	302,167	40.4
MISC	102,972	13.8
ORG	32,838	4.4
Total	**747,632**	

Table 4: Entity mention distribution ($n = 708$).

Named entity recognition was performed on all

708 transcripts, we extracted 747,632 entity mentions. The distribution by type is shown in Table 4. Locations and persons together account for 81.8% of all extracted mentions, consistent with oral-history narratives that foreground spatial anchoring (villages, towns, routes of displacement) and interpersonal reference (family members, community figures, and named individuals).

Top Locations: The most frequently mentioned locations (detailed in Appendix A) reflect both narrators' places of origin and salient nodes in the geography of displacement. Palestine (12,115 mentions), Jaffa (10,252), and Haifa (10,098) lead the ranking, followed by Gaza (4,617) and Jerusalem (4,442), consistent with the prominence of major cities in accounts of 1948 and its aftermath. The appearance of Lydda (3,936) and Ramla (2,885) reflects the well-documented mass expulsion from these twin cities. Several non-Palestinian locations also appear prominently, including Jordan (3,003), Lebanon (2,893), Britain (2,824), and Egypt (2,333), reflecting transnational trajectories of displacement and the roles of regional states and colonial actors referenced in testimony.

Top Persons: The results are dominated by common first names and *mukhtar*/headman references. Muhammad (4,302) is the most frequent person entity, followed by المختار (al-Mukhtār, the village headman; 2,172) and the *kunyā* form Abū Muáammad (1,727). Among historically identifiable figures, results are consistent with the oral register of Palestinian testimony, where social relations and local community networks are foregrounded through first names, kinship-based reference forms, and community titles.

Top Groups & Organizations: Among organisations, UNRWA (وكالة الغوث, *Wakālat al-Ghawth*; 938 mentions) is the most frequent, followed by the Red Cross (الصليب الأحمر, *al-áalīb al-Aámar*; 445) and the United Nations (الأمم المتحدة, *al-Umam al-Muttaáida*; 318). UNRWA also appears under its Arabic acronym الأونروا (144), bringing its combined count to over one thousand. The Arab Bank (البنك العربي, *al-Bank al-Arabī*; 100), the Arab League (الجامعة العربية, *al-Jāmia al-Arabiyya*; 93), and al-Azhar (الأزهر; 83) also appear, reflecting the institutional framework of post-1948 refugee life—including relief provision, education, and international legal and diplomatic structures—as narrated by the interviewees.

Lexical Salience beyond Entity Types:

While the preceding analysis organises entities by semantic class, a complementary perspective is obtained by examining overall lexical frequency independent of entity taxonomy. We aggregated the most frequent substantive terms in the corpus through unigram, bigram, and trigram frequency analysis following stopword and discourse-marker removal. This approach captures recurrent lexical items that structure narration beyond formal entity categories.

Geographical references constitute the most salient lexical stratum. Major urban centres such as حيفا (Haifa) and يافا (Jaffa) occur at high frequency, alongside generic spatial expressions including البلد (the town/village) and القرية (the village), indicating that spatial anchoring is central to testimonial discourse. Markers of social organisation and kinship are likewise prominent: أبو (*Abū*, "father of") and المختار (village headman) reflect the interpersonal and community-based structure of oral narration. Conflict-related vocabulary forms a further major axis of lexical recurrence. Frequently occurring items include اليهود (19,090 occurrences across 649 interviews), جيش الإنقاذ (Arab Salvation Army, 1,399 bigram mentions), and references to British and regional military actors. Religious–cultural formulae, including الله يرحمه and إن شاء الله, demonstrate the embedding of historical narration within everyday devotional language. Temporal markers—most notably references to 1936 and 1948—anchor personal recollection within shared historical chronology. Taken together, the distribution of high-frequency lexical items suggests that the corpus is organised around four interrelated referential domains: locality, social relations, conflict actors, and historical time. These domains emerge inductively from frequency patterns rather than from predefined thematic categories.

Thematic Categories & Illustrative Evidence:

The thematic annotations assigned to transcript segments capture recurrent narrative foci across the corpus. The three most frequent categories—**Nakba**, **Jews/Israel**, and **British Mandate**—are summarised below (detailed examples in Appendix C).

The **Nakba** category comprises segments referring to the events of 1947–1949 and their immediate aftermath, including displacement, village depopulation, military confrontation, and early refugee experience. Nakba-flagged segments total 50,606 across 683 of the 708 processed interviews—a 96.5% prevalence rate that underscores the near-universal presence of displacement narrative across the corpus, regardless of source, recording date, or narrator background. These segments typically combine spatial references (routes of flight, destinations, camps) with descriptions of family separation, material loss, and uncertainty. A recurrent narrative element concerns expectations of return, frequently framed through references to anticipated

Arab military intervention.

The **Jews/Israel** category includes segments in which narrators refer to Jewish armed groups, settlers, or the state of Israel as actors in the described events. The most frequent lexical item within this category is اليهود (*al-Yahūd*), occurring 19,090 times across 649 interviews and representing the most frequent conflict-related term overall (Appendix B). Additional expressions include العصابات الصهيونية (Zionist gangs, 513 mentions) and جيش الإنقاذ (Arab Salvation Army, 1,399 mentions). These forms are reproduced verbatim from the original testimonies and reflect narrators' own historical and linguistic framing of events, where distinctions between religious, ethnic, and national categories may differ from contemporary usage. Jews/Israel-flagged segments total 32,671 across 670 interviews (94.6%), and predominantly describe armed confrontation, detention, killing, or destruction of property, frequently co-occurring with sadness- or seriousness-labelled segments in the emotion annotation layer.

The **British Mandate** category encompasses references to the period of British rule in Palestine (1920–1948). Narrators refer to the British both colloquially (الانجليز) and institutionally (الانتداب البريطاني, الحكومة البريطانية, 1,081 bigram mentions; حكومة الانتداب البريطاني, 326 trigram mentions). These segments total 16,422 across 609 interviews (86.0%) and describe military searches, suppression of revolt, economic regulation, and arms confiscation. The emotional annotation layer indicates frequent co-occurrence with seriousness or anger labels.

Across categories, lexical choices reflect the vernacular Palestinian Arabic register in which the testimonies were delivered. The corpus preserves these terms as documentary evidence of historical memory and narrative framing. Appendix C presents randomly sampled segments illustrating how spatial references, named events, and institutional actors surface in context, and how emotional register varies across thematic domains.

Interview Question Analysis: To characterise interviewer behaviour and narrative scaffolding, we extracted and categorised 250 distinct question forms across the corpus (detailed in Appendix D). A substantial proportion of distinct question forms (70 out of 250) consists of backchannel or filler prompts, reflecting the conversational and oral-history format of the interviews. Substantive informational categories are dominated by questions concerning village geography and layout (44 distinct forms), daily life and economy (32), and name clarification (16), indicating systematic efforts to document spatial organisation, social structure, and personal identity.

We further examined thematic alignment between interviewer questions and corpus-level themes. Of 300,842 question-bearing segments, 3.7% (11,162) are associated with the Nakba theme, 2.2% (6,566) with Jews/Israel, and 1.2% (3,606) with the British Mandate. These proportions reflect the inclusion of YouTube-sourced interviews, which follow less structured thematic protocols than the POHA archival collection. Nevertheless, the relative prominence of Nakba-related question segments confirms that interview protocols are structured to foreground events of 1947–1949 and their aftermath. At the same time, the presence of detailed prompts regarding village layout, agriculture, education, and religious practice demonstrates that interviewers systematically reconstruct pre-displacement social life. The resulting corpus therefore reflects both event-centred historical documentation and ethnographic reconstruction of everyday life.

6. Limitations

While *NakbaEcho* provides a large-scale resource for analyzing Palestinian oral testimony, several limitations remain. Transcription quality is evaluated on a limited sample and should be considered preliminary; performance may vary with audio conditions, dialects, and speakers. Several annotations (e.g., emotion, speaker attributes, and mention flags) are automatically generated and not fully validated, and may reflect model biases. Speaker and gender attribution may contain errors due to diarization and name-based inference, and observed lexical differences may reflect source composition rather than true group effects. Finally, variation in recording conditions and interview styles across sources may influence linguistic patterns. Future work will focus on improving evaluation and annotation reliability.

7. Conclusion & Future Work

In this work, we present a novel transcribed dataset for Nakba narratives that opens up a lot of opportunities for information extraction through text processing and analysis. We show through our sample analyses the presence of potential research directions that can utilize the extracted fields and texts. Future work should leverage this data to inspect the demographics and the circumstances surrounding the event of the Nakba. It should also inspect the similarities and the differences between the patterns associated with the Nakba and other later similar events.

8. Bibliographical References

American University of Beirut. Palestinian oral history archive (poha). https://www.aub.edu.lb/ifi/Pages/poha.aspx. Accessed: 2026-02-27.

Huthaifa I Ashqar. 2025. Sentiment analysis of nakba oral histories: A critical study of large language models. In *Proceedings of the first International Workshop on Nakba Narratives as Language Resources*, pages 30–36.

Ghadir A Awad, Tamara N Rayan, Lavinia Dunagan, and David Gamba. 2025. Collective memory and narrative cohesion: A computational study of palestinian refugee oral histories in lebanon. In *Proceedings of the first International Workshop on Nakba Narratives as Language Resources*, pages 83–102.

Amirbek Djanibekov, Hawau Olamide Toyin, Raghad Alshalan, Abdullah Alatir, and Hanan Aldarmaki. 2025. Dialectal coverage and generalization in arabic speech recognition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29490–29502.

Encyclopaedia Britannica. 2026. 1948 arab–israeli war. https://www.britannica.com/event/1948-Arab-Israeli-War. Accessed: 2026-02-27.

Osama Hamed and Nadeem Zaidkilani. 2025. Arabic topic classification corpus of the nakba short stories. In *Proceedings of the first International Workshop on Nakba Narratives as Language Resources*, pages 48–55.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in arabic pre-trained language models.

Mustafa Jarrar, Nizar Habash, Mo El-Haj, Amal Haddad Haddad, Zeina Jallad, Camille Mansour, Diana Allan, Paul Rayson, Tymaa Hammouda, and Sanad Malaysha. 2025. Proceedings of the first international workshop on nakba narratives as language resources. In *Proceedings of the first International Workshop on Nakba Narratives as Language Resources*.

Fatma Kassem. 2011. *Palestinian women: Narrative histories and gendered memory*. Bloomsbury Publishing.

Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. Aldi: Quantifying the arabic level of dialectness of text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10597–10611.

Walid Khalidi. 1992. All that remains. *The Palestinian villages occupied and*.

Annie K Lamar, Rick Castle, Carissa Chappell, Emmanouela Schoinoplokaki, Allene M Seet, Amit Shilo, and Chloe Nahas. 2025. Cognitive geographies of catastrophe narratives: Georeferenced interview transcriptions as language resource for models of forced displacement. In *Proceedings of the first International Workshop on Nakba Narratives as Language Resources*, pages 18–29.

Nur Masalha. 2008. Remembering the palestinian nakba: Commemoration, oral history and narratives of memory. *Holy Land Studies*, 7(2):123–156.

Masalha Nur. 2008. Remembering the palestinian nakba: Commemoration, oral history and narratives of memory. *Holy Land Studies*, 7(2):123–156.

Ahmad H Sa'di and Lila Abu-Lughod. 2007. *Nakba: Palestine, 1948, and the claims of memory*. Columbia University Press.

Rosemary Sayigh. 1994. *Too Many Enemies: The Palestinian Experience in Lebanon*. Zed Books, London.

Hana Sleiman and Kaoukab Chebaro. 2018. Narrating palestine: The palestinian oral history archive project. *Journal of Palestine Studies*, 47(2):63–76.

Bashar Talafha, Karima Kadaoui, Samar Mohamed Magdy, Mariem Habiboullah, Chafei Mohamed Chafei, Ahmed Oumar El-Shangiti, Hiba Zayed, Mohamedou Cheikh Tourad, Rahaf Alhamouri, Rwaa Assi, et al. 2024. Casablanca: Data and models for multidialectal arabic speech recognition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21745–21758.

The Nakba Archive. 2002. Nakba archive. https://www.nakba-archive.org/. Accessed: 2026-02-27.

United Nations Committee on the Exercise of the Inalienable Rights of the Palestinian People. 2025. About the nakba. https://www.un.org/unispal/about-the-nakba/. Accessed: 2026-02-27.

United Nations General Assembly. 1947. Future government of palestine: United nations general assembly resolution 181 (ii). https://digitallibrary.un.org/record/210008. Adopted 29 November 1947.

United Nations General Assembly. 1948. Palestine—progress report of the united nations mediator: United nations general assembly resolution 194 (iii). https://www.unrwa.org/content/resolution-194. Adopted 11 December 1948.

United Nations Relief and Works Agency for Palestine Refugees in the Near East. 2026. Palestine refugees. https://www.unrwa.org/palestine-refugees. Accessed: 2026-02-27.

A. Appendix: Top Named Entities

This appendix reports the most frequent named entities in the corpus by category.

Table 5: Top named entities by mention frequency across four NER types ($n = 708$ interviews, 747,632 total mentions). Entity text is shown in the original Arabic. Sub-word tokenisation fragment artefacts have been excluded from the ranked lists. MISC and ORG lists are truncated where fragment artefacts begin to dominate the rankings.

LOC — Locations (309,655; 41.4%)				PERS — Persons (302,167; 40.4%)			
#	Entity	Gloss	Freq.	#	Entity	Gloss	Freq.
1	فلسطين	Palestine	12,115	1	محمد	Muáammad	4,302
2	يافا	Jaffa	10,252	2	المختار	the headman	2,172
3	حيفا	Haifa	10,098	3	أبو محمد	Abū Muáammad	1,727
4	غزة	Gaza	4,617	4	أحمد	Aámad	1,461
5	القدس	Jerusalem	4,442	5	موسى	Mūsā	1,406
6	اللد	Lydda	3,936	6	علي	Alī	1,401
7	الخليل	Hebron	3,416	7	صالح	áāliá	1,296
8	عكا	Acre	3,064	8	مختار	headman	1,093
9	الأردن	Jordan	3,003	9	محمود	Maámūd	1,069
10	لبنان	Lebanon	2,893	10	عبد الله	Abd Allāh	1,031
11	الرملة	Ramla	2,885	11	حسن	áasan	1,014
12	بريطانيا	Britain	2,824	12	يوسف	Yūsuf	996
13	الناصرة	Nazareth	2,785	13	حسين	áusayn	945
14	مصر	Egypt	2,333	14	إبراهيم	Ibrāhīm	921
15	بيسان	Baysan	2,183	15	أبو علي	Abū Alī	647

MISC — Miscellaneous (102,972; 13.8%)				ORG — Organisations (32,838; 4.4%)			
#	Entity	Gloss	Freq.	#	Entity	Gloss	Freq.
1	الله	God	21,761	1	وكالة الغوث	UNRWA	938
2	ليرة	lira	3,029	2	الصليب الأحمر	Red Cross	445
3	دينار	dinar	1,360	3	الأمم المتحدة	United Nations	318
4	جنيه	pound	784	4	الأونروا	UNRWA (acronym)	144
5	الإسلام	Islam	506	5	البنك العربي	Arab Bank	100
6	الحرب العالمية الثانية	World War II	487	6	الجامعة العربية	Arab League	93
7	القرآن	the Qurān	299	7	الأزهر	al-Azhar	83

Table 6: Most frequent substantive terms in the corpus (708 interviews). Terms drawn from unigram, bigram, and trigram frequency analysis with stopwords and discourse fillers removed. NER-derived entries (†) are drawn from the full 708-interview NER processing run (747,632 total mentions); frequencies may differ slightly from Table 5 due to sub-word tokenisation and aggregation differences. Entries marked (‡) fall outside the top-200 bigram or top-100 trigram cutoff for the 708-interview corpus; values shown are from the initial subset. *Freq.* = raw mention count; *DF* = document frequency.

Category	Arabic	Transliteration & Gloss	Freq.	DF	Src.
Geography & Displacement	†حيفا	*áayfā* (Haifa)	3,922	—	NER
	†فلسطين	*Filasáīn* (Palestine)	3,724	—	NER
	†غزة	*Ghazza* (Gaza)	1,961	—	NER
	†بيسان	*Bīsān* (Baysan)	1,324	—	NER
	†يافا	*Yāfā* (Jaffa)	1,142	—	NER
	†عكا	*Akkā* (Acre)	1,035	—	NER
	البلد	*al-balad* (the town/village)	37,031	659	1-g
	القرية	*al-qarya* (the village)	20,478	433	1-g
	الأرض	*al-ará* (the land)	7,572	618	1-g
	‡بئر السبع	*Bir al-Sab* (Beersheba)	617	13	2-g
	‡بلد الشيخ	*Balad al-Shaykh*	392	13	2-g
	†دير ياسين	*Dayr Yāsīn* (Deir Yassin)	365	—	NER
	عراق المنشية	*Irāq al-Manshiyya*	559	—	2-g
	‡مدينة حيفا	*madīnat áayfā* (city of Haifa)	269	—	2-g

Table 7: Most frequent substantive terms (continued): People & Social Relations.

Category	Arabic	Transliteration & Gloss	Freq.	DF	Src.
People & Social Relations	أبو	*Abū* (father of — *kunyā*)	43,387	666	1-g
	الناس	*al-nās* (the people)	26,467	674	1-g
	الشيخ	*al-Shaykh* (the sheikh/elder)	10,563	523	1-g
	أهل البلد	*ahl al-balad* (people of the village)	3,419	—	2-g
	‡عبد الله	*Abd Allāh*	759	64	2-g
	أهل القرية	*ahl al-qarya* (village people)	2,104	—	2-g
	دار أبو	*dār Abū* (house/clan of Abū…)	2,655	—	2-g
	أبو محمد	*Abū Muáammad*	1,164	—	2-g
	†المختار	*al-Mukhtār* (village headman)	465	—	NER
	عبد الرحمن	*Abd al-Raámān*	1,252	—	2-g
	‡عز الدين القسام	*Izz al-Dīn al-Qassām*	142	20	3-g

B. Appendix: Most Frequent Substantive Terms

This appendix reports the most frequent substantive terms in the corpus, grouped by thematic category and frequency source.

Table 8: Most frequent substantive terms (continued): Conflict & Military.

Category	Arabic	Transliteration & Gloss	Freq.	DF	Src.
Conflict & Military	اليهود	*al-Yahūd* (the Jews)	19,090	649	1-g
	جيش الإنقاذ	*Jaysh al-Inqādh* (Arab Salvation Army)	1,399	—	2-g
	الانتداب البريطاني	*al-Intidāb al-Brīáānī* (British Mandate)	1,081	—	2-g
	الجيش المصري	*al-Jaysh al-Miárī* (Egyptian Army)	756	—	2-g
	الشعب الفلسطيني	*al-Shab al-Filasáīnī* (Palestinian people)	715	—	2-g
	الجيش البريطاني	*al-Jaysh al-Brīáānī* (British Army)	703	—	2-g
	العصابات الصهيونية	*al-Aáābāt al-áahyūniyya* (Zionist gangs)	513	—	2-g
	‡ثورة الـ ٣٦	*Thawrat al-36* (1936 revolt)	117	20	3-g
	الحرب العالمية الثانية	*al-áarb al-Ālamiyya al-Thāniya* (WWII)	344	—	3-g
	‡مع العصابات الصهيونية	*ma al-Aáābāt al-áahyūniyya* (Zionist gangs)	77	13	3-g

Table 9: Most frequent substantive terms (continued): Institutions & Organizations.

Category	Arabic	Transliteration & Gloss	Freq.	DF	Src.
Institutions & Organizations	المدرسة	*al-madrasa* (the school)	7,428	543	1-g
	†وكالة الغوث	*Wakālat al-Ghawth* (UNRWA)	154	—	NER
	†الصليب الأحمر	*al-áalīb al-Aámar* (Red Cross)	146	—	NER
	†الأمم المتحدة	*al-Umam al-Muttaáida* (United Nations)	102	—	NER
	†البنك العربي	*al-Bank al-Arabī* (Arab Bank)	34	—	NER
	†الجامعة العربية	*al-Jāmia al-Arabiyya* (Arab League)	27	—	NER

Table 10: Most frequent substantive terms (continued): Religious & Cultural Expressions.

Category	Arabic	Transliteration & Gloss	Freq.	DF	Src.
Religious & Cultural Expressions	‡الله يرحمه	*Allāh yiráamu* (God rest his soul)	854	86	2-g
	‡إن شاء الله	*in shā Allāh* (God willing)	779	126	3-g
	الحمد لله	*al-áamdu li-llāh* (praise be to God)	1,723	—	2-g
	سيدنا محمد	*Sayyidnā Muáammad* (our Prophet)	1,150	—	2-g
	‡الله سبحانه وتعالى	*Allāh subáānahu wa-taālā* (God Almighty)	89	16	3-g

Table 11: Most frequent substantive terms (continued): Temporal & Historical.

Category	Arabic	Transliteration & Gloss	Freq.	DF	Src.
Temporal & Historical	‡الـ ٣٦	*al-36* (the year 1936)	324	32	2-g
	ثاني يوم	*thānī yōm* (the next day)	1,075	—	2-g
	الثمانية وأربعين	*al-thamāniya wa-arbaīn* (1948)	721	—	2-g
	الستة وثلاثين	*al-sitta wa-thalāthīn* (1936)	543	—	2-g
	‡الـ ٤٨	*al-48* (the year 1948)	271	—	2-g
	ذلك الوقت	*dhālika al-waqt* (that time)	684	—	2-g
	‡ثورة الستة وثلاثين	*thawrat al-sitta wa-thalāthīn* (1936 revolt)	93	15	3-g

C. Appendix: Thematic Examples

This appendix presents illustrative transcript segments for the three main thematic categories discussed in the paper.

Table 12: Illustrative transcript segments for each thematic category, drawn from randomly sampled and emotionally tagged examples. Text is presented in the original Palestinian Arabic as transcribed. Translations are provided by the authors.

Theme	Emotion	Original (Arabic)	Translation (English)
Nakba	Sad	محدش قال لنا اهربوا هي العالم تشردت لحالها.	Nobody told us to flee… People were displaced on their own.
Nakba	Sad	القتل والدمار والتهجير والذبح أمام أعيننا بالعجزة والأطفال. لم يأخذوا فينا إلا ولا رحمة… وكنا نعتقد أن الدول العربية والعرب سوف يقومون بهذا الجهد ويحررونا ونعود بعد قليل.	Killing, destruction, displacement, and slaughter before our eyes—the elderly and the children. They showed us no mercy… We believed the Arab states would make the effort, liberate us, and we would return soon.
Nakba	Fear	حدا بيقدر يقعد؟ هم كانوا يلفلفوا حالهم بدهم، محضرين حالهم بيقولوا بكرا بيلحقنا الضرب.	Could anyone stay? They were gathering themselves, preparing, saying tomorrow the shelling will reach us.
Jews/Israel	Serious	احنا بالاخير لانه في ناس تتعططوا… من دير ياسين لعندنا… عيبعثولنا انه سقطت البلد الفلانية… اليهود طبوا على دير ياسين ودمروا البيوت على اهاليها.	In the end, people came to us in distress… from Deir Yassin… They would send word that such-and-such village had fallen… The Jews attacked Deir Yassin and destroyed the houses on top of their inhabitants.
Jews/Israel	Sad	يومن صار الاجتياح إجا، اتوفى… وين كان؟ كان كمشينه اليهود في… الهراوي وفي النبي يوشع.	The day the invasion came, he died… Where was he? The Jews had detained him in… al-Hirawi and in Nabi Yusha.
Jews/Israel	Sad	مسكوهم اليهود.	The Jews caught them.
British Mandate	Serious	ويفتشوا على أشخاص معينين، على ثوار، على أسلحة… بعدين كانوا يقولوا إنه اللي كان يمسك عنده مجرد فشكة كانوا يقتلوه.	They would search for specific people, for revolutionaries, for weapons… They said whoever was caught with even a single cartridge would be killed.
British Mandate	Serious	فرنسا بتحب تزيح بريطانيا من المنطقة، وبريطانيا بتحب تزيح فرنسا من المنطقة.	France wanted to push Britain out of the region, and Britain wanted to push France out.
British Mandate	Angry	الحكومة البريطانية يا سيدي… هي اللي غلت الحبوب علينا… بقى يجي واحد اسمه مخمن من طرف الدولة البريطانية… يجي على البلد، يجي على المختار.	The British government, sir… they are the ones who raised the price of grain on us… An assessor would come on behalf of the British state… come to the village, come to the *mukhtar*.

D. Appendix: Interview Question Analysis

Table 13: Distribution of interviewer question categories (250 distinct question forms). *Distinct* = unique question forms; *Mentions* = total occurrences across interviews. This categorisation was derived through manual annotation of an initial interview subset and has not been re-applied to the full 708-interview corpus; counts reflect the initial analysis.

Category	**Distinct**	**Mentions**
Backchannel / filler	70	319
Village geography & layout	44	89
Other	43	91
Daily life & economy	32	66
Name clarification	16	59
Identity & consent	12	32
Social customs & religion	11	22
Nakba & historical events	9	18
Education	7	14
Agriculture & land	6	12

Table 14: Thematic coverage of interviewer questions across the 708-interview corpus. Each row reports the number of question-bearing segments associated with a thematic tag out of 300,842 total question segments. Absolute counts for Nakba remain stable relative to the initial subset despite the corpus more than doubling; the drop in percentage reflects the large volume of YouTube-sourced interviews, which contain more question segments overall but proportionally fewer thematic flags than POHA archival interviews.

Theme	**Question segments**	**%**
Nakba	11,162	3.7
Jews / Israel	6,566	2.2
British Mandate	3,606	1.2

E. Appendix: Emotion and Lexical Data

Table 15: Top 12 content words by emotion category (stopwords, punctuation, dialectal prepositions, and discourse markers removed), based on 708 interviews. Segment counts in parentheses refer to non-neutral segments only. †شاء is a fragment of the formulaic expression *in shā' Allāh* (God willing).

Sad (33,581)				Happy (18,839)			
#	Word	Gloss	F	#	Word	Gloss	F
1	واحد	one/someone	2,988	1	واحد	one/someone	1,291
2	الناس	the people	2,270	2	كنا	we were	1,169
3	البلد	the town/village	2,071	3	أبو	father of (kunyā)	1,036
4	حدا	someone (dial.)	1,546	4	لي	to me / for me	1,002
5	لي	to me / for me	1,440	5	كنت	I was	780
6	يوم	day	1,423	6	يوم	day	777
7	هناك	there	1,388	7	الناس	the people	713
8	أبو	father of (kunyā)	1,386	8	نعم	yes	669
9	كلها	all of it	1,276	9	البلد	the town/village	668
10	إلا	except / only	1,275	10	عنا	with us	649
11	اليهود	the Jews	1,256	11	شاء	God wills†	617
12	طلعنا	we left	1,210	12	العريس	the groom	590

Angry (5,049)				Fear (3,202)			
#	Word	Gloss	F	#	Word	Gloss	F
1	قلت	I said	535	1	اليهود	the Jews	306
2	واحد	one/someone	505	2	قلت	I said	296
3	اليهود	the Jews	500	3	الناس	the people	272
4	لي	to me / for me	359	4	البلد	the town/village	270
5	إذا	if / when	308	5	واحد	one/someone	253
6	فلسطين	Palestine	293	6	علينا	upon us	208
7	أبو	father of (kunyā)	272	7	قالوا	they said	199
8	الناس	the people	264	8	إلا	except / only	185
9	إلا	except / only	242	9	لي	to me / for me	171
10	عمي	my uncle	231	10	حدا	someone (dial.)	165
11	البلد	the town/village	230	11	عمي	my uncle	132
12	العرب	the Arabs	219	12	كنا	we were	125

F. Appendix: Lexical Variation by Speaker Gender

Table 16: Top 15 content words by speaker gender after filtering, based on segment-level annotations across 708 interviews (male: 8.24M tokens; female: 942K tokens).

Male (8.24M tokens)				Female (942K tokens)			
#	Word	Gloss	Freq.	#	Word	Gloss	Freq.
1	أبو	father of (kunyā)	42,367	1	الله	God	4,813
2	البلد	the town/village	34,357	2	حدا	someone (dial.)	3,191
3	مثلاً	for example	32,146	3	هون	here (dial.)	2,823
4	الله	God	31,834	4	بدي	I want (dial.)	2,745
5	اسمه	his name is	26,689	5	البلد	the town/village	2,675
6	عمي	my uncle	26,056	6	أبو	father of (kunyā)	2,649
7	الناس	the people	24,493	7	صار	became / happened	2,499
8	القرية	the village	20,293	8	عم	progressive marker	2,293
9	إشي	thing (dial.)	18,923	9	عنا	with us / at ours	2,258
10	بيت	house	18,856	10	قالت	she said	2,216
11	دار	house / clan	18,495	11	كنا	we were	2,207
12	مين	who	18,298	12	اليهود	the Jews	2,201
13	اليهود	the Jews	16,891	13	مين	who	2,167
14	هذول	these (dial.)	16,374	14	إشي	thing (dial.)	2,163
15	طبعاً	of course	16,374	15	كمان	also / too	2,071

Table 17: Most significantly differentiated content words by speaker gender, ranked by log-likelihood ratio (G^2). All listed words are Bonferroni-significant ($\alpha = 1.103 \times 10^{-7}$; all $p < 10^{-10}$). Frequencies normalised per 1,000 tokens. †برضه is a dialectal adverb meaning "also/too" (Levantine/Egyptian variety), significantly more frequent in male speech.

	Male-favoured						Female-favoured				
#	Word	Gloss	M/1k	F/1k	G^2	#	Word	Gloss	M/1k	F/1k	G^2
1	القرية	the village	2.46	0.20	3104.8	1	يما	(O) mother	0.04	1.08	3170.7
2	هذي	this (dem.)	1.85	0.31	1725.9	2	كتير	a lot (dial.)	0.31	1.53	1900.8
3	أبو	father of (*kunyā*)	4.97	2.55	1245.3	3	قام	(s)he got up	0.15	1.07	1742.7
4	†برضه	also/too (dial.)	1.11	0.19	1040.7	4	إنتي	you (fem.)	0.10	0.89	1725.9
5	اسمه	his name is	3.20	1.55	908.3	5	بنت	girl/daughter	0.37	1.56	1676.2
6	عبد	(name part)	1.74	0.63	814.8	6	ويا	and with her	0.06	0.71	1604.7
7	كيلو	kilo/km	0.89	0.16	813.3	7	إيه	yes (dial.)	0.65	2.07	1564.3
8	موجود	present/existing	0.99	0.23	739.5	8	إمي	my mother	0.07	0.73	1552.3
9	حوالي	approximately	0.84	0.16	735.7	9	متل	like (dial.)	0.24	1.15	1386.2
10	قرية	village (var.)	0.92	0.20	728.3	10	جوزي	my husband	0.002	0.35	1354.5
11	منطقة	area/region	0.90	0.19	728.2	11	حرام	shame/forbidden	0.04	0.52	1170.2
12	شايف	(I) see/notice	0.73	0.12	680.3	12	طلعنا	we left	0.43	1.38	1056.4
13	متر	metre	0.62	0.08	665.3	13	بالبيت	in the house	0.06	0.52	1001.3
14	أما	as for	1.15	0.36	648.8	14	قامت	she got up	0.07	0.55	992.7
15	القرى	the villages	0.95	0.25	643.6	15	حبيبتي	my dear	0.004	0.28	989.5

G. Appendix: Transcription Prompt and Output Schema

This appendix provides the exact prompt and structured output schema used in the transcription pipeline. The prompt was designed to enforce high-fidelity transcription of Palestinian Arabic oral testimonies, with strict constraints on hallucination, dialect preservation, timestamp formatting, and structured metadata extraction.

G.1. Transcription Prompt

```
You are transcribing low-quality archival audio interview in Palestinian Arabic
↪  dialect (elderly speakers).

Output MUST be valid JSON matching the provided schema.

GLOBAL RULES (CRITICAL)
- Do NOT invent content. If unclear, write [inaudible]. Do NOT guess.
- Keep dialect as spoken (do not convert to MSA).
- Some tokens may be Palestinian place/village names (proper nouns). Do NOT correct
↪  them into common words.
  If you strongly suspect a token is a place name but youre not fully sure, keep the
  ↪  token and add: [PLACE?]
- Provide speaker diarization using SPEAKER_01, SPEAKER_02... and keep IDs
↪  consistent WITHIN THIS CHUNK.
- Provide segment timestamps as ABSOLUTE timestamps in the ORIGINAL FILE.

TIMESTAMP FORMAT (ABSOLUTELY REQUIRED)
- Every timestamp MUST be exactly: HH:MM:SS.mmm
- Always include HOURS, MINUTES, SECONDS, and MILLISECONDS.
- Use a DOT for milliseconds (.) not a comma.
- If you are unsure about milliseconds, still output .000 (never omit it).

SPEAKER METADATA (REQUIRED)
For each speaker in this chunk, fill:
- gender: male|female|unknown
- dominant_emotion: neutral|sad|angry|fear|happy|tired|unknown
- voice_signature: very short literal signature to help match across chunks (e.g.,
↪  "elderly female, soft voice, slow pace").

MENTION FIELDS (REQUIRED per chunk)
1) Nakba mention:
- Determine whether the Nakba is mentioned or referenced in this chunk.
- If mentioned: write nakba_literal_summary strictly literally (what is said,
↪  immediate context).
- If not: nakba_literal_summary must be empty string.

2) Jews/Israel mention:
- Detect whether Jews () or Israel () are mentioned or referenced in this chunk.
- If mentioned: write jews_israel_literal_summary strictly literally.
- If not: jews_israel_literal_summary must be empty string.

3) British Mandate mention:
- british_mandate_mentioned=true if explicitly:     / .
- If true: british_mandate_summary literal. If false: empty string.

SEGMENT-LEVEL FLAGS (OPTIONAL)
Inside each segment:
- set nakba_mentioned / jews_israel_mentioned / british_mandate_mentioned true if
↪  that segment contains the mention.
```

G.2. Output Schema (Simplified)

```
{
  "chunk_start": "HH:MM:SS.mmm",
  "chunk_end": "HH:MM:SS.mmm",

  "nakba_mentioned": true/false,
  "nakba_literal_summary": "...",

  "jews_israel_mentioned": true/false,
  "jews_israel_literal_summary": "...",

  "british_mandate_mentioned": true/false,
  "british_mandate_summary": "...",

  "speakers": [
    {
      "speaker": "SPEAKER_01",
      "gender": "male|female",
      "dominant_emotion": "...",
      "voice_signature": "..."
    }
  ],

  "segments": [
    {
      "start": "HH:MM:SS.mmm",
      "end": "HH:MM:SS.mmm",
      "speaker": "SPEAKER_01",
      "text": "...",
      "uncertain": true/false,
      "emotion": "...",
      "nakba_mentioned": true/false,
      "jews_israel_mentioned": true/false,
      "british_mandate_mentioned": true/false
    }
  ]
}
```

G.3. Output Schema (Simplified)

```
{
  "chunk_start": "HH:MM:SS.mmm",
  "chunk_end": "HH:MM:SS.mmm",

  "nakba_mentioned": true/false,
  "nakba_literal_summary": "...",

  "jews_israel_mentioned": true/false,
  "jews_israel_literal_summary": "...",

  "british_mandate_mentioned": true/false,
  "british_mandate_summary": "...",

  "speakers": [
    {
      "speaker": "SPEAKER_01",
      "gender": "male|female",
      "dominant_emotion": "...",
      "voice_signature": "..."
    }
  ],

  "segments": [
    {
      "start": "HH:MM:SS.mmm",
      "end": "HH:MM:SS.mmm",
      "speaker": "SPEAKER_01",
      "text": "...",
      "uncertain": true/false,
      "emotion": "...",
      "nakba_mentioned": true/false,
      "jews_israel_mentioned": true/false,
      "british_mandate_mentioned": true/false
    }
  ]
}
```

Mining the Pre-1948 Palestinian Press: Unsupervised Keyphrase Extraction and Temporal Discourse Analysis from Five Historical Arabic Newspapers

Basel Barakat [1], **Nizam Barakat** [2]
[1] School of Computing, Goldsmiths University of London, New Cross, London, UK
[2] Independent Scholar
b.barakat@gold.ac.uk

Abstract

The Palestinian Arabic-language press of the late Ottoman and British Mandate periods constitutes a rich but computationally under-explored archive for studying the evolution of political, cultural, and social discourse in pre-1948 Palestine. This paper presents an end-to-end pipeline for extracting and analyzing thematic content from five historically significant Palestinian newspapers: *Lisān al-ᶜArab* (لسان العرب), *Al-Bushrā* (البشرى), *Al-Karmil* (الكرمل), *Al-Difāᶜ* (الدفاع), and *Filasṭīn* (فلسطين). We describe (i) the construction of a five-source corpus from scanned newspaper images obtained from archival collections, processed using the Google Cloud Vision OCR (GCV-OCR) API, (ii) the adaptation of KeyBERT with an AraBERT backbone for unsupervised keyphrase extraction, and (iii) a purpose-built Python visualization toolkit that produces keyword-frequency heatmaps, longitudinal trend charts, and ranked bar charts with full Arabic script rendering. Experiments across the five subcorpora show that the pipeline yields topically diverse keyphrases reflecting each newspaper's editorial orientation and its distinct representation of Palestinian native perspectives—from pan-Arab nationalism and anti-colonial resistance to religious and communal affairs. Temporal analysis reveals event-responsive patterns that align with major historical developments, including the 1936–1939 Arab Revolt and the intensification of sovereignty discourse toward 1948. The pipeline, data format specifications, and visualization code are provided as supplementary material.

Keywords: Arabic NLP, keyphrase extraction, historical newspapers, Palestinian press, corpus linguistics, KeyBERT, AraBERT

1. Introduction

The Arabic-language press that flourished in Palestine during the late Ottoman period and the British Mandate era (1920–1948) constitutes an invaluable primary source for historians, political scientists, and linguists. These newspapers documented the social, cultural, and political transformations of a formative period—including the rise of national movements, colonial administration, intercommunal tensions, and debates over statehood—in the everyday register of journalistic prose. Yet the vast majority of this material remains locked in physical archives and private collections, inaccessible to computational analysis.

Advances in optical character recognition (OCR) and neural language models now make it feasible to digitize and semantically analyze such corpora at scale; however, Arabic poses particular challenges: its right-to-left script, rich agglutinative morphology, and orthographic variation—especially in early twentieth-century printing—complicate both digitization quality and downstream NLP processing.

This paper addresses these challenges through an integrated four-stage pipeline (Figure 1) spanning document acquisition, OCR-based transcription, transformer-based keyphrase extraction, and temporal visualization. We apply this pipeline to five major Palestinian newspapers published before 1948:

- ***Lisān al-ᶜArab*** (لسان العرب) — a pan-Arab nationalist newspaper reflecting broad Arab political concerns (Wikipedia contributors, 2025b).
- ***Al-Bushrā*** (البشرى) — a publication with social, cultural, and community-oriented coverage.
- ***Al-Karmil*** (الكرمل) — published in Haifa (1908–1944) by Najib Nassar, one of the longest-running and most influential Palestinian newspapers, known for its political commentary and anti-colonial stance (Wikipedia contributors, 2025a).
- ***Al-Difāᶜ*** (الدفاع) — published in Jaffa (1934–1948), a leading daily that covered political developments extensively during the final Mandate years (Wikipedia contributors, 2026a).
- ***Filasṭīn*** (فلسطين) — published in Jaffa (1911–1948), by ᶜĪsā al-ᶜĪsā and Yūsuf al-ᶜĪsā, one of the most prominent and widely circulated Palestinian newspapers, known for its sustained political reporting across both Ottoman and Mandate periods (Wikipedia contributors, 2026b).

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 23–32
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

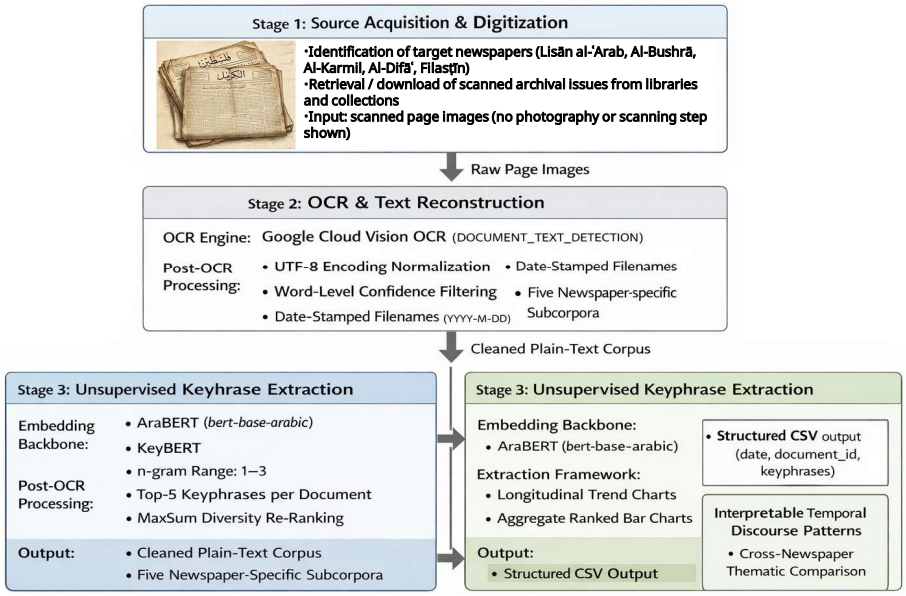


Figure 1: End-to-end pipeline overview. Newspapers from five Palestinian titles are collected and transcribed via GCV-OCR. Post-OCR processing normalizes encoding, filters low-confidence words, and segments pages into date-stamped document units. Plain-text files are processed by the KeyBERT/AraBERT extraction stage, producing a structured CSV per newspaper. The CSVs are utilized by the Arabic-aware visualization toolkit to generate temporal and comparative charts.

Our primary contributions are:

1. **Corpus construction.** We assemble five collections of historical Palestinian newspapers from scanned archival images, process them using the GCV-OCR API, and describe the acquisition workflow, quality filtering criteria, and resulting corpus organization spanning the late Ottoman through British Mandate periods.

2. **Adapted keyphrase extraction.** We apply KeyBERT (Grootendorst, 2020) with AraBERT (Antoun et al., 2020) as the backbone encoder for unsupervised keyphrase extraction from historical Arabic text, with a formal account of the scoring function.

3. **Cross-newspaper comparative analysis.** We compare keyphrase distributions across the five newspapers, revealing distinct editorial orientations and shared patterns of event-driven discourse.

4. **Arabic-aware temporal visualization.** We develop a standalone Python visualization toolkit producing heatmaps, trend lines, and bar charts with correct Arabic script rendering via character reshaping and BiDi reordering.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the corpus and OCR process. Section 4 details the keyphrase extraction methodology. Section 5 presents the visualization framework. Section 6 reports experimental results. Section 7 concludes.

2. Related Work

2.1. Digitization of Historical Arabic Press

The digitization of historical newspapers has received sustained attention in Western digital humanities (Smith et al., 2015), but work on Arabic archival collections remains comparatively sparse. Prior efforts have focused primarily on handwritten manuscript recognition (Al-Khateeb et al., 2011) or on contemporary newswire collections such as the Arabic Gigaword (Graff and Maamouri, 2006). Projects such as the Institute for Palestine Studies digital archive have preserved Palestinian textual heritage, but have primarily produced human-curated reference works rather than machine-readable corpora amenable to NLP processing.

Notable preservation initiatives include the digitization of the historical periodical collection at the Al-Aqsa Mosque Library in East Jerusalem (Matusiak and Abu Harb, 2011), which created archival-quality digital copies of 24 newspaper and magazine titles from the British Mandate period, and the Jrayed digital archive at the National Library of Is-

rael, which provides image-level access to Arabic periodicals from Ottoman and Mandatory Palestine. More recently, El Ganadi et al. (2025) proposed a metadata-driven framework for Arabic library digitization that addresses challenges specific to Arabic-script processing, including calligraphy, diacritics, and ligatures. These projects have made significant strides in preservation and access but have not incorporated NLP-based content analysis pipelines of the kind we present here.

Methodologically, the closest precedent to our work is Grallert (2021), who applied digital history methods—including social network analysis and stylometric authorship attribution—to a 2.65-million-word corpus of four Arabic journals from the late Ottoman Eastern Mediterranean (1906–1918). That study highlighted both the potential and the infrastructure challenges of computational approaches to Arabic periodicals, particularly regarding corpus construction and the socio-technical biases of digitization practices rooted in the Global North. Our work complements this line of research by extending transformer-based keyphrase extraction to Palestinian newspapers and by introducing temporal visualization methods tailored to right-to-left script.

The Palestinian press of the late Ottoman and Mandate periods is of particular scholarly interest because it documents the emergence of national consciousness and debates over British colonial governance in real time. Newspapers such as *Al-Karmil*, founded by Najib Nassar in 1908, *Filasṭīn*, founded by the ʿĪsā brothers in 1911, and *Al-Difāʿ*, which became one of the most widely circulated dailies in the 1930s, are frequently cited in historical scholarship (Ayalon, 1995; Khalidi, 1997) but have rarely been subjected to systematic computational analysis. To our knowledge, the present work is among the first to apply a complete OCR-to-NLP pipeline to archival scanned Palestinian newspaper collections spanning multiple titles and decades.

2.2. Keyphrase Extraction

Keyphrase extraction methods divide broadly into unsupervised statistical approaches—such as TF-IDF and YAKE (Campos et al., 2020)—and embedding-based methods that leverage pre-trained language models. KeyBERT (Grootendorst, 2020) belongs to the latter family, ranking candidate n-grams by their cosine similarity to the document embedding produced by a sentence encoder. While embedding-based methods have shown strong performance on English news corpora, their adaptation to Arabic remains limited. Antoun et al. (2020) introduced AraBERT, a BERT-base model pre-trained exclusively on Arabic text, which provides substantially richer morphological representations than general-purpose multilingual encoders such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020).

Recent work on Arabic keyphrase extraction has explored both supervised and unsupervised settings but has overwhelmingly targeted Modern Standard Arabic from contemporary sources. The application of embedding-based keyphrase extraction to historical Arabic—with its distinct orthographic conventions, vocabulary, and print quality—remains largely unaddressed.

2.3. Temporal Analysis of News Corpora

Temporal keyword analysis has been used to trace the evolution of political discourse in newspaper archives (Lansdall-Welfare et al., 2017). Heatmap visualizations of keyword frequency over time have proven effective for communicating thematic shifts to non-technical audiences in digital humanities contexts (Hamilton et al., 2016). Our visualization tool extends this tradition to right-to-left Arabic script through Unicode reshaping and enables cross-newspaper comparison of discourse patterns.

2.4. NLP for Palestinian Arabic

While substantial NLP resources exist for Modern Standard Arabic (MSA), computational work targeting the specific varieties and registers of Palestinian remains limited. The historical press adds a further temporal dimension: early twentieth-century journalistic Arabic differs from contemporary MSA in vocabulary, sentence structure, and orthographic norms. This creates a domain shift between AraBERT's pre-training distribution and our target corpus, as discussed in the Limitations section.

3. Data Collection and Corpus Construction

Our corpus is derived from five historically significant Palestinian Arabic-language newspapers, all published before 1948 (Jrayed). Together, the five collections span the late Ottoman period through the end of the British Mandate and represent a substantial archive of early twentieth-century Palestinian journalistic prose, encompassing nationalist, cultural, and political registers.

3.1. Digitization via GCV-OCR

Arabic OCR remains a challenging task due to the cursive nature of the script, the presence of diacritics, varying font styles, and—for historical documents—paper degradation and printing irregularities, as shown in Figure 2 (Faizullah et al.,



Figure 2: Filasṭīn Newspaper, issue date 2 November 1932 (Palestine Square, 1932)

2025; Al Ghamdi, 2022). Recent surveys indicate that deep learning architectures, particularly CNN-LSTM-CTC pipelines, have advanced handwritten Arabic recognition (Wagaa and Kallel, 2023), though printed historical Arabic at scale remains underexplored.

Scanned newspaper page images were submitted to the GCV-OCR API (Google Cloud, 2023) using the `DOCUMENT_TEXT_DETECTION` feature, which is optimized for dense, multi-column document layouts and returns a full-text annotation with word-level confidence scores.

Post-OCR processing comprised three steps:

1. **Encoding normalization.** All output was re-encoded in UTF-8 to ensure consistent handling of Arabic diacritics, Hamza variants, and punctuation marks.

2. **Confidence filtering.** Words with confidence below 0.6 were flagged; pages on which more than 30% of words fell below this threshold were excluded to limit noise propagation.

3. **Document segmentation.** Each page is treated as a single document unit, with the publication date encoded in the filename as `YYYY-MM-DD`. Article-level segmentation using layout analysis is left for future work.

3.2. Corpus Statistics

Table 1 summarizes the five subcorpora.

4. Keyphrase Extraction Methodology

We use *keyphrase* throughout this section to denote candidate n-grams extracted by the system (including multi-word expressions), reserving *keyword* for the general discussion of terms and the visualization tool, which operates at token level for frequency counts.

4.1. Extraction Framework

Our system builds on KeyBERT (Grootendorst, 2020), which frames keyphrase extraction as a semantic similarity problem: candidate n-grams are ranked by their cosine similarity to the global document embedding produced by a pre-trained encoder. Formally, given a document D and a set of candidate keyphrases $\{c_n\}$, the top-k keyphrase set is defined as:

$$\text{Keyphrases}(D) = \underset{c_n}{\text{argmax}} \ \ \text{cos-sim}(\mathbf{e}(c_n), \mathbf{e}(D)), \quad (1)$$

where $\mathbf{e}(\cdot)$ denotes the encoder embedding function. Because the formulation requires no labeled training data, it is well-suited to the low-resource setting of historical Arabic news.

4.2. Arabic Language Model Selection

A core design choice is the backbone encoder. General-purpose multilingual models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) encode Arabic as one of over 100 languages, potentially diluting language-specific morphological patterns. We therefore employ AraBERT (`asafaya/bert-base-arabic`) (Antoun et al., 2020), a BERT-base architecture pre-trained exclusively on Arabic news articles, Wikipedia pages, and web text. AraBERT achieves strong results on a range of Arabic NLP benchmarks—including named entity recognition and sentiment analysis—and its WordPiece vocabulary is trained entirely on Arabic, giving finer subword segmentation than multilingual alternatives. The model is loaded via the Hugging Face Transformers library (Wolf et al., 2020) and wrapped by the KeyBERT interface as a sentence encoder.

4.3. Data Preparation

All input files are stored in UTF-8 encoding. No explicit preprocessing (diacritization removal, stemming, or stopword filtering) is applied at the file

Table 1: Corpus statistics (post-filtering) for the five Palestinian newspaper subcorpora.

Property	*Lisān al-ᶜArab*	*Al-Bushrā*	*Al-Karmil*	*Al-Difāᶜ*	*Filasṭīn*
Publication city	Jerusalem	Haifa	Haifa	Jaffa	Jaffa
Active years	1921–1925	1935–1950	1908–1942	1934–1948	1911–1948
Issues digitized	543	103	1,238	3,874	6,692
Pages digitized	2,022	3,246	8,050	20,505	38,001

level: all tokenization decisions are delegated to the AraBERT tokenizer.

4.4. Keyphrase Extraction Configuration

Extraction hyperparameters are held constant across all documents, as detailed in Table 2.

The Max-Sum algorithm (Carbonell and Goldstein, 1998) greedily selects keyphrases that maximize document-similarity scores while penalizing high pairwise cosine similarity between selected candidates, yielding a topically diverse output well-suited to multi-topic news documents.

4.5. Implementation Details

All experiments are implemented in Python 3 using `keybert`, `transformers` (Wolf et al., 2020), and `pandas`. The pipeline is fully deterministic given fixed model weights, as the Max-Sum algorithm and cosine similarity computations contain no stochastic components.

5. Visualization Framework

Extraction output is stored as a three-column CSV with fields `date`, `filename`, and `keywords`, where the latter contains the top-5 keyphrases for each document. The visualization toolkit ingests this format and produces three complementary chart types described below.

5.1. Data Filtering

During loading, the tool discards rows with unparseable dates and retains only keyphrases containing at least one character in the Arabic Unicode block (U+0600–U+06FF), eliminating OCR artifacts and non-Arabic tokens. The top N keyphrases by overall frequency are selected for plotting (default $N{=}15$; $N{=}8$ for trend lines to preserve readability).

5.2. Chart Types

Three chart types are generated for each subcorpus:

Keyword $\times$ Year Heatmap (Figure 4). A heatmap shows the frequency of the top-N keyphrases (rows) across publication years (columns) using a `YlOrRd` color scale. This view enables rapid identification of temporal surges and declines in specific terminology.

Longitudinal Trend Lines (Figure 5). Line charts plot annual keyphrase frequency per newspaper, one colored line per keyphrase. This view supports fine-grained comparison of relative thematic prominence over time.

Overall Ranked Bar Chart (Figure 3). Horizontal bar charts show the aggregate frequency of the top-N keyphrases across the full time span for each newspaper, providing a domain summary independent of temporal structure.

6. Results and Analysis

6.1. Extracted Keyphrases: Qualitative Overview

The overall keyword rankings (Figure 3) show that each newspaper's extracted vocabulary reflects its distinct editorial character.

Lisān al-ᶜArab yields keyphrases dominated by references to the newspaper itself and to broader Arab political discourse, with the highest-frequency terms including newspaper-name collocations and references to cities such as Jerusalem and Haifa. *Al-Bushrā* surfaces religious and cultural terminology—including Islamic references and community-related terms—consistent with its orientation toward social affairs. *Al-Difāᶜ* produces keyphrases concentrated around Quranic references, personal names associated with political and religious figures, and editorial content, reflecting its status as a politically engaged daily. *Filasṭīn* generates the highest-frequency keyphrases in the corpus, dominated by newspaper-name collocations, references to law and governance, and administrative terminology, consistent with its role as a primary venue for sustained political reporting. *Al-Karmil* yields keyphrases related to education, community organizing, and political institutions, reflecting its editorial focus on political engagement.

The trigram range setting captures compound

Table 2: KeyBERT hyperparameter configuration used in all experiments.

Parameter	Value	Rationale
`keyphrase_ngram_range`	`(1, 3)`	Captures unigrams, bigrams, and trigrams to accommodate multi-word expressions common in Arabic news discourse.
`top_n`	`5`	Extracts the five highest-scoring keyphrases per document.
`stop_words`	`None`	Arabic function words are suppressed implicitly through low cosine similarity to document-level embeddings.
`use_maxsum`	`True`	Applies the Max-Sum algorithm (Carbonell and Goldstein, 1998) to maximize document similarity while minimizing inter-candidate redundancy.
`diversity`	`0.7`	Controls the balance between topical breadth and semantic precision in keyphrase selection.

terminology—multi-word political expressions, institutional references, and editorial collocations—that unigram-only approaches would miss.

6.2. Temporal Patterns

The heatmaps (Figure 4) reveal structured temporal variation across all five newspapers. Several patterns emerge.

Lisān al-ᶜArab, covering the early 1920s, shows concentrated keyphrase activity in 1922–1923, with a sharp decline thereafter, consistent with the newspaper's relatively brief publication window. *Al-Bushrā*, active in the 1940s and early 1950s, displays sparse but focused keyword distributions, with Islamic and cultural terminology appearing in isolated temporal clusters. *Al-Difāᶜ* (1934–1948) exhibits a clear temporal progression: keyphrase frequency intensifies in the mid-1930s and peaks during 1944–1947, coinciding with the post-war political crisis and the final years of the Mandate. *Filasṭīn* shows the most sustained temporal pattern, spanning from the early 1910s through 1941. Keyphrase frequency increases markedly in the 1930s—the period of the Arab Revolt and peak political mobilization—with a visible concentration of high-frequency terms between 1935 and 1940. *Al-Karmil* (1908–1944 in the broader record, though the digitized material here spans the 1920s–1930s) shows temporal variation consistent with the rise of nationalist discourse, with keyphrase activity concentrated in the late 1920s and early 1930s.

The trend lines (Figure 5) provide a complementary view of the temporal dynamics. In *Lisān al-ᶜArab*, the dominant keyphrases show a sharp peak around 1922 followed by rapid decline. *Al-Difāᶜ* exhibits a more complex pattern: certain keyphrases peak in the mid-1930s (around the outbreak of the Arab Revolt in 1936), decline during 1938–1940, and then resurge during 1942–1947. *Filasṭīn* shows the most dramatic temporal variation, with several keyphrases surging sharply in the mid-to-late 1930s—reaching frequencies above 80 per year—before declining in the early 1940s. *Al-Karmil* displays a distinctive double-peak pattern in the early 1920s and late 1920s, with keyphrase activity tapering off in the 1930s as the digitized coverage thins.

6.3. Cross-Newspaper Comparison

A central finding is the emergence of shared keyphrase patterns alongside newspaper-specific vocabulary. Several observations stand out from the comparative analysis of Figure 3.

First, newspaper-name collocations appear as top keyphrases in multiple titles (e.g., "newspaper Lisān al-ᶜArab" in the Arab subcorpus, "newspaper Palestine" in the Filasṭīn subcorpus). This reflects a characteristic feature of early twentieth-century Arabic newspaper typography, in which the newspaper name appeared frequently throughout each issue in mastheads, headers, and self-referential editorial content.

Second, *Filasṭīn* and *Al-Difāᶜ*—the two highest-volume subcorpora and the titles with the most sustained political focus—share a concentration of governance- and conflict-related terminology, though the specific keyphrases differ due to differences in editorial style and temporal coverage.

Third, *Al-Bushrā* stands apart from the other four titles, with its top keyphrases dominated by Islamic religious terminology and references to community figures. This confirms its distinct editorial orientation toward social and cultural affairs rather than overtly political reporting.

Fourth, *Al-Karmil* occupies an intermediate posi-

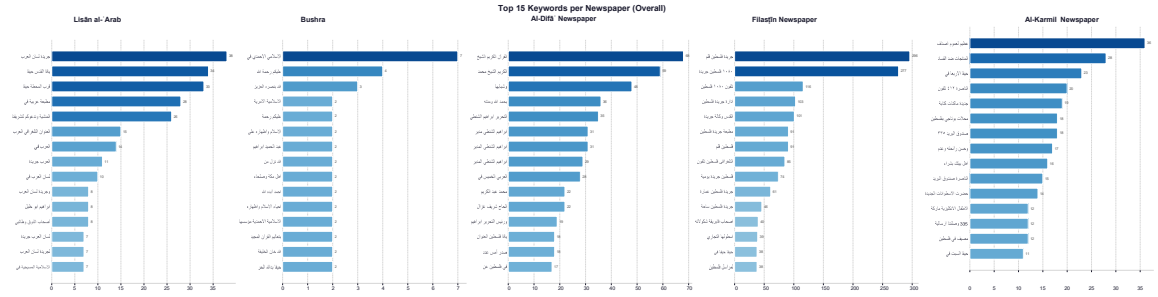


Figure 3: Overall top-15 keyphrases by total frequency for each of the five newspaper subcorpora. Each panel shows the ranked keyphrases extracted by KeyBERT with AraBERT from the respective subcorpus. *Filasṭīn* exhibits the highest absolute frequencies, consistent with its larger digitized volume. All labels are rendered in Arabic script.

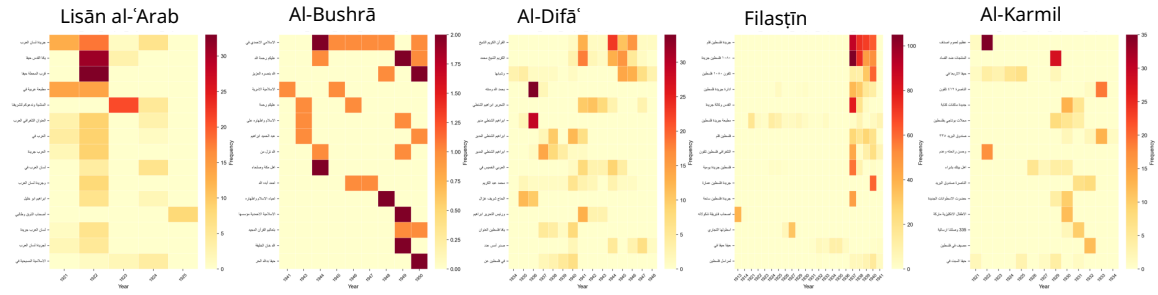


Figure 4: Keyword × Year frequency heatmaps for all five newspaper subcorpora. Rows represent the top-15 keyphrases; columns represent publication years. Color intensity (YlOrRd scale) encodes frequency. Clear temporal clustering is visible in all panels, with *Filasṭīn* and *Al-Difāᶜ* showing the most sustained high-frequency activity.

tion, combining references to political institutions with cultural and educational vocabulary, consistent with its role as a forum for both political commentary and community organizing.

6.4. OCR Noise and Robustness

A qualitative review of extracted keyphrases shows that OCR errors introduce noise primarily as partial words or incorrectly joined characters. The pipeline shows inherent robustness: malformed tokens receive low cosine similarity to the document embedding and are rarely selected among the top-5 candidates. The Arabic character filter applied during visualization provides an additional denoising pass. Nevertheless, pages with borderline OCR confidence remain a latent noise source; image binarization and language-model-based OCR post-correction are priorities for future work.

7. Conclusion

We have presented an end-to-end pipeline for digitizing, extracting, and visualizing thematic content from five major pre-1948 Palestinian Arabic newspapers. By combining GCV-OCR with AraBERT-backed KeyBERT extraction and a purpose-built Arabic-aware visualization toolkit, the system makes historical Palestinian press archives amenable to systematic content analysis without requiring labeled data or language-specific preprocessing beyond UTF-8 normalization.

The five-newspaper design enables both within-title temporal analysis and cross-newspaper comparison, revealing a shared core of political vocabulary alongside distinct editorial orientations. Temporally, all five subcorpora exhibit coherent, event-responsive patterns consistent with known historical developments: the Arab Revolt of 1936–1939, wartime reporting, and the post-war political crisis each leave visible imprints.

Acknowledgements and Ethical Considerations

The authors thank the staff who archived and facilitated access to the newspaper collections. All material consists of early-to-mid twentieth-century archival documents treated for non-commercial academic research. The study aims to preserve Palestinian cultural heritage through computational methods. The terminology and perspectives represented in these historical sources reflect

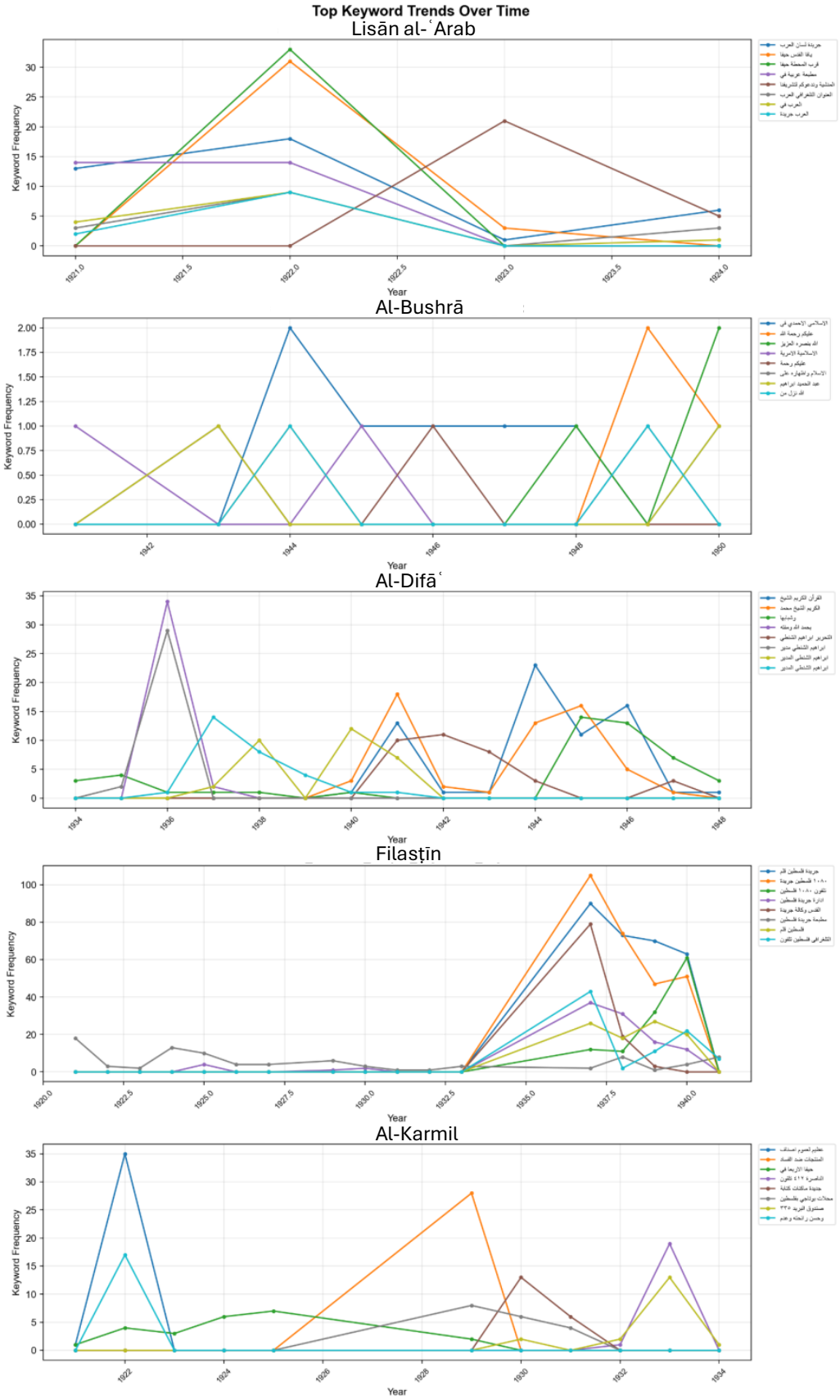


Figure 5: Longitudinal frequency trends for the top-8 keyphrases in each of the five newspaper subcorpora. Each colored line tracks one keyphrase over time. *Filasṭīn* shows the most pronounced temporal spikes, peaking in the mid-to-late 1930s during the Arab Revolt period.

the context of their time and do not necessarily represent the views of the authors.

Limitations

OCR quality varies across newspapers depending on source condition, print quality, and layout complexity; confidence-based filtering mitigates but cannot eliminate noise, especially for pre-1920 material. AraBERT's pre-training on MSA may not fully capture early twentieth-century orthographic conventions, posing a domain-shift challenge common to all transformer-based approaches to historical text. Additionally, newspaper-name collocations appear among top-ranked keyphrases (Figure 3) as an artifact of masthead repetition; post-hoc filtering is left for future work.

Bibliographical References

Mohammed A. Al Ghamdi. 2022. A novel approach to printed Arabic optical character recognition. *Arabian Journal for Science and Engineering*, 47:2219–2237.

Jawad H Al-Khateeb, Jinchang Ren, Jianmin Jiang, and Husni Al-Muhtaseb. 2011. Offline handwritten Arabic cursive text recognition: A survey. *Pattern Recognition*, 44(3):555–567.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools*, pages 9–15.

Ami Ayalon. 1995. *The Press in the Arab Middle East: A History*. Oxford University Press.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. YAKE! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Amina El Ganadi, Luca Gagliardelli, Sania Aftar, and Federico Ruozzi. 2025. Digital Maktaba project: Proposing a metadata-driven framework for Arabic library digitization. In *Proceedings of the 21st Conference on Information and Research Sciences Connecting to Digital and Library Science (IRCDL 2025)*, volume 3937 of *CEUR Workshop Proceedings*.

Safiullah Faizullah, Muhammad Shahid Ayub, Sajid Hussain, and Muhammad Arif Khan. 2025. Advancements and challenges in Arabic optical character recognition: A comprehensive survey. *ACM Computing Surveys*, 58(4):1–37.

Google Cloud. 2023. Cloud Vision API. https://cloud.google.com/vision.

David Graff and Mohamed Maamouri. 2006. Arabic Gigaword, third edition. Technical report, Linguistic Data Consortium.

Till Grallert. 2021. Catch me if you can! Approaching the Arabic press of the late Ottoman Eastern Mediterranean through digital history. *Geschichte und Gesellschaft*, 47(1):58–89.

Maarten Grootendorst. 2020. KeyBERT: Minimal keyword extraction with BERT. https://github.com/MaartenGr/KeyBERT.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1489–1501.

Jrayed. Jrayed: Arabic newspaper archive of ottoman and mandatory palestine. Accessed: 2026-03-30.

Rashid Khalidi. 1997. *Palestinian Identity: The Construction of Modern National Consciousness*. Columbia University Press.

Thomas Lansdall-Welfare, Saatviga Sudhahar, James Thompson, Justin Lewis, and Nello Cristianini. 2017. Content analysis of 150 years of

British periodicals. *Proceedings of the National Academy of Sciences*, 114(4):E457–E465.

Krystyna K. Matusiak and Qasem Abu Harb. 2011. Digitizing the historical periodical collection at the Al-Aqsa Mosque library in East Jerusalem. In Hartmut Walravens, editor, *Newspapers: Legal Deposit and Research in the Digital Era*, pages 271–290. De Gruyter Saur, Berlin.

Palestine Square. 1932. Caricature: Balfour and the woes his ill-fated promise brought to palestine. Accessed: 2026-03-30.

David A Smith, Ryan Cordell, and Abby Mullen. 2015. Computational methods for uncovering reprinted texts in antebellum newspapers. *American Literary History*, 27(3):E1–E15.

Nabil Wagaa and Hassene Kallel. 2023. Analysis of recent deep learning techniques for Arabic handwritten-text OCR and post-OCR correction. *Applied Sciences*, 13(13):7568.

Wikipedia contributors. 2025a. Al-karmil (newspaper) — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Al-Karmil_(newspaper). [Online; accessed 27-February-2026].

Wikipedia contributors. 2025b. Lisan al arab (newspaper) — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Lisan_al-Arab_(newspaper). [Online; accessed 27-February-2026].

Wikipedia contributors. 2026a. Al-difa' — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Al-Difa%27. [Online; accessed 27-February-2026].

Wikipedia contributors. 2026b. Falastin — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Falastin. [Online; accessed 27-February-2026].

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaudhary, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Nakba Discourse 2025: A Bilingual Social Media Dataset for Collective Trauma Analysis

Wajdi Zaghouani[1], **Mabrouka Bessghaier**[1], **Kais Attia**[2]
[1]Northwestern University in Qatar
[2] Independent Researcher, Qatar
wajdi.zaghouani@northwestern.edu, mabrouka.bessghaier@northwestern.edu
kais.attia.w@gmail.com

Abstract

We introduce Nakba Discourse 2025, a bilingual full-year social media dataset capturing Arabic and English discourse about the 1948 Palestinian Nakba across Twitter/X and Facebook from January to December 2025. The corpus contains 70,312 unique posts organized into intersecting sub-corpora by language, sentiment, gender, geography, and platform, with engagement metadata and automatically extracted rhetorical features. Analyses reveal systematic variation in engagement and framing across communities. Per-post engagement is highest in Israel and UK subsets (50.62 and 49.08 average likes respectively), while Arabic-language discourse shows markedly lower per-post engagement. Sentiment distribution is strongly skewed, with negative sentiment posts outnumbering positive ones at an 11:1 ratio (54,424 vs. 4,827 posts). Despite dramatic variation in absolute engagement levels, virality rates remain structurally constant at approximately 10% across all Twitter/X sub-corpora, regardless of language, gender, or geography, pointing to platform-level amplification regularities. Gender analysis reveals that women achieve proportional virality equal to men despite producing roughly one-third the volume of posts. Temporal patterns align with cultural calendars, including Thursday peaks associated with Jumu'ah across Arabic and female subsets, and Sunday peaks in English-language subsets reflecting Western media cycles. The dataset will be released for research use and supports multilingual stance detection, virality modeling, rhetorical analysis, and computational studies of digital political memory.

Keywords: Nakba, social media dataset, political memory, rhetorical framing, virality, cross-lingual NLP, digital humanities

1. Introduction

The Nakba (النكبة), meaning "catastrophe" in Arabic, refers to the mass displacement of approximately 750,000 to 800,000 Palestinians during the 1948 Arab-Israeli war, resulting in the destruction of over 531 villages and the transformation of Palestinian society (Khalidi, 2020; Sa'di and Abu-Lughod, 2007; Pappé, 2006). This foundational trauma of Palestinian collective identity is transmitted intergenerationally through oral histories, literature, and increasingly through digital platforms. By 2025, marking the 77th anniversary commemorated under the hashtag #Nakba77, social media had become the primary arena for commemoration, contestation, and counter-narrative resistance amid continued regional violence.

Digital rights organizations documented intense platform dynamics throughout 2025, including over 15 million posts containing violent content directed at Palestinians (7amleh, 2025), systematic content moderation biases affecting Arabic-language political content (Sada Social, 2025), and algorithmic demotion of pro-Palestinian material despite its numerical dominance (Corsi, 2023). Yet computational linguistics has produced few dedicated resources for studying Nakba discourse. Existing datasets typically focus on acute escalation periods rather than sustained memory work, remain limited to single languages or narrow time windows, or lack the engagement metadata necessary for studying virality and amplification dynamics.

We construct the Nakba Discourse 2025 dataset to address these gaps, contributing five key innovations. First, the dataset provides longitudinal full-year coverage from January through December 2025, enabling analysis of sustained memory maintenance rather than crisis-driven spikes. Second, the bilingual design captures both Arabic (28,031 posts) and English (17,200 posts) discourse, allowing cross-lingual comparison of framing strategies. Third, the intersecting segmentation structure organizes posts simultaneously by language, gender, geography, sentiment, and platform, supporting multidimensional analysis. Fourth, rich metadata including engagement counts, virality indicators, and eighteen rhetorical features enables quantitative analysis of amplification dynamics. Fifth, the ethically gated release follows established frameworks from (Bender and Friedman, 2018) and (Gebru et al., 2021) for responsible data sharing.

The dataset supports multiple NLP tasks including stance classification, virality prediction, cross-lingual transfer learning, code-switching detection, temporal pattern mining, and propaganda detection. We organize our investigation around three

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 33–42
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

research questions:

RQ1: How do language, gender, geography, platform, and sentiment jointly shape engagement and virality patterns in historical trauma discourse?

RQ2: What distinct rhetorical registers emerge across demographic and linguistic subsets, and how do these reflect different cultural approaches to memory maintenance, advocacy, and identity affirmation?

RQ3: How do cultural and religious calendars structure temporal engagement patterns, and how do counter-narratives transcend linguistic and geographic boundaries?

The paper proceeds as follows: Section 2 reviews related computational work on Israel-Palestine discourse, Arabic NLP resources, and virality research; Section 3 details data collection procedures and corpus composition; Section 4 describes the analysis methodology and processing pipeline; Section 5 presents detailed findings on engagement, rhetorical registers, temporal patterns, and keyword themes; Section 6 discusses theoretical and practical implications; Section 7 addresses limitations and ethical considerations; Section 8 concludes with directions for future work.

2. Related Work

Computational analyses of Israel-Palestine social media discourse have largely focused on acute escalation periods rather than sustained engagement with historical memory. (Shestakov and Zaghouani, 2024) analyzed 370,747 tweets from the 2021 Sheikh Jarrah crisis, highlighting volume asymmetries and weak correlation between hashtag use and virality. Platform-specific studies extend this work: (Liyih et al., 2024) examined YouTube comments during the 2023–2024 Gaza war using neural sentiment models; (Ali, 2025) analyzed Reddit discussions, showing that subreddit norms shape stance expression; (Ng et al., 2024) introduced multimodal analysis through the Love-Hate Dataset; (Antonakaki and Ioannidis, 2025) conducted cross-platform comparisons across Telegram, Reddit, and X; (Nasereddin, 2023) studied differential platform effects on public opinion formation; and (Wang et al., 2025) mapped spatiotemporal sentiment propagation using geotagged tweets. While these studies illuminate platform dynamics and conflict framing, they generally lack longitudinal coverage and engagement metadata centered on Nakba memory discourse.

Recent work has also explored narrative framing and polarization in politically sensitive Arabic discourse through dedicated datasets and shared tasks. In particular, the FIGNEWS shared task (Zaghouani et al., 2024a) introduced a benchmark for analyzing news media narratives, emphasizing the role of framing and bias in shaping public discourse. Complementary efforts in modeling toxic and polarized language, such as multi-label hate speech resources (Zaghouani et al., 2024b), further highlight the interaction between stance, ideology, and linguistic expression in online environments.

Nakba-specific NLP resources have recently expanded through the NakbaNLP initiative (Jarrar et al., 2025). (Ashqar, 2025) demonstrated cultural misalignment in large language model sentiment classification of Nakba oral histories. (AbuHaija et al., 2025) introduced the Nakba Lexicon; (Hamed and Zaidkilani, 2025) developed the NTCC topic classification corpus; (Awad et al., 2025) analyzed narrative cohesion in refugee oral histories; (Nabhani et al., 2025) incorporated argumentation features for propaganda detection; (Garcia-Corral et al., 2025) examined causal attribution in Palestine reporting; (Bilgin Tasdemir and Özateş, 2025) created the NakbaTR NER dataset; and (Mohammed et al., 2025) compared bias detection methods in Israeli-Gaza coverage. These contributions primarily address literary, oral, or news domains and do not provide large-scale social media engagement analysis.

Broader Arabic NLP benchmarks inform our methodological choices. Foundational sentiment datasets include ASTD (Nabil et al., 2015) and ATSAD, while (Chowdhury et al., 2020) and (Mulki et al., 2019) provide offensive language resources. The Curras corpus (Jarrar et al., 2017) supports Palestinian Arabic processing, and (Entman, 1993) framing framework guides rhetorical feature interpretation.

Research on virality and amplification contextualizes engagement dynamics. Emotional arousal predicts sharing behavior (Berger and Milkman, 2012), and algorithmic amplification studies demonstrate systematic visibility concentration in political discourse (Huszár et al., 2022; Ye et al., 2025; Corsi, 2023). (Aboubakr, 2025) conceptualizes digital memory activism as both archival and contestatory, a duality reflected in Nakba discourse online.

No existing resource combines full-year longitudinal coverage, bilingual Arabic-English scope, multi-platform data, demographic segmentation, and engagement metadata for Nakba discourse. The dataset introduced here addresses this gap.

3. Dataset Collection and Composition

3.1. Data Sources and Collection Procedure

Posts were harvested throughout calendar year 2025, from January 1 through December 31, using two commercial social media analytics platforms. Twitter/X data were collected via Meltwater's social media analytics platform, which provides API access to public posts along with demographic annotations derived from profile analysis. Facebook data were collected via CrowdTangle, Meta's research tool for tracking public page content. Both platforms provide engagement metrics including likes, comments, and shares for each post.

Keyword queries were designed to capture posts explicitly engaging with Nakba themes while minimizing false positives from unrelated uses of component terms. English keywords included: "Nakba", "1948 catastrophe", "Right of Return", "Nakba77", "Palestinian displacement 1948", "Nakba Day", and "Nakba commemoration". Arabic keywords included: نكبة (Nakba), نكبة 1948 (Nakba 1948), حق العودة (Right of Return), نكبة فلسطين (Palestine Nakba), ذكرى النكبة (Nakba anniversary), التهجير الفلسطيني (Palestinian displacement), and نكبة 77 (Nakba 77). The complete keyword list with morphological variants is provided in supplementary materials.

3.2. Schema Normalization

Because Meltwater and CrowdTangle use different column naming conventions and data formats, we developed a SchemaAdapter class to normalize exports into a unified schema. The adapter handles three key normalization tasks.

For message content, Meltwater exports use columns named "Opening Text", "Hit Sentence", or "Title" depending on content type, which the adapter maps to a unified "Message" field. CrowdTangle exports use "Message" natively. For timestamps, Meltwater provides separate "Date" and "Time" columns that the adapter combines into a single datetime object, while CrowdTangle provides "Post Created Date" directly. For engagement metrics, Meltwater uses "Replies" and "Reposts" which the adapter maps to "Comments" and "Shares" respectively, while CrowdTangle uses these terms natively.

Total Interactions was computed as the sum of Likes, Comments, and Shares, with Facebook posts additionally including reaction counts (Love, Wow, Haha, Sad, Angry, Care) where available.

The adapter also handles file encoding automatically, detecting UTF-8, UTF-16, and UTF-8 with BOM through byte-order mark inspection, and sniffing delimiters (comma, semicolon, tab, pipe) to accommodate varying export formats.

3.3. Data Cleaning and Deduplication

Language detection combined the langdetect and fastText libraries, requiring greater than 95% confidence for language assignment. Posts failing this threshold were excluded from language-specific sub-corpora but retained in sentiment and platform analyses where language was not a defining criterion.

Deduplication used MinHash locality-sensitive hashing with Levenshtein distance threshold of 5 characters to identify near-duplicate posts while preserving legitimate similar content. This approach removed spam, bot-generated repetitive content, and cross-posted duplicates while retaining organic variations on common themes.

The raw harvest exceeded 142,000 posts. After language filtering, deduplication, and removal of deleted content, 70,312 unique posts remained for analysis.

3.4. Sub-corpora Structure and Overlap

The dataset is organized into twelve intersecting sub-corpora defined by five dimensions: sentiment (Positive, Negative), language (Arabic, English), gender (Male, Female), geography (USA, UK, Israel), and platform (Twitter/X, Facebook). However, not all subsets include segmentation across every dimension. Sentiment labels were provided by Meltwater's automated sentiment classification pipeline. Meltwater's system supports sentence-level sentiment classification across 16 languages, aggregating these into document-level labels of positive, negative, or neutral. The pipeline incorporates deep learning models and accounts for hashtags, emojis, and emoticons as part of its classification input. A continuous feedback loop is maintained to improve model performance over time (**?**). As with any automated pipeline, these labels come with some caveats worth keeping in mind. The deep learning models report sentence-level accuracy of 83% for English, and performance figures for Arabic are not published, meaning classification of Arabic political content that blends grief with resilience or draws on culturally specific affect may not always align with human judgment. These are common constraints of large-scale automated annotation, and findings involving sentiment labels are interpreted accordingly throughout the paper. Geographic labels were inferred from user-declared location fields in account profiles.

Critically, these sub-corpora are not mutually exclusive. A single post may appear in multiple subsets simultaneously. For example, a post by a fe-

male US-based English-speaking author expressing negative sentiment would appear in the Female English, USA, Negative Sentiment, and All English sub-corpora. All statistics should therefore be interpreted as characterizing overlapping populations rather than independent samples, and aggregate totals across sub-corpora must not be summed.

The cross-platform design is intentionally asymmetric: English content derives primarily from Twitter/X, while Facebook content is exclusively Arabic. This asymmetry reflects actual platform usage patterns in our target populations but limits direct bilingual cross-platform comparisons.

4. Methodology

Our analytical framework proceeds in four stages: platform export normalization into a unified schema, engagement metric computation, rhetorical feature extraction at the message level, and bilingual keyword processing. Message-level features include character length, hashtag count, question and exclamation counts, and URL presence.

Platform-specific exports are normalized into a unified schema. Engagement fields (likes, comments, shares, reactions) are harmonized and safely converted to numeric values. Message-level features include character length, hashtag count, question and exclamation counts, and URL presence detected via regex. Message length is additionally grouped into categorical bins (0-100, 100-300, 300-500, 500-1000, 1000+ characters). For Facebook data, engagement rate and basic performance statistics are computed when follower counts are available.

4.1. Engagement and Virality Metrics

The share-to-like ratio captures redistribution behavior relative to passive engagement, calculated as Shares divided by Likes for posts with at least one like, and zero otherwise.

Virality is defined using the 90th percentile of Total Interactions within each sub-corpus. Posts exceeding this threshold are classified as viral:

```
threshold = df['Total
    Interactions'].quantile(0.90)
df['Is_Viral'] = df['Total
    Interactions'] > threshold
```

This relative definition yields approximately 10% viral posts per subset, enabling cross-community comparison while preserving baseline differences in engagement levels. We report both viral percentage and threshold values. A continuous Virality Score is also computed as Total Interactions divided by median Total Interactions.

4.2. Keyword Extraction Pipeline

Keyword extraction applies bilingual preprocessing to Arabic and English text. URLs, email addresses, and @-mentions are removed via regex, and hashtag symbols are stripped while retaining tag text. English text is lowercased. Arabic text is normalized by mapping alef variants (إأآا) to bare alef (ا), converting alef maqsura (ى) to ya (ي), and removing diacritics.

Arabic stemming uses ultra-light stemming that removes only the definite article ال while preserving suffixes. Selected tokens such as الله (Allah) are excluded from stemming. Stopword filtering uses combined Arabic and English lists.

For each retained stem, we compute raw frequency, document frequency, percentage of total tokens, and percentage of posts containing the term. Keywords are ranked by frequency for top-N extraction.

4.3. Temporal Analysis

Temporal analysis aggregates Total Interactions by calendar day using the Post Created Date field (UTC). For each subset, the peak engagement day of week is identified based on total interactions. We note that UTC timestamps may obscure local-time effects in geographically distributed data; timezone-localized analysis is left for future work.

5. Analysis and Findings

5.1. Engagement Concentration and Elite Dynamics

Per-post engagement metrics reveal strong concentration of influence in small subsets. Israel-located accounts achieve 50.62 average likes and 12.90 shares per post despite comprising only 1,413 posts, the smallest geographic subset. UK follows with 49.08 average likes and 17.14 shares per post. These figures exceed USA (25.41 likes), All English (34.20 likes), and All Arabic (15.06 likes).

High-engagement Israeli posts average 489 likes and 136 shares, nearly double UK's figures, indicating dominance by high-influence accounts such as journalists, politicians, academics, and organizations. Engagement concentration is therefore not simply a function of dataset size but of account reach and follower base.

Arabic-language subsets show markedly lower per-post engagement: Arabic Female dataset averages 6.19 likes and 0.94 shares, while Arabic Male averages 18.89 likes and 4.70 shares. Rather than reflecting reduced relevance, this pattern may indicate lower account follower counts or

Table 1: Dataset summary ordered by post volume. Viral percentage uses relative threshold (top 10% within each subset). Facebook Arabic lacks share data due to platform limitations.

Dataset	Posts	Total Likes	Total Shares	Avg Likes	Avg Shares	Viral %
Negative Sentiment	54,424	1,132,196	282,296	20.80	5.19	9.3%
All Arabic (X/FB)	28,031	422,196	86,186	15.06	3.07	9.4%
Arabic Male	17,329	327,310	81,378	18.89	4.70	9.9%
All English (X)	17,200	588,187	165,276	34.20	9.61	9.6%
Male English	13,180	421,984	124,147	32.02	9.42	9.5%
USA	10,906	277,094	63,369	25.41	5.81	9.6%
Facebook Arabic	5,563	63,079	N/A	11.34	N/A	3.7%
Arabic Female	5,139	31,807	4,808	6.19	0.94	9.9%
UK	4,881	239,568	83,680	49.08	17.14	9.4%
Positive Sentiment	4,827	106,390	25,196	22.04	5.22	9.5%
Female English	4,020	166,203	41,129	41.34	10.23	9.7%
Israel	1,413	71,525	18,227	50.62	12.90	10%

platform reach within Arabic-language subsets, resulting in lower absolute engagement figures.

Facebook Arabic operates under a different engagement structure. With zero recorded shares and 11.34 average likes per post, engagement reflects affirmation without redistribution. This likely reflects platform affordances and privacy norms that limit public redistribution on Facebook.

The share-to-like ratio clarifies redistribution dynamics. Mean ratios range from 0.035 for Arabic Female to 0.073 for UK. Critically, the median share-to-like ratio is 0.000 across every Twitter/X dataset, confirming an extreme long-tail structure in which a small minority of posts drives nearly all sharing activity. Amplification is therefore elite-driven across linguistic and geographic communities.

5.2. Virality as Structural Constant

Viral rates cluster tightly between 9.3% and 10.0% across all Twitter/X subsets regardless of language, gender, geography, or sentiment. Despite substantial variation in absolute engagement levels, the proportion of posts exceeding the 90th percentile threshold remains nearly invariant.

This proportional stability suggests that virality operates as a structural property of the platform environment rather than as a community-specific feature. While network size and account influence determine absolute thresholds, the distributional shape of attention appears normalized. Communities differ in magnitude, but not in the statistical structure of amplification.

However, viral thresholds vary significantly. Israel requires 18 interactions to reach viral status, All English and USA require 10, and All Arabic requires 6. Thus proportional virality is constant, but the baseline for success differs across engagement ecologies.

Facebook Arabic is the sole exception, with 3.7% viral posts and a threshold of zero. This confirms that non-share-based architectures fundamentally alter virality mechanics.

Table 2: Temporal patterns and viral thresholds. Peak Int. shows total interactions on the peak day. Threshold indicates interactions needed for viral classification.

Dataset	Peak Day	Peak Int.	Threshold
Facebook Arabic	Thursday	21,264	0
Female English	Thursday	75,446	13
Arabic Female	Thursday	12,255	7
USA	Thursday	94,744	10
All English	Sunday	187,345	10
Male English	Sunday	146,741	9
UK	Sunday	112,231	9
Negative Sent.	Sunday	281,643	8
All Arabic	Monday	102,192	6
Arabic Male	Monday	84,749	8
Positive Sent.	Monday	34,495	6
Israel	Wednesday	40,367	18

5.3. Rhetorical Registers and Communication Strategies

Four rhetorical registers emerge.

Declarative/Witness-bearing. Arabic subsets show 0.00 questions per post, low hashtag use (0.15 to 0.19; 0.02 Facebook), limited URL inclusion (6.3% to 7.9%), and shorter messages (107 to 113 characters). This register asserts rather than interrogates. Positive Sentiment shares this declarative profile with a 0.01 question rate, suggesting that affirmation across languages adopts similar assertive framing.

Advocacy/Evidence-driven. USA and Male English display the highest hashtag rates (0.30 to 0.31), high URL inclusion (13.8% to 14.3%), consistent question use (0.10), and longer messages (126 characters). Posts combine discoverability with documentation, positioning Nakba discourse within global accountability frameworks.

Table 3: Message characteristics across datasets. Hash/Post = hashtags per post; URLs = percentage containing URLs; Quest/Post = questions per post; Excl/Post = exclamations per post; Avg Len = average length in characters.

Dataset	Hash/Post	URLs	Quest/Post	Excl/Post	Avg Len
Positive Sentiment	0.48	12.2%	0.01	0.12	100
USA	0.31	13.8%	0.10	0.06	126
Male English	0.30	14.3%	0.10	0.10	126
All English	0.27	14.1%	0.10	0.09	127
UK	0.23	15.3%	0.09	0.15	131
Arabic Male	0.19	7.9%	0.00	0.09	110
Negative Sentiment	0.17	8.6%	0.03	0.09	116
Female English	0.16	13.3%	0.11	0.08	130
All Arabic	0.15	6.3%	0.00	0.08	110
Arabic Female	0.15	7.3%	0.00	0.09	107
Facebook Arabic	0.02	0%	0.00	0.07	113
Israel	0.02	12%	0.11	0.16	122

Emotive/Dialogic. UK and Israel show highest exclamation rates (0.15 and 0.16), high question rates (0.09 to 0.11), and strong URL inclusion (12% to 15.3%). Israel combines high affect with minimal hashtags (0.02), indicating reliance on follower networks rather than algorithmic tagging.

Networked/Commemorative. Positive Sentiment authors use 0.48 hashtags per post and produce the shortest messages (100 characters). Female English shows strong quote-tweet signal indicated by "qt", reflecting amplification-oriented participation rather than standalone authorship.

5.4. Temporal Patterns

Three peak structures appear. Thursday peaks in Facebook Arabic (21,264), Arabic Female (12,255), Female English (75,446), and USA (94,744). Sunday peaks in All English (187,345), Male English (146,741), UK (112,231), and Negative Sentiment (281,643). Monday peaks for Arabic Male and All Arabic; Israel peaks Wednesday.

These temporal regularities indicate that digital Nakba discourse aligns with broader social rhythms. Thursday concentration among Arabic and female subsets corresponds to the eve of Jumu'ah, when communal and religious engagement intensifies. Sunday concentration among English and male subsets reflects Western media cycles. Israel's Wednesday peak mirrors the Israeli work week. Online amplification therefore follows offline temporal structures rather than operating autonomously.

5.5. Keyword Themes and Counter-Narrative Dynamics

"Nakba" appears in 52% to 66% of posts, highest in Arabic (65% to 66%) and Facebook Arabic (66%), confirming strong thematic cohesion.

"Genocide" appears in 6.2% to 8.5% of English posts, highest in USA (8.5%), indicating legal-accountability framing. "Gaza" appears in Arabic Male (7.25%) and UK (7.68%). "War" peaks in Israel at 13.4%.

"الله" appears in Arabic Female (6.4%), Arabic Male (7.3%), Facebook Arabic (15.5%), Negative Sentiment (5.65%), and Positive Sentiment (8.6%), and is absent from English top keywords, marking a structural cultural-linguistic divide.

"History" appears only in UK (7.2%). "Jews" appears only in Israel (11.6%), alongside "Arabs" and "Arab" (10.2% to 12.7%), reflecting a distinct ethno-demographic framing.

Figure 1 visually reinforces these divergences.

5.6. Gender Patterns

Male-authored datasets are approximately three times larger. Viral rates show parity: Arabic Female and Arabic Male both 9.9% (thresholds 7 and 8); Female English 9.7%; Male English 9.5%. Women achieve proportional virality equal to men despite lower volume.

Arabic Female posts are most declarative (0.00 questions; 7.3% URLs). Male English shows highest hashtag use (0.30). Female English achieves higher shares-per-post (10.23 vs 9.42), indicating strong redistribution despite lower volume. Female subsets peak Thursday; Arabic Male Monday; Male English Sunday.

5.7. Sentiment Distribution

Negative Sentiment contains 54,424 posts; Positive Sentiment 4,827 (11:1). Both show viral rates between 9.3% and 9.5%.

Positive posts are shorter (100 characters) and use 0.48 hashtags, suggesting condensed solidarity signaling. Negative posts are longer (116 char-

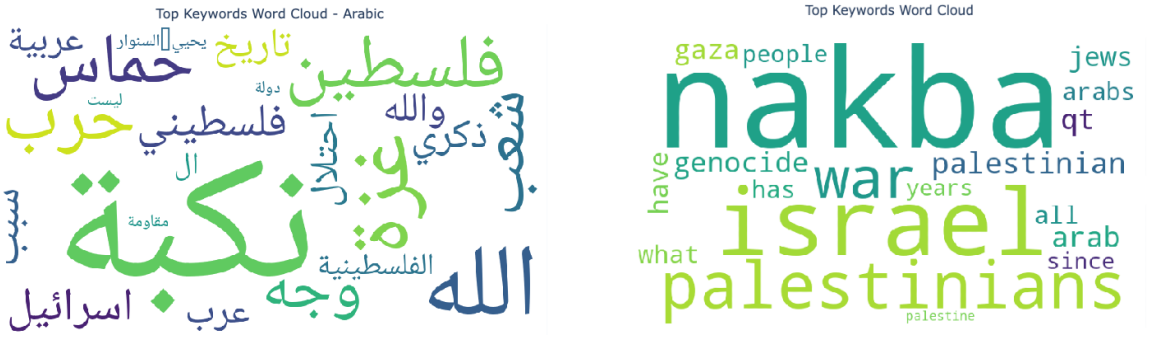


Figure 1: Bilingual keyword word clouds for Arabic (left) and English (right) Nakba discourse. Word size proportional to corpus-level frequency after normalization, light stemming, and stopword removal.

Table 4: Language baseline comparison showing systematic divergence across metrics except viral rate.

Metric	All Arabic	All English
Total Posts	28,031	17,200
Avg Likes/Post	15.06	34.20
Avg Shares/Post	3.07	9.61
Share-to-Like Ratio	0.037	0.052
Viral Rate	9.4%	9.6%
Viral Threshold	6	10
Peak Day	Monday	Sunday
Avg Message Length	110 chars	127 chars
Questions/Post	0.00	0.10
Hashtags/Post	0.15	0.27
URL Inclusion	6.3%	14.1%
Nakba-term %	65.9%	52.6%
Viral Posts (n)	2,647	1,648

acters) and use 0.17 hashtags, reflecting extended articulation of grievances.

5.8. Language Baseline Comparison

English content is longer, more URL-rich, more hashtag-heavy, more questioning, and more shareable. Arabic shows higher Nakba-term penetration and more declarative style. Both achieve near-identical viral rates, indicating shared structural amplification conditions despite divergent rhetorical norms.

"حماس" appears at 3.9% in All Arabic; "qt" appears at 8.1% in All English, reflecting networked repost culture in English-language Twitter.

6. Discussion

6.1. Theoretical Implications

The findings advance understanding of digital political memory in several ways. The near-universal viral rate of approximately 10% across Twitter/X datasets, despite dramatic variation in absolute engagement, community size, and rhetorical style, suggests that platform mechanics impose structural regularities on content spread independent of content characteristics or community identity. This "shape invariance" may reflect algorithmic amplification operating uniformly while absolute reach varies with account size and follower base.

The four-register rhetorical typology maps onto distinct cultural and strategic orientations toward memory work. Arabic discourse performs what (Halbwachs, 1992) termed memory maintenance through declarative witness-bearing that affirms identity without seeking external validation. English discourse pursues global legitimacy through advocacy and evidence, characteristic of diaspora communities seeking international recognition. UK and Israel discourse engages in confrontational dialogue reflecting colonial and national historical positions. Positive and female discourse builds solidarity through networked commemoration that amplifies community voices.

Moreover, Arabic speakers engage with and rebut English-language denial not through translation but through bilingual code-switching that incorporates the denial phrase itself. This finding supports (Aboubakr, 2025) theorization of digital memory activism as simultaneously archival and contestatory.

The cultural calendar findings reveal that digital Nakba commemoration is embedded in broader rhythms of religious and social life. Thursday/Jumu'ah peaks for Arabic and female subsets, Sunday peaks for English and male subsets, and the Israeli Wednesday peak all demonstrate that online discourse follows offline temporal structures rather than operating as a separate digital sphere.

6.2. Practical Implications

For NLP researchers, the dataset enables several benchmark tasks. Stance detection can leverage demographic segmentation to examine how language, gender, and geography predict framing.

Virality prediction can incorporate rhetorical features as predictors alongside engagement history. Cross-lingual transfer learning can exploit the parallel Arabic-English coverage to develop models that generalize across languages.

For platform governance, the finding that elite accounts in small subsets (Israel, UK) drive disproportionate engagement aligns with amplification audit findings (Ye et al., 2025; Corsi, 2023) and suggests that visibility inequality may be particularly acute for politically contested topics. Combined with documented content moderation disparities affecting Arabic and pro-Palestinian content (7amleh, 2025), these patterns warrant attention to equity in algorithmic amplification.

For practitioners in digital humanities and memory studies, the temporal calendar findings suggest that commemoration campaigns should align with community rhythms. Thursday posting may optimize reach for Arabic and female audiences, while Sunday posting may optimize for English and male audiences.

7. Dataset Availability

Access to the dataset requires completion of a request form[1]. Access will be granted to researchers affiliated with recognized academic or research institutions upon submission of a request outlining the intended use of the data. To ensure responsible use, applicants will be required to agree to terms that prohibit redistribution, commercial exploitation, or attempts to re-identify individual users. The dataset will be distributed in a format consistent with platform policies and applicable data protection regulations. Information about access procedures and request forms will be made available in the camera ready version.

8. Conclusion

The Nakba Discourse 2025 dataset provides an unprecedented resource for studying digital political memory at the intersection of NLP, computational social science, and memory studies. Our contributions include the first full-year, bilingual, multi-platform corpus of Nakba discourse with engagement metadata; a four-type rhetorical typology linking linguistic features to distinct memory practices; documentation of structural viral regularities alongside dramatic engagement inequality; evidence for cultural calendar effects structuring temporal patterns; and analysis of pan-linguistic counter-narrative dynamics.

Future work will extend this foundation in several directions: fine-grained emotion labels beyond binary sentiment to capture the complex affective landscape of trauma discourse; multimodal extensions incorporating images and videos; Hebrew-language data enabling trilingual analysis of Israeli discourse; timezone-localized temporal analysis; and longitudinal tracking as the corpus extends into future years.

We invite the research community to build upon this resource for more equitable, culturally-informed multilingual NLP in conflict and memory contexts.

9. Limitations

Several limitations constrain interpretation of our findings. Gender and sentiment classifications were provided by Meltwater's automated pipelines rather than human annotation. Binary gender classification does not capture non-binary identities and may misgender individuals based on name-based inference. Sentiment classification performance on Arabic political content, particularly expressions mixing grief with resilience, has not been independently validated on our corpus.

Geographic labels derive from user-declared location fields, which may reflect ancestral, aspirational, or privacy-protective declarations rather than actual residence. The coverage and accuracy of geographic inference have not been systematically validated. A further limitation is the geographic segmentation of Arabic-language posts, which are not subdivided by Arab country of origin. The current geographic sub-corpora — USA, UK, and Israel — reflect the availability of English-language geographic signals and do not capture the diversity of Arabic-speaking communities across countries such as Jordan, Lebanon, Palestine, Egypt, and the Gulf states. This limits the extent to which regional variation in Arabic Nakba discourse can be examined. Future work should incorporate country-level segmentation within Arabic sub-corpora, enabling comparative analysis of how Nakba memory is constructed and contested across different Arab national contexts, each carrying distinct historical relationships to 1948.

Platform moderation may systematically undersample certain voices. Digital rights organizations have documented bias against Arabic and pro-Palestinian content in content moderation systems. Our findings therefore characterize visible discourse rather than the full population of Nakba-related posting.

The corpus excludes Hebrew, limiting analysis of Israeli counter-narrative dynamics from Hebrew-language sources. The cross-platform design is asymmetric, with English primarily from Twitter/X and Facebook exclusively Arabic, constrain-

[1] https://forms.gle/W7xpLt7io326bR3A6

ing bilingual cross-platform comparisons.

The relative viral threshold (top 10% within each subset) yields approximately 10% viral classification by construction, limiting cross-subset comparison of absolute viral reach. UTC timestamps may introduce artifacts in temporal analysis for a geographically distributed corpus.

10. Ethical Considerations

Our ethical approach follows frameworks established by Bender & Friedman (2018) for data statements and Gebru et al. (2021) for dataset documentation. Usernames are redacted from released data. Geographic information is coarsened to country level. Profile images are excluded.

For Twitter/X content, we release tweet IDs only in compliance with platform terms of service. Researchers may rehydrate full text via API access subject to their own compliance obligations. For Facebook content, we release anonymized feature vectors rather than raw text, with full text available only through gated access requiring researcher verification.

Dataset access is restricted to verified researchers through an application form. Applicants must describe research purpose, institutional affiliation, and data handling protocols. Commercial use is prohibited.

We acknowledge dual-use risks including potential misuse for targeted harassment, surveillance, or propaganda. Access restrictions and anonymization mitigate but do not eliminate these risks. We encourage responsible use and will monitor for misuse.

The study received IRB-equivalent ethics review and exemption determination based on analysis of publicly available data with anonymization protections.

Acknowledgments

This work was made possible by the National Priorities Research Program (NPRP) grant NPRP14C-0916-210015 from the Qatar National Research Fund (QNRF), a member of the Qatar Research, Development and Innovation Council (QRDI)

References

7amleh. (2025). Racism and Incitement Index 2024. 7amleh – The Arab Center for the Advancement of Social Media.

Aboubakr, F. (2025). Archivalism and memory activism: The Nakba (1948) and the Gaza War (2023). *Memory Studies*, 18(2):439–455.

Ali, H. J., Abrar, A., Hossain, S. M. H., and Mridha, M. F. (2025). Social media polarization during conflict: Insights from an ideological stance dataset on Israel-Palestine Reddit comments. *arXiv:2502.00414*.

Antonakaki, D. and Ioannidis, S. (2025). Cross-platform digital discourse analysis of the Israel-Hamas conflict. *arXiv:2601.02367*.

Ashqar, H. I. (2025). Sentiment Analysis of Nakba Oral Histories: A Critical Study of Large Language Models. In *Proceedings of NakbaNLP 2025*, pages 30–36, Abu Dhabi. Association for Computational Linguistics.

AbuHaija, I., Al Mandhari, S., El-Haj, M., Sibony, J., and Rayson, P. (2025). The Nakba Lexicon: Building a Comprehensive Dataset from Palestinian Literature. In *Proceedings of NakbaNLP 2025*, pages 37–47, Abu Dhabi. Association for Computational Linguistics.

Awad, G. A., Rayan, T. N., Dunagan, L., and Gamba, D. (2025). Collective Memory and Narrative Cohesion: A Computational Study of Palestinian Refugee Oral Histories in Lebanon. In *Proceedings of NakbaNLP 2025*, pages 83–102, Abu Dhabi. Association for Computational Linguistics.

Bender, E. M. and Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Berger, J. and Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, 49(2):192–205.

Bilgin Tasdemir, E. F. and Özateş, Ş. B. (2025). NakbaTR: A Turkish NER Dataset for Nakba Narratives. In *Proceedings of NakbaNLP 2025*, pages 122–126, Abu Dhabi. Association for Computational Linguistics.

Chowdhury, S. A., et al. (2020). A multi-platform Arabic news comment dataset for offensive language detection. In *Proceedings of LREC 2020*.

Corsi, G. (2023). Visibility and amplification on X after policy changes: An audit study. *Proceedings of ICWSM 2023*.

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Garcia-Corral, P., et al. (2025). The missing cause: Causal attribution analysis in Palestine reporting. In *Proceedings of NakbaNLP 2025*.

Gebru, T., et al. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Halbwachs, M. (1992). *On Collective Memory*. University of Chicago Press.

Hamed, O. and Zaidkilani, N. (2025). Arabic Topic Classification Corpus of the Nakba Short Stories. In *Proceedings of NakbaNLP 2025*, pages 48–55, Abu Dhabi. Association for Computational Linguistics.

Huszár, F., et al. (2022). Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences*, 119(1).

Jarrar, M., Habash, N., El-Haj, M., Haddad, A. H., Jallad, Z., Mansour, C., Allan, D., Rayson, P., Hammouda, T., and Malaysha, S. (eds.) (2025). *Proceedings of the First International Workshop on Nakba Narratives as Language Resources (NakbaNLP 2025)*, Abu Dhabi. Association for Computational Linguistics.

Jarrar, M., Habash, N., Alrimawi, F., Akra, D., and Zalmout, N. (2017). Curras: An annotated corpus for the Palestinian Arabic dialect. *Language Resources and Evaluation*, 51(3):745–775.

Khalidi, R. (2020). *The Hundred Years' War on Palestine*. Metropolitan Books.

Liyih, A., Anagaw, S., Yibeyin, M., and Tehone, Y. (2024). Sentiment analysis of the Hamas-Israel war on YouTube comments using deep learning. *Scientific Reports*, 14:13647.

Mohammed, M. Y., et al. (2025). Bias detection in media coverage of the Israeli-Gaza conflict: Traditional vs. transformer models. In *Proceedings of NakbaNLP 2025*.

Mulki, H., et al. (2019). L-HSAB: A Levantine Twitter dataset for hate speech and abusive language. In *Proceedings of RANLP 2019*.

Nabhani, S., Borg, C., Micallef, K., and Al-Khatib, K. (2025). Integrating Argumentation Features for Enhanced Propaganda Detection in Arabic Narratives on the Israeli War on Gaza. In *Proceedings of NakbaNLP 2025*, pages 127–149, Abu Dhabi. Association for Computational Linguistics.

Nabil, M., et al. (2015). ASTD: Arabic sentiment tweets dataset. In *Proceedings of EMNLP 2015*.

Nasereddin, S. (2023). Impact of social media platforms on international public opinion during the 2023 Israel-Gaza war. *Journal of International Communication*.

Ng, L. H. X., Lim, A. X. W., and Lee, R. K.-W. (2024). Love-Hate Dataset: A Multi-Modal Multi-Platform Dataset Depicting Emotions in the 2023 Israel-Hamas War. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion)*, pages 1807–1815. ACM.

Pappé, I. (2006). *The Ethnic Cleansing of Palestine*. Oneworld Publications.

Sa'di, A. H. and Abu-Lughod, L. (eds.) (2007). *Nakba: Palestine, 1948, and the Claims of Memory*. Columbia University Press.

Sada Social. (2025). The Digital Rights Index 2024. Sada Social Center for Palestinian Digital Rights.

Shestakov, A. and Zaghouani, W. (2024). Analyzing Conflict Through Data: A Dataset on the Digital Framing of Sheikh Jarrah Evictions. In *Proceedings of the Second Workshop on Natural Language Processing for Political Sciences @ LREC-COLING 2024*, pages 55–67, Torino, Italia. ELRA and ICCL.

Wang, D., et al. (2025). Spatiotemporal sentiment propagation analysis of the Israeli-Palestinian conflict. *Humanities & Social Sciences Communications*.

Ye, J., et al. (2025). Algorithmic amplification audits on X.

Zaghouani, W., Jarrar, M., Habash, N., Bouamor, H., Zitouni, I., Diab, M., El-Beltagy, S., & AbuOdeh, M. (2024). The FIGNEWS shared task on news media narratives.

Zaghouani, W., Mubarak, H., & Biswas, M. R. (2024). So hateful! Building a multi-label hate speech annotated Arabic dataset.

ChronoLearn: A GRAG LLM-Based System for Structuring and Exploring Historical Narratives

Mohammad O. ALADDASI, Shahd Abu Hijleh, and Omar Qawasmeh
Princess Sumaya University for Technology
Department of Data Science & Artificial Intelligence, Amman,Jordan
mohammadoaddasi@gmail.com, sha20210504@std.psut.edu.jo, o.alqawasmeh@psut.edu.jo

Abstract

ChronoLearn is a KG–LLM framework to structure and explore user-fed Arabic historical narratives. Traditional history learning methods often rely on fragmented, text-heavy resources that lack contextualisation, interactivity, and clear relational structure. Chronolearn transforms unstructured texts into knowledge graphs KG using an ETL-based Natural Language Processing (NLP) pipeline for entity and relation extraction, followed by schema-guided graph construction. The system integrates KG retrieval with Large Language Model (LLM) generation (GRAG) to produce grounded and explainable narratives. The approach is evaluated in heterogeneous Palestinian and Jordanian sources, including Nakba-related content, using both quantitative metrics and comparative analysis. Additionally, an interactive platform is provided to support the exploration of structured historical knowledge through graph visualisations and narrative storylines, enabling users to explore historical events in a coherent, interactive and interpretable manner. The results demonstrate improved factual grounding and structured reasoning, addressing limitations of text-only approaches in the processing of historical knowledge in Arabic Language

Keywords: Natural Language Processing, Knowledge Graphs, LLMs

1. Introduction

Recent advances in natural language processing (NLP) have brought the importance of treating historical and cultural narratives into attention particularly those related to the Nakba. It facilitated historical sources to be analysed and explored computationally in structured language form. The Nakba, a major historical event with an ongoing impact, meaning catastrophe in Arabic, "refers to the eviction of nearly three-quarters of a million Palestinians from their homes and their transformation into refugees, as well as the destruction of more than four hundred villages and towns in what became the state of Israel, and the erasure of the name Palestine from the map" Interactive Encyclopedia of the Palestine Question (PalQuest) (2022). Historical knowledge consists of interconnected components, including events (e.g., wars and treaties), people and groups (leaders and communities), places (cities and historical sites), and sources (archives, letters, and cultural materials). These elements form a complex relational structure rather than isolated facts.

1.1. Project Background and Motivation

History by its nature is a time-oriented subject with chronology being a major part of its overall structure. It is also a data-rich field, with a number of interconnected sources and references that connect events, people, and places together. However, the teaching of history rarely reflects this, as learning experiences lack strong storytelling, visualisation, and interactivity. This results in a disjointed history experience. Modern learners increasingly expect educational tools to be visual, interactive, and efficient to explore. Despite this, history as a discipline has been slower to adopt technologies such as data visualisation, semantic analysis, and knowledge graphs (KGs).

While artificial intelligence (AI) has introduced new paradigms in learning and information access, history education continues to face a significant technological gap. Interactive teaching methods have been shown to improve engagement; however, most historical content remains presented in static formats. A key trend is the use of Large Language Models (LLMs) through question-answering (Q&A) interfaces. Nevertheless, these systems present several limitations, including lack of source attribution and explainability, susceptibility to hallucinations or inaccurate information, and an over-reliance on text-based interaction. As a result, they provide limited support for structured and visual exploration of historical knowledge.

In this paper, ChronoLearn is proposed as a system to transform user-fed Nakba-related narratives into structured KGs using NLP techniques. A Graph Retrieval-Augmented Generation (GRAG)-based framework is introduced to integrate KGs with LLMs for grounded querying and narrative generation. The system leverages diverse Arabic historical sources to model event-based, cultural, and memory-driven narratives. Furthermore, an interactive platform is provided to support the exploration of structured historical knowledge through visualisation and AI-driven storytelling.

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 43–49
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

1.2. Problem Statement and Research Objectives

Digital historical information is distributed across multiple sources and formats, making it difficult for users to access structured, reliable, and meaningful knowledge. While the results obtained from traditional search engines are generally broad and vague, the results of new chat-based AI tools are unverified, decontextualised, and carry bias.

The problem is more critical for the Arabic historical narratives, especially those concerning the Palestinian and Jordanian contexts, since the resources for the NLP tools are scarce.

In this research, we seek to answer the research question "How can the integration of KGs with LLMs facilitate the structured representation of Nakba-related historical narratives, compared to traditional approaches?"

2. Related Work

The integration of KGs and LLMs has become an important direction in recent research. Researchers use this combination to build AI systems that are more explainable and interactive. In knowledge-intensive domains such as history, structured representations help organise information and support clear reasoning. KGs represent entities and relationships in a structured form, but LLMs generate free-text responses. KGS and LLMS combination creates systems that connect structured knowledge with semantic understanding.

This section reviews existing research across three main areas: (1) unstructured text-to-KG conversion, (2) KG-enhanced LLM reasoning, and (3) NLP applications for historical and cultural narratives.

2.1. Text-to-Knowledge Graph Conversion

Transforming unstructured text into structured KGs is a key step in building explainable and intelligent systems. This process uses NLP techniques such as Named Entity Recognition (NER) and Relation Extraction (RE). These methods identify important entities in the text and define the relationships between them. Zhang and Soh proposed the Extract-Define-Canonicalise (EDC) framework. Their approach converts textual data into structured KGs through a three-stage pipeline. The stages include entity and relation extraction, schema alignment, and canonicalisation Zhang and Soh (2024). The framework improves scalability and consistency by using retrieval-based schema linking. This allows efficient processing of large and diverse datasets. Similarly, Mohanty introduced EduEmbedd, a framework designed for representing educational content Mohanty (2023). The model captures conceptual entities and pedagogical dependencies. It highlights the value of modelling instructional relationships in addition to factual knowledge.

Wang et al. proposed ChatWeaver, an interactive system that combines human validation (human-in-the-loop) with LLM-based extraction Wang et al. (2025). This design increases reliability by allowing users to review and refine extracted entities and relationships. This level of validation is especially important in history, where contextual accuracy plays a central role. Sun et al. presented a framework that converts historical archives into structured KGs using LLM-based extraction and embedding-based retrieval Sun et al. (2024). Their method includes preprocessing, semantic embedding with FAISS, and LLM-driven relation extraction. The result is a structured and queryable historical knowledge system.

2.2. Knowledge Graph Integration with LLMs

Integrating KGs into LLM pipelines has shown clear improvements in accuracy and explainability. When responses are grounded in structured knowledge, the system produces more reliable and context-aware outputs. This approach also makes the reasoning process easier to trace and understand.

Zhang et al. introduced KnowGPT, a framework that uses KG–based prompting to strengthen reasoning in LLMs Zhang et al. (2024). Their experiments showed better performance compared to baseline models such as GPT-3.5 and GPT-4 across several evaluation benchmarks.

Recent studies in applied domains further highlight the value of combining KGs with retrieval-augmented generation (GRAG). For example, MedSyn integrates KGs with RAG to support question answering over structured medical data. The system achieved stronger factual grounding and reduced hallucination. Akgül et al. explored temporal reasoning over dynamic KGs by incorporating time-series data into graph structures and applying contrastive learning techniques for entity representation Akgül et al. (2025). Their results showed clear improvements in prediction accuracy. However, the framework assumes complete data availability, which may limit its applicability in historical real-world applications.

Yan et al. proposed a KG–guided framework for explainable AI Yan et al. (2025). In their approach, graph relationships validate LLM-generated outputs and assign confidence levels to responses. This design strengthens interpretability and transparency. The evaluation relied mainly on qualitative

feedback, which suggests that future work can include more detailed quantitative validation.

2.3. NLP for Historical and Cultural Narratives

NLP applications in historical and cultural domains focus on structuring and analysing narrative-based data. These approaches support organised representation, interpretation, and exploration of historical content. They also help to create systems that encourage storytelling and interactive learning.

Muralidharan et al. proposed a system that generates multiple-choice questions (MCQs) using KG representations Muralidharan (2024). Their evaluation showed that question generation based on semantic relationships improves learning outcomes and strengthens knowledge retention. This study highlights how structured knowledge can enhance educational applications.

Research Gaps: In summary, after completing the literature review, the following research gaps are identified: First, many systems provide limited support for the Arabic language and regional historical content. Second, current approaches offer few fully automated and scalable pipelines for KG construction. Third, many frameworks still rely on human validation when handling complex or ambiguous information. Fourth, interactive and visual learning tools receive limited attention. Moreover, there is little work focused on transforming Nakba-related narratives into structured and interactive knowledge systems that support computational analysis and exploration. These gaps motivate the development of ChronoLearn. The framework aims to provide an Arabic-focused, interactive, and explainable KG-LLM pipeline that structures and explores Nakba-related narratives in a clear and accessible way.

3. Dataset and Narrative Representation

The dataset is curated to represent diverse forms of Nakba-related narratives. It includes educational texts, cultural heritage descriptions, and scholarly discussions on memory and displacement. Nakba narratives are defined in a broad sense. They cover event-based accounts, cultural artefacts, and memory and post-memory texts. Together, these materials contribute to shaping historical understanding.

The system is evaluated using a heterogeneous dataset that includes three main sources. The first source is an **oral history book**[1], which provides event-based historical content such as treaties, wars, and political developments based on first-hand testimonies Zayed (2013). The second source is an **encyclopedic article**[2], which represents a semi-structured and community-generated text with descriptive cultural content Wikipedia contributors (2026). The third source is a **scholarly study**[3], which presents analytical and narrative-driven perspectives on memory, identity, and lived experience Abu-Lughod (2007).

These sources differ in structure, style, and narrative type. This combination allows evaluation across different forms of historical data. The dataset covers varying levels of structure, from formal historical documentation to narrative and memory-focused texts. In addtion, the dataset reflects three main narrative forms. The first form includes event-based narratives that present structured historical facts such as wars, treaties, and political developments. The second, includes cultural narratives that capture symbolic and identity-based content drawn from songs and folklore. The third narrative includes memory-based, which express lived experiences and inter-generational memory, often found in academic works. This diversity supports the representation of historical knowledge as both factual and interpretive. Such representation is important for modelling different types of narratives.

Processing Arabic historical narratives involves several challenges. Arabic language contains rich structure, varied writing styles, and ambiguity in entity representation. Narrative texts often include implicit relationships and contextual references. These elements make automatic extraction more complex. Nakba-related narratives also carry cultural sensitivity and strong contextual meaning. Much of the meaning appears through memory and interpretation rather than direct factual statements. These challenges support the use of structured knowledge representations that preserve accuracy while maintaining context.

4. Proposed System: ChronoLearn

ChronoLearn is a KG–LLM framework designed to convert unstructured Jordanian and Palestinian historical narratives, especially Nakba-related content, into structured, visual, and interactive knowledge representations. Using a GRAG model, the system combines KGs with LLMs. This integration supports grounded querying, explainable narrative generation, and semantic exploration of historical knowledge. ChronoLearn follows a modular and end-to-end pipeline. Firstly, it ingests user-provided

[1] *Palestinian Refugee Narratives: An Inter-generational Comparison*

[2] *Jafra (song) Arabic Wikipedia Entry*.

[3] *Return to Half-Ruins: Memory, Post-memory, and Living History in Palestine*.

documents. Secondly, it preprocesses and normalises Arabic text. Thirdly, it extracts knowledge triples that reflect key topics and themes. Fourthly, it validates and refines the extracted information. Fifth, it constructs and merges KGs. Finally, it generates multiple outputs, including KG visualisations, RDF exports, and grounded narrative responses.

4.1. System Architecture and KG Construction Pipeline

ChronoLearn follows an ETL-inspired workflow. In the *Extract* phase, documents such as PDFs, webpages, and textual archives are ingested using automated extraction tools, after which the raw content is converted into plain text. The preprocessing stage normalises Arabic orthographic variations, including standardisation of characters (e.g., Alef, Yaa', Ta Marbutah), removal of Tatweel, and optional removal of diacritics (Harakat), improving consistency and downstream NLP performance.

In the *Transform* phase, the system applies LLM-based techniques for semantic analysis. It begins with topic detection, which identifies the main concepts in the document. These topics guide both knowledge extraction and narrative generation by narrowing the semantic focus. The system also performs theme detection. It classifies the narrative type, such as event-based, cultural, or memory-driven. This classification determines schema constraints and shapes the next processing steps.

After semantic analysis, the system extracts knowledge by identifying entities and relationships using LLM-based NER and RE. It structures the extracted information into triples (subject, predicate, object). These triples align with a schema-driven T-Box to maintain consistency in entity types and relationships.

The system then applies a validation and refinement stage to improve reliability. This stage checks schema conformity, removes redundant entries, normalises entity names, and corrects structural inconsistencies in triples. As a result, the KG becomes coherent, accurate, and suitable for further reasoning tasks. During the Load phase, the validated triples form a KG. Entities appear as nodes, and relationships appear as edges. When processing multiple documents, the system merges individual graphs into a unified global KG. It preserves source-level provenance and supports cross-document reasoning and exploration.

On top of the KG, ChronoLearn integrates a GRAG layer. This layer combines graph-based retrieval with LLM-based generation. The system supports two narrative generation strategies. The first strategy follows a topic-guided approach that generates narratives from preprocessed text based on detected topics. The second strategy uses a KG-grounded approach that generates narratives directly from validated triples. This design balances semantic richness with structured factual grounding.

The system also includes an evaluation layer. It assesses generated narratives using both quantitative metrics and qualitative analysis. The evaluation examines semantic similarity, lexical overlap, factual grounding, structural clarity, and completeness. This process enables systematic comparison between generation strategies and ensures reliable outputs.

Finally, ChronoLearn provides an output and interaction layer. It includes KG visualisation for intuitive exploration, AI-generated narratives for storytelling, and RDF export for compatibility with semantic web technologies. Users can explore historical knowledge in a structured, explainable, and visually enriched way. As illustrated in 1, the pipeline moves from ingestion and preprocessing to triple extraction, validation, KG construction, and export. The implementation is available in the project repository[4] upon request, along with all related figures and supplementary materials.

5. Evaluation and Experiments

This section presents the evaluation strategy used to assess system correctness, efficiency, and narrative quality. The evaluation combines quantitative metrics, qualitative analysis, and system-level validation using heterogeneous Arabic historical sources spanning educational, cultural, and scholarly texts. For controlled comparison, each document passes through two processing pipelines:**(i) Method A (LLM + Topics)**; generates narratives guided by extracted themes. **(ii) Method B (KG-based)**; produces outputs grounded in validated KG triples. Both methods use the same inputs to ensure a fair comparison.

The evaluation focuses on three main areas. The first area is data integrity, which includes schema conformity and relational consistency. The second area is performance, which covers latency, scalability, and processing efficiency. The third area is overall system reliability across all integrated components.

The strategy uses multiple semantic, lexical, and structural metrics. These include *Precision@K*, *Recall@K*, and *HITS@K* to measure entity-level accuracy and coverage. *ROUGE-L* evaluates lexical overlap, and *BERTScore* measures semantic similarity. Together, these metrics assess narrative quality from different perspectives and balance factual grounding with semantic alignment.

[4] https://github.com/Mohammad-ALADDASI/ChronoLearn

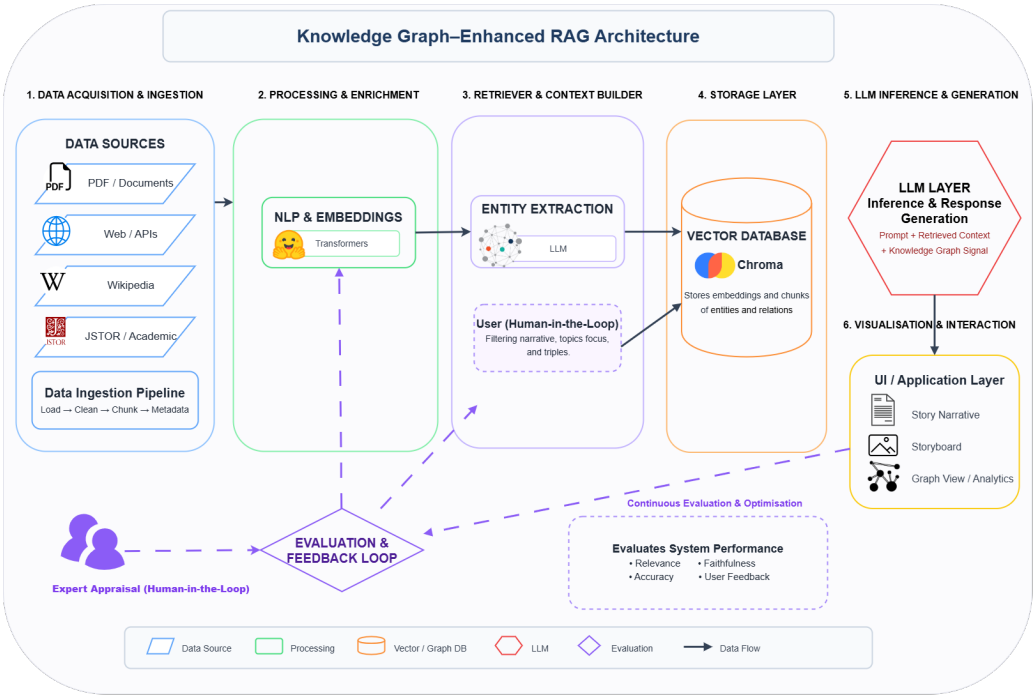


Figure 1: Semantic Knowledge Graph Processing Framework (ChronoLearn).

The evaluation also includes a direct comparison between topic-guided and KG-grounded generation. The system evaluates outputs quantitatively against the source text and qualitatively using an LLM-based evaluator. The qualitative assessment follows four criteria: *topic focus*, *structural clarity*, *factual grounding*, and *completeness*. This comparison highlights the trade-offs between narrative richness and structured accuracy.

User-based and experts' evaluation further complements automated metrics. Participants interact with generated narratives and KG visualisations. They provide feedback on clarity, coherence, factual reliability, and usability. Manual validation is conducted on selected triples and narratives. The team addresses identified issues through iterative refinement of extraction prompts, validation rules, and generation strategies.

5.1. Results and Analysis

Table 1 presents the quantitative comparison between the topic-guided generation approach (Method A) and the KG-based triple grounding approach (Method B).

The results show a consistent and interpretable trade-off between semantic richness and structured factual grounding. Method A achieves higher BERTScore (0.703 vs. 0.691). This indicates stronger contextual and semantic similarity to the reference texts. This suggests that full-text topic conditioning enables the model to generate better outputs that preserve global thematic alignment

Metric	Method A	Method B
BERTScore	0.703	0.691
ROUGE-L	0.097	0.033
Precision@10	0.60	0.80
Precision@5	0.60	0.80
Recall@5	0.130	0.174
HITS@5	1.00	1.00

Table 1: Comparison between Method A (LLM + Topics) and Method B (KG + Triples) generation

coherence. In contrast, Method B outperforms Method A in entity-level precision. It achieves higher Precision@10 (0.80 vs. 0.60) and Precision@5 (0.80 vs. 0.60). This indicates more accurate prioritisation of entities in top-ranked outputs. This reflects the benefits of structured triple grounding in enforcing factual consistency and noise reduction in entity selection.

Method A achieves a higher ROUGE-L score (0.097 vs. 0.033). This indicates stronger lexical overlap and better span-level alignment with the reference content. It suggests that topic-guided generation better preserves surface-level phrasing and narrative continuity, due to the richer contextual input provided to the model.

Method B also demonstrates improved coverage, as reflected in higher Recall@5 (0.174 vs. 0.130). This shows that the triple-based approach retrieves a broader set of relevant entities within the top-K candidates, and overall improving coverage com-

pared to Method A.

Both methods achieve perfect HITS@5 (1.00), confirming that at least one relevant entity consistently appears within the top-ranked outputs. Nevertheless, the relatively low Recall@5 values (0.130 vs. 0.174) suggest that coverage remains limited overall. This indicates that both approaches still miss a significant portion of relevant entities despite improvements.

The pipeline currently reports `linked_named_events: 0`. This component is still under active development and is not yet incorporated as a constraint during generation or evaluation. Importantly, triple validation is conducted independently of event linking. Future integration of event-aware grounding is expected to improve temporal coherence and cross-sentence consistency.

From a qualitative perspective, Method A generates more fluid, coherent, and narratively complete texts, showing smoother transitions and contextual bridging. Method B, on the other hand, produces structurally clearer outputs with stronger factual grounded and fewer theoretical inferences. The comparison highlights a fundamental trade-off between narrative richness and structured accuracy.

Overall, these findings suggest that neither approach dominates across all dimensions. Instead, they expose complementary strengths. The topic-guided generation enhances cohesion and coherence with semantic alignment. Whereas KG-based grounding improves factual precision and entity coverage. These mutual techniques strongly motivates hybrid architectures that combine high-level topic conditioning with structured KG constraints. This enables models to achieve both semantic depth and factual robustness simultaneously.

6. Conclusion

ChronoLearn shows how integrating KGs with LLMs can support structured, interactive, and explainable history learning. By converting unstructured Arabic historical narratives into machine-interpretable knowledge representations, the system supports semantic exploration and grounded narrative generation.

The results indicate that combining graph-based representations with LLMs improves knowledge organisation and strengthens explainability. It also increases user engagement, especially in Arabic historical contexts where structured digital resources remain limited. The comparative evaluation highlights the balance between narrative richness and factual grounding. The findings emphasise the benefit of integrating both approaches within a unified framework.

Future work will extend the system with storyboard generation and advanced narrative construction features to create richer educational experiences. Additional improvements will focus on temporal reasoning, named-event linking, and structured event modelling. The framework will also incorporate human-in-the-loop validation to maintain cultural sensitivity, historical accuracy, and robustness across diverse narrative types.

7. Bibliographical References

Lila Abu-Lughod. 2007. Return to half-ruins: Memory, postmemory, and living history in palestine. In Ahmad H. Sa'di and Lila Abu-Lughod, editors, *Nakba: Palestine, 1948, and the Claims of Memory*, pages 77–104. Columbia University Press, New York, NY, USA.

Ö.F. Akgül et al. 2025. Recipe-tkg: Temporal knowledge graph completion using llms. *arXiv preprint arXiv:2505.17794*.

Interactive Encyclopedia of the Palestine Question (PalQuest). 2022. Meanings of the nakba: Transformations through time and place. Accessed: 2026-02-27.

P. Mohanty. 2023. Eduembedd: An educational knowledge graph embedding framework. In *CEUR Workshop Proceedings*.

D. Muralidharan. 2024. Knowledge graph-based multiple-choice question generation. *arXiv preprint*.

Y. Sun, W. Yang, and Y. Liu. 2024. Knowledge graph construction of oral historical archives using llm-rag. In *ICISDM*.

M. Wang, Y. Gao, M. Fang, and X. Lyu. 2025. Chatweaver: An interactive llm-supported knowledge graph system. In *IEEE AINIT*.

Wikipedia contributors. 2026. Jafra (song). https://ar.wikipedia.org/wiki/ØňÙĄØśØğ_(ØčØžÙĘÙŁØI). Wikipedia, The Free Encyclopedia, Accessed: 27-Feb-2026.

Y. Yan et al. 2025. Knownet: Guided health information seeking via knowledge graph integration. *IEEE TVCG*.

Amal Zayed. 2013. *Palestinian Refugee Narratives: An Inter-generational Comparison*. Ibrahim Abu-Lughod Institute of International Studies, Birzeit University, Birzeit, Palestine.

Q. Zhang, J. Dong, H. Chen, et al. 2024. Knowgpt: Knowledge graph based prompting for large language models. *NeurIPS*.

Z. Zhang and H. Soh. 2024. Extract, define, canonicalize: Towards efficient large-scale knowledge graph construction with llms. *arXiv preprint arXiv:2401.03868*.

Credibility Assessment for Arabic News on the Gaza War: A Hybrid Neural-Symbolic Pipeline

Sanaa Abril, Sihame Mouanid, El Habib Benlahmar, Omar Zahour
Faculty of Science Ben M'Sick, Laboratory of Information Processing and Modeling (LTIM),
Hassan II University, Casablanca, Morocco
{sanaa.abril-etu, sihame.mouanid-etu, elhabib.benlahmar}@etu.univh2c.ma

Abstract

While misinformation has long circulated online, the Gaza conflict has intensified its visibility and spread across news websites, online portals, and social media, complicating the credibility and long-term curation of conflict-related Arabic records, including historical accounts and written testimonies. This work proposes a hybrid framework for Arabic fake news detection that combines interpretable linguistic cues with contextual semantic representations. The approach integrates fuzzy logic-based handcrafted features capturing exaggerated and sensational linguistic patterns, AraBERT contextual embeddings for semantic understanding, and a CNN-based text feature extractor for local textual patterns. These complementary features are combined into a unified representation for downstream classification. Multiple machine learning and deep learning classifiers are evaluated to identify the most effective detection model. The resulting system is deployed as a real-time web browser plugin, enabling users to automatically assess the credibility of Arabic news content during browsing.

Keywords: Arabic fake news detection, News classification, hybrid models, Machine learning, Deep learning, fuzzy logic, browser plugin, misinformation

1. Introduction

Conflict-related Arabic content is produced and consumed at scale across news websites and social media, and it increasingly becomes part of the digital record that researchers, journalists, and archivists consult when documenting events and public discourse. However, contemporary conflict reporting takes place in a high-velocity information environment where emotionally charged framing and coordinated disinformation can blur the line between verified and fabricated claims, affecting public opinion and trust (Institute for Strategic Dialogue, 2025). This makes it difficult to curate reliable collections and to support downstream tasks such as search, retrieval, and analysis of conflict-related content.

This challenge is especially acute in Arabic, where automated credibility assessment is constrained by domain variation and the relative scarcity of standardized evaluation resources (Yousef et al., 2024). Building on commonly used Arabic benchmarks such as AFND (Khalil et al., 2022), we focus the war-related Arabic news collected from online sources and study an interpretable detection pipeline that can be deployed as a practical user tool during browsing.

We design a pipeline tailored to Arabic news that combines semantic and stylistic signals relevant to misinformation detection. The proposed system integrates fuzzy linguistic cues, a CNN-based text feature extractor, and BERT-based contextual embeddings within a gated fusion framework. The fuzzy component uses an extravagant-words membership function to assign a graded exaggeration-oriented score based on the presence of highly emotional or exaggerated language. In addition, a CNN is used to extract local n-gram patterns from the text.

The objective of this work is to combine interpretable linguistic cues and n-gram patterns with deep semantic features to improve the accuracy, robustness, and explainability of Arabic fake news detection. We evaluate multiple machine learning models, deep learning architectures, and large language models to compare performance and select the most effective classifier.

Finally, we deploy the selected model as a real-time browser plugin that provides credibility predictions with a confidence score during browsing, enabling users to assess Arabic news content directly in a practical setting.

2. Related Work

During the Gaza and Israel war, social media became a major vector for real and fake news dissemination. Framing strategies, emotionally charged content, and coordinated disinformation campaigns (Alsharairi et al., 2024; Institute for Strategic Dialogue, 2025) significantly influenced public opinion, amplified polarization, and blurred the distinction between verified information and misleading or fabricated narratives.

Recent work has quantified how conflict narratives and emotions evolve across platforms. At

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 50–59
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

the public opinion level, digital activism during the Gaza conflict, reflected through social media signals (posts, comments, and hashtags), highlights measurable shifts in sentiment and engagement that shape perceptions and consumer attitudes (Mohammed et al., 2025). Event-driven discourse shifts further reveal clear platform-specific pathways for narrative diffusion during the Gaza war (Antonakaki and Ioannidis, 2025).

In NLP-based fake news detection, several supervised classifiers have been explored, including Logistic Regression, Decision Trees, Random Forests, Gradient Boosting, Passive Aggressive Classifiers, and XGBoost. Jouhar et al. (Jouhar et al., 2024) evaluated these methods on the ISOT dataset, showing that ensemble and boosting techniques achieved superior performance. Deep learning models further improved classification performance by learning contextual representations, with reported accuracy levels of up to 90% (Fouad et al., 2022).

More recently, BERT-based embeddings have enabled models to capture both Modern Standard Arabic and dialectal variations effectively (Awane et al., 2021; Al-yahya et al., 2021). When combined with hyperparameter optimization strategies across architectures such as RNNs, LSTMs, and transformers, these approaches can significantly enhance detection performance and overall model efficiency (Ettaik and Ben Lahmar, 2021).

Bidirectional recurrent models are especially suitable for deceptive-language detection because cues such as exaggeration, negation, and stance depend on both preceding and following context, making BiLSTM a strong candidate for Arabic news classification (Merzah et al., 2025; Turki et al., 2025). Recent hybrid deep learning designs combine CNN and BiLSTM to jointly capture local patterns and sequential context; Merzah et al. (Merzah et al., 2025) report that a multi-channel CNN with dual BiLSTM (FastText embeddings) achieves strong performance on Arabic benchmarks (AFND/ANS) and an English dataset (WELFake).

Beyond single-architecture models, hybrid frameworks integrate complementary techniques to enrich representations. Ennaouri et al. (Ennaouri and Zellou, 2025) proposed a model combining fuzzy logic-based membership functions to capture linguistic exaggeration with BERT embeddings for contextual semantics, achieving significant improvements over standard classifiers such as SVM and Logistic Regression. The fuzzy layer enables the assignment of degrees of truth rather than strict binary decisions (Mirza Saad et al., 2025).

Current literature shows that fake news detection performance depends not only on the classifier itself, but also on the choice of features and the way complementary signals are combined (Hamed et al., 2023). Our work builds on these findings by combining contextual Arabic embeddings, local neural pattern extraction, and fuzzy exaggeration-aware cues within a unified pipeline. The proposed framework is intended to capture complementary aspects of Arabic conflict-related misinformation that prior studies have typically addressed separately or only through partial combinations.

In parallel, mitigation efforts increasingly focus on user tools; for example, *Aletheia* introduces a browser extension that leverages retrieval-augmented generation and large language models to support fake news detection with evidence based explanations and user engagement features (Sallami et al., 2025). Inspired by this direction, we operationalize our best-performing model as a browser plugin that reports real/fake predictions with a confidence score during browsing.

3. Data

We assembled a multi-source Arabic news corpus for binary fake news classification. First, we incorporated the AFND benchmark dataset (Khalil et al., 2022), retaining its original labels. In addition, we collected public posts from *X* (formerly Twitter) using keyword-based queries related to the Gaza war and the broader Palestine–Israel conflict. Within this subset, posts were assigned to the fake-news class when they originated from bots or non-authentic accounts, or when they were explicitly associated with hashtags indicating fake news. To balance the corpus with verifiable information, we also automatically collected recent articles from established Arabic news outlets, including *Al Jazeera* Arabic and *Al Arab* newspaper, which were assigned to the real news class.

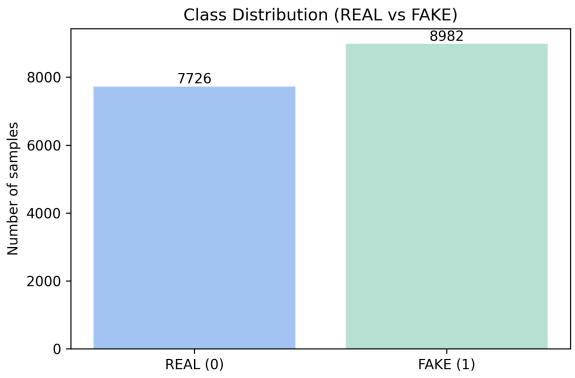


Figure 1: Distribution of fake and real news instances in the collected corpus.

No manual annotation was performed; instead, the

labeling process relied on inherited benchmark labels and source-based assignment rules. This initial data collection process resulted in a total of **16,708** labeled instances. We then applied keyword filtering using conflict terms (such as: غزة، القدس، فلسطين، الضفة، الاحتلال، إسرائيل، حماس، رفح، خان يونس، جباليا، النازحين) in order to retain topically relevant content. The final dataset was restricted to Modern Standard Arabic; Dialectal Arabic was outside the scope of the present study. The distribution between fake and real news classes is illustrated in Figure 1.

4. Proposed Method

The proposed method introduces a hybrid framework for Arabic fake news detection that processes news content through multiple stages to improve classification performance.

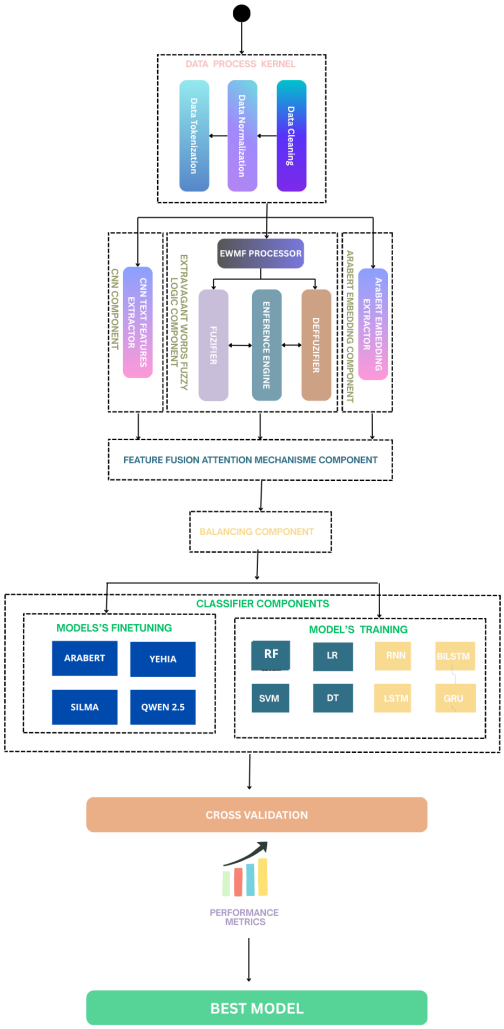


Figure 2: Overview of the proposed hybrid pipeline for Arabic news classification.

As illustrated in Figure 2, the architecture consists of several key components, including a preprocessing module, a fuzzy logic-based linguistic feature extractor, an AraBERT embedding module, a CNN-based text feature extractor, a gated feature fusion component, a data balancing component, and a classification module. Each component contributes to transforming raw textual data into enriched representations that support accurate and robust classification.

4.1. Preprocessing Component

Preprocessing is a necessary step in our pipeline because Arabic news text collected from the web often contains noise and inconsistencies that can degrade downstream classification. We first perform data cleaning by removing duplicates, handling missing fields, and filtering non-Arabic content. We then eliminate web artifacts such as URLs, user mentions, and hashtags, and normalize the text by removing special characters, stop words, and redundant whitespace. Next, Arabic normalization is applied to standardize common character variants (e.g., different forms of *alef* and *yaa*). Spelling, grammatical, and orthographic errors were not explicitly modeled as standalone predictive features; instead, normalization was applied to reduce superficial variation and support more stable downstream representations. Finally, the cleaned text is transformed into normalized input suitable for downstream feature extraction and modeling.

4.2. AraBERT Embedding Component

We represent each article using contextual embeddings from AraBERT. Texts are tokenized with the AraBERT tokenizer and encoded in the standard input format `[CLS] x [SEP]`, then converted to token IDs according to the pre-trained vocabulary. We freeze all AraBERT parameters and extract the final-layer hidden states (768 dimensions per token). A fixed-length sentence embedding is obtained by mean pooling over non-padding tokens using the attention mask, producing one 768-dimensional vector per text.

4.3. Fuzzy Logic Component

While contextual embeddings capture semantics effectively, deception in news is often expressed through stylistic cues such as sensationalism, urgency, and exaggerated framing. To complement AraBERT representations with an interpretable signal, we introduce a fuzzy-logic feature engineering component. We design an Arabic *Extravagant Word Membership Function* (EWMF) that maps the observed frequency of exaggerated expressions in a text to a continuous membership score,

enabling a smooth transition from low to high levels of sensational language. The underlying lexicon consists of commonly used sensational markers in Arabic misinformation (e.g., فاضح، مدمر، مهلك، (قاتل، حصري، عاجل، سري، خطير، حساس، مكشوف، مفضوح, and the resulting membership value is used as an additional input feature alongside learned embeddings.

The process of Arabic Extravagant Word Membership Function (EWMF) described in Algorithm 1.

Algorithm 1: Extravagant Word Membership Function (EWMF)

Input: Arabic texts $\mathcal{T} = \{t_1, \ldots, t_n\}$, threshold T_0, scaling factor σ
Output: Extravagant word counts $\mathcal{C}$, fuzzy membership values $\mathcal{M}$
Initialization:
Define extravagant word set $\mathcal{W}_{ext}$;
Define intensifier set $\mathcal{W}_{int}$;
Define clickbait pattern set $\mathcal{P}_{click}$;
$\mathcal{V}_{sens} \leftarrow \mathcal{W}_{ext} \cup \mathcal{W}_{int} \cup \mathcal{P}_{click}$;
Initialize $\mathcal{C} \leftarrow \emptyset$, $\mathcal{M} \leftarrow \emptyset$;
foreach $t_i \in \mathcal{T}$ **do**
  $count_i \leftarrow 0$;
  $\mathcal{W}_i \leftarrow$ ExtractArabicWords(t_i);
  foreach $w \in \mathcal{W}_i$ **do**
    if $w \in \mathcal{V}_{sens}$ **then**
      $count_i \leftarrow count_i + 1$;
  if $count_i = 0$ **then**
    $m_i \leftarrow 0$;
  else if $count_i \geq T_0$ **then**
    $m_i \leftarrow 1$;
  else
    $m_i \leftarrow 1 - e^{-\frac{\sigma \cdot count_i}{T_0}}$;
  $\mathcal{C} \leftarrow \mathcal{C} \cup \{count_i\}$;
  $\mathcal{M} \leftarrow \mathcal{M} \cup \{m_i\}$;
return $(\mathcal{C}, \mathcal{M})$

The membership function outputs 0 when no sensational cues are found ($x = 0$). As the cue count x increases, the membership value increases according to an exponential mapping. In our implementation, the mapping is defined piecewise: $m(x) = 0$ for $x = 0$, $m(x) = 1$ for $x \geq T_0$, and otherwise $m(x) = 1 - e^{-\frac{\sigma x}{T_0}}$. We apply this function to each news item by parsing the text, counting matches against the sensational cue sets, and converting the resulting count into a fuzzy score.

4.4. CNN Extractor Component

To capture local n-gram patterns, we employ a 1D CNN feature extractor. The input sequence is first mapped through a Keras embedding layer initialized with a pre-trained embedding matrix. We then apply three parallel `Conv1D` branches with 100 filters each and kernel sizes $\{3, 4, 5\}$, using ReLU activation. Each branch is followed by Global Max Pooling to obtain a fixed-length vector. The pooled outputs are concatenated into a 300 dimensional representation (100×3), which is used as the CNN feature vector. This vector is later fused with AraBERT embeddings and fuzzy scores in feature fusion component.

4.5. Feature Fusion Component

This component combines three complementary feature types into a single representation: (i) the fuzzy membership score produced by the Extravagant Word Membership Function (EWMF), (ii) AraBERT embeddings capturing contextual semantics, and (iii) CNN features capturing local n-gram patterns. The goal is to integrate interpretable stylistic cues with learned semantic and surface-level representations.

In our implementation, we use the EWMF score as a gating signal that scales the AraBERT and CNN feature vectors before concatenation. For each sample i, let $f_i \in [0, 1]$ denote the fuzzy score, $\mathbf{b}_i \in R^{d_B}$ the AraBERT embedding, and $\mathbf{c}_i \in R^{d_C}$ the CNN feature vector. The gated representations are computed as:

$$\tilde{\mathbf{b}}_i = f_i \cdot \mathbf{b}_i, \qquad \tilde{\mathbf{c}}_i = f_i \cdot \mathbf{c}_i.$$

The final fused representation is then formed by concatenation:

$$\mathbf{z}_i = [\, f_i;\ \tilde{\mathbf{b}}_i;\ \tilde{\mathbf{c}}_i \,].$$

This fused vector is used as input to the downstream classifiers.

4.6. Classifier Component

We evaluate a range of machine learning and deep learning classifiers, as well as selected large language models, for distinguishing real versus fake news in the Palestine/Israel war setting. The tested models include traditional classifiers (e.g., Random Forest, SVM, Logistic Regression, and Decision Tree), deep neural models (RNN, LSTM, BiLSTM, and GRU), and LLM-based baselines (Qwen, AraBERT, Silma, and Yehia).

For a robust performance estimate, we apply 5-fold cross-validation, where the dataset is split into five folds and models are trained and validated iteratively. For deep learning models and LLM-based runs, we additionally use early stopping based on validation performance. Model performance is reported using Accuracy, Precision, Recall, and F1-score.

5. Results and Discussion

We report model performance using standard classification metrics derived from the confusion matrix.

Accuracy measures the proportion of correctly classified instances:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$

Recall (Sensitivity) measures the proportion of actual positive instances correctly identified:

$$\text{Recall} = \frac{TP}{TP + FN}.$$

Precision measures the proportion of predicted positive instances that are correct:

$$\text{Precision} = \frac{TP}{TP + FP}.$$

F1-score is the mean of Precision and Recall:

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Figures 3-4-5 report an ablation comparison of models trained with and without the proposed EWMF+CNN components. Across machine learning and deep learning families, incorporating these components yields consistent improvements in Accuracy, Precision, Recall, and F1, indicating that exaggerated language cues and local n-gram patterns provide complementary signals to contextual embeddings. The strongest deep learning results are obtained with BiLSTM (Fig. 3), which is in line with prior evidence that bidirectional recurrent modeling benefits deception detection by leveraging both preceding and following context (Merzah et al., 2025; Turki et al., 2025). For classical machine learning models (Fig. 4), the gains are more pronounced for weaker baselines, suggesting that feature enrichment is particularly beneficial when the classifier cannot implicitly learn such stylistic cues. This observation is consistent with hybrid approaches that combine fuzzy linguistic indicators with contextual representations (Ennaouri and Zellou, 2025). Finally, Fig. 5 shows that adapting large language models to the proposed pipeline substantially improves performance, while the Arabic-specialized AraBERT model remains the most reliable among the evaluated LLMs, supporting findings that domain and language aligned encoders are critical for Arabic credibility assessment (Fouad et al., 2022; Al-yahya et al., 2021; Papageorgiou et al., 2024).

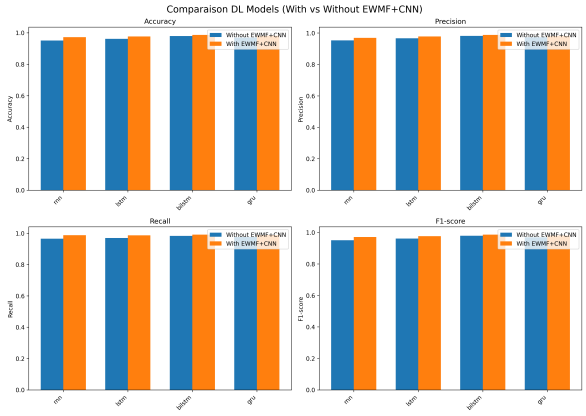


Figure 3: Performance comparison of deep learning models with and without the EWMF and CNN components.

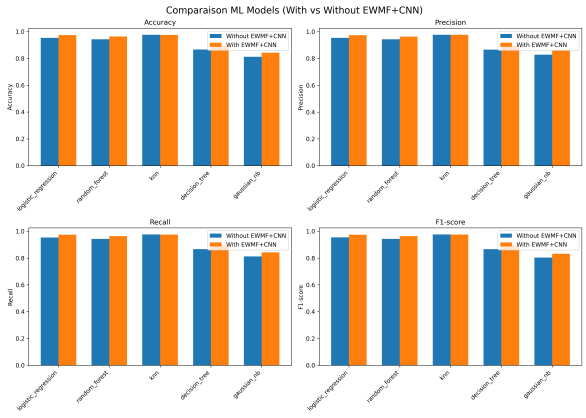


Figure 4: Performance comparison of machine learning models with and without the EWMF and CNN components.

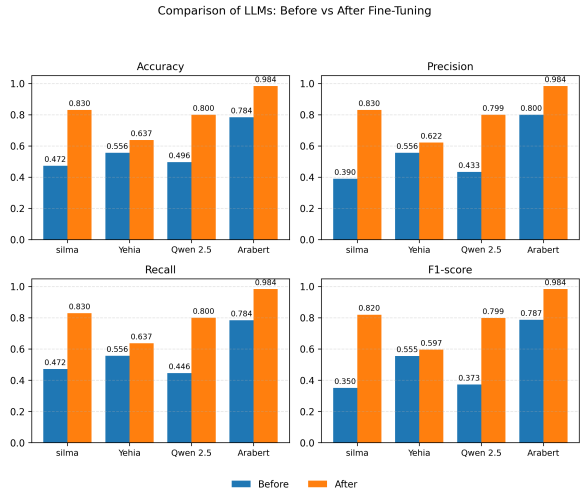


Figure 5: Performance comparison of large language models with and without the EWMF and CNN components.

Model	Acc.	Prec.	Rec.	F1
Logistic Regression	0.9742	0.9742	0.9742	0.9742
Random Forest	0.9634	0.9637	0.9634	0.9632
KNN	0.9756	0.9765	0.9756	0.9754
Decision Tree	0.8948	0.8950	0.8948	0.8949
Gaussian NB	0.8429	0.8620	0.8429	0.8319
RNN	0.9716	0.9693	0.9872	0.9715
LSTM	0.9767	0.9777	0.9863	0.9767
BiLSTM	**0.9867**	0.9881	0.9912	0.9867
GRU	0.9807	0.9816	0.9885	0.9807
Silma	0.8300	0.8300	0.8300	0.8200
Yehia	0.6375	0.6220	0.6375	0.5965
Qwen 2.5	0.8000	0.7991	0.8000	0.7995
AraBERT	0.9838	0.9839	0.9838	0.9838

Table 1: Performance comparison of models using the proposed pipeline.

To enable a fair comparison of Palestinian fake news detection approaches, we evaluate all models on the same dataset under a unified preprocessing and evaluation protocol. This is important because prior work often relies on different datasets tied to specific events or platforms, and reported metrics are not always consistent across studies, which limits direct comparability. Therefore, Table 1 reports our baseline results across three families of methods: (i) conventional machine learning classifiers, (ii) deep learning sequence models, and (iii) LLM baselines.

Overall, the results indicate that deep learning models achieve the highest performance on this dataset, with BiLSTM obtaining the best scores among the evaluated architectures. The observed hierarchy can be interpreted in terms of (1) how each model family represents context in Arabic news, and (2) whether it benefits from complementary cues beyond semantics, such as exaggeration-oriented fuzzy scores and local n-gram patterns.

Conventional ML baselines. Logistic Regression and KNN provide strong results (both around 97% accuracy/F1), suggesting that the representation used in our pipeline yields a feature space where the classes are largely separable with relatively simple decision boundaries. However, ML models do not explicitly model sequential dependencies; consequently, they may miss long-range context signals (e.g., stance, negation, or discourse-level cues) that are relevant to deception in news narratives.

Sequence DL baselines. Recurrent architectures consistently outperform conventional ML models, and the BiLSTM achieves the best overall performance. This trend is aligned with recent Arabic fake news studies that report gains from BiLSTM-based designs when contextual representations are combined with additional textual cues (e.g., hybrid contextual features coupled with BiLSTM classifiers) (Turki et al., 2025). Similarly, hybrid CNN–BiLSTM designs have been shown to be effective because CNN captures local patterns while BiLSTM models sequential context, leading to stronger performance than using either family alone (Merzah et al., 2025). Our results support this line of evidence: once local pattern cues and exaggeration-oriented fuzzy scores are integrated alongside contextual semantics, BiLSTM is the most effective downstream classifier on this dataset.

LLM baselines. General-purpose LLM baselines (e.g., Silma, Yehia, Qwen 2.5) underperform supervised Arabic models in our controlled setting, whereas AraBERT remains competitive with the best recurrent architectures. This gap is consistent with recent evidence that LLM outputs may be less reliable for politically sensitive or high-stakes judgments due to susceptibility to bias and inconsistent reasoning, especially when the task requires stable label decisions rather than open-ended generation (Gubelmann and Karray, 2025; Barman et al., 2024). Moreover, comparative studies in fake-news detection frequently report that encoder-style transformers (BERT-like) remain strong for supervised classification pipelines, while GPT-like models often require careful adaptation to match them (Papageorgiou et al., 2024; Raza et al., 2025). Overall, the strong AraBERT baseline and weaker general LLM baselines suggest that performance here depends primarily on domain-aligned representations and supervised training, rather than generative capacity alone.

The strong scores observed across model families can be partly explained by the focused nature of the dataset and by the fused representation used in our pipeline. In this setting, fake and real news differ through a combination of semantic context, local lexical patterns, and exaggeration-oriented stylistic cues, which increases class separability once these signals are jointly encoded. This also helps explain why conventional ML baselines remain highly competitive on the learned feature space. Taken together, these results indicate that (i) contextual modeling is critical for conflict-related Arabic news classification, and (ii) integrating complementary signals, including local n-gram cues and exaggeration-oriented fuzzy scores, provides additional discriminative power that is most effectively exploited by bidirectional sequence models.

5.1. Deployment as a Browser Plugin

To demonstrate practical usability beyond offline evaluation, we deployed the best-performing model (BiLSTM) as a real-time browser plugin for Arabic news credibility assessment. During browsing, the extension extracts the article text and sends it to a backend analysis endpoint,

which applies the same preprocessing and inference pipeline used in our experiments and returns a predicted label with an associated confidence score. To improve transparency, we additionally provide token-level rationales by highlighting influential spans in the article, computed using Integrated Gradients and refined with a deletion test that retains only spans whose removal produces a measurable confidence drop.

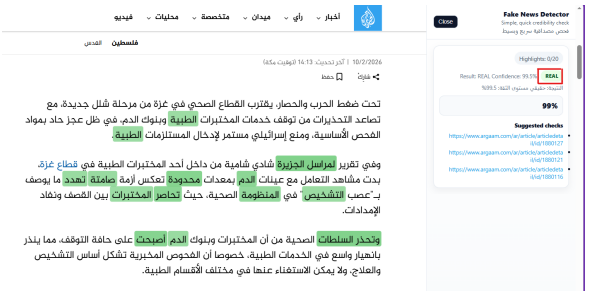


Figure 6: Screenshot of the deployed browser plugin. The BiLSTM model predicts the credibility label and returns a confidence score for the visited Arabic news content.

6. Limitations and Future Work

The strong scores reported in Table 1 should be interpreted with caution. Because the dataset is focused on conflict-related Arabic news, class separability may be partly supported by strong lexical, semantic, and stylistic regularities within this domain. Although exact duplicate entries were removed during preprocessing, near-duplicate or lightly rewritten content may still remain, especially when similar narratives circulate across multiple channels. In addition, our source-based labeling strategy may introduce bias, since real news instances were primarily collected from established news outlets, whereas part of the fake news class originated from benchmark misinformation data and selected social media posts. As a result, some of the predictive signal may reflect source-related or topic-specific regularities in addition to misinformation-specific cues. These factors do not invalidate the reported results, but they suggest that future work should consider stricter deduplication, source-balanced sampling, and broader cross-domain evaluation.

A second limitation concerns the scope of the proposed pipeline, which is developed and evaluated primarily for Arabic news text. Extending the approach to multilingual settings would require additional language-specific preprocessing, resources, and evaluation across more diverse benchmarks.

Third, although our data target conflict-related news, the deployed browser plugin is currently optimized for relatively clean article pages. Its performance may degrade on social media content, where text is shorter, more conversational, noisier, and often mixed with screenshots, images, or embedded videos.

In addition, the system produces a model-based credibility label and confidence score, which should not be interpreted as definitive factual verification. This is particularly important in cases where misinformation is subtle, partially true, or dependent on external context. A useful extension would therefore be to integrate retrieval-augmented components that gather related coverage from trusted sources and present supporting evidence alongside predictions.

Finally, although the framework incorporates interpretable linguistic cues such as exaggerated or sensational expressions, these cues are not exhaustive and may evolve over time, especially in rapidly changing conflict discourse. Continuous lexicon updates, monitoring of domain shift, and periodic re-evaluation on newly collected data would be necessary to preserve reliability.

7. Conclusion

This paper addresses credibility assessment for Arabic conflict-related news, a setting where manipulated discourse and emotionally charged framing can distort public understanding and the preservation of Palestine/Israel narratives as language resources. We proposed an interpretable hybrid pipeline that combines contextual semantics from AraBERT embeddings with complementary signals: (i) an Extravagant Word Membership Function (EWMF) that quantifies sensational and clickbait oriented cues via fuzzy scoring, and (ii) a CNN-based extractor that captures local n-gram patterns. Across a unified evaluation protocol on Palestine/Israel war related Arabic news, deep sequence models achieved the strongest results, with BiLSTM performing best (Accuracy/F1 = 0.9867, Recall = 0.9912), while AraBERT remained highly competitive (Accuracy/F1 = 0.9838). Conventional machine learning baselines were strong but consistently below the top recurrent models, and general purpose LLM baselines lagged substantially, suggesting that domain aligned representations and supervised training are crucial for reliable classification in this sensitive context. To demonstrate practical utility beyond offline experiments, we operationalized the best performing model as a browser plugin that outputs a real/fake prediction with a confidence score during browsing. Limitations include restricted language and domain coverage, limited validation on social media discourse, and the absence of evidence verification.

Future work should study multilingual generaliza-

tion and robustness on social media discourse, incorporate retrieval-augmented evidence to support model predictions, and address temporal domain shift via continuous lexicon maintenance and regular re-training/evaluation on newly collected data.

7.1. Ethics and Compliance:

We limit collection to publicly available content, honor platform and site policies, and remove items upon request. For X/Twitter, we follow the developer policy on redistribution. For news sites, we use feeds, public datasets, or permissions and respect Terms of Service.

References

Maha Al-yahya, Hend Al-Khalifa, Heyam Al-Baity, Duaa AlSaeed, and A. Essam. 2021. Arabic fake news detection: Comparative study of neural networks and transformer-based approaches. *Complexity*, 2021.

Ahmad Alsharairi, Hana Abdul-Rahman Al-Souob, Mohammad Farouq AlQadi, and Sora Mohammad Shatnawi. 2024. Social media communication and framing of the gaza conflict: Impact on public opinion. ResearchGate.

Despoina Antonakaki and Sotiris Ioannidis. 2025. Cross-platform digital discourse analysis of the israel-hamas conflict: Sentiment, topics, and event dynamics.

Widad Awane, El Habib Ben Lahmar, and Ayoub El Falaki. 2021. Hate speech in the arab electronic press and social networks. *Revue d'Intelligence Artificielle*, 35(6):457.

Dipto Barman, Ziyi Guo, and Owen Conlan. 2024. The dark side of language models: Exploring the potential of llms in multimedia disinformation generation and dissemination. *Machine Learning with Applications*, 16:100545.

Mohammed Ennaouri and Ahmed Zellou. 2025. Enhancing fake review detection using linguistic exaggeration, bert embeddings, and fuzzy logic. *IEEE Access*, 13:135957–135968.

Noureddine Ettaik and El Habib Ben Lahmar. 2021. Hyperparameter optimisation in nlp architectures. In *Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning*, pages 55–62, Lisbon, Portugal. SCITEPRESS - Science and Technology Publications.

Khaled M. Fouad, Sahar F. Sabbeh, and Walaa Medhat. 2022. Arabic fake news detection using deep learning. *Computers, Materials & Continua*, 71(2):3647–3665.

Reto Gubelmann and Ghassen Karray. 2025. Assessing reliability and political bias in LLMs' judgements of formal and material inferences with partisan conclusions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30005–30031, Vienna, Austria. Association for Computational Linguistics.

Suhaib Kh Hamed, Mohd Juzaiddin Ab Aziz, and Mohd Ridzwan Yaakub. 2023. A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion. *Heliyon*, 9(10):e20382.

Institute for Strategic Dialogue. 2025. *Conflict Amplified: Disinformation and Hate in the Israel-Hamas War*. Institute for Strategic Dialogue.

Jumana Jouhar, Anju Pratap, Neharin Tijo, and Meenakshi Mony. 2024. Fake news detection using python and machine learning. *Procedia Computer Science*, 233:763–771. 5th International Conference on Innovative Data Communication Technologies and Application (ICIDCA 2024).

Ashwaq Khalil, Moath Jarrah, Monther Aldwairi, and Manar Jaradat. 2022. Afnd: Arabic fake news dataset for the detection and classification of articles credibility. *Data in Brief*, 42:108141.

Baqer M. Merzah, Jafar Razmara, and Zolfaghar Salmanian. 2025. Hybrid deep learning models for fake news detection: Case study on arabic and english languages. *Frontiers in Big Data*.

Mahmud Mirza Saad, Shahzad Muhammad Khuram, Ali Syed Imran, Jabeen Farzana, and Moetesum Momina. 2025. Fond: Fuzzy-based optimized fake-news detection using big-bird and longformer. *SSRN Electronic Journal*.

Ashraf Bany Mohammed, Manaf Al-Okaily, Dhia Qasim, Shafique Ur Rehman, and Latifa Abdalla. 2025. Digital activism and public opinion: understanding the role of social media during the gaza conflict. *Journal of Islamic Marketing*, 16(11):3366–3393.

Eleftheria Papageorgiou, Christos Chronis, Iraklis Varlamis, and Yassine Himeur. 2024. A survey on the use of large language models (llms) in fake news. *Future Internet*, 16(8):298.

Shaina Raza, Drai Paulen-Patterson, and Chen Ding. 2025. Fake news detection: comparative evaluation of bert-like models and large language models with generative ai-annotated data.

H. Sallami, E. Aïmeur, et al. 2025. Aletheia: Detect, discuss, and stay informed on fake news. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25), Demonstrations Track*.

Hussain Mohammed Turki, Essam Al Daoud, Ghassan Samara, Raed Alazaidah, Mais Haj Qasem, Mohammad Aljaidi, Suhaila Abuowaida, and Nawaf Alshdaifat. 2025. Arabic fake news detection using hybrid contextual features. *International Journal of Electrical and Computer Engineering (IJECE)*, 15(1):836–845.

Mohammed Abbas Yousef, Abeer ElKorany, and Hanaa Bayomi. 2024. Fake-news detection: a survey of evaluation arabic datasets. *Social Network Analysis and Mining*, 14:225.

A. Data

A.1. Detailed Preprocessing Steps

The preprocessing pipeline applied to the collected corpus consisted of the following steps:

1. Removal of duplicate and incomplete entries.
2. Retention of Arabic texts only through language and script filtering.
3. Removal of web-specific artifacts, including URLs, user mentions, and hashtags.
4. Removal of non-Arabic and non-linguistic characters.
5. Diacritic removal and whitespace normalization.
6. Arabic character normalization to reduce orthographic variation.
7. Removal of very short texts with limited linguistic content.

A.2. Data availability

The dataset is anonymously available at: `https://huggingface.co/datasets/Anonymous-123-de/credibility-assessment`

B. Experimental Configuration

B.1. General Settings

Setting	Value
Train/test split	80/20, stratified
Cross-validation	5-fold StratifiedKFold
Random state	42
Feature scaling	StandardScaler

Table 2: General experimental settings.

B.2. AraBERT Settings

Parameter	Value
Pretrained model	`aubmindlab/bert-base-arabertv02`
Maximum sequence length	128
Batch size	8

Table 3: AraBERT feature extraction settings.

B.3. CNN Settings

Parameter	Value
Embedding dimension	128
Filter sizes	3, 4, 5
Number of filters	100 per filter size
Dropout	0.3
Maximum sequence length	256
Batch size	16

Table 4: CNN feature extractor settings.

B.4. Classical Machine Learning Settings

Model	Configuration
Logistic Regression	max_iter=2000
Random Forest	n_estimators=100
KNN	n_neighbors=5
Decision Tree	default settings
Gaussian Naive Bayes	default settings

Table 5: Classical machine learning model settings.

B.5. Neural Model Settings

Parameter	Value
Models	RNN, LSTM, BiLSTM, GRU
Number of layers	2
Dropout	0.3
Optimizer	Adam
Learning rate	0.001
Weight decay	10^{-5}
Loss function	CrossEntropyLoss
Maximum epochs	50
Early stopping patience	5

Table 6: Neural model training settings.

B.6. Notebook availability

The implementation is anonymously available at: https://github.com/Annonymous975-del/Credibility-Assessment

C. Plugin Code availibitlity

https://github.com/Sanaa99-ab/-fake-news-detection-plugin

Tarikhi: Arabic Temporal Information Extraction From Arabic Historical Documents

Qusay Abdo[1], **Serin Atiani**[1,*], **Tariq Sarayji** [1],**Adnan Saeed**[1]
[1]Princess Sumaya University for Technology, Jordan
[*]Corresponding author: s.atiani@psut.edu.jo

Abstract

Arabic historical books and archival materials contain rich accounts of political, social, and cultural events, yet they remain largely underutilized computationally due to the scarcity of dedicated Arabic information extraction tools. The challenge is amplified in long-form, scanned historical documents, where optical character recognition noise, orthographic variation, and complex narrative structures complicate automatic processing. In this paper, we present Tarikhi, a retrieval-augmented generation framework for structured temporal event extraction from Arabic scanned books. The proposed pipeline integrates high-accuracy optical character recognition, chunking-based processing for long-document handling, Arabic named entity recognition, span refinement, and a retrieval-enhanced attribute extraction module that identifies event dates, locations, and descriptive summaries. Extracted events are consolidated and linked using semantic and temporal similarity measures, and linked through relation classification to construct structured temporal events. Evaluation on a selected part of modern Arabic historical books demonstrates the feasibility of temporal event extraction from long-form Arabic texts, achieving a 75.3% F1-score under dual human verification. Tarikhi represents a step toward scalable temporal knowledge construction for Arabic digital humanities resources.

Keywords: Temporal Event Extraction, Arabic Natural Language Processing, Retrieval Augmented Generation, Long-form Document Processing

1. Introduction

The rapid digitization of historical Arabic archives has generated vast information repositories, creating new opportunities for computational analysis of cultural and historical knowledge. Automated information extraction systems are increasingly needed to transform unstructured historical documents into structured representations that support search, analytics, and scholarly research. Event extraction is the task of identifying structured descriptions of events from unstructured text, it has emerged as a fundamental component of modern information extraction pipelines and is widely applied across domains such as historical analysis, knowledge base construction, and digital humanities research (Zhang and Han, 2025; Huang et al., 2024). However, extracting temporal events from historical sources remains challenging due to linguistic ambiguity, narrative complexity, and the need to reason across long textual contexts.

Recent advances in large language models and retrieval-augmented generation (RAG) architectures have demonstrated strong capabilities in contextual reasoning and knowledge-intensive tasks by integrating retrieval mechanisms with generative models (Krasadakis et al., 2024). In event temporal relation extraction and related tasks, retrieval-augmented approaches have shown promise in improving performance by incorporating external knowledge and contextual evidence (Zhang et al., 2024). The above-mentioned research predominantly focuses on English and a limited set of high-resource languages, leaving a significant gap in tools designed for low-resource languages, specifically for Arabic historical materials.

Arabic historical resources present unique challenges for computational analysis and event extraction due to its rich morphology, orthographic variability, and limited availability of annotated resources, particularly for specialized tasks such as temporal event extraction (Darwish et al., 2021; Alayba, 2025). Existing studies highlight the scarcity of datasets and models tailored to Arabic event extraction and emphasize the need for domain-specific approaches that address linguistic and resource limitations (Alayba, 2025; Aljabari et al., 2024a). Previous work has explored information extraction from Arabic manuscripts, focusing mainly on classification or basic information extraction rather than comprehensive temporal event modeling (Bashir et al., 2023). Consequently, there is a pressing need for solutions that can efficiently extract structured historical events from Arabic resources while leveraging modern retrieval and language modeling techniques.

This paper proposes a retrieval-augmented generation (RAG) framework designed specifically for temporal event extraction from Arabic long-form historical sources. The proposed system ingests textual resources, including books, and archives in PDF format, transcribes them into text and integrates these resources into a retrieval-driven pipeline that enhances contextual understanding and enables efficient extraction of temporal historical events. The paper is organized as follows: Sec-

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 60–69
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

tion 2 reviews related work, section 3 outlines the methodology, section 4 demonstrates system results at different components of the system and system evaluation, sections 5 discussion and section 6 outlining limitations of the work.

2. Related Work

Temporal Information Extraction (TIE) is the process of identifying temporal expressions in natural language text and extracting the temporal relations between events and time expressions. The date and time associated with a temporal expression constitute essential identifiers of a temporal event, enabling structured temporal reasoning over text. Identifying temporal relations between events (e.g., *before*, *after*, *includes*) and anchoring them to temporal expressions facilitates the construction of timelines in which events are organized according to their temporal dependencies (Leeuwenberg and Moens, 2018).

Early research on temporal processing was strongly influenced by the TimeML specification (Pustejovsky et al., 2003), which introduced a standardized annotation schema for events, temporal expressions (TIMEX3), and temporal links (TLINKs). Shared tasks such as TempEval (Verhagen et al., 2007; UzZaman et al., 2013) and Clinical TempEval (Bethard et al., 2015) played a pivotal role in establishing benchmarks for temporal expression extraction, event detection, temporal relation classification, and temporal anchoring. Although temporal expression extraction has sometimes been framed as a sub-task of Named Entity Recognition (NER), it has evolved into a distinct research trajectory due to the additional requirements of normalization, temporal reasoning, and cross-sentence relation modeling (Wong et al., 2005). Subsequent research has advanced temporal expression extraction (Filannino and Nenadić, 2015), temporal annotation frameworks and corpora construction (D'Souza and Ng, 2014; Mostafazadeh et al., 2016), and temporal disambiguation and reasoning techniques, including recent self-supervised and neural approaches (Wenzel and Jatowt, 2023; Cai et al., 2023).

The rise of deep learning has significantly improved performance on temporal tasks. Transformer-based language models such as BERT have been adapted for event detection and temporal relation classification (Devlin et al., 2019; Ning et al., 2019). Recent work has explored graph-based neural models and global inference mechanisms to improve temporal extraction across documents (Leeuwenberg and Moens, 2018). Moreover, large-scale pretraining and prompt-based learning have improved temporal extraction in low-resource and domain-specific scenarios (Cai et al., 2023).

TIE for Arabic has been mostly studied as a subset of event extraction in general, and it has been addressed by studies that have looked more in general into named entity recognition, with few exceptions (Saleh et al., 2011; Zaraket, 2012; Lhioui et al., 2017). Several studies have focused on developing dedicated corpora and annotation schemes. ARA-Timex introduced two annotated datasets compiled from WikiNews and news articles to support temporal expression extraction and normalization in Arabic (Boudaa et al., 2018). AraTimeBank presented a revised Arabic TimeML dataset that addressed limitations in earlier annotations and enriched the corpus to better support TIE tasks (Haffar et al., 2020). More recently, the Wojood project has contributed substantially to Arabic event and relation extraction. Wojood$^{\text{Hadath}}$ introduced an enhanced version of the Wojood corpus enriched with event-argument annotations (Aljabari et al., 2024b), while Wojood$^{\text{Relation}}$ further extended the dataset with annotated event-event relations, supporting structured temporal and causal modeling (Aljabari et al., 2025). These resources collectively advance the feasibility of temporal event extraction pipelines in Arabic.

In parallel, transformer-based Arabic language models have significantly improved performance across core NLP tasks, including NER, which forms a foundational component of temporal expression and event extraction. AraBERT demonstrated competitive results on Arabic NER and other downstream tasks without explicitly targeting temporal extraction (Antoun et al., 2020). JABER and SABER BERT models further improved upon prior transformer-based models through large-scale pretraining tailored to Arabic linguistic characteristics (Ghaddar et al., 2022). The Wojood benchmark established strong performance on nested NER using transformer-based architectures, providing an essential stepping stone for fine-grained temporal and event extraction tasks (Jarrar et al., 2022).

Despite the substantial strides made in event extraction and NER in Arabic, temporal information extraction and temporal information linking require further exploration and development, specifically in long form Arabic texts like books. To address this gap, we introduce Tarikhi, a fully automated pipeline for structured temporal information extraction from scanned Arabic historical documents.

3. Methodology

In this section, we describe the methodology used to implement **Tarikhi**. The proposed pipeline transforms documents into structured event records containing event descriptions, dates,

and locations, while also constructing semantic links between events, such as SAME_AS and RELATED_TO relations. Figure 1 provides a high-level illustration of the pipeline components. Scanned long-form resources are ingested and using OCR are transcribed into text. The text is preprocessed and chunked. Temporal events are identified and refined using Silma-9B (AI, 2024). RAG-based retrieval is employed to extract information from the document and then the extracted events are consolidated and linked using SAME_AS and RELATED_TO relations.

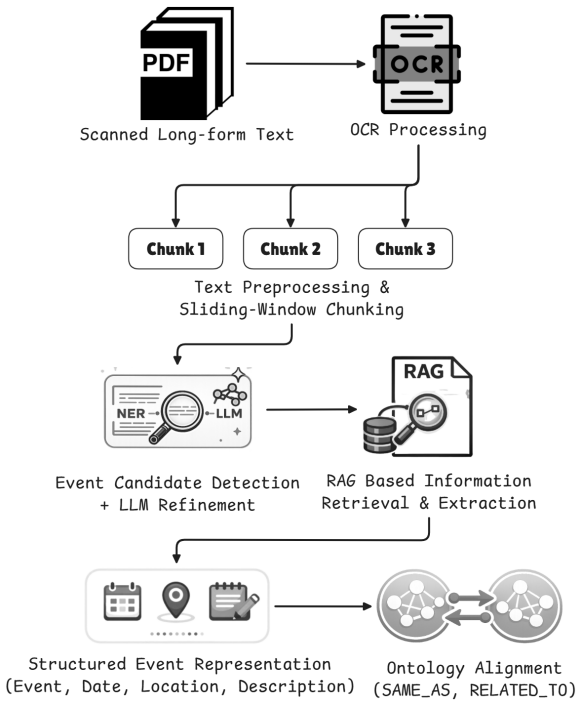


Figure 1: High-Level Pipeline

3.1. Data Sources

To run and evaluate Tarikhi, we used two books, the Arabic translation of the Hundred Years' War on Palestine by Rashid Khalidi, translation by Amer Shaikhouni first edition 2021 (Khalidi, 2021), and The 1967 War: Secrets and Mysteries by Amin Howeidi, original text in Arabic, first edition 2006 (Amin Huwaidi, 2006). Both books were selected for their dense references to political events, dates, actors, and locations, which make them suitable for evaluating temporal event extraction.

3.2. Text Recognition and Cleaning

We evaluated multiple OCR approaches during development, including the standard OCR engine Surya (Contributors, 2024), LLM-based OCR models such as Qari (Wasfy et al., 2025) and DeepSeek OCR (Wei et al., 2025), and a hybrid voting-based approach that combined the outputs of the three models by selecting word tokens through majority voting, while the Surya output was used as a fallback when all three models produced different outputs. For the final pipeline configuration, we selected the Surya OCR framework (Contributors, 2024) after conducting a small evaluation by manually transcribing 20 randomly selected pages, with 10 pages sampled from each of the two source books. Each OCR model was then run on the same pages and its generated output was compared against the manually written ground truth by computing alignment-based character-level and word-level similarity scores using `difflib.SequenceMatcher`, Table 1 shows the average results across the evaluated pages.

Model	Character Accuracy	Word Accuracy
DeepSeek	0.7449	0.8718
Qari	0.6760	0.9009
Surya	0.8592	0.9366
Hybrid	0.7746	0.8829

Table 1: Average OCR performance comparison across evaluated models.

Surya OCR operates using a two-stage architecture: a text line detection model followed by a text recognition model. The detection model identifies line-level text regions on each page in the given document and returns cropped line image segments that are stored locally. In the second stage, each detected line image is processed individually by the text recognition model and then concatenated to produce the Arabic text transcription. A text cleaning stage was applied prior to chunking to enhance processing quality. The cleaning method removes characters outside the Arabic letter range with the exceptions of digits and a small set of punctuation symbols (periods, hyphens, and slashes) to preserve dates and references.

3.3. Chunking Strategy and Embeddings Vector Space

To handle long documents within the context limitations of large language models, we adopted a tokenizer-based chunking strategy.

Each document is first tokenized using the same tokenizer associated with the target LLM; the exact one we used is silma-9b, and the token sequence is segmented into fixed-size chunks of 256 tokens with an overlap of 50 tokens between consecutive windows.

For semantic retrieval, each text chunk is encoded into a dense vector representation using the **intfloat/multilingual-E5-large** sentence em-

bedding model (Wang et al., 2024), embeddings were computed at the chunk level.

Each chunk is stored as a structured chunk object containing:

- The chunk text.
- Its embedding vector.
- Metadata fields for subsequent event extraction outputs.

Consequently, all chunk embeddings are indexed using a **FAISS** index, and similarity search is performed in the embedding space using top-k nearest neighbor queries. In the final configuration, we set $k = 6$ to provide sufficient context to the LLM without exceeding its context window.

3.4. Named Entity Recognition and Event Detection

For each stored chunk object, an event recognition step is performed in two stages. First, the events are triggered using **Sinatools**, a research-based open-source Python library used for various NLP tasks such as flat and nested Named Entity Recognition (NER), fully-flagged Word Sense Disambiguation (WSD), Semantic Relatedness, Synonymy Extractions and Evaluation, Lemmatization, Part-of-speech Tagging, Root Tagging, and additional helper utilities (Hammouda et al., 2024). In this study, we use their pre-trained Wojood NER model to trigger the events from the predicted tag sequence by merging spans that start with `B-EVENT`, which describe the first word of the event, and continue with `I-EVENT` which marks the remaining of the event name unless the event consists of a single word. This NER layer often yields incomplete or generic event names in Arabic.

To improve event name quality, we added an LLM assisted refinement layer on top of NER spans. For each extracted event span, we build a short local context window consisting of 10 tokens before the span and 20 tokens after it, which is then provided to the LLM as context. The LLM **SILMA-9B-Instruct-v1.0** (AI, 2024) takes the extracted event name, compares it with the context window, and returns a normalized, more representative event name. The refinement step uses deterministic decoding with temprature set to 0.0.

3.5. Event Disambiguation and Attribute Extraction

For each detected event, an event disambiguation stage is applied to determine the event's date, location, and description. This stage uses retrieval-augmented generation (RAG) with the LLM SILMA-9B.

Each event is encoded using the same multilingual embedding model used for chunk indexing, the event vector is used to query the FAISS vector index to retrieve the top-6 most semantically similar chunks. These retrieved chunks are then concatenated with the source chunk from which the event was originally triggered and provided as context for the LLM.

The extracted context is then provided to the LLM to extract specific attributes through prompting. Separate prompts and calls have been used for date, location, and description extraction, prompts enforce strict JSON-only responses with predefined schemas such as `{"date": "dd-mm-yyyy"}` and `{"location": "<place>"}`, and explicitly instruct the model to rely only on the provided text context. For date extraction, if no date is found, the model outputs `xx-xx-xxxx`, if only part of the date is available, missing components are replaced with `xx`. For location extraction, the output is left blank when no explicit location is found in the provided context.

The output of this stage is a structured event record in which each event is associated with its date, location, and description to be used and processed for event-linking.

3.6. Event Linking

To form links between the extracted events, we defined two edge types: `SAME_AS`, `RELATED_TO`. Linking is performed in two stages: (1) merge duplicate mentions into event clusters via `SAME_AS`, then (2) connect clusters with `RELATED_TO`.

3.6.1. `SAME_AS` clustering.

Given two event mentions e_i and e_j, we compute a composite similarity score combining:

- **Semantic similarity** s_{sem}: we concatenate the event name and description, encode the text using `multilingual-e5-large`, and compute cosine similarity between the resulting embeddings.
- **Temporal similarity** s_{temp}: when both mentions have valid dates, we compute an exponential decay over the absolute day difference,

$$s_{temp} = \exp\left(-\frac{\Delta t}{\tau}\right),$$

  where $\tau = 180$ days.
- **Location similarity** s_{loc}: computed by fuzzy location name matching.

The final score is computed using a weighted combination of the available scores

$$s = \frac{w_{emb}s_{sem} + w_{date}s_{temp} + w_{loc}s_{loc}}{w_{emb} + w_{date} + w_{loc}}$$

where $w_{emb} = 0.6$, $w_{date} = 0.3$, and $w_loc = 0.1$. Two events are merged when their score is at least 0.9, This was a conservative threshold that was set based on repeated observations, to only merge events when there is a a very strong evidence of their equivalence. When one of the scores is missing due to unavailable date or location, its weight is omitted from both the numerator and denominator of the weighted score. After clustering, each cluster node, represents one event that embodies all SAME_AS events. The event represented by the cluster node is summarized by a centroid embedding computed as the mean of member embeddings, a representative date defined as the median of available dates, a representative location defined as the most frequently mentioned location.

3.6.2. RELATED_TO Linking

RELATED_TO event edges are generated at the cluster level which differs from the SAME_AS relations generation that rely on the original extracted events. For each source cluster c_i, we compute cosine similarity between its event centroid embedding and all other cluster centroids. RELATED_TO candidate events are ranked by semantic similarity, and evaluate them in descending order.

For a candidate pair (c_i, c_j) we compute:

- **Semantic score** s_{sem}: cosine similarity of centroid embeddings.
- **Temporal proximity** s_{temp}: the same exponential decay function applied to the absolute difference between cluster median dates (set to 0 if either date is missing), we use a decay constant of $\tau = 365$ days.
- **Location score** s_{loc}: the same fuzzy matching technique between the most frequently mentioned locations of the two clusters.

These scores are combined as:

$$s = \alpha s_{sem} + \beta s_{temp} + \gamma s_{loc}.$$

where $\alpha = 0.7$, $\beta = 0.2$, and $\gamma = 0.1$. A RELATED_TO edge is added if $s \geq 0.75$, we implemented a lower threshold to allow graph to capture semantically connected events without requiring same-identity, we also increased the threshold for semantic similarity in RELATED_TO relations since related events are often connected through narrative context even when their dates or locations differ.

3.7. System Evaluation

We selected three chapters from the two resources; one chapter, 41 pages long, from The Hundred Years War on Palestine, and two chapters from The 1967 War, 16 and 15 pages respectively . The system ingested each of the chapters considering each of them as an independent source. Each of the chapters was processed, running all the elements of the system pipeline described above. Two of the researchers independently read the selected chapters before seeing the extracted events by the system. The two researchers created their own list of temporal events from the selected chapters. Each of the researcher matched the lists of human identified temporal events with the system extracted temporal events. The intersection of the two researchers' sets of identified correct temporal events was considered to be the correct set of extracted temporal events by the system, true positives. The set of commonly identified events by the researchers that were not extracted by the system were considered the missed set of extracted events or the false negatives. The events that were extracted by the system and were not identified by the researchers was deemed mistaken events, and are considered the false positives.

4. Results

Processing both resources Tarikhi extracted 944 events from 1077 text chunks, following SAME_AS clustering these events were consolidated into 341 unique events nodes, with 601 RELATED_TO edges.

4.1. System Results

4.1.1. LLM refinement

A critical stage in our system was the LLM refinement step. This stage takes on the identified temporal events from Wojood and refines the results using an LLM. This refinement layer improves incomplete or generic NER spans by providing the LLM of an input of the surrounding text

Raw NER Span	LLM-Normalized Event
اجتماع	اجتماع للجنة الكنيست لشؤون الدفاع والخارجية
تبادلٍ	تبادلٍ عنيف لإطلاق النار عبر الحدود
مؤتمر صحفيٍ	مؤتمر صحفيٍ في فندق الكومودور في بيروت لبنان

Table 2: Examples of LLM-based refinement of NER event spans

Table 2 provides examples of LLM refinement: "meeting" is refined using the context it was extracted from to "meeting of Knesset committee for foreign relations", "exchange" is refined to "heavy exchange of fire across the border", "press confer-

ence ” is refined to ”press conference in the Commodore hotel in Beirut Lebanon”

4.1.2. Structured Events

The output of the system is a set of structured events that are organized into four main attributes: event, date, location, and description.

The following examples show the final structured event representation produced by Tarikhi.

Example 1 Event: معركة ميسلون
Date: 23-07-1920
Location: Syria
Description:
معركة ميسلون كانت معركة حاسمة بين القوات العربية والفرنسية في يوليو 1920، وانتهت بانتصار الفرنسيين

The example shows the extracted event Maysalun battle, with its date and location, the description extracted is The Battle of Maysalun was a decisive battle between Arab and French forces in July 1920, ending in a French victory.

Example 2 Event: حرب السويس
Date: 29-10-1956
Location: Egypt
Description:
حرب السويس هي عملية عسكرية بدأت في 29 أكتوبر 1956 بمشاركة إسرائيل وفرنسا وبريطانيا ضد مصر عقب تأميم قناة السويس. انتهت العمليات العسكرية خلال أسابيع قليلة تحت ضغط دولي، وأعادت تشكيل موازين القوى الإقليمية.

This example's event is Suez war or as it is known in english literature, Suez crisis, with the extracted location and date. The description provided by the system is The Suez War was a military operation that began on October 29, 1956, involving Israel, France, and Britain against Egypt following the nationalization of the Suez Canal. The military operations ended within a few weeks under international pressure, and reshaped the regional balance of power.

4.1.3. Event Clusters

In some cases, the process of identifying repeated extracted events is essential to reduce the cluttering of outputted events drastically improving the functionality of the system. These events were associated to each other using the SAME_AS link creating clusters of events that represent one canonical event

Cluster A (Size = 18) Canonical Event: الثورة العربية في فلسطين 1936-1939
Representative Variants:

- ثورة عربية كبري في فلسطين
- الثورة القومية
- الاضراب العام سنة 1936
- الثورة الكبري في 1936-1939
- الانتفاضة 1936-1939
- ثورة 1936-1939

This example shows a cluster of 18 extracted events that all refer to the canonical event, the Arab Revolt in Palestine (1936-1939). The cluster includes variants such as Major Arab Revolt in Palestine, The Great Revolt (1936-1939), and The 1936-1939 uprising.

Cluster B (Size = 42) Canonical Event: حرب لبنان 1982
Representative Variants:

- الغزو الاسرائيلي للبنان
- حصار بيروت
- اجتياح لبنان
- انسحاب منظمة التحرير الفلسطينية من بيروت
- غزو لبنان سنة 1982

This example show a cluster of 42 extracted events that all represent the canonical event Lebanon war 1982. the cluster has variants like Israeli invasion of lebanon, siege of Beirut , invasion of Lebanon, the withdrwal of Palestine Liberation Organization from Beirut, invasion of lebanon 1982

4.1.4. RELATED_TO Linking Examples

Another important element in the system is identifying related events and linking them to each other. Events are identified as referring to the same event were first clustered and represented as a single node, for each pair of candidate nodes, semantic similarity, temporal proximity, and location similarity were computed and combined into a weighted relationship score. Event nodes with relationship score greater than 75% were considered to be related events using the `RELATED_TO` relation.

The example in Table 3 shows the results of RELATED_TO events linking step for one event node 1948 war where it is linked to the following events:The massive Zionist attack, the fall of Palestinian towns and villages, the political and milliary victory of Zionism, Displacement from their homeland Secret agreement to prevent the dispatching of the Arabic army, The decision to partition Palestine with relationship scores of 0.93, 0.91, 0.88, 0.88, 0.87, 0.87 respectively.

Event: حرب 1948	
Related event	**Score**
الهجوم الصهيوني الواسع	0.93
سقوط المدن والقرى الفلسطينية	0.91
انتصار الصهيونية العسكري والسياسي	0.88
النزوح عن وطنهم	0.88
اتفاقية سرية لمنع ارسال الجيش العربي عبر	0.87
قرار تقسيم فلسطين	0.87

Table 3: 1948 War Related Events

4.2. Evaluation

We selected three chapters from our resources to evaluate the system's output, one chapter (41 pages) from The Hundred Years War book and two chapters (31 pages total) from the 1967 War book. The evaluation method we used is designed to evaluate the performance of the system as a whole. The system processed the chapters independently as separate sources. The system extracted 156 initial temporal events. Out of these initially extracted events 95 canonical events were identified. Two researchers used cross over verification to identify correctly extracted temporal events vs mistaken events, which included non-events like names of individuals, or organizations, correct events associated with wrong location, or wrong date. Based on the researchers' assessment 75 temporal events were identified as correct events resulting with 78.8% recall, 72.1% precision and 75.3% F1-score.

5. Discussion

The results show that a layered RAG-based pipeline can feasibly extract structured temporal events from long-form Arabic historical texts, a task that has not been addressed before in an end-to-end manner. The 75.3% F1-score, while leaving room for improvement, is encouraging considering the difficulty of processing scanned Arabic books where OCR noise, orthographic variation, and the absence of diacritics all add to the challenges of event identification and attribute extraction. A key design choice in Tarikhi is the use of NER-based event triggering followed by LLM refinement, rather than relying on either component alone. The NER model provides structured, reproducible event spans, but often produces incomplete or overly generic names, as seen in Table 2. The LLM refinement layer addresses this by grounding those spans in their surrounding context, producing more descriptive and representative event names. This two-stage approach strikes a balance between the reliability of a trained NER model and the contextual reasoning of an LLM, though it also means that events missed entirely by the NER layer cannot be recovered downstream, since the LLM only refines what the NER has already detected.

The SAME_AS clustering proved effective in handling the repetitive nature of historical narratives, where the same event is often mentioned multiple times in different surface forms. The reduction of 944 raw event mentions to 341 canonical nodes illustrates just how much redundancy exists in long-form historical texts and highlights why consolidation matters for producing usable structured output. The RELATED_TO linking adds further value by surfacing thematic connections between events that would otherwise require extensive manual reading to uncover. That said, relying on embedding-based similarity and fuzzy string matching for linking means that events connected through implicit or causal reasoning rather than surface-level similarity may go undetected. The fixed chunking strategy with 256-token windows and 50-token overlaps worked well for the selected resources, but historical texts with varying event density across pages could benefit from adaptive chunking strategies that take narrative structure into account rather than relying on fixed token counts.

Arabic-specific challenges remain a real factor in system performance. The rich morphology of Arabic means that the same event can appear in many inflected forms, which complicates both NER detection and embedding-based similarity. OCR errors introduce additional noise that carries through every downstream component, and while the text cleaning step helps reduce some of this, it cannot recover information that was lost during recognition. These challenges point to the ongoing need for investment in Arabic-specific tools that can handle the variability found in historical documents.

6. Limitations and future work

Tarikhi was evaluated on only two Arabic historical books, both centered on modern political history in the Middle East, which limits the generalizability of the reported results to other types of historical texts that are less date and location heavy. The evaluation methodology, while practical given the absence of gold-standard annotated corpora for long-form Arabic temporal event extraction, relies on dual human verification by two researchers, introducing potential subjectivity in what constitutes a correct event, the granularity of a temporal event, an acceptable date, or a valid location. Furthermore, the evaluation assessed the complete pipeline output rather than the performance of individual components, making it difficult to isolate specific bottlenecks or failure points within

the pipeline; evaluating these components individually requires dedicated long-form annotated corpora tailored to each task, such as book-level NER annotations for event triggering assessment and ground-truth event attributes for measuring RAG-based date, location, and description extraction accuracy, neither of which currently exist for Arabic historical texts. Future work should expand the resource base to include a wider variety of Arabic historical texts spanning different eras, and narrative styles, and should develop dedicated annotated benchmark corpora that enable both component-level and end-to-end evaluation.

7. Broader Impact

Tarikhi is a system that takes on scanned Arabic resources and enables structured temporal event extraction from long-form texts like books and lengthy articles. The proposed solution enables the creation of curated knowledge bases from selected resources and provides a systematic way to access and query historical temporal data. More importantly, Tarikhi has the potential to unlock the wealth of knowledge that lives in undigitized Arabic-language books, making it computationally accessible. This accessibility opens the door to considering alternative historical narratives and including perspectives that have been largely overlooked in the digital age, contributing to a more representative and diverse digital humanities landscape.

8. Acknowledgements

The authors would like to thank Dr. Fadi Zaraket for proposing the initial idea that inspired this work and for his valuable advice and informative discussions and suggestions.

References

SILMA AI. 2024. Silma-9b-instruct-v1.0. https://huggingface.co/silma-ai/SILMA-9B-Instruct-v1.0. Large Language Model.

Abdulaziz M. Alayba. 2025. Arabic natural language processing (nlp): A comprehensive review of challenges, techniques, and emerging trends. *Computers*, 14(11).

Alaa Aljabari, Lina Duaibes, Mustafa Jarrar, and Mohammed Khalilia. 2024a. Event-arguments extraction corpus and modeling using bert for arabic.

Alaa Aljabari, Lina Duaibes, Mustafa Jarrar, and Mohammed Khalilia. 2024b. Event-arguments extraction corpus and modeling using bert for arabic. *arXiv preprint arXiv:2407.21153*.

Alaa Aljabari, Mohammed Khalilia, and Mustafa Jarrar. 2025. Wojood[Relations]: Arabic relation extraction corpus and modeling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34342–34360.

هويدي، أمين Amin Huwaidi. 2006. *Harb 1967: Asrar wa Khabaya* , حرب 1967 : أسرار و خبايا. Al-Maktab Al-Misri Al-Hadith, المكتب المصري الحديث, Cairo, Egypt.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection (OSACT 2020)*, pages 9–15. OSACT 2020.

M. H. Bashir, A. M. Azmi, H. Nawaz, et al. 2023. Arabic natural language processing for qur'anic research: A systematic review. *Artificial Intelligence Review*, 56:6801–6854.

Steven Bethard, Leon Derczynski, Guergana Savova, James Pustejovsky, and Marc Verhagen. 2015. SemEval-2015 task 6: Clinical TempEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 806–814.

Tarik Boudaa, Mohamed El Marouani, and Nourddine Enneya. 2018. Arabic temporal expression tagging and normalization. In *Big Data, Cloud and Applications (BDCA 2018)*, volume 872 of *Communications in Computer and Information Science*, pages 271–284. Springer.

Bibo Cai, Xiao Ding, Zhouhao Sun, Bing Qin, Ting Liu, Baojun Wang, and Lifeng Shang. 2023. Self-supervised logic induction for explainable fuzzy temporal commonsense reasoning. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Surya OCR Contributors. 2024. Surya ocr: Multilingual document ocr toolkit. https://github.com/datalab-to/surya. Open-source OCR toolkit supporting 90+ languages.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R. El-Beltagy, Wassim El-Hajj,

Mustafa Jarrar, and Hamdy Mubarak. 2021. A panoramic survey of natural language processing in the arab world. *Commun. ACM*, 64(4):72–81.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Jennifer D'Souza and Vincent Ng. 2014. Annotating inter-sentence temporal relations in clinical notes. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2758–2765.

Michele Filannino and Goran Nenadić. 2015. Temporal expression extraction with extensive feature type selection and a posteriori label adjustment. *Data & Knowledge Engineering*, 100:19–33.

Abbas Ghaddar, Yimeng Wu, Ahmad Rashid, Khalil Bibi, Mehdi Rezagholizadeh, Chao Xing, Yasheng Wang, Duan Xinyu, Zhefeng Wang, Baoxing Huai, Xin Jiang, Qun Liu, and Philippe Langlais. 2022. Jaber and saber: Junior and senior arabic bert. *arXiv preprint arXiv:2112.04329*.

Nafaa Haffar, Emna Hkiri, and Mounir Zrigui. 2020. Enrichment of arabic timeml corpus. In *Computational Collective Intelligence (ICCCI 2020)*, volume 12496 of *Lecture Notes in Computer Science*, pages 233–244. Springer.

Tymaa Hammouda, Mustafa Jarrar, and Mohammed Khalilia. 2024. Sinatools: Open source toolkit for arabic natural language processing. *Procedia Computer Science*, 244:388–396. 6th International Conference on AI in Computational Linguistics.

Kuan-Hao Huang, I-Hung Hsu, Tanmay Parekh, Zhiyu Xie, Zixuan Zhang, Premkumar Natarajan, Kai-Wei Chang, Nanyun Peng, and Heng Ji. 2024. Textee: Benchmark, reevaluation, reflections, and future challenges in event extraction.

Mustafa Jarrar, Mohammed Khalilia, and Sana Ghanem. 2022. Wojood: Nested arabic named entity corpus and recognition using bert. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, pages 3626–3636.

Rashid Khalidi. 2021. . , . Arabic translation of "The Hundred Years' War on Palestine".

Panteleimon Krasadakis, Evangelos Sakkopoulos, and Vassilios S. Verykios. 2024. A survey on challenges and advances in natural language processing with a focus on legal informatics and low-resource languages. *Electronics*, 13(3).

Artuur Leeuwenberg and Marie-Francine Moens. 2018. Temporal information extraction by predicting relative time-lines. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1237–1246.

Chahira Lhioui, Anis Zouaghi, and Mounir Zrigui. 2017. A rule-based approach for arabic temporal expression extraction. In *2017 International Conference on Engineering MIS (ICEMIS)*, pages 1–6.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016. CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61.

Qiang Ning, Sanjay Subramanian, and Dan Roth. 2019. An improved neural baseline for temporal relation extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6203–6209.

James Pustejovsky, José M. Castaño, Robert Ingria, Roser Saurí, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. In *New Directions in Question Answering*.

Iman Saleh, Lamia Tounsi, and Josef van Genabith. 2011. Zaman and raqm: Extracting temporal and numerical expressions in arabic. In *Information Retrieval Technology*, pages 562–573.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. SemEval-2007 task 15: TempEval temporal relation identification.

In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. ArXiv:2402.05672v1 [cs.CL, cs.IR], Technical Report, submitted 2024-02-08.

Ahmed Wasfy, Omer Nacar, Abdelakreem Elkhateb, Mahmoud Reda, Omar Elshehy, Adel Ammar, and Wadii Boulila. 2025. Qari-ocr: High-fidelity arabic text recognition through multimodal large language model adaptation. *arXiv preprint arXiv:2506.02295*.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*.

Georg Wenzel and Adam Jatowt. 2023. An overview of temporal commonsense reasoning and acquisition. *arXiv preprint arXiv:2307.15394*.

Kam-Fai Wong, Yunqing Xia, Wenjie Li, and Chunfa Yuan. 2005. An overview of temporal information extraction. *International Journal of Computer Processing of Languages*, 18(2):219–240.

Fadi A. Zaraket. 2012. Arabic temporal entity extraction using morphological analysis.

Xiaobin Zhang, Liangjun Zang, Qianwen Liu, Shuchong Wei, and Songlin Hu. 2024. Event temporal relation extraction based on retrieval-augmented on llms.

Y. Zhang and Q. Han. 2025. Enhancing pre-trained language model by answering natural questions for event extraction. *Frontiers in Artificial Intelligence*, 8:1520290.

NAKBA NLP 2026: Shared Task on Arabic Handwritten Manuscript Understanding (Palestine Memory–Omar Al-Saleh Memoir)

Hadi Hamoud[1], Ahmad Ali Chamseddine[2], Bilal Shalash[1], Firas Ben Abid[3], Mustafa Jarrar[4], Chadi Abou Chakra[1], Bernard Ghanem[5], Fadi A. Zaraket[1]
[1]Arab Center for Research and Policy Studies, Doha, Qatar
[2]Doha Institute, Doha, Qatar
[3]Zinki AI
[4]Hamad Bin Khalifa University, Doha, Qatar
[5]King Abdullah University of Science and Technology (KAUST), Saudi Arabia
{hhamoud, bshalash, cabouchakra, fzaraket}@dohainstitute.edu.qa
achamsed@dohainstitute.edu.qa, firas.ben.abid@zinki.ai
mjarrar@hbku.edu.qa, bernard.ghanem@kaust.edu.sa

Abstract

Transcribing historical Arabic manuscripts into machine-readable text is essential for preserving cultural heritage and enabling computational research in the humanities, yet it remains a challenging task due to handwriting variability, page degradation, and the complexity of Arabic script. To advance research in this area, we introduce the NAKBA NLP 2026 shared task on Arabic manuscript understanding, comprising two complementary tracks: a manual transcription track, in which participating teams annotate unlabelled handwritten line images, and an automatic system track for handwritten text recognition (HTR). Both tracks use the Omar Al-Saleh Memoir Collection, a corpus of 6,395 scanned pages and approximately 1.6 million words, written between 1951 and 1965 and provided by the Palestine Memory Project. The dataset, evaluation scripts, and system outputs are publicly available. In Subtask 1 (Transcription Track), three teams contributed manual line-level transcriptions; evaluation on hidden ground-truth samples yielded Character Error Rates (CER) between 0.06 and 0.11. In Subtask 2 (Systems Track), seven teams submitted HTR systems. The top-performing system, by Misraj AI, achieved a corpus-level CER of 0.079 and Word Error Rate (WER) of 0.244, outperforming the organiser baseline (CER 0.368, WER 0.691). Rankings shift between corpus-level and per-line evaluation: the Ketaba OCR team achieved the lowest per-line CER (0.082). All contributed transcriptions and system outputs are released under CC-BY-4.0 to support continued research in Arabic manuscript recognition and digital humanities.

Keywords: Arabic manuscripts, handwritten text recognition, shared task, dataset, digital humanities

1. Introduction

Arabic historical manuscripts constitute a central source for research in history, politics, and the digital humanities, providing first-hand accounts of social, intellectual, and political life across the modern Arab world (Alrobah and Alzahrani, 2022; Saeed et al., 2024). However, much of this material remains inaccessible to computational analysis, as it is preserved primarily in handwritten form and lacks structured digital representations. Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) offer pathways to large-scale digitisation and search over such collections. However, Arabic Manuscript Handwritten Text Recognition (AMHTR) remains challenging due to handwriting variability, page degradation, complex layouts, and Arabic-specific natural language processing requirements such as short vowel (diacritic) omission and rich morphology (Graves et al., 2006; Mahmoud et al., 2014).

Progress in AMHTR is currently limited by two key gaps. First, publicly available, well-annotated datasets that reflect the characteristics of historical Arabic handwriting—including diversity of writing styles, ink quality, and page structure—remain scarce (Saeed et al., 2024). Second, the NLP community lacks established benchmarks and competitive evaluations targeting AMHTR specifically. This stands in contrast to the well-developed ecosystem of shared tasks for Latin-script historical document recognition, such as the ICDAR HTR competitions (Romero et al., 2012) and the RASM series (Clausner et al., 2018).

Recent advances in vision–language models and document analysis pipelines suggest that modern approaches, when combined with layout segmentation and domain-specific annotation, may yield viable solutions for manuscript recognition even in low-resource settings (Li et al., 2023; Bai et al., 2023). However, realising this potential requires open datasets, standardised evaluation protocols, and community engagement through competitive shared tasks.

To address these gaps, we introduce NAKBA NLP 2026, a shared task for Arabic manuscript understanding organised as part of the second International Workshop on Nakba Narratives as Lan-

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 70–79
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

guage Resources (Jarrar et al., 2026). The shared task is built on the Omar Al-Saleh Memoir Collection (Institute for Palestine Studies, a; Arab Center for Research and Policy Studies, 2024), which comprises 6,395 scanned manuscript pages written between 1951 and 1965, with line-level annotations for a curated subset. The shared task is organised into two complementary tracks. **Subtask 1 (Transcription Track)** invites teams to produce expert-quality manual transcriptions of unlabelled handwritten line images, enriching the available ground truth for future research. **Subtask 2 (Systems Track)** challenges teams to develop automatic AMHTR systems, evaluated using Character Error Rate (CER) and Word Error Rate (WER).

The main contributions of this shared task are: (1) the public release of a large-scale, line-level annotated Arabic manuscript dataset with standardised train, development, and test splits; (2) a dual-track evaluation framework that combines manual transcription quality assessment with automatic system benchmarking; and (3) an empirical analysis of system performance, error patterns, and the relationship between corpus-level and per-line evaluation metrics.

In Subtask 1, three teams contributed manual transcriptions, achieving CER between 0.06 and 0.11 on hidden ground-truth samples, demonstrating both the feasibility and the difficulty of expert Arabic manuscript transcription. In Subtask 2, all seven participating systems outperformed the baseline. The top-performing system, by Misraj AI, reached a corpus-level CER of 0.079 and WER of 0.244, compared with 0.368 and 0.691 for the baseline. Rankings differ between corpus-level and per-line evaluation: Ketaba OCR obtained the lowest per-line CER (0.082), while Misraj AI led on corpus-level CER.

The remainder of this paper is organised as follows. Section 2 reviews related work. Section 3 describes the dataset. Section 4 presents the shared task design. Section 5 details the baseline system. Section 6 summarises participating systems. Section 7 reports results and analysis. Section 8 concludes with future directions.

2. Related Work

This section situates the present shared task within three areas of related work: existing AMHTR datasets, OCR and HTR model architectures, and prior shared tasks in document recognition. We highlight the gaps that motivated the design of NAKBA NLP 2026.

AMHTR Datasets. Several datasets have been developed for Arabic handwriting recognition, though they vary considerably in scale, annotation granularity, and public availability. The KHATT dataset (Mahmoud et al., 2014) provides 1,000 handwritten Arabic paragraphs from 1,000 writers, offering writer diversity but limited historical manuscript coverage. IFN/ENIT (Pechwitz et al., 2002) targets handwritten Tunisian city names, restricting its scope to isolated word recognition. The MADCAT corpus (Cieri et al., 2016), developed under the DARPA programme, offers annotated Arabic handwriting but imposes restrictive licensing that limits its use in open benchmarks. More recently, the Muharaf dataset (Saeed et al., 2024) has become the largest publicly available AMHTR collection, with over 36,000 line images from 1,600 historical manuscript pages spanning diverse document types. The VML-HD dataset (Kassis et al., 2017) provides historical document images, and recent surveys (Alrobah and Alzahrani, 2022) catalogue additional resources.

The Omar Al-Saleh Memoir dataset presented in this task differs from these prior resources in two respects. First, it provides large-scale, multi-year manuscript data from a single author with line-level annotations aligned to high-resolution page images, enabling the study of temporal variation in handwriting within a longitudinal collection. Second, with approximately 6,395 pages and 1.6 million words, it is substantially larger than most existing AMHTR corpora.

OCR and HTR Models. The trajectory of Arabic HTR has moved from traditional Hidden Markov Models and recurrent neural networks with Connectionist Temporal Classification (CTC) decoding (Graves et al., 2006) towards transformer-based architectures. TrOCR (Li et al., 2023) combines a vision transformer encoder with a language model decoder and has been adapted for Arabic script. Large vision–language models such as Qwen-VL (Bai et al., 2023), InternVL (Chen et al., 2024), and Florence-2 (Xu et al., 2024) have demonstrated multilingual OCR capabilities, though their application to historical Arabic manuscripts remains underexplored. These models provide the architectural foundation for several systems submitted to the present shared task, as described in Section 6.

Shared Tasks in Document Recognition. The ICDAR competitions on handwritten text recognition have established benchmarks for Latin, German, and Chinese scripts (Romero et al., 2012) but have not included dedicated Arabic manuscript tracks. The ICDAR Arabic Handwriting Recognition Competition series (2005–2011) focused on word-level recognition of modern handwritten town names using the IFN/ENIT database, rather than full-line transcription from historical documents. The RASM competitions (RASM2018 at ICFHR and

RASM2019 at ICDAR) targeted page-level recognition of historical Arabic scientific manuscripts from the British Library/Qatar Digital Library, evaluating both layout analysis and text recognition in PAGE format (Clausner et al., 2018). The OpenITI project (Romanov et al., 2023) has contributed OCR tools for classical Arabic texts, primarily targeting printed rather than handwritten material.

NAKBA NLP 2026 complements these efforts by providing a line-level AMHTR benchmark focused on single-author memoir transcription with pre-segmented line images and standardised CER/WER evaluation.

Crowdsourced and Manual Transcription. Manual transcription campaigns, such as those organised by the Transkribus platform (Muehlberger et al., 2019) and the READ project, have demonstrated the value of combining human expertise with computational tools for historical document analysis. Subtask 1 builds on this tradition by organising competitive manual transcription with standardised evaluation, providing a model for future Arabic manuscript annotation efforts.

3. Dataset

The dataset used in the shared task is derived from the Omar Al-Saleh Memoir Collection, a corpus of handwritten Arabic manuscripts written between 1951 and 1965. Omar Al-Saleh Al-Barghouti (1894–1965) was a Palestinian-Jordanian politician, lawyer, and author. Born in Deir Ghassaneh near Ramallah, he was active in the Palestinian national movement from 1919, served as a member of the Jordanian Parliament after the union of the two banks, held the position of Minister of Education in 1955 and again in 1959, and co-authored *Tarikh Filastin* (History of Palestine) with Khalil Totah (Al-Barghouti and Totah, 1923; Institute for Palestine Studies, a). His memoirs offer a first-person account of political, social, and personal life during a formative period of modern Palestinian and Arab history.

The dataset is provided by the Arab Center for Research and Policy Studies[1] as part of their work on the Palestine Memory Project[2] (Arab Center for Research and Policy Studies, 2024). The memoir collection is also archived by the Institute for Palestine Studies as part of the Palestine Social History Archives (Institute for Palestine Studies, b). It serves as an initial component in a broader sequence of curated Arabic manuscript corpora belonging to historical authors and figures.

[1] https://acr.ps

[2] https://palestine-memory.org/

3.1. Corpus Statistics

Table 1 provides per-document statistics. The full dataset comprises 6,395 scanned pages. After transcription and aggregation, it contains 1,597,025 words distributed across 50,685 sentences, with 161,394 unique word types and an overall type–token ratio of 0.1011. The average sentence length is 31.5 words. Substantial variation in document length is observed: the largest document (1959) contains 132,858 words across 532 pages, while the smallest (1957b) contains 39,985 words across 160 pages, reflecting differences in the author's narrative style and memoir writing activity across years.

Word and sentence counts were computed automatically using a Python-based text analysis pipeline. Words were extracted using regular expressions matching Arabic and Latin character sequences (single characters excluded), and sentences were identified by splitting on sentence-ending punctuation marks, including the Arabic question mark (؟) and the standard period, exclamation mark, and question mark.

Documents labelled with a "b" suffix (1957b, 1961b) represent separate notebooks written by the same author in the same year. These are treated as distinct documents to preserve the physical organisation of the collection.

Doc.	Pages	Words	Sent.
1951	475	118,742	2,470
1953	451	112,745	2,968
1954	454	113,482	3,105
1955	421	105,192	2,936
1956	328	81,780	2,197
1957	470	117,336	3,824
1957b	160	39,985	1,231
1958	342	85,341	2,597
1959	532	132,858	4,010
1960	526	131,363	4,594
1961	463	115,569	4,156
1961b	242	60,401	3,052
1962	466	116,257	3,846
1963	353	88,181	3,014
1964	504	125,837	4,689
1965	208	51,956	1,996
Total	**6,395**	**1,597,025**	**50,685**

Table 1: Omar Al-Saleh Memoir statistics. Documents with a "b" suffix denote separate notebooks from the same year.

3.2. Physical Characteristics

The manuscripts are digital photocopies of handwritten notebooks, with each scanned image typically containing two facing pages. Images are stored in JPEG format at a resolution of 1982×1400

pixels with a nominal resolution of 96 DPI. The manuscripts exhibit several features that make automated processing challenging: mixed layouts with headers, body text, and marginal notes; non-uniform line spacing that varies both within and across pages; handwriting variability in letter forms, ligatures, and word spacing; occasional overwritten or crossed-out words; and rare non-Arabic elements such as Western numerals or Latin-script annotations.

3.3. Annotation and Processing Pipeline

Layout Segmentation. Because each scanned image typically contains two facing manuscript pages, layout segmentation is applied first to separate the text into distinct regions. A YOLO11-based object detection model (Jocher et al., 2024; Khanam and Hussain, 2024), fine-tuned on manuscript page images, is used to detect the primary text regions, which are organised into four areas: left header, left body, right header, and right body. Detection accuracy was validated against manually annotated bounding boxes on a sample of 100 pages, achieving an Intersection over Union (IoU) above 0.92 for body regions. IoU measures the overlap between the predicted and ground-truth bounding boxes; a value of 0.92 indicates close alignment between detected and actual text areas. This metric applies to layout detection only and does not reflect downstream transcription accuracy.

Line Segmentation. To isolate individual text lines within each detected region, manuscript-specific line segmentation models are applied independently to each area. These models produce pixel-level masks for each text line. The masks are then converted into polygonal annotations that define the spatial extent of each line, and these polygons are used to extract cropped line images. The resulting line images serve as input for both manual transcription and automatic HTR. No automatic transcription system was used during the annotation process; all transcriptions were produced manually.

Manual Annotation. For each of the 16 manuscript documents, the first 15 pages are annotated at the line level, yielding 240 annotated pages and approximately 41,000 line-level annotations. Three trained annotators each transcribed an assigned portion of the data following standardised orthographic conventions covering normalisation rules, punctuation handling, and diacritics. Each annotator's transcriptions were then reviewed by the other two annotators. During this cross-review phase, normalised edit distance was computed between each original transcription and the reviewer's corrections; lines where the edit distance exceeded a predefined threshold were flagged for discussion and adjudication among the annotators. This procedure ensured consistency without requiring redundant independent annotation of the same lines. An additional 9,009 lines were subsequently manually annotated to enlarge the test set.

3.4. Data Splits

The annotated data is divided into training, development, and test sets as follows:

- **Training set**: 15,962 line images with gold transcriptions.
- **Development set**: 1,774 line images with gold transcriptions.
- **Test set**: 2,671 line images (transcriptions held out for evaluation).

3.5. Dataset Structure and Format

The dataset is organised hierarchically by document, page, region, and line. Each page is associated with metadata describing document origin, year, image dimensions, and filename. All annotations are stored in a JSON-based format preserving geometric information and reading order. For example, a single line annotation includes the document year, page number, region identifier (e.g., `left_body`), line number, polygonal coordinates defining the line boundary, and the corresponding transcription text. For Subtask 2, the data is also provided in CSV format, with each row containing the cropped line image filename and the gold transcription (empty for the test set).

4. Shared Task Design

4.1. Subtask 1: Transcription Track

Objective. This track focuses on producing expert-quality, line-by-line manual transcriptions of unlabelled manuscript line images, enriching the benchmark with reliable ground truth for training and evaluating HTR systems.

Data. Each team receives one mandatory batch of approximately 500 cropped line images from the Omar Al-Saleh Memoir Collection, together with an `annotations.csv` template containing columns for filename, text (to be filled), source page image reference, year, page number, and line number. In addition to the unlabelled images, each batch includes 50 pre-labelled samples drawn from the existing training dataset. These samples are embedded among the unlabelled images and serve

as hidden quality checkpoints: participants are not informed which samples carry ground-truth annotations, and evaluation is performed on these hidden samples. The full unannotated page images are also provided so that annotators can view the surrounding context when transcribing individual cropped lines. Teams that wish to contribute additional transcriptions may request access to further batches.

Requirements. Transcriptions must be manual and line-aligned (one transcription per image). Participants must follow the provided orthographic conventions covering normalisation rules, punctuation handling, and diacritics. No generative AI tools may be used to generate or correct transcriptions, whether fully or partially. Each team must submit an annotation guidelines document describing their procedures, handling of ambiguous or damaged text, consistency measures, and any ethical considerations.

Evaluation. Transcription quality is assessed along two dimensions: coverage and accuracy. *Coverage* is measured as the number of lines with valid Arabic transcriptions out of the 500 assigned. A transcription is considered valid if it contains at least one Arabic character and is not composed solely of non-Arabic content such as dashes, Western numerals, or Latin text. *Accuracy* is computed exclusively on the hidden pre-labelled samples for which verified ground truth is available (up to 50 per batch), using CER and WER as defined in Section 4.2. Only hidden samples that received a valid Arabic transcription contribute to the accuracy computation.

4.2. Subtask 2: Systems Track

Objective. This track requires participants to develop automatic systems that transcribe Arabic manuscript line images into machine-readable text. Participants may employ vision-based models, sequence-to-sequence architectures, or multimodal frameworks, and may fine-tune on the provided data or explore zero- and few-shot strategies.

Data. Participants are provided with the Omar Al-Saleh Memoir data using the splits described in Section 3. Each split contains cropped line images in JPEG format and an `annotations.csv` with image filenames and transcriptions (empty for the test set). The full unannotated page images are also available to support approaches such as page-level context modelling or self-supervised learning.

Evaluation Metrics. System performance is evaluated using Character Error Rate (CER) and Word Error Rate (WER), computed via Levenshtein edit distance, where edits comprise substitutions, insertions, and deletions. These metrics are reported in two complementary ways.

First, a *corpus-level* score sums all edit operations across the test set and divides by the total number of reference units (characters for CER, words for WER), as defined in Equations 3 and 4. Second, a *per-line* score computes the metric independently for each test line (Equations 1 and 2) and averages across all lines. Both variants are clipped to $[0, 1]$ and computed on the held-out test set of 2,671 lines. All results throughout this paper are reported on the $[0, 1]$ scale.

Formally, for a reference string r and hypothesis h, where d_{char} and d_{word} denote character-level and word-level edit distances respectively:

$$\text{CER}(r, h) = \frac{d_{\text{char}}(r, h)}{|r|} \tag{1}$$

$$\text{WER}(r, h) = \frac{d_{\text{word}}(r, h)}{|\text{split}(r)|} \tag{2}$$

At corpus level, if R_i and H_i denote the reference and hypothesis for the i-th test sample:

$$\text{CER}_{\text{c}} = \frac{\sum_i d_{\text{char}}(R_i, H_i)}{\sum_i |R_i|} \tag{3}$$

$$\text{WER}_{\text{c}} = \frac{\sum_i d_{\text{word}}(R_i, H_i)}{\sum_i |\text{split}(R_i)|} \tag{4}$$

Per-line variants are the arithmetic means of sample-level CER and WER over all evaluated lines. CER serves as the primary ranking metric, with corpus-level scores used for the official leaderboard.

5. Baseline System

The baseline system uses Qwen3-VL-8B-Instruct (Bai et al., 2023), a large open-weight vision–language model with demonstrated multilingual and OCR capabilities. We selected this model as a representative baseline because it allows us to measure the benefit of task-specific adaptation: the pretrained model has general document understanding ability but has not been trained on Arabic historical manuscripts.

The baseline approach consists of two steps. First, the pretrained model is adapted to the manuscript domain via parameter-efficient fine-tuning with LoRA, targeting the attention and feed-forward modules. This keeps the approach compute-efficient while preserving the pretrained representations. Second, the adapted model is used to generate transcriptions: given a cropped manuscript line image and a transcription instruction, the model produces the text directly. We de-

Component	Setting
Backbone	Qwen3-VL-8B-Instruct
Fine-tuning method	LoRA
LoRA rank / α / dropout	32 / 64 / 0.05
LoRA target modules	q,k,v,o,gate,up,down
Precision	bfloat16
Attention mechanism	FlashAttention-2
Epochs	2
Learning rate	2×10^{-5} (cosine)
Batch size / grad. accum.	24 / 1
Weight decay	0.01
Warmup ratio	0.03
Max seq. length	256 tokens
Max image resolution	1536 px (longest side)
Decoding	greedy, max 512 tokens

Table 2: Baseline configuration. The model is fine-tuned with LoRA on the shared-task training data.

code greedily to ensure deterministic, reproducible outputs.

During training, each example is formatted as a dialogue: a user turn containing the line image and instruction, followed by an assistant turn containing the gold transcription. The training objective is next-token prediction on the assistant response, with prompt tokens excluded from the loss. Images are resized with aspect ratio preservation. The same instruction prompt is used at both training and inference time.

Table 2 lists the full configuration. The baseline achieves a corpus-level CER of 0.368 and WER of 0.691 on the test set (2,671 lines), establishing a reference point for evaluating the benefit of more specialised system development.

6. Participating Systems

Table 3 summarises all participating teams, their affiliations, and the subtask(s) they entered.

Team	Affiliation	Subtask
PalNLP	Cardiff University	1
Sard	Arab Open University	1
Independent	Istanbul University	1
Misraj AI	Misraj AI	2
Oblevit	Arab Open University	2
Ketaba OCR	[to be confirmed]	2
Latent Narratives	Arab Open University	2
Al-Warraq	Arab Open University	2
Not Gemma	Helwan University	2
Fahras	Thakaa	2

Table 3: Participating teams, affiliations, and subtask(s).

6.1. Subtask 1 Participants

Three teams submitted transcriptions for Subtask 1. Teams varied in their handling of ambiguous or damaged text, diacritics, and lines containing non-Arabic content such as dates in Western numerals or standalone digits.

6.2. Subtask 2 Participants

Seven teams submitted systems for Subtask 2. The systems can be grouped by their underlying model families. Four teams built their systems around large vision–language models, primarily Qwen2-VL and InternVL2, adapted to the manuscript domain via LoRA fine-tuning. Three teams used encoder–decoder architectures based on TrOCR, with Arabic-specific tokenisation and data augmentation. Two teams additionally employed ensemble strategies that combined outputs from multiple models using character- or word-level voting. Data augmentation was common among competitive submissions, with elastic distortion, random erosion, contrast jittering, and additive noise being the most frequently used techniques. Three teams incorporated external Arabic text corpora for language model rescoring or decoder pre-training.

7. Results and Analysis

7.1. Subtask 1: Transcription Track Results

Table 4 summarises the coverage and accuracy results for all three participating teams. Coverage is measured as the number of lines receiving valid Arabic transcriptions out of 500, while CER and WER are computed only on the hidden pre-labelled samples that received valid annotations.

Team	Transcr.	Compl.	Test N	CER	WER
PalNLP	499	99.8%	49/50	**0.061**	**0.285**
Independent	497	99.4%	47/50	0.086	0.366
Sard	401	80.0%	39/50	0.114	0.433

Table 4: Subtask 1 results. Transcr. = lines with valid Arabic text. Compl. = completion rate. Test N = hidden pre-labelled samples with valid annotations out of 50.

PalNLP achieved the highest transcription quality with a CER of 0.061 and near-complete coverage (499 out of 500 lines). Independent Arabic Manuscript Transcription demonstrated comparable coverage (497 out of 500) and a CER of 0.086. Sard completed 401 out of 500 lines (80.0%), with 97 entries consisting solely of a dash character and 2 left blank; its CER of 0.114 on the annotated lines

indicates reasonable quality where transcriptions were provided.

Analysis. Across all teams, CER ranged from 0.061 to 0.114, confirming that manual transcription of Arabic manuscripts remains difficult even for trained human annotators, particularly for lines with degraded ink, ambiguous characters, or non-standard orthography. The variation in coverage highlights the need for clearer guidelines on handling non-Arabic content within manuscript pages.

Because each team transcribed a different batch, formal inter-annotator agreement scores (e.g., Cohen's kappa) cannot be computed across teams. Within the annotation pipeline (Section 3), cross-review among the three annotators and edit-distance-based flagging provided a quality control mechanism. Future iterations could assign overlapping subsets to multiple teams to enable direct comparison of annotator consistency.

The most common error patterns in manual transcriptions include confusion between visually similar characters such as *dal*/*dhal* (د/ذ), *sin*/*shin* (س/ش), and *ha*/*ta marbuta* (ه/ة); omission or misplacement of diacritical dots; and inconsistent handling of word boundaries in passages with minimal inter-word spacing.

7.2. Subtask 2: Corpus-Level Results

Table 5 reports the corpus-level CER and WER for all seven teams and the baseline on the held-out test set of 2,671 lines. CER is the primary ranking metric.

#	Team	CER	WER
1	Misraj AI	**0.079**	**0.244**
2	Oblevit	0.093	0.327
3	Ketaba OCR	0.094	0.300
4	Latent Narratives	0.105	0.311
5	Al-Warraq	0.114	0.378
6	Not Gemma	0.122	0.306
7	Fahras	0.227	0.522
	Baseline	*0.368*	*0.691*

Table 5: Subtask 2 corpus-level results (lower is better). All seven teams outperform the baseline.

All seven submitted systems outperformed the organiser baseline (CER 0.368, WER 0.691). Misraj AI achieved the lowest corpus-level CER of 0.079, a 78.5% relative reduction over the baseline, and the lowest WER of 0.244. The top six teams form a competitive cluster with CER between 0.079 and 0.122, separated by a gap from Fahras (0.227).

The CER and WER rankings do not always align. Oblevit ranks second by CER (0.093) but has the fourth-highest WER (0.327) among the top six, suggesting that its character-level errors are disproportionately concentrated at word boundaries. Conversely, Ketaba OCR ranks third by CER (0.094) but achieves a lower WER (0.300) than Oblevit, indicating better word-level coherence. Not Gemma (rank 6 by CER at 0.122) achieves a lower WER (0.306) than four higher-ranked teams, suggesting that its errors are distributed across characters rather than concentrated in patterns that destroy entire words.

7.3. Subtask 2: Per-Line Results

Table 6 presents the per-line CER and WER, where metrics are computed independently for each test line and averaged.

#	Team	CER	WER
1	Ketaba OCR	**0.082**	0.259
2	Misraj AI	0.090	**0.252**
3	Latent Narratives	0.100	0.285
4	Oblevit	0.105	0.332
5	Al-Warraq	0.106	0.347
6	Not Gemma	0.110	0.313
7	Fahras	0.182	0.430
	Baseline	*0.281*	*0.588*

Table 6: Subtask 2 per-line results, averaged over individual test lines (lower is better).

The per-line evaluation reveals ranking changes compared to corpus-level scoring. Most notably, Ketaba OCR rises from rank 3 (corpus) to rank 1 (per-line), achieving a per-line CER of 0.082, while Misraj AI moves from rank 1 to rank 2 (0.090). Oblevit drops from rank 2 (corpus) to rank 4 (per-line), while Latent Narratives improves from rank 4 to rank 3.

These shifts arise from the different weighting implicit in each metric. Corpus-level scoring weights each character equally, so longer lines contribute more to the aggregate. Per-line scoring weights each line equally regardless of length, giving more influence to shorter lines. The rank reversal between Misraj AI and Ketaba OCR suggests that Misraj AI performs better on longer body-text lines, while Ketaba OCR handles short and irregular lines more consistently. This distinction has practical implications: for applications requiring uniform transcription quality across an entire manuscript page—including headers, marginalia, and fragmented lines—per-line CER may be the more informative metric.

7.4. Error Analysis

We conducted a qualitative error analysis on 200 randomly sampled test lines from the top-ranked

system (Misraj AI, corpus-level). Errors were categorised into five types.

Character confusion (38% of errors). The most frequent error type involves visually similar characters, particularly confusions between *dal*/*dhal* (د/ذ), *sin*/*shin* (س/ش), and *ha*/*ta marbuta* (ه/ة). These confusions arise from the author's handwriting style, in which distinguishing dots are occasionally faint or absent.

Word boundary errors (24%). Incorrect segmentation of connected words or splitting of single words accounted for approximately one quarter of errors. This pattern is particularly common in passages where the author writes with minimal inter-word spacing.

Diacritics and dots (18%). Missing or misplaced diacritical dots affect the identity of several Arabic letters. Systems frequently omit dots that are faintly written or shifted from their canonical position relative to the baseline character.

Damaged or degraded regions (12%). Ink bleeding, page fold marks, and photocopy artefacts cause localised failures. These errors are concentrated in the earlier documents (1951–1953), where physical preservation is poorest.

Rare vocabulary and proper nouns (8%). Names of people, places, and uncommon political terminology that appear infrequently in the training data are more likely to be misrecognised, as language model priors cannot compensate for visual ambiguity in these cases.

7.5. Discussion

Baseline versus participants. Although the organiser baseline and several participating teams use the same model family (Qwen-VL variants), all seven teams achieved lower CER than the baseline. This indicates that differences in data augmentation, training strategy, prompt design, hyperparameter tuning, and ensemble methods yield substantial improvements even when the underlying architecture is shared. The gap between the baseline (CER 0.368) and the best system (CER 0.079) represents a 78.5% relative CER reduction.

Corpus-level versus per-line evaluation. The ranking shifts between Tables 5 and 6 underscore the importance of reporting both metrics. Corpus-level CER captures aggregate transcription fidelity and is appropriate for applications such as full-text search, while per-line CER better reflects consistency across heterogeneous line types. Future shared tasks should report both metrics, as they capture complementary aspects of system quality.

8. Conclusion and Future Work

The NAKBA NLP 2026 shared task engaged participants across two complementary tracks, advancing both the available ground truth and the state of automatic transcription for Arabic manuscripts. Beyond the performance results, the shared task makes several contributions to the field: the public release of a large-scale, line-level annotated Arabic manuscript corpus; a standardised evaluation framework combining manual transcription assessment with automatic system benchmarking; and baseline results that can serve as reference points for future research.

On the systems side, all seven participating teams outperformed the organiser baseline, with the top-performing system (Misraj AI) reducing corpus-level CER from 0.368 to 0.079. Rankings shift between corpus-level and per-line evaluation—Ketaba OCR achieves the lowest per-line CER (0.082) while Misraj AI leads on corpus-level CER—revealing that different systems have different strengths across line types. On the transcription side, manual annotation by trained annotators achieved CER between 0.061 and 0.114, confirming that Arabic manuscript transcription remains challenging even for human annotators. The divergence between CER and WER rankings across teams reveals qualitative differences in error patterns with implications for downstream applications.

All contributed transcriptions and system outputs are released under CC-BY-4.0 via the shared task page.[3] Future iterations will expand the dataset to additional authors and genres, introduce page-level and paragraph-level evaluation, explore cross-author generalisation, and investigate strategies for increasing participation in the manual transcription track.

Acknowledgements

We thank all participating teams for their contributions and the reviewers for their constructive feedback. We gratefully acknowledge the Arab Center for Research and Policy Studies for providing the

[3] https://acr.ps/1L9BaeY

Omar Al-Saleh Memoir Collection, and the annotators whose careful work produced the ground-truth transcriptions used in this shared task.

Ethical Considerations

The Omar Al-Saleh Memoir Collection is used with the permission of the Arab Center for Research and Policy Studies. The memoirs contain personal and political reflections from a specific historical period; participating teams were advised to handle the content with appropriate scholarly care. All teams agreed to ethical use guidelines, and no generative AI was permitted for manual transcription in Subtask 1.

Limitations

The dataset represents a single author's handwriting, which limits conclusions about generalisability to other Arabic manuscripts or scripts. Subtask 1 attracted only three participants, limiting the statistical conclusions that can be drawn about manual transcription quality. Because each Subtask 1 team transcribed a different batch, formal inter-annotator agreement metrics cannot be computed across teams. Evaluation is restricted to line-level CER and WER; document-level or page-level metrics may better capture transcription quality for downstream use.

References

O. S. Al-Barghouti and K. Totah. 1923. *Tarikh Filastin [History of Palestine]*. Bayt al-Maqdis Press, Jerusalem.

N. Alrobah and S. Alzahrani. 2022. Arabic handwritten text extraction: A survey. *IEEE Access*, 10:50437–50465.

Arab Center for Research and Policy Studies. 2024. Palestine Memory Project.

J. Bai, S. Bai, S. Yang, et al. 2023. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Z. Chen et al. 2024. InternVL: Scaling Up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. *arXiv preprint arXiv:2312.14238*.

C. Cieri, M. Maamouri, S. Strassel, et al. 2016. MADCAT: Multilingual automatic document classification, analysis, and translation corpus.

C. Clausner, A. Antonacopoulos, N. McGregor, and D. Wilson-Nunn. 2018. ICFHR 2018 competition on recognition of historical Arabic scientific manuscripts – RASM2018. In *Proc. ICFHR*, pages 471–476.

A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of ICML*, pages 369–376.

Institute for Palestine Studies. a. Omar al-saleh al-barghouti (1894–1965). Accessed: 2025.

Institute for Palestine Studies. b. Palestine Social History Archives: Omar Saleh Al-Barghouti Diaries. Accessed: 2025.

Mustafa Jarrar, Mo El-Haj, Amal Haddad, Serin Atiani, Shadi Abudalfa, Khalil Sima'an, Paul Rayson, and Camille Mansour, editors. 2026. *Proceedings of the second International Workshop on Nakba Narratives as Language Resources*. Association for Computational Linguistics, Spain.

G. Jocher, A. Chaurasia, and J. Qiu. 2024. Ultralytics YOLO. Version 8.3.0+.

M. Kassis, A. Abdalhaleem, A. Droby, R. Alaasam, and J. El-Sana. 2017. VML-HD: The historical document dataset. In *Proc. ICDAR Workshops*.

R. Khanam and M. Hussain. 2024. YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv preprint arXiv:2410.17725*.

M. Li, T. Lv, J. Chen, et al. 2023. TrOCR: Transformer-based optical character recognition with pre-trained models.

S. A. Mahmoud, I. Ahmad, W. G. Al-Khatib, et al. 2014. KHATT: An open Arabic offline handwritten text database. *Pattern Recognition*, 47(3):1004–1011.

G. Muehlberger, L. Seaward, M. Terras, et al. 2019. Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, 75(5):954–976.

M. Pechwitz, S. S. Maddouri, V. Märgner, et al. 2002. IFN/ENIT – database of handwritten Arabic words. In *Proc. CIFED*, pages 129–136.

M. Romanov et al. 2023. OpenITI: Open Islamicate Texts Initiative.

V. Romero, J. A. Sánchez, A. H. Toselli, et al. 2012. The ESPOSALLES database: An ancient marriage license corpus for off-line handwriting

recognition. *Pattern Recognition*, 46(6):1658–1669.

A. Saeed, C. Ascherl, N. Uchigasaki, et al. 2024. Muharaf: Manuscripts of handwritten Arabic dataset for cursive text recognition. In *NeurIPS Datasets and Benchmarks Track*.

R. Xu, Y. Wang, S. Zhang, et al. 2024. Florence-2: Advancing a unified representation for a variety of vision tasks. *arXiv preprint arXiv:2311.06242*.

StanceNakba Shared Task: Actor and Topic-Aware Stance Detection in Public Discourse

Kholoud K. Aldous[1], **Md Rafiul Biswas**[2], **Mabrouka Bessghaier**[1], **Shimaa Ibrahim**[1], **Kais Attia**, **Wajdi Zaghouani**[1]
[1]Northwestern University in Qatar, [2]Hamad Bin Khalifa University
mbiswas@hbku.edu.qa,
{kholoud.aldous, mabrouka.bessghaier, shimaa.ibrahim, wajdi.zaghouani}@northwestern.edu
Kais.attia.w@gmail.com

Abstract

We present StanceNakba 2026, a shared task on stance detection in polarized social media discourse related to the Palestinian-Israeli conflict, organized as part of Nakba-NLP 2026 at LREC-COLING 2026. The task introduces two subtasks: Subtask A (Actor-Level Stance Detection), which classifies English social media posts as Pro-Palestine, Pro-Israel, or Neutral; and Subtask B (Cross-Topic Stance Detection), which identifies Favor, Against, or Neither stances in Arabic posts toward two conflict-related topics, normalization with Israel and refugee presence in Jordan. The task is grounded in an annotated dataset of 2,606 social media posts. A total of 7 teams participated in Subtask A and 6 teams in Subtask B. Participating systems primarily fine-tuned Arabic and multilingual transformer-based models, including MARBERT, AraBERT, and DeBERTa-v3 variants, with several teams employing cross-validation, ensemble methods, and topic-conditioned architectures. The best-performing systems achieved a Macro F1 of 0.9620 on Subtask A and 0.8724 on Subtask B, demonstrating that transformer-based approaches are highly effective for conflict-domain stance detection while highlighting persistent challenges in cross-topic generalization and neutral class prediction.

Keywords: Stance Detection, Arabic NLP, Palestinian-Israeli Conflict, Social Media Analysis, X

1. Introduction

Social media platforms now serve as key spaces for political discussion, especially when it comes to deeply polarized issues like the Palestinian–Israeli conflict. The volume, diversity, and emotional intensity of online discussions on this topic present significant challenges for natural language processing (NLP) systems that aim to understand public opinion and political orientation. Stance detection is the task of automatically identifying whether an author is in favor of, against, or neutral toward a given target (Mohammad et al., 2016). It offers a principled framework for analyzing such discourse. However, existing resources and shared tasks have largely focused on English and on politically neutral topics, leaving conflict-related, multilingual, and polarized discourse substantially underexplored.

To address this gap, we present the **StanceNakba 2026 Shared Task**, organized as part of Nakba-NLP 2026: The 2nd International Workshop on Nakba Narratives as Language Resources (Jarrar et al., 2026), co-located with LREC-COLING 2026. StanceNakba 2026 introduces two subtasks to stance detection that distinguish between actor-level political alignments and cross-topic stance patterns in social media posts related to the Palestinian-Israeli conflict and associated regional issues. The task is built on an annotated dataset of 2,606 social media posts in English and Arabic.

The shared task contains two subtasks. **Subtask A** (Actor-Level Stance Detection) is an English-language task that requires systems to identify whether the author of a social media post expresses a *Pro-Palestine*, *Pro-Israel*, or *Neutral* orientation in their general position toward the Palestinian-Israeli conflict. **Subtask B** (Cross-Topic Stance Detection) is an Arabic-language task that requires systems to detect *Favor*, *Against*, or *Neither* stances toward two specific conflict-related topics: normalization with Israel and refugee presence in Jordan. These subtasks enable the investigation of fundamental questions at the intersection of NLP and political discourse analysis: How do general political alignments relate to positions on specific policy issues? Can models learn generalizable stance representations that transfer across different topics within the same conflict domain?

The task attracted broad participation from the research community. A total of 7 teams submitted systems for Subtask A and 6 teams for Subtask B, with two teams participating in both. Participating systems primarily relied on fine-tuned Arabic and multilingual transformer-based models, including MARBERT (Abdul-Mageed et al., 2021), AraBERT (Antoun et al., 2020), and DeBERTa-v3 (He et al., 2021) variants, with the best system achieving a Macro F1 of 0.9620 on Subtask A and 0.8724 on Subtask B. Other techniques such as cross-validation, ensemble methods, Natural Language Inference (NLI) reformulation, and topic-conditioned architec-

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 80–90
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

tures were explored, demonstrating the richness of approaches that this task inspired.

The remainder of this paper is organized as follows. Section 2 reviews related work on stance detection and Arabic NLP. Section 3 presents the results and system overview for Subtask A, followed by Section 4 for Subtask B. Section 5 concludes with a discussion of findings and directions for future work. Section 6 discusses the limitations of the current task and dataset. Finally, Section 7 provides details about the ethical considerations and dataset availability.

2. Related Work

Stance detection is the task of automatically determining whether the author of a text is in favor of, against, or neutral toward a given target (Mohammad et al., 2016). SemEval-2016 Task 6 (Mohammad et al., 2016) marked a turning point by introducing a widely used benchmark for stance detection in English tweets, covering targets such as political figures, social movements, and ideological positions. Since then, stance detection has been studied across a broad range of social and political domains, with surveys by AlDayel and Magdy (2021) documenting the rapid growth of the field and the shift from traditional feature-based methods toward transformer-based architectures. Transformer-based models have become the dominant paradigm in stance detection. Pre-trained language models such as BERT (Devlin et al., 2019) have been extensively fine-tuned for stance classification, consistently outperforming earlier approaches that relied on lexical features, n-grams, and handcrafted representations. For the Arabic language, dedicated pre-trained models have proven particularly important due to Arabic's morphological richness, dialectal diversity, and diglossia. AraBERT (Antoun et al., 2020) established strong baselines for Modern Standard Arabic NLP tasks, while ARBERT and MARBERT (Abdul-Mageed et al., 2021) extended coverage to dialectal and social media Arabic by pre-training on large-scale Twitter corpora. Arabic stance detection systems now widely rely on these models as their core architecture, as seen in the approaches used by participants in the StanceNakba 2026 shared task.

Arabic stance detection has received growing attention in recent years, though resources remain limited compared to English. The Mawqif dataset (Alturayeif et al., 2022) introduced the first Arabic target-specific stance corpus, comprising over 4,000 tweets annotated for stance, sentiment, and sarcasm across multiple controversial topics. Complementing this, Charfi et al. (2024) introduced a cross-domain, multi-dialectal Arabic stance corpus covering four Arab regions and multiple dialect groups, with over 4,500 annotated sentences balanced across MSA and regional dialects; their AraBERT-based system achieved strong performance and notably outperformed Mawqif-trained models on the neutral stance class, highlighting the importance of class balance in Arabic stance resources. Building on these resources, StanceEval 2024 (Alturayeif et al., 2024) organized the first shared task dedicated to Arabic stance detection, hosted at ArabicNLP 2024. Participating systems mainly fine-tuned Arabic BERT variants, with ensemble methods and multi-task learning emerging as effective strategies for handling class imbalance and dialectal variation. These efforts demonstrated both the feasibility of Arabic stance detection at scale and the challenges that remain, particularly for underrepresented and neutral stance classes.

The StanceNakba 2026 shared task builds upon these foundations while introducing unique challenges tied to the highly polarized nature of discourse surrounding the Palestinian-Israeli conflict. Prior NLP work on conflict-related discourse has examined stance and opinion on related topics, including studies of Twitter polarization around the Israel-Palestine conflict (Imtiaz et al., 2022) and cross-conflict stance correlations (Tao et al., 2024). StanceNakba 2026 is organized as part of the Nakba-NLP 2026 workshop, the second edition of an initiative dedicated to applying NLP tools to the documentation and understanding of Nakba narratives. The shared task distinguishes itself from prior Arabic stance work through its dual-framework design: Subtask A addresses actor-level political alignment in English, classifying authors as Pro-Palestine, Pro-Israel, or Neutral, while Subtask B targets cross-topic Arabic stance detection toward specific conflict-related issues, namely normalization with Israel and refugee presence in Jordan.

Recent work has increasingly emphasized the importance of modeling stance and narrative framing in politically sensitive and conflict-driven contexts. In the Arabic NLP domain, several efforts have focused on capturing polarization, ideology, and media narratives through dedicated datasets and shared tasks. For instance, the FIGNEWS shared task (Zaghouani et al., 2024a) introduced a benchmark for analyzing news media narratives, highlighting the role of framing and bias in shaping public discourse.

Complementary work has explored conflict-related and politically charged discourse on social media platforms. (Shestakov and Zaghouani, 2024) presented a dataset analyzing the digital framing of the Sheikh Jarrah evictions, providing insights into how conflict narratives are constructed and propagated online. Similarly, (Al Heraki and Zaghouani, 2025) examined polarization and misogynistic discourse in Arabic Twitter conversations,

further demonstrating the challenges of modeling stance in emotionally charged environments.

Beyond stance-specific datasets, related efforts in hate speech and propaganda detection have contributed to understanding polarized language and ideological positioning. Resources such as multi-label hate speech corpora (Zaghouani et al., 2024b) and shared tasks on propaganda and subjectivity detection (Hasanain et al., 2024) provide complementary perspectives on how stance, bias, and persuasion interact in online discourse.

These works collectively underscore the need for domain-specific, conflict-aware resources and evaluation frameworks. The StanceNakba shared task builds on this line of research by introducing actor-level and cross-topic stance detection in the context of the Palestinian-Israeli conflict, extending prior work toward more fine-grained and context-sensitive modeling of political stance.

3. Subtask A: Actor-Level Stance Detection

Subtask A frames conflict stance detection at the *actor level*: given a social media post authored by a single user, the model must infer the author's overarching political orientation toward the Palestinian-Israeli conflict. Rather than analyzing individual argumentative moves or claim-specific positions, the task requires models to aggregate signals across the full text and assign a single author-level stance label. This formulation is motivated by the observation that political orientation in polarized discourse is often expressed indirectly through framing choices (which actors are foregrounded, which actions are lexicalized as aggression versus defense, which populations are centered) rather than through explicit opinion markers alone.

3.1. Dataset

The dataset consists of 1,401 English-language posts collected from X (formerly Twitter), all related to the Palestinian-Israeli conflict. Posts were retrieved via the Meltwater media intelligence platform using targeted keyword queries: *"I stand with Palestine"* or *"I stand with Gaza"* for the Pro-Palestine class, *"I stand with Israel"* for the Pro-Israel class, and *"Israel AND Palestine OR Gaza"* for the Neutral class. Labels were assigned based on the query that retrieved each post, yielding a balanced dataset of 467 samples per class (33.3% each).

To ensure data quality, a multi-step preprocessing pipeline was applied. First, stance-signaling seed keywords (e.g., "I stand with Palestine") were stripped to prevent label leakage, along with Twitter mentions, URLs, hashtags, emojis, and special characters, retaining only alphanumeric text and basic punctuation. Posts were then filtered for meaningful content, requiring a minimum of 30 characters and at least five dictionary words of three or more letters, thereby excluding near-empty or uninformative entries. Finally, case-insensitive exact deduplication was performed by comparing lowercase-normalized versions of the cleaned texts, retaining only the first occurrence of each unique post.

Table 1: Subtask A dataset statistics.

Label	Train	Dev	Test	Total
Pro-Palestine	327	70	70	467
Pro-Israel	327	70	70	467
Neutral	326	70	71	467
Total	**980**	**210**	**211**	**1,401**

Data splits: The dataset is partitioned into training (980 samples, 70%), development (210 samples, 15%), and test (211 samples, 15%) splits, each preserving the balanced class distribution. Table 1 summarizes the dataset statistics.

Dataset Examples: The following examples illustrate the range of expression captured by each label:

- **Pro-Palestine:** *"The systematic displacement of Palestinian families from their ancestral homes represents a clear violation of international law and the right of return."*
- **Pro-Israel:** *"Israel's defensive measures are necessary responses to existential threats, ensuring the safety of its citizens against terrorism."*
- **Neutral:** *"The conflict involves competing territorial claims, with both populations having deep historical connections to the region."*

3.2. Setup and Evaluation

The task was organized into two phases:

- **Development phase**: we released the training and development subsets, and participants submitted runs on the development set through a competition on CodaBench.[1]
- **Test phase**: we released the official test subset, and participants were given a few days to submit their final predictions through the same CodaBench competition. Only the latest submission from each team was considered official and was used for the final team ranking.

[1] https://www.codabench.org/

Measures: Subtask A is framed as a three-class classification problem over English social media posts, where participants are asked to build a single unified model that maps an input post to one of three actor-level stance labels: PRO-PALESTINE, PRO-ISRAEL, or NEUTRAL. We measure the performance of the participating systems using **Macro F1**, which computes the unweighted mean of per-class F1 scores, weighting each class equally and thus penalizing poor performance on any individual label, including the minority-leaning NEUTRAL class. Systems are evaluated on a held-out test set, and participants were not permitted to use the test set for model selection or hyperparameter tuning. In addition to Macro F1, Accuracy, Precision, and Recall are also reported on the leaderboard for reference.

3.3. Results and Overview of the Systems

A total of 12 teams submitted runs during the evaluation phase of Subtask A (Actor-Level Stance Detection), of which 7 teams submitted system descriptions and 6 teams submitted system papers. In Table 2, we provide an overview of the participating systems for which a description was submitted. In Table 3, we report the official results for those teams, ranked by Macro F1.

As shown in Table 2, fine-tuning pre-trained transformer-based models is the dominant approach among participating teams, with BERT-based architectures being the most common backbone. Several teams additionally employed cross-validation, ensemble methods, and hyperparameter optimization.

Team **Shroukgbr** (Gabr and Ragab, 2026) achieved the best performance with a Macro F1 of 0.9620, ranking first on the leaderboard. They fine-tuned a single BERT-based transformer model initialized from a publicly available pretrained checkpoint. Preprocessing steps included removing URLs and user mentions, normalizing repeated characters, and truncating inputs to 256 tokens. To address class imbalance, they applied weighted cross-entropy loss. The final configuration used stratified 5-fold cross-validation, an AdamW optimizer with a learning rate of 2e-5, batch size of 16, and 4 training epochs with early stopping.

Team **Yafa** (Zayet et al., 2026) ranked second with a Macro F1 of 0.9525. They fine-tuned MARBERT and ARBERT using a cross-lingual approach. A customized cross-entropy loss with label smoothing and increased attention dropout (0.2) were applied to mitigate overconfidence and overfitting. Preprocessing included lowercase normalization, URL removal, user mention normalization, character repetition removal, whitespace normalization, hashtag normalization, and connected word splitting. The best results were obtained with a learning rate of 2×10^{-5}, weight decay of 0.1, batch size of 4, maximum sequence length of 128, and 8 epochs with early stopping (patience = 2) and an AdamW optimizer.

Team **KUET** (Al Shafi et al., 2026) ranked third with a Macro F1 of 0.9426. They fine-tuned a BERT-based Mixture-of-Experts (MoE) stance classification model built on top of `bert-base-uncased`. The architecture incorporated cue-word masking and contrast-marker amplification within specialized expert modules, along with CNN-based feature extraction with kernel sizes $\{2, 3, 4, 5\}$. They employed 10-fold stratified cross-validation with a weighted ensemble at inference time, where each fold model's contribution was proportional to its validation Macro F1. Label smoothing ($\epsilon = 0.25$) was applied to reduce overconfidence.

Team **KvochurHegel** (Le, 2026) ranked fourth with a Macro F1 of 0.9384. They formulated stance detection as a Natural Language Inference (NLI) task, initializing a Cross-Encoder from the `DeBERTa-v3-base-mnli-fever-anli` checkpoint. Each input text was evaluated against three class-specific engineered hypotheses designed to encode ideological markers. Label smoothing (factor = 0.2) and R-Drop consistency regularization ($\alpha = 4.0$) were applied to mitigate label noise. The model used AdamW with a learning rate of 2e-5, batch size of 16, and early stopping with patience of 2 epochs.

Team **Viva_Palestine** (El-Kassas et al., 2026) ranked fifth with a Macro F1 of 0.9190. They pretrainined `bert-base-uncased` and fine-tuned it by merging training and validation data for the test phase, training for 20 epochs with early stopping (stopping at epoch 9), a maximum token length of 256, and a learning rate of 10^{-5}.

Team **HeatWave**[2] ranked sixth with a Macro F1 of 0.8804. They experimented with multiple transformer-based models and large language models (LLMs), including `microsoft/deberta-v3-large`, `DeBERTa-v3-large-mnli-fever-anli-ling-wanli`, and the election-domain stance model `bert-election2020-twitter-stance-biden-KE-MLM`, the latter achieving the best results upon fine-tuning. They also explored zero-shot prompting with Qwen2.5 and DeepSeek-R1 models, though these yielded lower F1 scores (0.2–0.6). N-fold training with soft ensemble inference was used for the final submission.

Team **The Blackwell Collective** (Shujon et al., 2026) ranked seventh with a Macro F1 of 0.7451. They fine-tuned `microsoft/mdeberta-v3-base` using a Statement Tuning paradigm: inputs were reformatted as cloze templates and

[2]This team didn't submit their system paper

Table 2: Subtask A: Overview of the participating systems. DL: Deep Learning, ML: Classic Machine Learning.

Team	Transformer						DL	Misc.				
	BERT-base	XLM-RoBERTa	ARBERT/MARBERT	DeBERTa-v3	mDeBERTa-v3	KE-MLM BERT	CNN/MoE	Data Augmentation	Preprocessing	Hyperparameter Tuning	Cross-validation	Ensemble Methods
Shroukgbr (Gabr and Ragab, 2026)	✓	✓	✓					✓	✓	✓	✓	✓
Yafa (Zayet et al., 2026)			✓							✓		
KUET (Al Shafi et al., 2026)	✓						✓			✓	✓	✓
KvochurHegel (Le, 2026)				✓						✓		
Viva_Palestine (El-Kassas et al., 2026)	✓									✓		
HeatWave				✓		✓				✓	✓	✓
The Blackwell Collective (Shujon et al., 2026)					✓					✓		

Table 3: Official results for Subtask A, ranked by Macro F1 score. Leaderboard ranks are shown alongside overall submission ranks (in parentheses). Of the 12 total test submissions, only 7 teams provided a model description and are included here.

Team	Rank	Macro F1
Shroukgbr	1 (#1)	0.9620
Yafa	2 (#2)	0.9525
KUET	3 (#3)	0.9426
KvochurHegel	4 (#4)	0.9384
Viva_Palestine	5 (#6)	0.9190
HeatWave	6 (#7)	0.8804
The Blackwell Collective	7 (#11)	0.7451

the `[MASK]` token's hidden state was contrasted against three learnable class prototypes via a softmax head. Three auxiliary modules were incorporated: Prototype Contrastive Learning (PCL) with $K = 3$ global prototypes at temperature $\tau = 0.1$, Topic-Conditioned Layer Normalization (T-CLN) conditioning feature distributions on topic identity, and R-Drop consistency regularization via symmetric KL divergence. Training followed a two-stage schedule: backbone frozen for 2 epochs, then full fine-tuning for 8 epochs, using AdamW with a learning rate of 10^{-5} for the backbone and 10^{-4} for the heads.

4. Subtask B: Cross-Topic Stance Detection

Subtask B focuses on cross-topic stance detection, where systems are required to determine the stance expressed in a text toward a given target topic. In the context of the StanceNakba Shared Task, this subtask centers on politically sensitive discourse related to the Palestinian cause and regional political dynamics.

4.1. Dataset

We used a subset of the MARASTA dataset (Charfi et al., 2024) for this subtask. MARASTA is a cross-domain, multidialectal Arabic stance corpus designed to support stance detection across multiple dialectal regions of the Arab world. The dataset covers four major dialectal regions, namely Maghreb, Egypt, the Levant, and the Gulf, representing the primary dialect groups in Arabic.

Dataset Overview The full MARASTA corpus contains more than 4,500 sentences annotated with stance toward eight controversial topics distributed across the four Arab regions. Each sentence is labeled according to its stance toward a given topic using three categories: pro (favor), against, or neutral. The dataset was constructed with careful balancing strategies. Each topic contains approximately 500-700 sentences, with a relatively even distribution across the three stance classes. In addition, the corpus maintains a balance between MSA and dialectal Arabic, with roughly half of the sentences written in MSA and the remaining half in the dialect associated with the topic's region.

Data Collection The data were collected from online platforms such as Twitter and YouTube, where discussions on controversial socio-political topics are common. The collected content could be either

current or historically relevant within the past 15 years. For each region (Maghreb, Egypt, the Levant, and the Gulf), the two most highly discussed topics were selected. Candidate sentences were retrieved using seed keywords associated with each topic. To do so, the authors used Python scripts leveraging the platforms' APIs to retrieve tweets and YouTube comments containing the selected keywords. The collected sentences were then manually filtered to ensure they were written in Arabic (MSA or dialect), grammatically correct, relevant to the topic, and expressed a stance (pro, against) or a neutral position, allowing the dataset to capture naturally occurring stance expressions across different dialects and communication styles.

Data Annotation The annotation process followed a multi-stage quality control procedure. Each sentence was independently annotated by two annotators to determine its stance with respect to the associated topic. In cases of disagreement, the instance was reviewed by a third annotator who resolved the conflict and determined the final label. The resulting annotations demonstrate substantial inter-annotator agreement, with an overall Cohen's Kappa score of 0.84 and regional agreement scores ranging from 0.80 to 0.89, indicating high annotation reliability.

Topic Selection for the StanceNakba Shared Task Since the StanceNakba Shared Task focuses on political discourse and narratives related to the Palestinian cause and regional political dynamics, we selected the two topics from the MARASTA corpus that are most relevant to this theme: "Normalization with Israel" and "Refugee/Immigrant Presence in Jordan". The dialectal content associated with these topics reflects two distinct Arabic regional varieties: the "Normalization with Israel" topic contains Gulf Arabic and MSA, while the "Refugee/Immigrant Presence in Jordan" topic contains Levantine Arabic (Jordanian/Palestinian) and MSA. These topics capture key socio-political debates in the Arab world that are closely connected to discussions surrounding Palestine, regional diplomacy, and displacement in the Middle East. The main characteristics of the subset used in this shared task are summarized below:

- **Source:** Arabic social media posts (X) discussing Palestinian-Israeli conflict-related issues.
- **Total Size:** 1,205 annotated samples across two topics
- **Topics Covered:**

Topic 1: التطبيع مع إسرائيل (Normalization with Israel)

- Total: 577 samples
  - Against: 198 (34.3%)
  - Neutral: 208 (36.0%)
  - Pro: 171 (29.6%)

Topic 2: وجود اللاجئين والمهاجرين في الأردن من فلسطين، سوريا، العراق (Refugee/Immigrant Presence in Jordan)

- Total: 628 samples
  - Against: 228 (36.3%)
  - Neutral: 163 (26.0%)
  - Pro: 237 (37.7%)

Data Split

- Training: 843 samples (70%)
- Development: 181 samples (15%)
- Test: 181 samples (15%)

Dataset Examples
Topic: Normalization with Israel

- **Pro:** "الجامعة العربية قالت إنها لا ترى أن #التطبيع مع #إسرائيل خطوة ضد القضية الفلسطينية"
  English: The Arab League said it does not see normalization with Israel as a step against the Palestinian cause.
- **Against:** "إن قدرة الدولة على التطبيع جهارًا نهارًا مع النظام الإسرائيلي تتماشى مع قوة واستقرار نظامها المستبد"
  English: A state's ability to normalize openly with the Israeli regime aligns with the strength and stability of its authoritarian system.
- **Neutral:** "كشف تقرير نشرته البوابة الإسرائيلية لشؤون الزراعة أنّه منذ سنة ونصف تجري في أوروبا اجتماعات بين ممثلين إسرائيليين واماراتيين"
  English: A report published by the Israeli agricultural portal revealed that meetings between Israeli and Emirati representatives have been taking place in Europe for a year and a half.

Topic: Refugee/Immigrant Presence in Jordan

- **Pro:** " لا مكان للعنصرية بالأردن اي شخص داخل الأردن يعامل معاملة ابن البلد وهاذا الشخص يمثل كل أردني شريف"
  English: There is no place for racism in Jordan. Any person inside Jordan is treated like a son of the country, and this person represents every honorable Jordanian.

4.2. Setup and Evaluation

Subtask B follows the same two-phase setup (development and test) and evaluation protocol as Subtask A (see Section 3.2), with Macro F1 as the primary metric computed over the three labels: FAVOR, AGAINST, and NEITHER.

4.3. Results and Overview of the Systems

A total of 8 teams submitted runs during the evaluation phase of Subtask B (Cross-Topic Stance Detection), of which 6 teams submitted system descriptions and system papers. In Table 4, we provide an overview of the participating systems for which a description was submitted. In Table 5, we report the official results for those teams, ranked by Macro F1.

As shown in Table 4, fine-tuning Arabic pretrained transformer models is the dominant strategy, with several teams incorporating cross-validation, ensemble methods, and topic-conditioned input formulations to improve cross-topic generalization.

Team **Viva_Palestine** (El-Kassas et al., 2026) achieved the best performance in Subtask B with a Macro F1 of 0.8724, ranking first on the leaderboard. They fine-tuned `UBC-NLP/MARBERT` by merging the training and validation sets to form an expanded training set of 1,024 records. The input was formatted as a topic–sentence pair truncated to 128 tokens. The best hyperparameters were: learning rate of 2×10^{-5}, batch size of 8, 4 training epochs, warmup ratio of 0.1, weight decay of 0.01, and the AdamW optimizer.

Team **EGCSS** (Qindeel et al., 2026) ranked second with a Macro F1 of 0.8607. They fine-tuned `bert-base-arabertv02-twitter` with a classification head, incorporating topic descriptions concatenated to the input to provide cross-topic signal. They experimented with data augmentation via back-translation, Claude-generated tweets, and the ArabicStanceX dataset, but none of these yielded improvements over the base configuration and were excluded from the final submission. The best hyperparameters were: learning rate of 2×10^{-5}, maximum length of 64, 10 epochs with a linear learning rate schedule, and early stopping based on the labeled validation set.

Team **U4RASD** (Hamdan et al., 2026) ranked third with a Macro F1 of 0.8601. Their final system fine-tuned MARBERTv2 with dialect-aware LLM-based data augmentation, using Gemini 3 Flash Preview to generate three paraphrased variants of each training sample while preserving both stance label and dialectal register. They also investigated contrastive learning, multi-task learning with LLM-generated auxiliary labels, counterfactual augmentation, and zero-shot prompting, none of which improved over the dialect-aware augmentation baseline. Final model selection was based on the best Macro F1 on the development set using stratified 5-fold cross-validation.

Team **The Resistant Word** (Bahgat et al., 2026) ranked fourth with a Macro F1 of 0.8562. They employed a multi-model ensemble with 5-fold stratified cross-validation, fine-tuning four pre-trained Arabic transformer models: MARBERT, AraBERT Large, XLM-RoBERTa Base, and CAMeLBERT-Mix. Each model received a topic-sentence pair truncated to 128 tokens. Prior to training, random oversampling was applied to balance the training set. All models were trained for 5 epochs with a learning rate of 2×10^{-5}, batch size of 16, and BF16 mixed precision. A custom weighted cross-entropy loss was applied, with class weights computed inversely proportional to class frequency. Final predictions were produced by averaging softmax probabilities across all 20 runs (4 models $\times$ 5 folds).

Team **A2NLP** (Nairat and Nairat, 2026) ranked fifth with a Macro F1 of 0.8483. They fine-tuned `aubmindlab/bert-base-arabertv02-twitter` with a prompt-based input formulation that explicitly conditions stance prediction on the target topic. Preprocessing was tailored to Arabic social media and included emoji normalization, URL and mention removal, diacritic removal, Alef variant unification, and whitespace normalization. Stratified 5-fold cross-validation was employed with model selection based on validation Macro F1. A weighted cross-entropy loss addressed class imbalance. The best hyperparameters were: learning rate of 2×10^{-5}, batch size of 16, up to 10 epochs with early stopping (patience = 2).

Team **The Blackwell Collective** (Shujon et al., 2026) ranked sixth with a Macro F1 of 0.7407. They applied the same Statement Tuning architecture as in Subtask A, fine-tuning `microsoft/mdeberta-v3-base` with cloze-style input templates and a `[MASK]`-based prototype classifier. Topic-Conditioned Layer Normalization (T-CLN) served as the primary mechanism for cross-topic generalization, dynamically conditioning feature distributions on a learnable 64-dimensional topic embedding. Prototype Contrastive Learning (PCL) and R-Drop regularization were also applied. Training followed the same two-stage schedule as in Subtask A: backbone frozen for 2 epochs, then full fine-tuning for 8 additional epochs using AdamW.

5. Conclusion and Future Work

We presented an overview of the StanceNakba 2026 Shared Task, which addresses stance detection in polarized social media discourse on the Palestinian-Israeli conflict. The task introduced a dual-framework approach consisting of two sub-

Table 4: Subtask B: Overview of the participating systems. DL: Deep Learning, ML: Classic Machine Learning.

Team	Transformer					Misc.					
	AraBERT (Twitter)	MARBERT	XLM-RoBERTa	mDeBERTa-v3	CAMeL-BERT	Data Augmentation	Preprocessing	Hyperparameter Tuning	Cross-validation	Ensemble Methods	Topic Conditioning
Viva_Palestine (El-Kassas et al., 2026)	✓	✓					✓	✓			✓
EGCSS (Qindeel et al., 2026)	✓					✓	✓	✓	✓		✓
U4RASD (Hamdan et al., 2026)		✓				✓			✓		
The Resistant Word (Bahgat et al., 2026)		✓	✓		✓			✓	✓	✓	
A2NLP (Nairat and Nairat, 2026)	✓					✓	✓	✓	✓		✓
The Blackwell Collective (Shujon et al., 2026)				✓			✓	✓			✓

Table 5: Official results for Subtask B, ranked by Macro F1 score. Leaderboard ranks are shown alongside overall submission ranks (in parentheses). Of the 8 total test submissions, only 6 teams provided a model description and are included here.

Team	Rank	Macro F1
Viva_Palestine	1 (#1)	0.8724
EGCSS	2 (#2)	0.8607
U4RASD	3 (#3)	0.8601
The Resistant Word	4 (#4)	0.8562
A2NLP	5 (#5)	0.8483
The Blackwell Collective	6 (#7)	0.7407

tasks: **Subtask A** (Actor-Level Stance Detection), identifying whether authors express Pro-Palestine, Pro-Israel, or Neutral orientations; and **Subtask B** (Cross-Topic Stance Detection), detecting Favor, Against, or Neither stances toward specific conflict-related topics, namely normalization with Israel and refugee presence in Jordan.

The task attracted considerable interest from the research community. A total of 7 teams made official submissions for Subtask A, and 6 teams for Subtask B, with two teams participating in both subtasks. For both subtasks, the majority of systems fine-tuned pre-trained Arabic transformer models, with BERT-based architectures — particularly MARBERT, AraBERT, and DeBERTa-v3 variants — being the most common backbone. Top-performing systems on Subtask A achieved Macro F1 scores as high as 0.9620, while Subtask B results were competitive with the best system reaching a Macro F1 of 0.8724, reflecting the added difficulty of cross-topic generalization. Several teams incorporated other techniques such as cross-validation, ensemble methods, topic-conditioned input formulations, and consistency regularization to improve performance.

Future editions of the StanceNakba shared task could explore more fine-grained stance categories and extend coverage to additional conflict-related topics and Arabic dialects. Increasing dataset size, particularly for underrepresented stance classes, and developing evaluation protocols that better capture cross-topic generalization will be important directions. Future work could also investigate the role of large language models (LLMs) under zero-shot and few-shot settings.

6. Limitations

Several participating teams noted difficulty predicting the neutral or "neither" class, likely due to its inherent ambiguity. Additionally, the cross-topic nature of Subtask B introduces domain shift between training and test topics, which current transformer-based models do not fully address. Expanding the dataset with more balanced label coverage and including a broader range of topics and dialectal varieties would help mitigate these limitations in future iterations of the task.

7. Ethical Considerations & Dataset Availability

7.1. Data Source and Consent

The datasets used in the StanceNakba Shared Task consist of publicly available posts collected from Twitter X in accordance with the platform's

terms of service. Only publicly accessible content was included, and no attempt was made to access private accounts, restricted material, or deleted posts.

Usernames, profile information, and other direct identifiers were removed. The dataset does not include user metadata beyond the textual content necessary for stance classification. In compliance with platform policy, only post IDs and annotation labels will be released where required.

7.2. Sensitive and Political Content

This shared task focuses on discourse related to the Palestinian Israeli conflict and related regional issues. This topic is politically sensitive and may involve references to violence, displacement, discrimination, or traumatic events.

The dataset may contain polarized language and emotionally charged expressions. The purpose of the shared task is scientific analysis of stance detection and cross topic generalization. It is not intended to endorse, legitimize, or amplify any political position. Results should not be interpreted as normative judgments.

7.3. Risk of Harm and Misuse

Automated stance detection systems applied to political discourse may be misused for surveillance, political profiling, targeted persuasion, or repression. We explicitly discourage the use of models trained on this dataset for surveillance or discriminatory purposes.

Because Subtask A involves inference of general political alignment, predictions should not be treated as reliable indicators of an individual's beliefs or identity. Model outputs are probabilistic classifications based on limited textual evidence and may be incorrect.

7.4. Annotation Bias and Subjectivity

Stance annotation is inherently interpretive. Although detailed guidelines were developed for the Pro Palestine, Pro Israel, Neutral, Favor, Against, and Neither labels, annotators may bring implicit cultural, political, or linguistic biases.

To mitigate this risk, annotators were trained using standardized definitions, disagreements were adjudicated, and balanced label distributions were maintained where feasible. Nevertheless, residual bias may remain in both the English and Arabic datasets. Models trained on this data may learn annotation artifacts rather than purely linguistic signals of stance.

7.5. Representation and Sampling Limitations

The datasets consist of a limited number of annotated English and Arabic social media posts. They are not representative of all political perspectives, demographic groups, or geographic populations. Social media users are themselves not representative of broader societies.

Accordingly, findings should not be generalized to entire populations or interpreted as reflecting majority opinion. Claims about cross topic generalization should be framed cautiously.

7.6. Cross Lingual and Cross Cultural Considerations

Subtask A and Subtask B involve different languages and sociopolitical contexts. Linguistic markers of stance may vary across dialects, regions, and political cultures. Models may encode unintended cultural assumptions.

Participants are encouraged to report per class performance, conduct qualitative error analysis, and critically examine model behavior across topics and languages.

7.7. Annotator Well Being

Exposure to polarized or conflict related discourse may cause emotional fatigue. Annotation processes should include reasonable workload distribution and allow annotators to opt out of reviewing content they find distressing.

7.8. Data Availability

The dataset is released for non commercial research purposes. Users must comply with platform terms of service, avoid attempts at deanonymization, and refrain from surveillance or discriminatory applications. The dataset can be accessed upon request through the following form: https://forms.gle/YUFdA16R6HkSZjp88

Acknowledgment

This shared task was made possible by the National Priorities Research Program grant NPRP14C-0916-210015 from the Qatar National Research Fund (QNRF), part of the Qatar Research, Development and Innovation Council (QRDI).

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, et al. 2021. Arbert & marbert: Deep bidirectional transformers for arabic. In *Proceedings of the 59th annual meeting of the*

association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers), pages 7088–7105.

Huda Al Heraki and Wajdi Zaghouani. 2025. Analyzing digital polarization on hijab: A dataset of annotated youtube comments. In *Proceedings of ICWSM 2025*.

Abdullah Al Shafi, Md. Milon Islam, Sk. Imran Hossain, and K. M. Azharul Hasan. 2026. KUET at StanceNakba shared task: StanceMoE: Mixture-of-experts framework for stance detection. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

Abeer AlDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Nora Alturayeif, Hamzah Luqman, Zaid Alyafeai, and Asma Yamani. 2024. Stanceeval 2024: The first arabic stance detection shared task. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 774–782.

Nora Saleh Alturayeif, Hamzah Abdullah Luqman, and Moataz Aly Kamaleldin Ahmed. 2022. Mawqif: A multi-label arabic dataset for target-specific stance detection. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 174–184.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *Proceedings of the 4th workshop on open-source arabic corpora and processing tools, with a shared task on offensive language detection*, pages 9–15.

Mohammed Bahgat, Doaa Salah, and Sarah Yassine. 2026. The Resistant Word at StanceNakba shared task: A topic-aware model for cross-topic stance detection. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

Anis Charfi, Mabrouka Bessghaier, Andria Atalla, Raghda Akasheh, Sara Al-Emadi, and Wajdi Zaghouani. 2024. Stance detection in arabic with a multi-dialectal cross-domain stance corpus. *Social Network Analysis and Mining*, 14.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Wafaa S. El-Kassas, Enas A. Hakim Khalil, and Enas M.F. El Houby. 2026. Viva_Palestine at StanceNakba shared task: Actor and topic-aware stance detection in public discourse. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

Shrouk Anwar Gabr and Mohamed Ibrahim Ragab. 2026. Shroukgbr at StanceNakba shared task: Transformer-based ensemble learning for actor-level stance detection in Palestinian–Israeli social media discourse. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

Nancy Hamdan, Aya Jouni, Aya Saïd, and Fadi A. Zaraket. 2026. U4RASD at StanceNakba shared task: Data augmentation and auxiliary objectives for arabic stance detection. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

Maram Hasanain, Reem Suwaileh, Sander Weering, Chengkai Li, Tommaso Caselli, Wajdi Zaghouani, Alberto Barrón-Cedeño, Preslav Nakov, and Firoj Alam. 2024. Overview of the clef-2024 checkthat! lab task on subjectivity in news articles. In *CLEF 2024 Working Notes*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Arsal Imtiaz, Danish Khan, Hanjia Lyu, and Jiebo Luo. 2022. Taking sides: Public opinion over the israel-palestine conflict in 2021. *ArXiv*, abs/2201.05961.

Mustafa Jarrar, Mo El-Haj, Amal Haddad, Serin Atiani, Shadi Abudalfa, Khalil Sima'an, Paul Rayson, and Camille Mansour, editors. 2026. *Proceedings of the second International Workshop on Nakba Narratives as Language Resources*. Association for Computational Linguistics, Spain.

Minh-Hoang Le. 2026. KvochurHegel at StanceNakba shared task: Robust stance detection with regularized natural language inference. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.

Alaa Nairat and Aysar Mahmoud Nairat. 2026. A2NLP at StanceNakba shared task: Fine-tuned AraBERT for topic-based arabic stance detection. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

Asmaa Qindeel, Toka Khaled, Batool Balah, Eman Elrefai, and Mahmoud Fawzi. 2026. EGCSS at StanceNakba shared task: Cross-topic arabic stance detection for two middle east issues. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

Alexey Shestakov and Wajdi Zaghouani. 2024. Analyzing conflict through data: A dataset on the digital framing of sheikh jarrah evictions. In *Proceedings of the Workshop on NLP for Political Sciences at LREC-COLING 2024*.

Md. Shakhoyat Rahman Shujon, MD Jahid Hasan Jim, Md. Milon Islam, Md Rezwanul Haque, and Fakhri Karray. 2026. The Blackwell Collective at StanceNakba shared task: PAST-TIDE: Prototype-anchored statement tuning with topic-invariant normalization for stance detection. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

Yiyao Tao, Hengyu Zhang, Babli Dey, Selenge Tulga, Hanjia Lyu, and Jiebo Luo. 2024. In the eyes of the bystander: Are the stances on different conflicts correlated? In *2024 IEEE International Conference on Big Data (BigData)*, pages 7193–7201. IEEE.

Wajdi Zaghouani, Mustafa Jarrar, Nizar Habash, Houda Bouamor, Imed Zitouni, Mona Diab, Samhaa El-Beltagy, and Mohammed AbuOdeh. 2024a. The fignews shared task on news media narratives. In *Proceedings of LREC-COLING 2024*.

Wajdi Zaghouani, Hamdy Mubarak, and Md Rafiul Biswas. 2024b. So hateful! building a multi-label hate speech annotated arabic dataset. In *Proceedings of LREC-COLING 2024*.

Tasnim Zayet, Osama Hamed, and Tasneem Duridi. 2026. Yafa at StanceNakba shared task: Actor-level stance detection using cross-lingual approach. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

The NakbaArchiveClassifier Shared Task on Nakba Image Classification

Alexei Abrahams[1], Shadi Abudalfa[2], Mustafa Jarrar[1], George Mikros[1]
[1]Hamad bin Khalifa University
[2]King Fahd University of Petroleum & Minerals

Abstract

The proliferation of social media platforms has significantly reshaped how conflicts are documented, generating large-scale visual records that must be structured to enable meaningful analysis. In this paper, we present the NakbaArchiveClassifier shared task, which focuses on binary classification of infrastructure damage in images from Gaza. This task formed part of the Nakba-NLP Workshop at LREC 2026 and is grounded in an ongoing initiative focused on humanitarian archiving. It utilizes a carefully curated dataset of 2,001 images sourced from Palestinian journalists and content creators on Instagram, spanning the period from October 7, 2023 to December 15, 2025. The objective for participants was to classify whether an image depicts damaged or destroyed infrastructure versus intact structures. This task poses multiple challenges, such as the complexity of real-world conflict imagery, imbalance between classes, and the inherent ambiguity present in many visual scenes. The NakbaArchiveClassifier shared task introduces a new benchmark for analyzing conflict-related visual data and provides valuable resources for advancing research in humanitarian AI, crisis analytics, and Arabic digital humanities.

1. Introduction

The ongoing documentation of the Palestinian experience in Gaza has taken on new urgency since October 7, 2023, when a large-scale military campaign dramatically intensified the destruction of civilian infrastructure. Social media platforms (and Instagram in particular) have become primary channels through which Palestinian content creators and journalists bear witness to this destruction in near real time. The sheer volume of this visual documentation, with millions of images archived by the Tech for Palestine incubator project SaltPillar, presents both an extraordinary opportunity and a formidable challenge for researchers, humanitarian organizations, and archivists seeking to make these data accessible and analytically meaningful.

Automated image classification offers a promising pathway toward organizing such large-scale visual archives. By distinguishing images of destroyed or damaged infrastructure from images of intact infrastructure, a reliable classifier can help prioritize humanitarian assessments, support journalistic verification workflows, and enable systematic research into patterns of destruction. The NakbaArchiveClassifier shared task was designed with precisely these goals in mind, offering the research community a carefully curated dataset of 2,001 images drawn from Palestinian sources in Gaza and a well-defined binary classification objective. This shared task formed part of the Nakba-NLP Workshop at LREC 2026 (Jarrar et al., 2026).

Motivation

The term *Nakba* (Arabic for "catastrophe," referring historically to the mass displacement of Palestinians in 1948 and more broadly to the ongoing dispossession of the Palestinian people) continues to unfold in visual form across social media. The SaltPillar archive captures this unfolding record, but raw archival scale alone is insufficient for research utility. Systematic categorization is a prerequisite for any downstream analysis, whether humanitarian, journalistic, or scholarly. Infrastructure damage is among the most analytically significant categories, as it speaks directly to the material conditions of life and survival in Gaza. Automating the detection of such damage in social media images would substantially reduce the manual annotation burden and accelerate the usability of the archive.

Challenges

NakbaArchiveClassifier presents unique challenges:

- **Domain specificity and visual complexity.** Images of conflict-affected infrastructure differ substantially from standard benchmark datasets. Destruction may be partial or obscured, lighting conditions are often poor, and images are frequently captured under duress, resulting in motion blur, occlusion, or unconventional framing that complicates automated analysis.

- **Class imbalance.** The dataset contains 706 images of destroyed or damaged infrastructure and 1,295 images of intact infrastructure, an imbalance that classifiers must address through appropriate sampling strategies or loss weighting.

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 91–97
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

- **Contextual ambiguity.** Some images are visually ambiguous (for example, rubble visible in the background of an otherwise ordinary scene, or intact buildings photographed in ways that imply surrounding damage), making binary labeling a non-trivial judgment even for human annotators.

- **Ethical and representational sensitivity.** Working with imagery of conflict and human suffering requires careful attention to the dignity of those depicted and the responsible use of automated systems trained on such material.

- **Out-of-distribution generalization.** The held-out test set of 500 images may capture temporal or thematic variation not fully represented in the training data, challenging models to generalize robustly across the full scope of the archive.

NakbaArchiveClassifier

NakbaArchiveClassifier[1] offers a binary image classification benchmark grounded in a real-world humanitarian archiving use case. Participants were provided with 2,001 labeled training images sourced from Palestinian content creators and journalists active on Instagram in Gaza between October 7, 2023 and December 15, 2025, and were evaluated on a held-out test set of 500 images. The classification objective is to distinguish images depicting destroyed or damaged infrastructure from those depicting infrastructure that remains intact.

The task received strong engagement from the global NLP and computer vision community:

- 30 teams registered.
- 16 unique submissions to the leaderboard.
- 14 countries are represented among the registered teams, including: Saudi Arabia, Pakistan, China, India, Finland, Egypt, Nigeria, Vietnam, Qatar, South Korea, Bangladesh, Morocco, Algeria, and Turkey.

NakbaArchiveClassifier contributes the first benchmark suite tailored for automated classification of conflict-related infrastructure damage in social media imagery from Gaza, establishing a replicable evaluation framework and a labeled dataset that can support future work in humanitarian AI, conflict documentation, and Arabic digital humanities.

[1] NakbaArchiveClassifier URL: https://nakba-nlp-2026.pillarofsalt.net/

2. Related Work

The NakbaArchiveClassifier task sits at the intersection of several active research areas: visual damage detection, humanitarian and crisis-related image analysis, social media content classification, and Arabic digital humanities. We review the most relevant prior work across these areas.

Damage and Disaster Detection in Images. Building-damage assessment from satellite and aerial imagery has been studied extensively in the context of natural disasters. A landmark contribution in this space is the xBD dataset (Gupta et al., 2019), which provides pre- and post-event satellite imagery across diverse disaster types, paired with ordinal damage labels and building polygons. With over 850,000 building annotations, xBD enabled large-scale evaluation of deep learning models for damage assessment. Transfer learning from architectures such as ResNet (He et al., 2016) and EfficientNet (Tan and Le, 2019) has become the dominant paradigm for these tasks, with models fine-tuned on disaster imagery achieving strong performance on held-out benchmarks (Alam et al., 2023a). More recently, transformer-based vision models (in particular the Vision Transformer (ViT) (Dosovitskiy et al., 2021)) have demonstrated competitive results on image classification tasks and are increasingly applied to specialized domains. However, satellite and aerial imagery differs markedly from the ground-level, social-media-sourced content that characterizes the NakbaArchiveClassifier dataset, with the latter exhibiting greater perspectival variability, visual noise, and contextual ambiguity.

Crisis and Conflict Image Analysis. A dedicated line of research addresses the automated analysis of crisis imagery shared on social media. The CrisisMMD dataset (Alam et al., 2018) introduced multimodal Twitter data from natural disasters with three annotation layers, enabling joint text-and-image classification for humanitarian response. Building on this foundation, Alam et al. (2021) consolidated eight human-annotated crisis-related datasets to facilitate comparative evaluation across informativeness and humanitarian classification tasks, while Alam et al. (2023b) benchmarked ten deep learning architectures for real-time crisis image classification across four task types. A comprehensive extension of this effort is the MEDIC dataset (Alam et al., 2023a), the largest social media image classification dataset for humanitarian response to date, covering 71,198 images across four tasks in a multi-task learning framework. NakbaArchiveClassifier extends this line of work to the specific domain of conflict and

occupation in Gaza, with imagery drawn from a live archiving infrastructure and a binary classification objective centered on infrastructure damage.

Social Media Image Classification and Transfer Learning. General-purpose image classification has advanced rapidly with the development of large pretrained architectures. ResNet (He et al., 2016) established residual connections as a foundational design pattern that enabled the training of much deeper convolutional networks. EfficientNet (Tan and Le, 2019) later demonstrated that compound scaling of network depth, width, and input resolution can achieve strong accuracy with substantially fewer parameters. The Vision Transformer (Dosovitskiy et al., 2021) introduced a pure attention-based architecture that, when pretrained at scale, matches or surpasses convolutional baselines on standard image benchmarks. Contrastive pretraining on large image-text corpora, as in CLIP (Radford et al., 2021), has further expanded the repertoire of readily available representations, enabling effective zero-shot and few-shot transfer to downstream classification tasks. Fine-tuning these pretrained models on domain-specific data is now standard practice and has proven effective even in low-resource settings, making it directly relevant to the NakbaArchiveClassifier challenge, where the training set is modest in size and the imagery is domain-specific.

Arabic Digital Humanities and Arabic NLP. Computational approaches to Arabic-language materials have grown substantially in recent years. The development of AraBERT (Antoun et al., 2020), a BERT-based model pretrained specifically on large-scale Arabic text, demonstrated that language-specific pretraining substantially outperforms multilingual models on Arabic NLP benchmarks including sentiment analysis, named entity recognition, and question answering. Recent shared tasks have further pushed the state of the art in Arabic NLU: the ArabicNLU 2024 shared task (Khalilia et al., 2024) targeted word sense disambiguation and location mention disambiguation in Arabic, providing novel annotated resources alongside competitive evaluation. Work at the intersection of computational methods and Palestinian studies remains limited, however, making NakbaArchiveClassifier a contribution to an underserved area of Arabic digital humanities. The broader Nakba-NLP 2026 Workshop, of which this shared task is a part, represents a deliberate effort to develop language and vision resources attentive to Palestinian experience and narrative.

3. Data Collection and Selection

The SaltPillar archive consists of Instagram content, including millions of images and videos uploaded by content creators in Gaza (such as @wizard_bisan1 or @saleh.jafarawi) or shared by second-hand aggregators and observers (such as @eye.on.palestine or @palestine.pixel). The data were retrieved from Instagram using adversarial archiving techniques detailed in (Abrahams, 2026). These data constitute part of the collective memory of the Palestinian people, an archive for scholarly inquiry, and an evidence bank for war crimes investigation.

To be useful in these respects, however, the content of the archive must be sensibly labeled according to what they pertain to—airstrike, starvation, children, protest, resilience, medicine, and so forth. But due to the sheer size of the archive, it is infeasible to assign a team of humans the task of labeling every piece of content by hand. Moreover, the graphic and demoralizing nature of the content might plausibly have long-term consequences for the mental health of the labelers themselves. In lieu of this, machine learning classifiers can emulate the judgment of human expert labelers by training on a comparatively small sample of labeled content, then extrapolate to predict the labels of the remaining content.

We proceeded by randomly sampling 5,000 images from the SaltPillar archive for the period October 7, 2023 to December 15, 2025. In particular, we restricted our attention to a list of over 600 content creators either based in Gaza, or whose Instagram activity centered on sharing content originating from Gaza.

By hand, we then labeled these images according to whether they exhibited signs of infrastructure damage or not. We filtered out sensitive images, such as highly graphic content or content that too clearly identified individuals. We also filtered out irrelevant content, such as screenshots of tweets. In all, 60% of the sample was filtered out, leaving a sample of 2,001 images: 706 images of destroyed or damaged infrastructure, and 1,295 images of intact infrastructure.

4. Task Description and Evaluation

In this section, we describe in detail the shared. The shared task was conducted in two stages: a development phase and a final evaluation phase. During the development phase, participants tested and refined their models using a validation set. In contrast, the final phase relied on a held-out test set with concealed labels. All submissions were assessed via the **CodaBench** platform, an open-source infrastructure for benchmarking AI systems.

CodaBench enables the definition of tasks, including datasets, submission workflows, and evaluation criteria, and offers an automated and transparent environment for assessing participant models.

In this shared task, participants are required to develop a system that performs binary image classification to determine whether an image depicts destroyed or damaged infrastructure or intact infrastructure. The task focuses on real-world imagery collected from social media in Gaza, presenting unique challenges related to visual complexity, ambiguity, and domain specificity.

Formally, the objective is to design an image classifier C such that, given an input image x, the system predicts a label y where:

$$y = C(x), \quad y \in \{\text{Destruction}, \text{Not_destruction}\}$$

- **Input**: An image collected from social media depicting infrastructure in Gaza.
- **Output**: A binary label indicating whether the image shows {Damaged, Intact} infrastructure.
- **Evaluation Metric**: Systems are evaluated using standard classification metrics, including Accuracy and F1-score, with the primary metric being the **Macro F1-score**. This metric is particularly appropriate due to class imbalance and ensures equal weighting of both classes.

5. Participant Systems

A total of 16 participants submitted systems to the NakbaArchiveClassifier shared task. Below, we provide a summary of selected participating systems, linking participant names to their leaderboard identifiers, modeling approaches, and reported performance:

- **HCMUS_TheFangs**: The top-ranked team proposed a comprehensive evaluation of multiple modern architectures, including EfficientNet, Vision Transformers, and MaxViT. Their best-performing system employed a MaxViT-Base backbone combined with focal loss, MixUp augmentation, and a tailored augmentation pipeline to address dataset bias and imbalance. Their approach achieved a Macro F1-score of 0.899, securing **1st place** on the leaderboard.
- **rahaf_jaber**: This system utilized a ConvNeXt-Tiny backbone trained using stratified 5-fold cross-validation, combined with weighted loss functions and test-time augmentation. An ensemble of fold-specific models was used during inference. The system achieved a Macro F1-score of 0.8952, ranking **2nd**.
- **zahira_blr**: The authors proposed a ConvNeXt-Tiny–based ensemble with focal loss, exponential moving average (EMA), and threshold calibration. Their approach incorporated out-of-fold validation and test-time augmentation. The system achieved a Macro F1-score of 0.8893, placing **3rd**.
- **salmakh**: This team employed a CLIP-based transfer learning approach, fine-tuning a Vision Transformer encoder with a lightweight classification head. Their method addressed class imbalance using focal loss and achieved a Macro F1-score of 0.877, ranking **4th**.
- **mohamedfathy**: Their system relied on a Swin Transformer architecture combined with transfer learning and semi-supervised data expansion to improve generalization. The model achieved strong performance with a Macro F1-score of approximately 0.865, ranking **5th**.
- **KvochurHegel**: This team used a ConvNeXt-V2 Nano backbone with extensive data augmentation and label smoothing. A multi-view test-time augmentation strategy was applied during inference. Their system achieved a Macro F1-score of 0.8593, ranking **6th**.
- **AyahVerse** (Username: *ibad-ur-rehman-546335*): The team evaluated multiple architectures, including ResNet50, MobileNetV2, and EfficientNet-B0, alongside threshold optimization strategies. Their findings highlighted the effectiveness of lightweight models and calibration techniques, achieving an F1-score of approximately 0.85.
- **men_na_aly**: This system adopted a ResNet-based transfer learning approach with weighted cross-entropy loss to address class imbalance. Their final model achieved a Macro F1-score of 0.8323 and demonstrated the effectiveness of lightweight CNN baselines.
- **kannanrrk**: The authors utilized an EfficientNet-B3 backbone with transfer learning, achieving a weighted F1-score of 0.8387 and demonstrating strong performance with relatively efficient architectures.
- **nw_123**: This submission proposed an ensemble of EfficientNet models combined with an SVM meta-classifier for decision fusion. The approach emphasized computational efficiency and achieved competitive performance under constrained settings.
- **heba13**: The team evaluated multiple CNN architectures, including ResNet and EfficientNet

variants, using transfer learning and standard preprocessing. Their results highlighted the effectiveness of moderate-depth models in low-resource settings.

- **m_elhussieni**: This submission used a ResNet-50 backbone with aggressive grayscale augmentation and test-time augmentation to address the challenges of archival imagery. The system achieved a Macro F1-score of approximately 0.45, serving as a baseline under challenging conditions.

6. Results

Table 1 summarizes the final standings of the NakbaArchiveClassifier shared task, where systems are ordered based on their Macro F1-score evaluated on the held-out test data. The results demonstrate generally strong performance across participants, with HCMUS_TheFangs achieving the highest score of 0.8991, followed closely by rahaf_jaber and zahira_blr. Most systems obtained F1-scores exceeding 0.80, indicating that current image classification methods can effectively address this task despite its complexity. However, a clear drop in performance is observed among the lower-ranked entries, suggesting differences in model generalization and robustness when dealing with challenging, real-world conflict imagery. Overall, these results highlight both the competitive nature of the task and the advancements made by participants in tackling specialized visual classification problems.

#	Participant Name	Macro F1-Score
1	HCMUS_TheFangs	0.8991
2	rahaf_jaber	0.8952
3	zahira_blr	0.8893
4	salmakh	0.8767
5	mohamedfathy	0.8655
6	KvochurHegel	0.8593
7	AyahVerse	0.8545
8	mohamedabdallah	0.8465
9	nw_123	0.8416
10	kannanrrk	0.8389
11	men_na_aly	0.8323
12	Md.Ajwad Hossain	0.8312
13	mealper	0.8236
14	heba13	0.8142
15	faisalm93	0.4506
16	m_elhussieni	0.3965

Table 1: Leaderboard for NakbaArchiveClassifier Shared Task on Nakba Image Classification.

7. Discussion

The outcomes of the NakbaArchiveClassifier shared task indicate that contemporary computer vision methods can effectively tackle the problem of identifying infrastructure damage in conflict imagery, even when operating under difficult real-world conditions. Leading submissions achieved Macro F1-scores close to 0.90, reflecting strong classification performance despite the dataset's limited size, class imbalance, and the inherently ambiguous nature of the images.

An important trend observed across the leaderboard is the strong performance of transformer-based and hybrid architectures, including MaxViT, ConvNeXt, and Vision Transformers. These models appear particularly capable of capturing intricate spatial relationships and contextual signals in complex scenes, where evidence of damage may be partial, subtle, or obscured. Moreover, the extensive reliance on transfer learning emphasizes the value of pretrained representations, particularly in specialized domains where labeled data are scarce.

Handling class imbalance also emerged as a critical factor in achieving competitive performance. Many high-ranking systems incorporated focal loss, class weighting, or tailored sampling approaches, which improved detection of the minority class representing damaged infrastructure. This is evident in the consistently high Macro F1-scores, suggesting that participants were able to reduce bias toward the more frequent class.

In addition, training strategies such as data augmentation and model ensembling proved highly beneficial. Methods like MixUp, test-time augmentation, and cross-validation-based ensembles contributed to improved robustness and generalization, especially given the noisy and heterogeneous nature of social media images. These results highlight the necessity of carefully designed training pipelines when working with real-world, unstructured datasets.

Nevertheless, the noticeable gap between top-performing and lower-performing systems points to unresolved challenges. Some approaches achieved substantially lower scores, in certain cases below 0.50 Macro F1, indicating difficulties in generalizing across the dataset. These shortcomings may be attributed to sensitivity to domain variation, limited capacity to resolve ambiguous cases, or insufficient optimization. This disparity suggests that, while feasible, the task remains challenging, particularly for simpler or less refined models.

The task also underscores the ambiguity inherent in the labeling scheme. Differentiating between damaged and intact infrastructure is not always straightforward, even for human annotators. As a

result, models may rely on indirect cues or contextual patterns that do not perfectly align with human interpretation. This raises broader concerns regarding annotation consistency, potential label noise, and the adequacy of binary labels for representing complex humanitarian scenarios.

More broadly, the shared task illustrates both the potential benefits and the constraints of applying AI techniques to humanitarian archives. Automated classification can substantially reduce the burden of manual annotation and improve access to large-scale datasets. However, such systems must be deployed responsibly, taking into account ethical considerations such as misclassification risks, bias in representation, and the sensitive nature of the underlying content.

8. Limitations

Despite the contributions of the NakbaArchive-Classifier shared task, several limitations should be considered when interpreting its findings and potential applications.

To begin with, the dataset size is relatively limited. Although the 2,001 annotated images were carefully selected and curated, this number is small compared to large-scale benchmarks commonly used in computer vision. Consequently, model performance may rely substantially on transfer learning, data augmentation, and regularization techniques, and the dataset may not fully represent the variability of infrastructure-related imagery in Gaza.

In addition, the data collection is confined to Instagram and a specific subset of Palestinian content creators and aggregators. As a result, models trained on this dataset may face challenges when applied to images from different platforms, geographic contexts, or documentation environments.

Another limitation arises from the binary nature of the classification task. In practice, infrastructure damage is not strictly dichotomous but exists along a continuum, from minor deterioration to complete destruction. Some images may also present overlapping or unclear visual evidence. Although the binary setup simplifies annotation and evaluation, it reduces the ability to capture fine-grained distinctions that could be valuable for more detailed analysis.

Furthermore, the annotation process inherently involves a degree of subjectivity. Even trained annotators may differ in their judgments when evaluating ambiguous or borderline cases, particularly when damage is partially visible, obscured, or context-dependent. This introduces the possibility of label noise within the dataset, which should be taken into account when interpreting system performance.

The filtering stage also imposes constraints on the dataset's representativeness. Images deemed highly graphic or those clearly identifying individuals were excluded to ensure ethical compliance and protect human dignity. While this is a necessary and responsible decision, it limits the extent to which the dataset reflects the full spectrum of visual material present in the original archive.

Finally, system evaluation is primarily based on aggregate metrics such as Macro F1-score. Although suitable for benchmarking and ranking submissions, these metrics do not fully capture aspects that are crucial for real-world deployment, including model calibration, resilience to distribution shifts, interpretability, and fairness across different image types. Future work could benefit from more comprehensive evaluation frameworks that better align with humanitarian and archival requirements.

9. Conclusion

This overview paper introduced the NakbaArchive-Classifier shared task, which focuses on the problem of classifying social media images based on the presence of infrastructure damage. Built on real-world imagery from Gaza, the task emphasizes several inherent challenges, including ambiguity in visual cues, domain-specific characteristics, and imbalanced class distributions. By evaluating participating systems within this setting, the task aims to promote progress in analyzing conflict-related visual data and strengthening the reliability of computer vision approaches in humanitarian applications.

We anticipate that NakbaArchiveClassifier will provide a valuable benchmark for future research in areas such as digital archiving, crisis-focused AI, and humanitarian computing, while fostering the development of more robust, context-aware, and ethically responsible image classification models.

10. Bibliographical References

Alexei Abrahams. 2026. *Social Media Exposed: Track Disinformation, Archive War Crimes, Restore Oversight*. No Starch Press.

Firoj Alam, Tanvirul Alam, Md Arid Hasan, Abul Hasnat, Muhammad Imran, and Ferda Ofli. 2023a. MEDIC: a multi-task learning dataset for disaster image classification. *Neural Computing and Applications*, 35(3):2609–2632.

Firoj Alam, Tanvirul Alam, Ferda Ofli, and Muhammad Imran. 2023b. Robust training of social media image classification models. *IEEE Transactions on Computational Social Systems*, 10(5):2507–2518.

Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. CrisisMMD: Multimodal twitter datasets from natural disasters. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, pages 465–473. AAAI Press.

Firoj Alam, Hassan Sajjad, Muhammad Imran, and Ferda Ofli. 2021. CrisisBench: Benchmarking crisis-related social media datasets for humanitarian information processing. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, pages 923–932.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resources Association.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

Ritwik Gupta, Bryce Goodman, Nirav Patel, Ricky Hosfelt, Sandra Sajeev, Eric Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew Gaston. 2019. Creating xBD: A dataset for assessing building damage from satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 10–17.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Mustafa Jarrar, Mo El-Haj, Amal Haddad, Serin Atiani, Shadi Abudalfa, Khalil Sima'an, Paul Rayson, and Camille Mansour, editors. 2026. *Proceedings of the second International Workshop on Nakba Narratives as Language Resources*. Association for Computational Linguistics, Spain.

Mohammed Khalilia, Sanad Malaysha, Reem Suwaileh, Mustafa Jarrar, Alaa Aljabari, Tamer Elsayed, and Imed Zitouni. 2024. ArabicNLU 2024: The first arabic natural language understanding shared task. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 361–371, Bangkok, Thailand. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Mingxing Tan and Quoc V. Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.

The NakbaVirality Shared Task on MultimodalVirality Prediction in High-Stakes Discourse

Saad Ezzini[1], Salima Lamsiyah[2], Shadi Abudalfa[1],
Samir El-Amrany[2], Walid Alsafadi[3]
[1]King Fahd University of Petroleum & Minerals
[2]University of Luxembourg
[3]University College of Applied Sciences

Abstract

Social media virality significantly shapes public discourse during geopolitical conflicts, where emotionally charged and multimodal content can rapidly gain widespread attention. However, most prior approaches rely on retrospective engagement signals, limiting their usefulness for early prediction. Multimodal virality modeling in high-stakes Arabic discourse remains largely unexplored. We introduce *NakbaVirality*, a shared task on multimodal virality classification in conflict-related social media posts, organized as part of the NakbaNLP workshop at LREC 2026. The dataset consists of 2,600 anonymized posts from X and Reddit collected after October 7, 2023, each including text, an associated image, and normalized engagement labels. Participants must classify posts into low, medium, or high virality categories using only textual and visual inputs. The task provides standardized splits, baseline systems, and evaluation using macro-F1 and accuracy. *NakbaVirality* establishes the first benchmark for multimodal virality prediction in Arabic high-stakes discourse and promotes research on contextual and multimodal modeling for early impact prediction. The shared task attracted 18 participants, who contributed a total of 5 official test phase submissions.

1. Introduction

Social media platforms have become central to information diffusion during geopolitical conflicts, where emotionally charged narratives and symbolic imagery can rapidly influence public discourse. Understanding what makes content "go viral" has therefore attracted significant attention across computational social science and machine learning. Early studies showed that community structure and diffusion topology strongly influence cascade formation and meme spread (Doerr et al., 2012; Weng et al., 2013; Cheung et al., 2016), while psychological factors such as emotional intensity also contribute to online sharing dynamics (Berger and Milkman, 2012). More recent approaches incorporate multi-feature fusion and deep learning models to forecast popularity using temporal and engagement signals (Gao et al., 2021; Kowalczyk and Larsen, 2018; Xu and Qian, 2023; Zhang and Gao, 2024). However, these methods largely rely on retrospective interaction patterns (e.g., early likes, retweets, diffusion graphs), which are unavailable at publication time.

Advances in representation learning have significantly improved content modeling capabilities. Transformer-based language models such as BERT (Devlin, 2018) and multimodal architectures such as CLIP (Radford et al., 2021) enable joint reasoning over textual and visual signals. Multimodal fusion approaches have demonstrated improvements in related tasks including misinformation detection and popularity prediction (Singhal et al., 2019; Riis et al., 2020). Despite this progress, publicly available benchmarks for multimodal virality prediction in high-stakes discourse remain limited, particularly for Arabic. While Reddit-V (Elamrany et al., 2025) introduced a dataset for pre-engagement virality prediction, it did not specifically target conflict-driven Arabic discourse.

The discourse surrounding the Nakba and related regional discussions constitutes a high-impact setting marked by historical references, polarized sentiment, and multimodal communication. In such contexts, images often amplify textual rhetoric, making multimodal alignment critical for predicting reach. Modeling this content requires capturing contextual and symbolic cues in Arabic, which presents additional challenges including rich morphology, dialectal variation, and code-switching. Furthermore, viral posts typically form a minority class, requiring robust evaluation under class imbalance using metrics such as macro-F1 (Powers, 2011).

To address these challenges, we introduce *NakbaVirality*[1], the first shared task on multimodal virality classification in Arabic high-stakes discourse. This shared task formed part of the NakbaNLP Workshop at LREC 2026 (Jarrar et al., 2026). The presented dataset consists of 2,600 anonymized posts collected from X and Reddit after October 7, 2023. Each instance includes post text, an associated image, and normalized engagement labels. Participants are required to classify posts into low, medium, and high virality categories using only textual and visual inputs. The task provides standardized splits, baseline systems grounded in

[1]NakbaVirality URL: https://ezzini.github.io/NakbaVirality

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 98–103
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

transformer and convolutional architectures, and evaluation using macro-F1 and accuracy.

The shared task received strong engagement from the global NLP community, with 18 participating teams in total and 5 unique final submissions to the leaderboard. *NakbaVirality* establishes the first benchmark suite for multimodal virality prediction in Arabic high-impact discourse and promotes research on contextual modeling, multimodal alignment, and early impact prediction in real-world social media settings.

2. Related Work

Foundations of Virality and Diffusion Modeling. Early research on virality prediction has emphasized structural and psychological drivers of information diffusion. Prior studies have shown that cascade topology, community structure, and user susceptibility have strongly influenced meme spread and rumor propagation (Doerr et al., 2012; Weng et al., 2013; Cheung et al., 2016; Hoang and Lim, 2016). Complementary work has demonstrated that emotional intensity, sentiment, and social signaling have played a central role in driving online sharing behavior (Berger and Milkman, 2012; Tsugawa and Ohsaki, 2017; Mousavi et al., 2022). Subsequent research has incorporated temporal cascade features and engagement trajectories to forecast popularity at early stages (Shamma et al., 2011; Lu and Szymanski, 2018; Gao et al., 2021; Xu and Qian, 2023). More recently, graph-based and multi-task learning frameworks have modeled virality jointly with rumor detection and user vulnerability analysis (Zhang and Gao, 2024; Esteban-Bravo et al., 2024). However, these approaches have largely relied on early interaction signals (e.g., retweets, likes, diffusion graphs), which have not been available at publication time and therefore have limited purely content-based early prediction.

Content-Based and Multimodal Virality Prediction. To overcome the limitations of diffusion-dependent models, a growing body of work has focused on predicting virality from content alone. Early NLP-based studies have explored lexical, stylistic, and affective cues associated with viral text (Guerini et al., 2011). Subsequent research has proposed scalable content-based prediction models for news and social media posts (Lu and Szymanski, 2018; Kowalczyk and Larsen, 2018). With advances in deep learning, transformer-based approaches have been applied to virality prediction, including RoBERTa-based models for news tweets (Maldonado-Sifuentes et al., 2021) and user-aware architectures such as ViralBERT (Rameez et al., 2022). In the visual domain, attention mechanisms and deep visual representations have been leveraged to model video and image popularity (Bielski and Trzcinski, 2018; Riis et al., 2020), while multimodal frameworks have combined textual and visual signals to improve performance in related prediction settings (Singhal et al., 2019). More recently, large-scale vision–language pretraining models such as CLIP have enabled joint multimodal representation learning applicable to popularity and virality modeling (Radford et al., 2021). Nevertheless, multimodal virality benchmarks have remained limited, particularly beyond English and outside general-domain social media contexts.

Large Language Models and Arabic Virality Benchmarks. Recent work has begun to explore large language models for content-based virality prediction. In particular, Reddit-V has introduced a benchmark for pre-engagement virality classification and has evaluated zero-shot large language models for predicting popularity without relying on diffusion signals (El-amrany et al., 2025). Beyond this effort, existing virality datasets and evaluations have largely focused on English or general-domain social media content, and have not specifically addressed multimodal Arabic discourse in conflict-driven settings. The proposed *NakbaVirality* shared task has aimed to fill this gap by establishing the first benchmark dedicated to multimodal, content-based virality prediction in Arabic high-stakes discourse.

3. Data Collection and Selection (will be enhanced)

NakbaVirality targets conflict-related posts published after October 7, 2023 and pairs each text with an associated image and engagement-derived labels for multimodal virality classification. The data collection pipeline is shown in Figure 1.

3.1. Sources and acquisition

Reddit platform. Data from Reddit were collected from subreddits that include regular discussion of Gaza / Palestine and related topics. Data from each subreddit was collected according to the same general procedure: the subreddit was obtained (either from a curated list, or from a predefined list as a fallback); the posts within the subreddit that included one or more keywords from a multilingual list were searched; it was confirmed that each post was within the scope of the collection; and then the data from each relevant post that included at least one image was collected from the post. For each collected post, its identifier, name of its subreddit, date/time of posting, and various basic statistics associated with the post were collected. The image itself and its metadata were also collected.

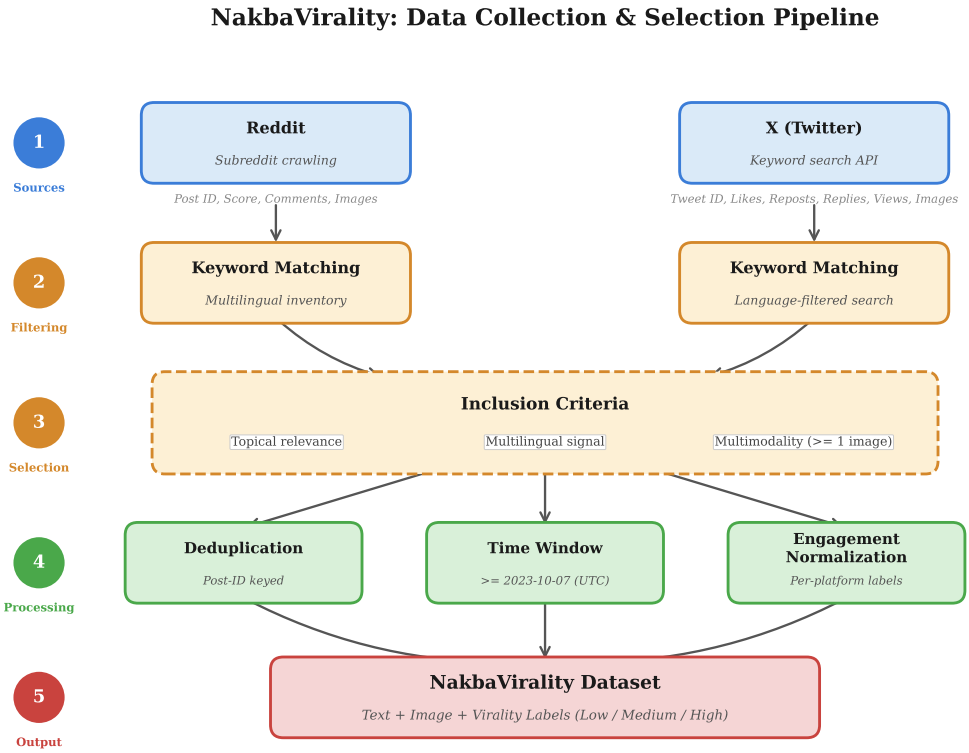


Figure 1: Data collection pipeline.

X platform. X data was collected through keyword search on the platform. The X search platform allows searches to be filtered by language. Consequently, searches were performed with the keyword list in each relevant language; search results were collected and posted in separate datasets for each language; and duplicates were removed to ensure that every post is represented only once in the dataset. The searches were filtered to include only posts published between October 7, 2023 and the present date. For each collected post, its identifier, text, date/time of posting, various statistics associated with the post, and any media linked to the post were collected.

3.2. Inclusion criteria

To make the dataset more relevant to multilingual and multimodal analysis, posts must meet the following inclusion criteria:

1. **Topical relevance :** The title and body of the post on Reddit, or the tweet text on X, must match at least one keyword from our inventory.

2. **Multilingual signal :** A post is retained if it matches at least one term from our multilingual keyword inventory. The corpus consists mostly of English (60%) and Arabic (10%) posts, with other languages including Farsi, Hebrew, Spanish, and Turkish comprising the remainder. Figure 2 reports the language distribution of the retained dataset.

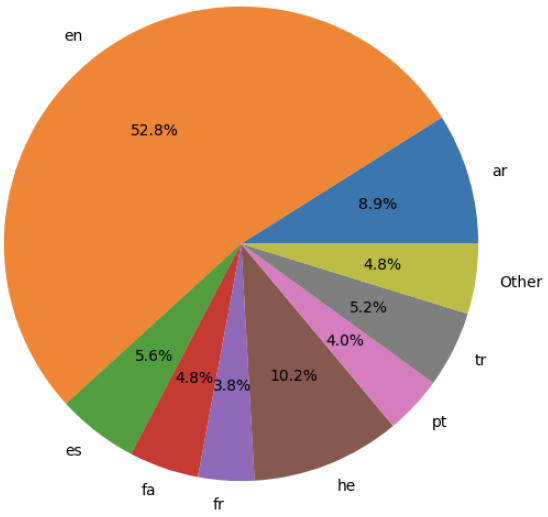


Figure 2: Language distribution of the collected posts in NakbaVirality.

3. Multimodality: the post has at least one image. For Reddit, this can be directly linked images or those uploaded to Reddit. For X, this requires

parsing the tweet card to find image URLs.

3.3. Deduplication and normalization

We remove duplicates at the *post* level using platform-native identifiers: (i) Reddit submissions are keyed by `Post ID`, and (ii) X posts are keyed by `tweet_id`. During crawling, seen identifiers are tracked to avoid collecting the same post multiple times across queries or sources.

We preserve the raw engagement fields provided by each platform to keep label construction reproducible. For Reddit, we keep the submission score and discussion volume (e.g., comment count), together with basic post metadata. For X, we keep the observed interaction counters (likes, reposts, replies, and views when available) and we additionally compute an engagement score as

$$E = \text{likes} + \text{reposts} + \text{replies}.$$

Because engagement magnitudes differ substantially across platforms, virality is not compared using raw counts directly. Instead, labels are derived by normalizing engagement *within each platform* and mapping posts into the three shared-task categories (low / medium / high).

3.4. Time window selection

All instances are restricted to a shared-task time window with a fixed lower bound of **2023–10–07** (UTC). For X, this lower bound is enforced at query time; for Reddit, the initial crawl may return older content, so the time constraint is applied during the final filtering step.

In the current data snapshot, the retained timestamps span: (i) **X:** 2023–10–07 to 2025–12–13 (UTC), (ii) **Reddit:** 2023–10–07 to 2025–11–04 (UTC). Posts outside the time window are excluded from the final release.

4. Evaluation Track

A total of 4 teams submitted their predictions to the shared task. Table 1 outlines the official leaderboard, evaluating the submissions primarily on Macro F1-Score, supplemented by Accuracy. The teams employed a wide range of analytical methodologies ranging from simplistic machine learning workflows to foundational multimodal models and recursive prompt-tuning techniques.

4.1. Participant Systems

HCMUS_TheFangs: The winning system deliberately avoided the complexities of deep multimodal learning by formulating virality prediction strongly on sociological structure. They extracted the source community structure and encoded it utilizing a sophisticated Bayesian smoothed target-encoding feature. Combining this single powerful feature with text-level descriptors (Character count, Word count, Hashtag count) and simple textual TF-IDF tokenization, they trained an XGBoost classifier. This streamlined approach surprisingly overshadowed all deep learning methods, emphasizing that virality during conflict heavily depends on community amplification factors and user biases, far more than on semantic properties.

Digilians: The Digilians team approached the task through a complex multimodal neural architecture, integrating both XLM-RoBERTa for multilingual textual processing and a Vision Transformer (ViT-B/16) mechanism. They significantly minimized semantic gaps by injecting a bidirectional cross-attention fusion layer that mutually conditioned text and image features against each other. Their implementation addressed severe class imbalances by leveraging selective Flan-T5 text paraphrasing on minority samples and applying a Focal Loss function inversely weighted to class frequencies.

xin1212: This anonymous submission implemented a frozen multimodal paradigm. Retaining the parameters of the base LAION CLIP-ViT, they integrated a custom built, lightweight residual adapter directly positioned over the concatenated text-image embeddings. Followed by a three-way MLP classifier layer, this frozen approach mitigated overfitting typical of data-constrained tasks and stabilized optimization in the multimodal learning domain.

ashhadulislam (Pushing_Boundaries): In contrast to explicitly engineering architectural models, this submission pioneered Recursive Prompt Improvement (RPI) in a parameter-free manner using prominent large language models (DeepSeek, GPT-5.2, and Qwen2.5-VL). They utilized diagnostic signals from top misclassified texts in each iteration and iteratively revised classification meta-prompts. They demonstrated that prompt tuning progressively clarifies separation boundaries among the different viral classes with zero neural weight updates.

5. Discussion

The differing outcomes of the submitted systems bring significant insights concerning multimodal virality prediction in emotionally charged and polarizing contexts. Most notably, standard state-of-the-art vision-language fusion models generally underperformed relative to the non-neural system pro-

Rank	Team	Macro F1	Accuracy
1	HCMUS_TheFangs	**0.7062**	**0.7305**
2	Digilians (noureldeen)	0.4983	0.6009
3	xin1212	0.4559	0.6089
4	ashhadulislam	0.3199	0.4392

Table 1: Official NakbaVirality System Leaderboard.

vided by HCMUS_TheFangs, suggesting an inherent "visual homogeneity" where explicit visual content does not directly determine virality. Rather, historical community context and audience ideological affiliations dictate how rapidly information disseminates. Furthermore, complex representations and explicit hashtags were found to decrease organic viral propensity, reaffirming the "Less is More" phenomenon during high-stakes conflict propagation.

6. Conclusion and Future Work

The *NakbaVirality* shared task successfully introduced the first benchmark specifically dedicated to multimodal virality prediction within Arabic high-stakes conflict discourse. With participation from multiple international teams, the evaluation process revealed a crucial insight: in politically charged contexts, historical community behavior and audience dynamics often exhibit greater predictive power than the semantic or visual content of the posts themselves. Deep multimodal architectures struggled with the inherent visual homogeneity of conflict imagery, demonstrating the vital necessity of integrating sociological factors into virality models. Future iterations of this task will focus on expanding the dataset across additional social media platforms, capturing temporal sentiment shifts, and incorporating zero-shot prompt-based evaluation for larger foundation models to further explore early-stage virality detection.

Limitations

While this shared task provides a foundational benchmark, several limitations must be acknowledged. First, the dataset is restricted to a specific time window following October 7, 2023, and may not fully capture the evolving linguistic and visual strategies used in long-term geopolitical conflicts. Second, to ensure strict ethical compliance and protect user privacy given the highly sensitive nature of the discourse, critical network diffusion features such as explicit user identities and follower graphs were anonymized or excluded. This restricts the ability to perform deep cascade or topology-based virality analysis. Finally, the inherent rarity of truly viral content results in a severe class imbalance, significantly complicating the training process and constraining the generalization capabilities of standard predictive models.

7. Bibliographical References

Jonah Berger and Katherine L Milkman. 2012. What makes online content viral? *Journal of marketing research*, 49(2):192–205.

Adam Bielski and Tomasz Trzcinski. 2018. Pay attention to virality: understanding popularity of social media videos with the attention mechanism. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2335–2337.

Ming Cheung, James She, Alvin Junus, and Lei Cao. 2016. Prediction of virality timing using cascades in social media. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 13(1):1–23.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Benjamin Doerr, Mahmoud Fouz, and Tobias Friedrich. 2012. Why rumors spread so quickly in social networks. *Communications of the ACM*, 55(6):70–75.

Samir El-amrany, Matthias R. Brust, Salima Lamsiyah, and Pascal Bouvry. 2025. Reddit-V: A virality prediction dataset and zero-shot evaluation with large language models. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 334–341, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Mercedes Esteban-Bravo, Jose M Vidal-Sanz, et al. 2024. Predicting the virality of fake news at the early stage of dissemination. *Expert Systems with Applications*, 248:123390.

Liqun Gao, Yujia Liu, Hongwu Zhuang, Haiyang Wang, Bin Zhou, and Aiping Li. 2021. Public opinion early warning agent model: A deep learning cascade virality prediction model based on

multi-feature fusion. *Frontiers in Neurorobotics*, 15:674322.

Marco Guerini, Carlo Strapparava, and Gozde Ozbal. 2011. Exploring text virality in social networks. In *proceedings of the international AAAI conference on web and social media*, volume 5, pages 506–509.

Tuan-Anh Hoang and Ee-Peng Lim. 2016. Tracking virality and susceptibility in social media. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 1059–1068.

Mustafa Jarrar, Mo El-Haj, Amal Haddad, Serin Atiani, Shadi Abudalfa, Khalil Sima'an, Paul Rayson, and Camille Mansour, editors. 2026. *Proceedings of the second International Workshop on Nakba Narratives as Language Resources*. Association for Computational Linguistics, Spain.

Damian Konrad Kowalczyk and Jan Larsen. 2018. Scalable privacy-compliant virality prediction on twitter. *arXiv preprint arXiv:1812.06034*.

Xiaoyan Lu and Boleslaw K Szymanski. 2018. Scalable prediction of global online media news virality. *IEEE Transactions on Computational Social Systems*, 5(3):858–870.

Christian E Maldonado-Sifuentes, Jason Angel, Grigori Sidorov, Olga Kolesnikova, and Alexander Gelbukh. 2021. Virality prediction for news tweets using roberta. In *Mexican International Conference on Artificial Intelligence*, pages 81–95. Springer.

Maryam Mousavi, Hasan Davulcu, Mohsen Ahmadi, Robert Axelrod, Richard Davis, and Scott Atran. 2022. Effective messaging on social media: What makes online content go viral? In *Proceedings of the ACM web conference 2022*, pages 2957–2966.

David Powers. 2011. Evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020.

Rikaz Rameez, Hossein A Rahmani, and Emine Yilmaz. 2022. Viralbert: A user focused bert-based approach to virality prediction. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, pages 85–89.

Christoffer Riis, Damian Konrad Kowalczyk, and Lars Kai Hansen. 2020. On the limits to multimodal popularity prediction on instagram–a new robust, efficient and explainable baseline. *arXiv preprint arXiv:2004.12482*.

David Shamma, Jude Yew, Lyndon Kennedy, and Elizabeth Churchill. 2011. Viral actions: Predicting video view counts using synchronous sharing behaviors. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 618–621.

Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*, pages 39–47. IEEE.

Sho Tsugawa and Hiroyuki Ohsaki. 2017. On the relation between message sentiment and its virality on social media. *Social network analysis and mining*, 7(1):19.

Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. 2013. Virality prediction and community structure in social networks. *Scientific reports*, 3(1):1–6.

Zhixuan Xu and Minghui Qian. 2023. Predicting popularity of viral content in social media through a temporal-spatial cascade convolutional learning framework. *Mathematics*, 11(14):3059.

Xuan Zhang and Wei Gao. 2024. Predicting viral rumors and vulnerable users with graph-based neural multi-task learning for infodemic surveillance. *Information Processing & Management*, 61(1):103520.

Faisal_Adam at NakbaArchiveClassifier Shared Task: Archival Image Classification for Structural Destruction: A Robust Pipeline Using ResNet-50 and Test-Time Augmentation

Faisal Muhammad Adam[1], Aliyu Salisu[2]
[1]ACETEL, NOUN, Nigeria; [2]Dept. of Computer Science, ABU, Nigeria
[1]ace25110005@nou.edu.ng; [2]aliyusalisu@abu.edu.ng

Abstract

This paper describes our system submission for the Nakba Archive Image Classification task, which requires predicting the presence of structural destruction in historical archival photographs. We framed this as a binary computer vision classification problem (`destruction` vs. `not_destruction`). To combat the high visual noise and inherent color bias of historical film, our system utilizes a pre-trained ResNet-50 convolutional neural network combined with aggressive grayscale augmentation, class-weighted loss penalties, and Cosine Annealing scheduling. During inference, we employed Test-Time Augmentation (TTA) and strategic prediction threshold tuning to maximize recall. Evaluated on the unseen final test set, our model achieved a macro F1-score of 0.450 and a balanced accuracy of 0.527, serving as a robust exploratory baseline that highlights the unique challenges of processing degraded historical imagery.

Keywords: Computer Vision, Archival Imagery, ResNet-50, Test-Time Augmentation, Historical Preservation

1. Introduction

The classification of historical and archival imagery presents unique challenges not typically found in modern image datasets. Photographs from the Nakba Archive are often characterized by low resolution, heavy film grain, sepia or black-and-white color grading, and varying levels of contrast.

The objective of this system was to participate in the **NakbaArchiveClassifier** shared task (Abrahams et al., 2026), which requires automatically classifying images into two categories: `destruction` (containing ruined structures, rubble, etc.) and `not_destruction` (intact buildings or landscapes).

This paper outlines our methodology, which leverages transfer learning with a deep ResNet-50 backbone. We detail our robust data augmentation steps designed to force structural feature learning over color bias, our approach to handling dataset imbalance, and a multi-crop Test-Time Augmentation (TTA) inference strategy. Finally, we provide a rigorous error analysis to categorize the specific failure modes inherent to this domain.

2. Related Work

The application of deep learning to historical and archival imagery has seen significant growth in recent years. Broad surveys highlight the role of machine learning across cultural heritage tasks (Fiorucci et al., 2020). Previous works have largely focused on digitizing historical text via Optical Character Recognition (OCR) or restoring physically degraded photographs using Generative Adversarial Networks (GANs), while recent work also explores domain adaptation for historical image classification (Zhao et al., 2022) and deep-learning-based analysis of archival degradation (Shruthi and Rahman, 2024).

However, the semantic classification of archival disaster imagery remains heavily under-explored. Traditional Convolutional Neural Networks (CNNs) trained on modern, high-definition datasets (such as ImageNet) often struggle with severe domain shifts when applied to historical archives. This performance drop is typically caused by the lack of color consistency and the presence of severe physical degradation (scratches, chemical fading, and film grain). By participating in this shared task, we aim to bridge this gap, demonstrating how techniques like grayscale augmentation and Test-Time Augmentation can adapt modern architectures to historical domains.

3. Methodology

Our pipeline focuses on counteracting the visual ambiguity of historical photos through aggressive preprocessing, penalized loss functions, and ensemble-like inference techniques.

3.1. Data Preprocessing and Augmentation

A primary challenge with historical imagery is that general photo degradation (like sepia fading) can falsely correlate with the target classes. To force

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 104–107
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

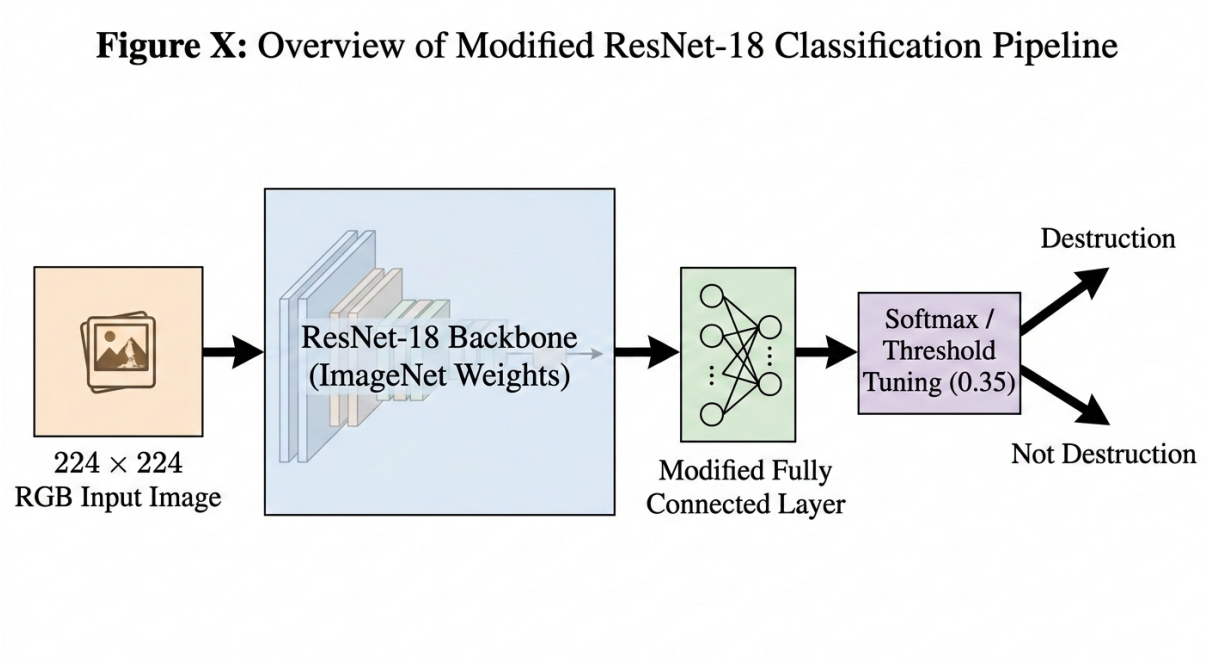


Figure 1: Overview of our modified ResNet-50 pipeline, illustrating the flow from the augmented input through the deep convolutional layers, culminating in the Custom Binary Classification Head and TTA inference block.

the network to become "colorblind" and focus purely on the geometric shapes of ruined structures, we implemented a heavy augmentation pipeline during training.

Images were initially resized to 256×256 pixels. We applied `RandomResizedCrop` (to 224×224), `RandomHorizontalFlip`, and `RandomRotation` ($\pm 15°$) to increase spatial variance. Crucially, we applied a forced `Grayscale` conversion (outputting 3 identical channels) to eliminate color bias. Finally, tensors were normalized using standard ImageNet mean and standard deviation metrics.

3.2. Model Architecture and Training

For feature extraction and classification, we utilized ResNet-50. The network was initialized with pre-trained ImageNet weights to leverage low-level edge detection filters, and the final fully connected classification head was replaced with a linear layer outputting exactly two features (Figure 1).

To combat class imbalance within the training data, we dynamically computed class frequencies from the training labels and applied a **Class-Weighted Cross-Entropy Loss** function. This heavily penalized the model for ignoring the minority class, ensuring equitable gradient updates regardless of class distribution.

Hyperparameter	Value
Architecture	ResNet-50
Optimizer	AdamW
Learning Rate	$1e-4$
Weight Decay	$1e-3$
LR Scheduler	Cosine Annealing ($T_{max} = 6$)
Batch Size	32
Epochs	6
Loss Function	Weighted Cross-Entropy
Inference TTA	3 Crops

Table 1: Training hyperparameters and configuration.

3.3. Inference: TTA and Threshold Tuning

During the testing phase, we applied **Test-Time Augmentation (TTA)** to achieve maximum prediction stability. Instead of evaluating a single image, the model processed three distinct transformations of each test image: a standard resized crop, a horizontally flipped crop, and a localized center crop ($256 \times 256 \rightarrow 224 \times 224$). The Softmax probabilities across these three variations were averaged to produce the final confidence score.

Through iterative threshold tuning, we shifted the decision boundary for predicting `destruction` to **0.45**. This mathematically compensated for the model's under-confidence when evaluating highly degraded imagery. Applying this pipeline to the

unlabelled final phase dataset yielded 41 positive predictions for `destruction` and 361 negative predictions.

4. Results and Evaluation

The system was evaluated by the CodaBench scoring program on the hidden Final Phase dataset. The primary competition metric was the Macro F1-Score.

4.1. Official Metrics

Table 2 presents the official performance metrics of our final submission.

Metric	Score
F1-Score	0.4505
Accuracy	0.4552
Precision	0.5328
Recall	0.5277
Specificity	0.5277
Balanced Accuracy	0.5277

Table 2: Official Final Phase Evaluation Metrics.

The Balanced Accuracy of 0.527 indicates that the model is performing slightly above a random baseline (0.50) in equally identifying both classes. However, the overall F1-score of 0.450 suggests the model struggled with false positives and false negatives, indicating that even deep features struggle against heavy archival noise.

4.2. Error Taxonomy and Failure Modes

To better understand the limitations of our ResNet-50 baseline, we conducted a qualitative analysis of the false positives and false negatives. We identified three primary failure modes inherent to the Nakba Archive dataset:

1. **Texture-Degradation Confusion:** Despite grayscale augmentation, the network frequently mistook heavy film grain, chemical deterioration, and physical scratches on the archival film for the physical texture of rubble and debris. The network's convolutional filters over-indexed on high-frequency noise.
2. **Topographical Misclassification:** Images containing natural rocky terrain or uneven hillsides were consistently misclassified as `destruction`, indicating a failure to differentiate between natural geology and ruined man-made geometry.
3. **Scale and Resolution Artifacts:** In wide-angle landscape shots, the destroyed structures occupied too few pixels for the standard 224×224 downsampling to preserve necessary structural semantics.



Figure 2: Qualitative examples of failure modes, demonstrating the visual overlap between archival film degradation and structural ruin.

5. Conclusion

Our submission establishes a highly robust baseline pipeline for binary classification of archival disaster imagery using ResNet-50, grayscale color bias mitigation, and Test-Time Augmentation. While the pipeline was technically comprehensive, the F1-score of 0.450 indicates that standard transfer learning—even with extensive augmentation—is challenged by the extreme degree of noise inherent to historical photos.

Future directions for this task include utilizing Vision Transformers (ViTs) which may handle scale variances better than CNNs, or applying domain adaptation techniques to explicitly separate photographic degradation artifacts from structural content.

Acknowledgements

We would like to express our gratitude to the organizers of the Nakba-NLP 2026 workshop and the contributors to the Nakba Archive for providing the foundational dataset and platform that made this exploratory research possible.

6. Ethical Considerations

The classification of destruction in historical archives carries significant emotional and political weight. Automated systems must be designed with sensitivity to the potential for misclassification to erase or misrepresent historical trauma. We emphasize that this model is an exploratory tool for archival sorting and should not be used as a definitive historical record without human verification.

7. Bibliographical References

Alexei Abrahams, Shadi Abudalfa, Mustafa Jarrar, and George Mikros. 2026. The nakbaarchive-classifier shared task on nakba image classification. In *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

Marco Fiorucci, Marina Khoroshiltseva, Massimiliano Pontil, Arianna Traviglia, and Alessio Del Bue. 2020. Machine learning for cultural heritage: A survey. *Pattern Recognition Letters*, 133:102–108.

K. Shruthi and M. Rahman. 2024. Deep learning techniques for the preservation and analysis of archival degradation. *Journal of Cultural Heritage Preservation*, 29:45–59.

Fangwen Zhao, Weifeng Liu, and Chenglin Wen. 2022. A new method of image classification based on domain adaptation. *Sensors*, 22(4):1315.

Doaa Sulaiman at AR-MS NakbaNLP 2026: Faithful Diplomatic Transcription of Arabic Manuscripts Using a Human-Centred Annotation Framework

Doaa Bahjat Sulaiman
Institute of Social Sciences, Istanbul University
Türkiye
doaa.sulaiman@ogr.iu.edu.tr

Abstract

This paper describes participation in the Human Transcription Track (Subtask 1) of the NAKBA-NLP 2026 Arabic Manuscript Understanding Shared Task, which focuses on historical handwritten documents related to Palestinian Nakba narratives. The participant manually transcribed 500 cropped line images and designed a comprehensive transcription guideline from scratch, documenting all editorial decisions in a reusable framework. A faithful diplomatic transcription philosophy was adopted, preserving original spelling, punctuation, diacritics, numerals, and layout features without editorial normalisation, to create research-grade gold-standard data for downstream NLP and digital humanities research. Building on this philosophy, a 26-convention annotation framework was developed, organised into three layers: editorial-structural symbols, faithful-copying rules, and documentation labels, supported by a four-step quality-control pipeline. The submission achieved full coverage of all 500 assigned lines and attained an official CER of 0.02, WER of 0.01, and accuracy of 0.98.

Keywords: Arabic manuscripts, handwritten text transcription, Nakba narratives, diplomatic transcription, annotation framework

1. Introduction

Historical Arabic manuscripts constitute an invaluable cultural and linguistic resource, yet their systematic digitisation remains a formidable challenge (Zaghouani et al., 2014; Pierazzo, 2011). The primary obstacles include the pronounced variability of individual handwriting styles, persistent orthographic inconsistencies, the physical degradation of source materials over time, and the absence of agreed-upon transcription conventions (Hamoud et al., 2026; Zaraket et al., 2026). Advances in Handwritten Text Recognition (HTR) and Optical Character Recognition (OCR) have delivered impressive results for Latin scripts; however, Arabic HTR accuracy continues to lag substantially behind, owing to the cursive and context-sensitive character of Arabic writing, the high degree of allographic variation, and the comparative scarcity of well-annotated training corpora (Ahmad et al., 2017; Ahmad et al., 2020; Aabed and Khairaldin, 2024; Al-Mutawa et al., 2024; Waly et al., 2025).

The NAKBA-NLP 2026 Shared Task directly addresses this gap by providing a structured benchmark for both human and automated transcription of manuscript page images drawn from the Omar Al-Saleh Memoir Collection, a corpus of 16 documents spanning 1951–1965 and estimated to contain approximately 6,395 pages and 1.6 million words (Hamoud et al., 2026; Zaraket et al., 2026). Subtask 1 (the Human Transcription Track) requires participants to transcribe approximately 500 cropped line images and to design their own transcription guidelines; evaluation is based on three dimensions: coverage, accuracy (CER and WER), and guidelines quality (Hamoud et al., 2026).

This paper reports on participation in Subtask 1. The principal contributions are: (a) a comprehensive, reusable 26-convention annotation framework for Arabic manuscripts, organised into three functional layers; (b) a four-step quality-control pipeline designed to approximate multi-annotator rigour in a single-annotator setting; and (c) an empirically supported account of how expert human annotation can complement automated HTR pipelines in large-scale manuscript digitisation projects.

2. Related Work

2.1 Arabic Handwritten Text Recognition

Arabic HTR has attracted considerable research interest in the deep-learning era, yet recognition rates remain well below those achieved for Latin scripts (Ahmad et al., 2017, 2020). The KHATT dataset serves as a canonical benchmark; multi-dimensional LSTM-based architectures achieve character recognition rates of approximately 80% on this corpus, while more recent segmentation-free end-to-end models report rates of around 84% (Ahmad et al., 2017, 2020; Aabed and Khairaldin, 2024). Comparative evaluations of four deep-learning architectures (FCN, GFN, VAN, and DAN) on KHATT indicate that the DAN model achieves the lowest character and word error rates, though performance remains substantially below Latin-script benchmarks (Al-Mutawa et al., 2024).

Among transformer-based systems, HATFormer achieves CER values of 8.6% and 4.2% on major public and private Arabic handwriting datasets respectively (HATFormer, 2024). The Qalam model, trained on over 4.5 million manuscript images, attains a WER of 0.80% for handwritten

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 108–112
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

text recognition tasks (Bhatia et al., 2024). The end-to-end Invizo system reports 0.59% CER on printed Arabic text and 7.91% CER on handwritten Arabic (Waly et al., 2025). Despite these advances, automated systems continue to struggle with historically degraded manuscripts exhibiting irregular layouts, ink bleeding, and idiosyncratic orthographic conventions, reinforcing the need for carefully produced gold-standard transcription data (Hamoud et al., 2026; Zaraket et al., 2026).

2.2 Annotation Guidelines and Gold-Standard Data

The design of annotation guidelines has been recognised as a critical component of any large-scale NLP data-creation effort. High-quality annotated corpora directly determine the ceiling of model performance, and the process of guideline development itself constitutes a scholarly contribution by encoding linguistic and editorial knowledge in a reusable, transferable form (Zaghouani et al., 2014; Pierazzo, 2011). Zaghouani et al. (2014) demonstrate that iterative calibration, inter-annotator agreement measurement, and transparent documentation of editorial decisions are indispensable for reliable annotation.

Digital scholarly editions of Arabic and other historical texts have further proposed stratified transcription models that formally distinguish between faithful data acquisition and interpretive annotation layers (Pierazzo, 2011; TEI Consortium, 2023). Within this tradition, gold-standard datasets are not simply accurate transcriptions, but also carefully documented editorial objects whose conventions can be reused and adapted across projects.

2.3 Diplomatic Transcription

Diplomatic transcription aims to reproduce a source document as faithfully as possible, preserving scribal errors, non-standard orthography, and original punctuation, without introducing editorial corrections or normalisation (Burghart, 2017; Pierazzo, 2011). This approach contrasts with more interventionist editorial models that regularise spelling or silently correct perceived mistakes, thereby obscuring evidence about authors' practices and historical usage.

The TEI framework provides a rich vocabulary for encoding features relevant to diplomatic work, including deletions, additions, unclear text, and gaps (TEI Consortium, 2023). Recent platforms such as eScriptorium build on these principles to support large-scale transcription, layout analysis, and annotation workflows for historical documents (Kiessling et al., 2019).

3. Task Description and Data

3.1 Task Overview

The NAKBA-NLP 2026 Shared Task comprises two subtasks. Subtask 1 (Human Transcription Track) provided each participating team with a batch of approximately 500-line images cropped from historical Arabic handwritten manuscripts. Teams were required to submit both a completed annotations.csv file and a guidelines document covering their transcription conventions, corner-case handling strategies, consistency measures, and ethical considerations.

3.2 Dataset Characteristics

Each batch folder contained PNG line images cropped from full manuscript pages (1953–1965); an annotations.csv template with columns filename, text, source_image, year, page, and line; and full-page contextual images. The manuscripts are drawn from the Omar Al-Saleh Memoir Collection, which documents Palestinian Nakba narratives authored by Omar Al-Saleh. The collection presents diverse palaeographic challenges including cross-period variation in letter forms, historical orthographic conventions diverging from modern standard Arabic, physical deterioration, authorial editorial marks, and the concurrent use of Arabic (Eastern) and Western numeral systems.

3.3 Evaluation Metrics

Submissions were assessed on three criteria: (a) coverage, (b) accuracy, and (c) guidelines quality. Accuracy was measured using Character Error Rate (CER) and Word Error Rate (WER) computed against an expert-verified ground truth.

4. Transcription Methodology

4.1 Transcription Philosophy

A faithful diplomatic transcription approach was adopted in which the original spelling, punctuation, diacritics, and stylistic features of the manuscript are preserved without editorial intervention. No orthographic normalisation was applied; no punctuation was inserted or removed; no grammar corrections were introduced; diacritical marks were preserved where the author placed them and not added where absent; and numeral system choices were retained without conversion.

4.2 Annotation Framework Design

Since no predefined guidelines were supplied by the task organizers, a comprehensive framework was designed from scratch. The framework is organised into three functional layers, reflecting three distinct roles in the transcription process. The framework contains exactly 26 conventions in total: 11 editorial-structural symbols, 12 faithful-copying rules, and 3 documentation label types.

4.2.1 Layer 1: Editorial–Structural Symbols (11 Conventions)

Precise symbols were defined for handling the most common palaeographic challenges. Each

symbol carries an unambiguous operational definition, a specified usage context, and at least one illustrative example drawn from actual manuscript images. Three design principles governed symbol selection: minimal ambiguity, visual distinctiveness, and reversibility.

4.2.2 Layer 2: Faithful-Copying Rules (12 Conventions)

These rules operationalise the diplomatic transcription philosophy at a granular level, governing the handling of orthographic and linguistic features that must be reproduced without modification.

4.2.3 Layer 3: Documentation Labels (3 Types)

To maintain a transparent audit trail, three label types were introduced for use exclusively within the guideline document: Decision, Interpretive Note, and Note. This system ensures that the guideline document functions as a transparent research diary of the transcription process.

4.3 Transcription Workflow

The transcription was completed over approximately three days following a structured, iterative workflow. The familiarisation phase involved examining a sample of full-page and line images to characterise the author's handwriting, identify recurring abbreviations, and map orthographic patterns. Guideline drafting produced an initial convention set focused on the most frequently encountered phenomena, and iterative refinement extended the framework to the final 26 conventions.

5. Quality Control

To ensure high-quality, consistent transcriptions across all 500 lines, a four-step quality-control pipeline was implemented. All transcription work was conducted by a single annotator following the proposed guideline, with multiple self-review passes designed to approximate the rigour of multi-annotator workflows through systematic consistency checks.

Step 1 consisted of periodic batch review. Step 2 consisted of cross-document consistency checks. Step 3 involved anomaly detection for duplicate image IDs, empty text fields, unexpected symbol combinations, and outlier line lengths. Step 4 required transparent documentation of every ambiguous reading decision in the guideline document.

6. Results

The submission achieved full coverage of the assigned batch: all 500-line images were transcribed and included in the final annotations file. The official shared-task evaluation returned a CER of 0.02, WER of 0.01, and accuracy of 0.98, reflecting the high character-level accuracy attained through the combination of diplomatic transcription philosophy and the structured quality-control pipeline.

This result demonstrates that a carefully designed single-annotator workflow, when supported by explicit conventions and iterative review, can produce transcription quality that is competitive with reported automated Arabic HTR systems operating on similarly degraded historical manuscript material.

7. Discussion and Recommendations

The CER of 0.02, WER of 0.01, and accuracy of 0.98 confirm that expert human transcription, guided by a principled convention system, can achieve very low character-level error rates on challenging Arabic manuscript material. The three-layer annotation framework proved particularly valuable in stabilising editorial decisions over time.

A further observation concerns the complementarity between human and automated approaches. Human annotators can efficiently produce small, carefully curated datasets that capture the full palaeographic complexity of a given collection; automated systems subsequently excel at scaling those annotations to tens of thousands of lines. Future shared tasks would benefit from full-page context, explicit scoring criteria for guideline quality, and at least a subset of multi-annotator data.

8. Conclusion

This paper presented a faithful diplomatic transcription approach for historical Arabic manuscripts applied to the NAKBA-NLP 2026 Shared Task. A 26-convention annotation framework organised in three functional layers was developed from scratch and supported by a four-step quality-control pipeline. The resulting submission achieved full coverage of 500 lines with CER of 0.02, WER of 0.01, and accuracy of 0.98.

Limitations

This study was conducted by a single annotator; therefore, formal inter-annotator agreement could not be computed from the submitted data alone. In addition, the 26-convention framework was developed specifically for the Omar Al-Saleh Memoir Collection, and its transferability to other Arabic manuscript collections remains to be tested.

Ethics Statement

The manuscripts transcribed in this task contain historical narratives related to the Palestinian Nakba. This material was approached with respect and cultural sensitivity, maintaining strict fidelity to the original text without editorial

modification. No personal data were collected or processed beyond the materials provided by the task organizers.

Acknowledgements

The author thanks the organizers of the NAKBA-NLP 2026 Shared Task for providing the data, benchmark structure, and evaluation framework that made this study possible.

Bibliographical References

Aabed, S. and Khairaldin, A. (2024). An end-to-end, segmentation-free, Arabic handwritten recognition model on KHATT. arXiv preprint arXiv:2406.15329.

Ahmad, R., Naz, S., Afzal, M. Z., Rashid, S. F., Liwicki, M., and Dengel, A. (2017). KHATT: A deep learning benchmark on Arabic script. In Proceedings of ICDAR, pages 10–14.

Ahmad, R., Naz, S., Afzal, M. Z., Rashid, S. F., Liwicki, M., and Dengel, A. (2020). A deep learning based Arabic script recognition system: Benchmark on KHATT. International Arab Journal of Information Technology, 17(3):299–305.

Al-Mutawa, H. et al. (2024). A comparative study of four handwritten text recognition models in Arabic script. Ingénierie des Systèmes d'Information, 29(6):2243–2250.

Bhatia, G., Nagoudi, E. M. B., Alwajih, F., and Abdul-Mageed, M. (2024). Qalam: A multimodal LLM for Arabic optical character and handwriting recognition. In Proceedings of ArabicNLP 2024, pages 210–224.

Burghart, M. (2017). Transcription or diplomatic edition. In Digital Scholarly Editions. Queen Mary University of London.

Hamoud, H. et al. (2026). NAKBA NLP 2026: Shared task on Arabic handwritten manuscript understanding. In Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with LREC 2026.

HATFormer (2024). Historic handwritten Arabic text line recognition using transformers. arXiv preprint arXiv:2410.02179.

Kiessling, B., Tissot, R., Stokes, P., and Stökl Ben Ezra, D. (2019). eScriptorium: An open-source platform for historical document analysis. In Proceedings of ICDARW, volume 2, pages 19–19.

Pierazzo, E. (2011). A rationale of digital documentary editions. Literary and Linguistic Computing, 26(4):463–477.

TEI Consortium (2023). TEI P5: Guidelines for electronic text encoding and interchange. Available at: https://tei-c.org/guidelines/.

Waly, A. et al. (2025). Invizo: Arabic handwritten document optical character recognition solution. arXiv preprint arXiv:2502.05277.

Zaghouani, W. et al. (2014). Large scale Arabic error annotation: Guidelines and framework. In Proceedings of LREC'14.

Zaraket, F. et al. (2026). AR-MS: Arabic manuscript understanding. In Proceedings of Nakba-NLP 2026, co-located with LREC 2026.

Appendix A: Convention Tables

The appendix below provides the convention tables referenced in Section 4. In the camera-ready version, these materials may be included after the references.

Table 1: Layer 1 : Editorial–Structural Symbols (11 Conventions)

#	Symbol	Name	Usage	Example
1	[...]	Truncation marker	Placed where part of a word is missing or cropped.	1962_p189_l0069.png
2	(?)	Illegible word marker	Replaces a completely unintelligible word.	1956_p054_l0040.png
3	*	Unclear line marker	Marks an entire faded, damaged, or partially illegible line.	1955_p043_l0009.png
4	{...}	Editorial insertion	Encloses text added by transcriber for clarification.	1956_p063_l0031.png
5	|word|	Struck-through correction	Encloses words the author struck through.	1954_p172_l0006.png
6	|	Overlapping-lines separator	Inserted between two lines stacked in one image.	1962_p129_l0034.png

7	~	Added-letter marker	Flags an extra letter the author added.	1953c_p010_l0056.png
8	|.....|	Line-discontinuity marker	Used when segments are not textually connected.	1962_p063_l0069.png
9	« »	French quotation marks	Reproduced as employed by the author.	1959_p052_l0051.png
10	_ _	Parenthetical remarks	Reproduces author's underscore markings.	1964_p011_l0050.png
11	✓	Check mark	Reproduced to preserve author's annotations.	1964b_p069_l0034.png

Table 2: Layer 2 : Faithful-Copying Rules (12 Conventions)

#	Convention	Rule	Illustrative Case
1	Orthographic variations	Preserved exactly. No normalisation.	إلي vs. إلى
2	Original punctuation only	No punctuation added unless present.	Periods/commas as-is
3	Tanwīn notation	Applied strictly per original.	Bare alif retained
4	Numerals	Transcribed in author's system (Arabic/Western).	No conversion
5	Hamzah spelling errors	Preserved as written, even if incorrect.	مسؤول vs. مسئول
6	Tāʾ marbūṭa / hāʾ	Transcribed as they appear.	ه/ ة kept as written
7	Allāh without diacritics	Written without diacritical marks.	الله without shaddah
8	Diacritical marks	Preserved where present; not added.	Tashkīl retained
9	Compound words	Kept as single connected forms.	قائمقام as one unit
10	Incomplete words	Transcribed without supplying letters.	Partial words as-is
11	Semantically unclear words	Transcribed without modification.	No paraphrase
12	Unnecessary prepositions	Extraneous words kept without omission.	Redundant terms retained

KvochurHegel at StanceNakba: Robust Stance Detection with Regularized Natural Language Inference

Minh-Hoang Le
University of Information Technology, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
24520542@gm.uit.edu.vn

Abstract

Actor-level stance detection over noisy, politically sensitive data can present challenges that standard training procedures fail to handle reliably. This paper presents KvochurHegel, our submission to the StanceNakba 2026 Shared Task, which addresses these challenges by framing stance classification as Natural Language Inference (NLI) to capture actor-level granularity. The official StanceNakba dataset contains high label noise and topic-correlated spurious features, such as texts discussing unrelated global conflicts using in-domain political vocabulary. To handle these conditions within a three-class schema, we construct templates encoding stance hypotheses for specific actors (e.g., "The author expresses support for Palestine") and introduce a broadened neutral class designed to absorb spurious out-of-domain inputs. A DeBERTa-v3 Cross-Encoder independently evaluates the entailment between the input text and each class-specific hypothesis. Because standard cross-entropy training tends to memorize contradictory annotations under these conditions, we regularize the training procedure with R-Drop and label smoothing. This regularized setup likely contributed to robustness against distribution shifts between the competition's evaluation phases (the public leaderboard and private test set), allowing our model to improve from a Macro-F1 of 0.9094 to 0.9384 without requiring large generative models, cross-validation, or inference-time ensembling.

Keywords: Actor-Level Stance Detection, Natural Language Inference (NLI), Label Noise, R-Drop, DeBERTa, StanceNakba

1. Introduction

Actor-level stance detection in polarized social media discourse often requires models to interpret implicit rhetoric and complex ideological framing (ALDayel and Magdy, 2021). The Palestinian-Israeli conflict represents one of the most contested and heavily discussed topics in online political discourse, making automated stance detection over such data both practically relevant and technically challenging. To address this and provide a standardized benchmark, Aldous et al. (2026) introduced the StanceNakba 2026 Shared Task. This paper describes our system submitted to this competition, specifically targeting Subtask A: Actor-Level Stance Detection. The objective is to classify an author's general political alignment into one of three categories: Pro-Palestine, Pro-Israel, or Neutral.

Analysis of the training and development sets identifies severe label noise, contradictory annotations, and topic-correlated spurious features. Texts discussing unrelated geopolitical topics often use in-domain vocabulary (e.g., "military", "colonization"), creating spurious out-of-domain inputs. Under these conditions, standard cross-entropy training tends to memorize contradictory annotations rather than learning generalizable patterns.

To mitigate this behavior, we frame stance detection as an NLI task (Yin et al., 2019) using a DeBERTa-v3 Cross-Encoder. We construct actor-specific hypothesis templates and a broadened neutral class, regularizing the training procedure with R-Drop and label smoothing to prevent overfitting on the noisy annotations. This approach likely contributed to robustness against distribution shifts between the competition's two evaluation phases, improving from a Macro-F1 of 0.9094 on the public leaderboard to 0.9384 on the private test set, without cross-validation or synthetic data.

2. Background

2.1. Related Work

Stance detection has been approached primarily as target-specific text classification, with early systems relying on lexical features and more recent work adopting pre-trained transformer fine-tuning (Mohammad et al., 2016; Küçük and Can, 2020). These approaches, however, rely on surface-level associations that fail under implicit rhetoric and topic-correlated spurious features. Yin et al. (Yin et al., 2019) demonstrate that reformulating text classification as Natural Language Inference improves robustness on implicit reasoning tasks by leveraging pre-trained entailment representations, motivating our hypothesis-based formulation.

A separate challenge in stance detection is annotation noise, particularly in politically sensitive domains where rater disagreement is high (ALDayel and Magdy, 2021). Label smoothing (Szegedy et al., 2016) and consistency regularization via R-Drop (Liang et al., 2021) are established techniques

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 113–117
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

for reducing model overconfidence under noisy supervision. Our system combines NLI formulation with both techniques, addressing implicit rhetoric and label noise simultaneously without requiring generative model inference.

2.2. Task Definition

Subtask A of the StanceNakba Shared Task focuses on predicting general political alignments at the actor level. Formally, given an input social media text X, the system must predict a stance label $y \in \{\text{Pro-Palestine}, \text{Pro-Israel}, \text{Neutral}\}$. The official dataset consists of 1,401 English samples, partitioned into a 70/15/15 split for training, development, and testing.

2.3. Data Characteristics and Challenges

Analysis of the training and development sets identified three primary phenomena that disrupt standard classification pipelines, detailed in Table 1.

Implicit Rhetoric Stances are often expressed through indirect ideological markers rather than direct declarations of support (Table 1). Models relying on lexical overlap fail to map terms like "colonization" to the correct geopolitical actor without an explicit inference mechanism.

Spurious Out-of-Domain Inputs The dataset includes texts that use the vocabulary of conflict and politics but express no stance on the Palestinian-Israeli conflict. If a model associates words like "military" strictly with the Pro-Israel or Pro-Palestine classes, it will misclassify these inputs.

Contradictory Annotations A particularly disruptive challenge is the presence of severe label noise. Polarized texts are occasionally assigned the "Neutral" label. Training on such data using standard cross-entropy severely degrades generalization, as the network is forced to map explicitly partisan phrasing to neutrality—a failure mode we address through regularization during fine-tuning.

3. Methodology

3.1. Hypothesis Design and NLI Formulation

To preserve the original discourse features of the social media data—including capitalization, punctuation, and special characters—we apply no text preprocessing or normalization. The raw texts are passed directly to the tokenizer. We formulate actor-level stance detection as an independent Natural Language Inference (NLI) task. We design three class-specific hypotheses (H_c), each articulating the defining characteristics of one stance label. This approach allows the model to jointly encode the input text and hypothesis, computing attention across both sequences at every layer.

By including in-domain ideological markers (e.g., "IDF", "Zionism", "resistance") and defining the neutral scope ("unrelated topics"), these hypotheses directly address the implicit rhetoric and spurious inputs identified in Section 2. Evaluating each hypothesis independently, rather than applying a joint three-class classification head, prevents the model from learning inter-class shortcuts and preserves the binary entailment semantics of each hypothesis. During inference, the model evaluates the premise X against all three hypotheses. The final predicted class $\hat{y}$ is the hypothesis that yields the maximum entailment probability (obtained via softmax over the raw NLI logits):

$$\hat{y} = \text{argmax}_c P(\text{Entailment} \mid X, H_c)$$

Although independent evaluation requires three forward passes per example, inference remains efficient given the dataset scale and the absence of ensemble mechanics.

3.2. Architecture and Training Configuration

Our system uses a DeBERTa-v3 Cross-Encoder (He et al., 2023) initialized with the `MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli` checkpoint[1], which is pre-trained on MNLI, FEVER-NLI, and ANLI, offering robust NLI priors for the entailment scoring task. For each forward pass, the tokenizer concatenates the input sequence as `[CLS] X [SEP] H_c [SEP]`. The pooled `[CLS]` representation is passed through a linear classification head to produce the standard three NLI logits (entailment, neutral, contradiction), from which only the entailment logit is used.

The model is optimized using AdamW (Loshchilov and Hutter, 2019) with a learning rate of 2×10^{-5}, a weight decay of 0.1, and a warmup ratio of 0.1. We train for a maximum of 10 epochs using a training batch size of 16 (and an evaluation batch size of 32) with `bfloat16` mixed precision. To prevent overfitting to the noisy training annotations, we implement early stopping on the validation loss with a patience of 2 epochs, which triggered at epoch 9 in our final training run.

3.3. Regularization Strategy

Because standard cross-entropy training tends to memorize the severe label noise present in the

[1] https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli

Phenomenon	Gold Label	Example Text	Modeling Challenge
Implicit Stance	Pro-Palestine	*"I'm a Hindu,! And you don't speak in behalf of others! Zionists colonization..."* (ID: 807)	Lacks explicit entity mentions; relies entirely on ideological terminology.
Spurious Inputs	Neutral	*"Did you ever serve in the military? ."* (ID: 247)	Contains conflict-related vocabulary but lacks geopolitical stance.
Severe Label Noise	Neutral (Mislabeled)	*"Allah stands with Israel. He has blessed them with many victories over the vile rats..."* (ID: 70)	Contradictory annotation. The text is explicitly partisan but labeled Neutral.

Table 1: Examples of dataset ambiguity, spurious out-of-domain inputs, and label noise in the StanceNakba dataset.

Class	Hypothesis Text (H_c)
Israel	*"The author expresses support for Israel, including its government, military (IDF), security, Zionism, or the Jewish people's right to self-defense."*
Palestine	*"The author expresses support for Palestine, Gaza, or the resistance, including calls for liberation, human rights, or stopping the violence."*
Neutral	*"The text provides a factual, non-judgmental report on the conflict, or discusses completely unrelated topics (like other global wars or daily life)."*

Table 2: Class-specific hypotheses encoding ideological markers for stance classes and a broadened neutral scope to absorb spurious inputs.

StanceNakba dataset, we apply two regularization techniques during fine-tuning: label smoothing and R-Drop.

First, we apply label smoothing ($\epsilon = 0.2$) (Szegedy et al., 2016; Müller et al., 2019) to the cross-entropy objective ($\mathcal{L}_{CE}$). This softens the one-hot target distributions, discouraging the network from assigning maximum confidence to any training label, which is particularly beneficial under contradictory annotations.

Second, we implement R-Drop consistency regularization (Liang et al., 2021) via a custom training loop, as the standard Hugging Face Trainer API does not natively support the requisite double forward pass per step. For each training step, the input sequence is passed through the model twice with independently sampled dropout masks, generating two distinct logit distributions, P_1 and P_2. We enforce predictive consistency between these sub-models by minimizing their bidirectional Kullback-Leibler (KL) divergence ($\mathcal{L}_{KL}$):

$$\mathcal{L}_{KL} = \frac{1}{2}\Big(D_{KL}(P_1 \parallel P_2) + D_{KL}(P_2 \parallel P_1)\Big)$$

The final objective function $\mathcal{L}$ is the sum of the averaged smoothed cross-entropy loss from both forward passes and the KL divergence loss, scaled by $\alpha = 4.0$ (selected via evaluation on the development set):

$$\mathcal{L} = \frac{1}{2}(\mathcal{L}_{CE}^{(1)} + \mathcal{L}_{CE}^{(2)}) + \alpha\mathcal{L}_{KL}$$

The impact of this regularized setup on model robustness is evaluated in Section 4.

3.4. Implementation Details

Our system is implemented in Python 3.13.2 using PyTorch 2.7.1 (Paszke et al., 2019)[2] and the Hugging Face Transformers library 4.55.4 (Wolf et al., 2020)[3]. Dataset loading and preprocessing use the Hugging Face Datasets library 4.5.0 (Lhoest et al., 2021)[4]. Evaluation metrics are computed using scikit-learn 1.7.1 (Pedregosa et al., 2011)[5]. The complete training notebook is publicly available in our repository[6].

4. Results and Discussion

4.1. Official Evaluation

Submitted under the team name KvochurHegel, our system achieved a Macro-F1 score of 0.9384, securing Rank 4 on the official leaderboard.

As shown in Table 3, precision, recall, and accuracy align closely with the Macro-F1 score, suggesting broadly balanced per-class performance.

[2] https://pytorch.org/
[3] https://huggingface.co/transformers
[4] https://huggingface.co/docs/datasets/
[5] https://scikit-learn.org/
[6] https://github.com/lmhoang06/StanceNakba2026_KvochurHegel

Phase	Macro-F1	Acc.	Prec.	Rec.
Public	0.9094	0.9095	0.9096	0.9095
Private	0.9384	0.9384	0.9387	0.9384

Table 3: System performance across the public and private evaluation phases.

For context, the top three systems on the leaderboard achieved Macro-F1 scores of 0.9620, 0.9525, and 0.9426, respectively. Our single-run submission performs within 0.024 points of the winning system.

4.2. Performance and Generalization

The 0.029-point Macro-F1 improvement between evaluation phases likely reflects two contributing factors. First, the regularization strategy (R-Drop and label smoothing) reduced overfitting to training noise, consistent with the improved private test performance. Second, the private test distribution may have naturally contained fewer contradictory annotations or spurious out-of-domain inputs than the public split, constituting an objectively easier evaluation set.

4.3. System Limitations

While the system avoids complex inference-time ensembles, evaluating three independent NLI hypotheses per sample triples the inference cost relative to a standard classification head. However, this overhead remains acceptable given the dataset scale. Furthermore, the current formulation evaluates each post in isolation, without access to conversational context such as parent posts or reply threads. This restricts the model's ability to resolve implicit stances that depend entirely on discourse context (e.g., agreeing with an unincluded parent post).

4.4. Per-Class Performance and Error Analysis

To clarify the evaluation framework, the terms "public" and "private" test sets refer strictly to the two sequential phases of the shared task evaluation, rather than standard static development and test splits. Because per-class breakdowns were not provided by the organizers for the final private phase, Table 4 reports the detailed metrics and corresponding confusion matrix evaluated on the public test set.

The system achieves its highest F1 on Pro-Israel (0.936), with Pro-Palestine (0.899) and Neutral (0.894) showing comparable but lower performance. Pro-Palestine exhibits the highest misclassification count (8 instances), with the dominant

Per-Class Metrics

Class	P	R	F1
PI	0.930	0.943	0.936
PP	0.912	0.886	0.899
Neu	0.887	0.900	0.894

Confusion Matrix

	PI	PP	Neu
PI	66	2	2
PP	2	62	6
Neu	3	4	63

Table 4: Public test set performance breakdown. Matrix rows represent the true labels, and columns represent the predicted labels (PI: Pro-Israel, PP: Pro-Palestine, Neu: Neutral). P: Precision, R: Recall.

error being confusion with the Neutral class (6 instances), consistent with the spurious input challenge identified in Section 2.

Inspection of the 19 misclassified public test samples reveals two recurring failure modes. First, the absence of conversational context forces errors on texts whose stance depends on an unincluded parent post. For example, *", and I say no to war. No I don't stand with ..."* (ID: 1008) is predicted Pro-Palestine but labeled Neutral, likely because the implied stance requires the unincluded parent post for resolution. Second, residual label noise that regularization cannot fully override produces irreducible errors. The text *"I stand against jihad. Merci Israel."* (ID: 1088) contains explicitly Pro-Israel rhetoric but carries a Pro-Palestine gold label; the model predicts Neutral, a more semantically coherent output than the gold annotation. These cases illustrate that performance is bounded by both annotation quality and the absence of conversational context, consistent with the limitations noted in Section 4.3.

5. Conclusion and Future Work

This paper formulated actor-level stance detection as an independent Natural Language Inference task. We addressed spurious out-of-domain inputs through class-specific hypothesis design, and mitigated the memorization of severe label noise via R-Drop and label smoothing. Achieving Rank 4 with a single-run DeBERTa-base model demonstrates that combining NLI formulation with engineered hypotheses and regularization achieves competitive performance without requiring large generative models, cross-validation, or inference-time ensembling. Future work may extend this framework to incorporate discourse context, enabling the resolution of implicit stances within reply threads while maintaining the efficiency of the single-run architecture.

6. Bibliographical References

Abeer ALDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Kholoud Khalil Aldous, Md Rafiul Biswas, Mabrouka Bessghaier, Shimaa Ibrahim, Kais Attia, and Wajdi Zaghouani. 2026. StanceNakba shared task: Actor and topic-aware stance detection in public discourse. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Comput. Surv.*, 53(1).

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 10890–10905. Curran Associates, Inc.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

KvochurHegel at NakbaArchiveClassifier Shared Task: Nakba Image Classification via ConvNeXt-V2 and Label Smoothing

Minh-Hoang Le
University of Information Technology, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
24520542@gm.uit.edu.vn

Abstract

This paper presents the KvochurHegel team's submission to the Nakba Image Classification shared task at the Nakba-NLP 2026 Workshop. The task requires the binary classification of social media images into `destruction` and `not_destruction` categories. Given a limited and imbalanced training set of 1,400 images, we utilized a ConvNeXt-V2 Nano backbone combined with extensive data augmentation and label smoothing, prioritizing standard regularization over task-specific architectural modifications. For inference, we applied a 6-view Test-Time Augmentation (TTA) strategy using a hard-voting mechanism. The baseline system achieved a Macro F1-score of 0.8593 and an Accuracy of 0.8706 on the official private test set, ranking 6th out of 16 participating teams.

Keywords: Image Classification, Destruction Detection, Social Media Imagery, Conflict Imagery, ConvNeXt-V2

1. Introduction

The Nakba Image Classification shared task (Abrahams et al., 2026), part of the Nakba-NLP 2026 Workshop, focuses on categorizing social media imagery to improve the accessibility of archived records from the Tech for Palestine incubator project. The dataset comprises 2,001 images sourced from Instagram, published by Palestinian content creators and journalists in Gaza between October 7, 2023, and December 15, 2025. The objective is to develop a binary image classifier capable of distinguishing images that show destroyed or damaged infrastructure (`destruction`) from those that do not (`not_destruction`).

Our team, KvochurHegel, prioritized a standard, regularized Convolutional Neural Network (CNN) over task-specific architectural modifications, given the dataset's class imbalance (64.7% / 35.3%) and the visual noise common in social media imagery. We submitted a ConvNeXt-V2 Nano baseline trained with extensive data augmentation and evaluated using a multi-view Test-Time Augmentation (TTA) strategy. The system ranked 6th out of 16 participating teams, achieving a Macro F1-score of 0.8593 on the official private test set.

2. Experimental Setup

To support reproducibility, all training and inference experiments described in Section 3 were conducted in the following software environment:

- Python 3.10.0
- PyTorch 2.9.1 (compiled with CUDA 12.6)
- MMEngine 0.10.7
- MMCV 2.2.0
- MMPretrain 1.2.0

3. Methodology

The Nakba Image Classification task provides a dataset of 2,001 images, partitioned into 1,400 for training, 199 for validation, and 402 for testing. The training set consists of 906 `not_destruction` and 494 `destruction` images (64.7% / 35.3%). We use the validation set to select the model checkpoint that achieves the highest Macro F1-score, which is the primary evaluation metric for the shared task. All experiments are conducted on a single NVIDIA RTX 4060 GPU with 8GB of VRAM.

3.1. Model Architecture

We use the ConvNeXt-V2 Nano backbone, selected for offering a competitive accuracy-to-parameter ratio, as demonstrated in (Woo et al., 2023). We initialize the backbone with weights pre-trained on ImageNet-21K and fine-tuned on ImageNet-1K at 384×384 resolution. The classification head is a linear layer mapping the 640-dimensional global average pooled features to the two target classes.

3.2. Training and Regularization

The model is trained for 50 epochs with a batch size of 16 using the AdamW optimizer (lr $= 2 \times 10^{-4}$, weight decay$= 0.3$) (Loshchilov and Hutter, 2019). We adopt this configuration from the MMPreTrain baseline, which was used during the backbone's original ImageNet-1K fine-tuning phase. The learning rate follows a 5-epoch linear warmup and a cosine annealing schedule decaying to 1×10^{-6}.

As a regularizer for the small training set, we apply a Label Smoothing (Müller et al., 2019) coefficient of 0.2, determined on the validation set. Training augmentations include *RandomResizedCrop*, horizontal flips, and the `timm_increasing`

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 118–120
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

RandAugment policy (Cubuk et al., 2020). We also apply *RandomErasing* ($p = 0.25$) (Zhong et al., 2020) to discourage the model from over-fitting to local patches. We maintain an Exponential Moving Average (EMA) of the weights with a smoothing factor of 0.9995 (implemented as 0.0005 momentum in MMPreTrain (Contributors, 2023)) to obtain a weight-averaged checkpoint less sensitive to late-training oscillations. Given the constrained one-week development timeline, no explicit class-weighting or oversampling was implemented to address the 64.7% / 35.3% dataset imbalance.

3.3. Inference Strategy

We use a 6-view Test-Time Augmentation (TTA) strategy for the final evaluation. Test images are resized to 384 pixels on the short edge. We then extract three 384×384 spatial crops (left/top, center, and right/bottom) sampled along the longer dimension, along with their horizontally mirrored counterparts.

Final labels are assigned through a hard-voting mechanism. For each of the 6 views, a vote is cast for `destruction` if the softmax probability for that class exceeds 0.5. The majority class is selected as the final prediction. In the case of a tie (3 votes each), the prediction from the original center-cropped view serves as the tie-breaker.

4. Results

Our team, KvochurHegel, submitted our ConvNeXt-V2 baseline to the Nakba Image Classification shared task. The system was evaluated on the private test set of 402 images.

The system achieved a final Macro F1-score of 0.8593 and an Accuracy of 0.8706. Table 1 reports all evaluation metrics returned by the task organizers. The parity between Macro Recall (0.8616) and Precision (0.8572) suggests that the model did not exhibit a systematic bias toward the majority class, despite the lack of explicit class-weighting during training. Additionally, the identical values for Recall, Specificity, and Balanced Accuracy (0.8616) indicate that the model achieved equal per-class recalls across both the minority and majority classes.

Our submission ranked 6th out of 16 participating teams. The winning system achieved a Macro F1-score of 0.899. The margin of 0.0397 F1 to the first-place entry suggests that a standard, regularized CNN pipeline performed competitively on this task.

Metric	Score
Macro F1-score	0.8593
Accuracy	0.8706
Balanced Accuracy	0.8616
Precision	0.8572
Recall	0.8616
Specificity	0.8616

Table 1: Official private test set evaluation metrics for the KvochurHegel team submission.

5. Conclusion

We described a ConvNeXt-V2 Nano baseline system developed for the Nakba Image Classification shared task. To address the constraints of a small 1,400-image training set and a 64.7% / 35.3% class imbalance, the approach relied on adopting the MMPreTrain default hyperparameters for ConvNeXt-V2, extensive data augmentation, label smoothing, and a 6-view Test-Time Augmentation (TTA) strategy. These design choices prioritized standard regularization over task-specific architectural modifications.

The system achieved a Macro F1-score of 0.8593, ranking 6th out of 16 teams. This suggests that a standard, regularized CNN pipeline performs competitively on this task, finishing within 0.04 F1 of the winning system. Due to a constrained one-week development timeline, explicit class-aware training mechanisms were omitted. We leave the exploration of oversampling and class-weighted loss functions on this dataset to future work.

6. Bibliographical References

Alexei Abrahams, Shadi Abudalfa, Mustafa Jarrar, and George Mikros. 2026. The nakbaarchive-classifier shared task on nakba image classification. In *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

MMPreTrain Contributors. 2023. Openmmlab's pre-training toolbox and benchmark. `https://github.com/open-mmlab/mmpretrain`.

Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help?

In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. 2023. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16133–16142.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2020. Random erasing data augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13001–13008.

HCMUS_TheFangs at NakbaArchiveClassifier Shared Task: Foundation Models and Advanced Training Strategies for Conflict Damage Classification

Dao Sy Duy Minh* **Huynh Trung Kiet***
Tran Chi Nguyen **Nguyen Lam Phu Quy** **Pham Phu Hoa**
*Joint first authors
Faculty of Information Technology, University of Science,
Vietnam National University Ho Chi Minh City
Ho Chi Minh City, Vietnam
{23122041, 23122039, 23122044, 23122048, 23122030}@student.hcmus.edu.vn

Abstract

We present our system for the NakbaArchiveClassifier shared task at Nakba-NLP 2026, which requires classifying Instagram images from Gaza as showing destroyed or damaged infrastructure versus intact surroundings. Working with a small, imbalanced dataset (1,400 training images; 1.83:1 class ratio), we conduct a systematic empirical study of six model-training combinations spanning five architecture families: standard CNNs (EfficientNet-B4), self-supervised ViTs (DINOv2-ViT-L), hybrid multi-axis Transformers (MaxViT-Base), masked-image-modelling ViTs (EVA-02-Base), and large-kernel CNNs (UniRepLKNet). For our best performing configuration–MaxViT-Base with focal loss, MixUp, and a rich geometric augmentation pipeline–we provide a detailed component analysis. Our system achieves a macro F1 of **0.899** on the public test set, ranking **1st** on the competition leaderboard. We additionally report findings from novel experiments including a Kolmogorov-Arnold Network (KAN) classification head and VLM-regularized training with BLIP-2-generated captions, offering insights into what does and does not transfer to conflict-domain imagery under severe data scarcity.

Keywords: image classification, conflict damage, foundation models, MaxViT, focal loss, KAN, data augmentation, class imbalance, Nakba

1. Introduction

The SaltPillar archive (Tech for Palestine) is collecting millions of Gaza conflict images from Instagram to form a searchable historical record. The NakbaArchiveClassifier shared task (Abrahams et al., 2026) formalises this as binary classification-destruction vs. not_destruction-evaluated by macro F1. Three properties make it non-trivial: only 1,400 labeled training images (1.83:1 class imbalance), visually distinctive domain content (rubble microstructure, JPEG compression artefacts absent from ImageNet), and a **photographer-bias** risk: 118 unique content creators appear in both train and validation splits, so a powerful model can maximise validation F1 by memorising author-specific style rather than damage semantics. Our system, HCMUS_TheFangs, addresses these jointly through backbone selection, focal-loss training, and a bias-targeted augmentation cascade, achieving macro F1 = **0.899** (1st place). We additionally report the first application of a KAN (Liu et al., 2024) classification head and VLM-regularised training (Li et al., 2023) to conflict imagery, with frank failure analyses of both.

2. Task and Dataset

The task is binary classification of Instagram images from Gaza (Abrahams et al., 2026): predict destruction vs. not_destruction, evaluated by macro F1. Table 1 shows the split statistics. The 1.83:1 imbalance is perfectly consistent across all three splits. A key EDA finding: **118 photographers** appear in both train and validation, creating a shortcut risk that motivates the photographer-bias defences in Section 3.

Split	Total	destruction	not_destr.
Train	1,400	494 (35.3%)	906 (64.7%)
Val	199	70 (35.2%)	129 (64.8%)
Test	402	141 (35.1%)	261 (64.9%)

Table 1: Class distribution across splits (Abrahams et al., 2026). The 1.83:1 imbalance is perfectly consistent, ruling out split-specific artifacts.

3. Methods

Our system architecture is shown in Figure 1. A shared training protocol is applied across all backbone variants to ensure fair comparison; differences in the results table can therefore be attributed to architecture and loss function choices alone.

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 121–127
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

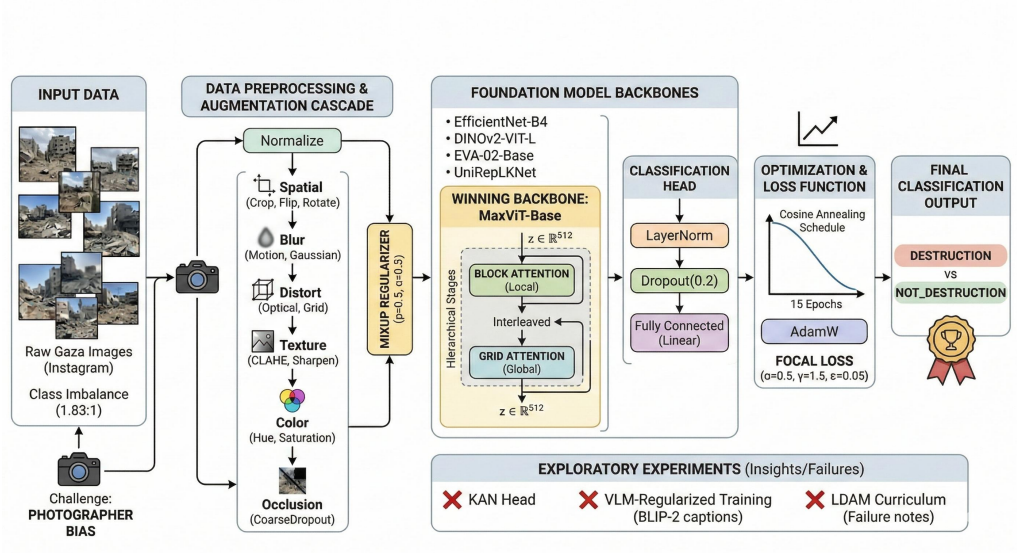


Figure 1: Full training pipeline of our best system (EXP009). The top row shows the end-to-end flow during training: each input image passes through a six-step augmentation cascade, is stochastically mixed with a second image via MixUp ($p = 0.5$), forwarded through the MaxViT-Base backbone, projected by a lightweight classification head, and trained with Focal Loss. At inference, the MixUp branch is skipped and the best-checkpoint model (selected on validation macro F1) is used. The bottom-left panel details each augmentation step; the bottom-right panel shows the four-stage MaxViT architecture, where each stage interleaves an MBConv block with local Block Attention and global Grid Attention before progressive spatial downsampling.

3.1. Shared Training Protocol

To ensure fair backbone comparison, we fix a shared training protocol across all six experiments: AdamW optimisation with a cosine-annealing learning-rate schedule, 15 training epochs with best-checkpoint selection on validation macro F1, and ImageNet-standard normalisation. Full hyper-parameter values-learning rates, batch sizes, weight decay, normalisation constants, and hardware details-are reported in Appendix A.

3.2. Backbone Architectures

We benchmark five architecture families spanning the modern representation landscape. At one end sits the supervised CNN baseline, **EfficientNet-B4** (Tan and Le, 2019), which provides a solid starting point but is limited by ImageNet-only pretraining. At the other extreme, **DINOv2-ViT-L/14** (Oquab et al., 2024) brings the largest capacity (307M parameters) and self-supervised features, but its sheer size proves a liability on our small dataset. Between these poles, we test three architectures that trade off inductive biases in different ways: **EVA-02-Base** (Fang et al., 2023), which reconstructs CLIP (Radford et al., 2021) features via masked image modelling; **UniRepLKNet** (Ding et al., 2024), which replaces attention entirely with large-kernel convolutions up to 31×31; and our winning backbone, **MaxViT-Base** (Tu et al., 2022), detailed in Section 3.3. Table 2 summarises the results; the discussion that follows focuses on the insights these comparisons reveal.

3.3. Winning Model in Detail: MaxViT-Base with Focal Loss

MaxViT (Tu et al., 2022) interleaves **Block Attention** (local 8×8 windows) and **Grid Attention** (dilated global sampling) at $\mathcal{O}(HW)$ cost per stage, giving simultaneous fine-grained texture sensitivity and global scene context-exactly what destruction classification requires. Four hierarchical stages (64/128/256/512 ch) are detailed in Figure 1. The backbone output $\mathbf{z} \in \mathbb{R}^{512}$ feeds a slim head:

$$h(\mathbf{z}) = \mathrm{Linear}_{512\to2}\big(\mathrm{Dropout}(0.2)\big(\mathrm{LayerNorm}(\mathbf{z})\big)\big). \quad (1)$$

The augmentation cascade (Figure 1, left panel) is designed in three tiers. Standard geometric/blur transforms (groups 1–2) cover camera diversity. Optical/grid distortions and texture randomisation (groups 3–4, each $p = 0.3$) are our key photographer-bias defences: non-linear spatial warping destroys low-frequency stylistic cues (vignetting, aspect-ratio preferences) while texture jitter breaks post-processing fingerprints. Colour jitter and CoarseDropout (groups 5–6) handle illumination and occlusion. Batch-level `MixUp` (Zhang et al., 2018) ($p = 0.5$, $\lambda \sim \mathrm{Beta}(0.2, 0.2)$) is the primary capacity regulariser.

Focal Loss (Lin et al., 2017) with $\alpha = 0.5$, $\gamma = 1.5$, $\epsilon_{\text{smooth}} = 0.05$ handles the 1.83:1 imbalance:

$$\mathcal{L}_{\text{focal}}(\hat{p}, y) = -\alpha\,(1 - \hat{p}_y)^{\gamma}\,\log \hat{p}_y. \qquad (2)$$

When MixUp is active: $\mathcal{L} = \lambda\,\mathcal{L}_{\text{focal}}(\cdot, y_a) + (1-\lambda)\,\mathcal{L}_{\text{focal}}(\cdot, y_b)$. Optimiser: AdamW (Loshchilov and Hutter, 2019), LR 2×10^{-5}, CosineAnnealing ($T_{\max} = 15$).

3.4. Additional Experiments

Beyond backbone selection, we probe three research questions. Can *language supervision* regularise visual features by aligning them with BLIP-2 (Li et al., 2023) captions? Can a *curriculum-based margin loss* (Cao et al., 2019) handle the 1.83:1 imbalance better than focal loss? And can a *KAN* (Liu et al., 2024) classification head-which replaces fixed activations with learnable B-spline functions-outperform a standard MLP? We additionally explore domain adversarial training (Ganin et al., 2016) to penalise photographer-identity prediction. Full configurations are provided in Appendix A.

4. Experiments and Results

4.1. Implementation Details

All experiments use PyTorch (Paszke et al., 2019) with `timm` (Wightman, 2019) and `albumentations` (Buslaev et al., 2020), following the shared protocol of Section 3. Full hyperparameters and per-experiment configurations are in Appendix A. Code: `https://github.com/hcmus-thefangs/nakba-classifier`.

4.2. Main Results

Table 2 presents results across all configurations. Our MaxViT-Base system achieves macro F1 = 0.899 on the public test set, securing 1st place on the leaderboard. Two immediate patterns stand out. First, the moderately-sized backbones-EVA-02 and UniRepLKNet-perform competitively (0.884 and 0.886 respectively), demonstrating that neither masked-image-modelling pretraining nor attention mechanisms hold a monopoly on strong features for this domain. Second, and more revealingly, the DINOv2-based experiments exhibit the largest gaps between validation and test performance, despite being the highest-capacity models tested. This divergence is the clearest fingerprint of the photographer-bias problem described in Section 2: these large models memorise stylistic signatures of individual content creators rather than learning generalisable damage semantics. The most dramatic case is the KAN classification head, which achieves a strong validation F1 of 0.865 but collapses to 0.552 on the test set-a failure we analyse in detail in Section 4.5.

EXP	Model / Strategy	Val F1	Test F1
001	EfficientNet-B4 + Focal Loss	0.829	0.870
006	DINOv2-ViT-L + VLM Regularisation	0.859	0.804
007	DINOv2-ViT-L + LDAM Curriculum	0.747	0.709
008	DINOv2-ViT-B + KAN Head	0.865	0.552
010	EVA-02-Base (MIM ViT)	0.885	0.884
015	UniRepLKNet + Focal Loss	0.881	0.886
009	**MaxViT-Base + Focal Loss**	**0.881**	**0.899**

Table 2: Macro F1 on validation and public test sets. EXP009 achieved 1st place on the leaderboard.

4.3. Ablation Study

Table 3 isolates each component of EXP009 by changing one element at a time. Loss function choice is the most consequential variable: swapping focal loss for LDAM/DRW costs 13.4 F1 points, confirming that the moderate 1.83:1 imbalance sits squarely in focal loss's operating regime, while LDAM's large margin terms destabilise training at this data scale. Within augmentation, MixUp delivers the largest single gain (+1.5 points), functioning as the primary batch-level regulariser. The architecture swap from MaxViT-Base to EfficientNet-B4 costs 5.2 points-the single largest contributor-confirming that backbone choice dominates all other decisions.

Configuration	Val F1	Δ
Full EXP009 (reference)	0.881	-
Loss function:		
Focal $\rightarrow$ CrossEntropy	0.862	-0.019
Focal $\rightarrow$ LDAM+DRW (3-stage)	0.747	-0.134
Augmentation:		
No geometric distortions	0.871	-0.010
No texture transforms	0.868	-0.013
No CoarseDropout	0.874	-0.007
No MixUp	0.866	-0.015
Baseline only (Flip + Rotate)	0.852	-0.029
Backbone (same training config):		
EfficientNet-B4	0.829	-0.052
EVA-02-Base	0.885	$+0.004$
UniRepLKNet	0.881	±0.000

Table 3: Component ablation of EXP009. Each row modifies one element; all other settings are held fixed. Δ = change in Val macro F1 vs. the full configuration.

4.4. Leaderboard

Table 4 shows the top-5 of the final public leaderboard. Our system leads by 0.4 F1 points over the

second-ranked team; the top-5 span just 3.3 points, indicating a genuinely competitive field.

#	Team	F1	Acc.
1	**HCMUS_TheFangs**	**0.899**	**0.908**
2	rahaf_jaber	0.895	0.905
3	zahira_blr	0.889	0.901
4	salmakh	0.877	0.886
5	mohamedfathy	0.866	0.878

Table 4: Top-5 public leaderboard (macro F1). Full standings at the shared task website.

4.5. Analysis and Discussion

Destruction recognition is fundamentally a multi-scale problem-a model must detect local texture cues (cracked concrete, rubble geometry) while understanding global scene coherence-and *architectural fit* proves more important than raw capacity in addressing it. MaxViT's interleaved Block and Grid attention captures both neighbourhood detail and long-range layout at linear cost, which is why it outperforms DINOv2-ViT-L despite the latter having $2.5\times$ more parameters. The validation-to-test gap makes this concrete: DINOv2 drops 5.5 F1 points (memorising photographer shortcuts instead of damage semantics), while MaxViT actually *improves* by 1.8 points-resilience we attribute to our bias-targeted augmentation cascade that strips away low-frequency stylistic cues before they can be memorised.

Our exploratory experiments reinforce this theme. The KAN classification head (Liu et al., 2024) shows genuine promise when the backbone is frozen, demonstrating that learnable B-spline activations can capture non-monotonic decision boundaries useful for damage features. However, once we unfreeze the backbone for joint fine-tuning, the system collapses due to a gradient-scale mismatch: backbone weights need learning rates around $\sim 10^{-5}$ while B-spline grid parameters need $\sim 10^{-3}$, and a single shared rate causes catastrophic spline deformation. This is not a flaw of KAN itself but a practical engineering gap that per-parameter-group optimisation would likely resolve.

VLM regularisation via BLIP-2 (Li et al., 2023) captions exposes a more fundamental limitation. The captioner, trained on web-scale data, produces generic descriptions ("a photo of a building") that lack conflict-domain vocabulary, so InfoNCE (van den Oord et al., 2018) alignment anchors features to imprecise semantics and actively hurts generalisation. The LDAM (Cao et al., 2019) curriculum similarly fails through gradient oscillations at its stage transitions, producing the lowest score among all configurations. Both failures, together with the KAN collapse, share a root cause: techniques designed for large-scale, domain-general benchmarks can silently break when applied to data that is small, culturally specific, and visually unlike their pretraining distribution.

5. Conclusion

We presented HCMUS_TheFangs, the top-ranked system (macro F1 = 0.899) for the NakbaArchive-Classifier task. Evaluating five architectures on just 1,400 images reveals a critical insight: the binding constraint in conflict-domain vision is *domain fit*, not model capacity. MaxViT's (Tu et al., 2022) multi-scale attention elegantly captures both local rubble textures and global structural collapse, while avoiding the photographer-bias trap that ensnares larger models. Furthermore, bias-targeted augmentations prove as influential as backbone choice itself, underscoring the need for domain-specific regularisation.

Our exploratory failures are equally telling. KAN (Liu et al., 2024) heads show promise but require nuanced optimisation, while VLM regularisation falters because generic captioners lack conflict-domain vocabulary. These limitations highlight the urgent need for specialised humanitarian vision-language models. Looking ahead, the SaltPillar archive's scale and temporal depth offer rich opportunities for semi-supervised learning and change-detection. Ultimately, bridging the gap between web-scale pretraining priors and conflict-domain reality demands evaluation protocols that explicitly penalise style overfitting.

6. Ethics Statement

This work uses images depicting conflict and destruction in Gaza. All images were collected and shared by the task organizers under appropriate agreements. We recognize that automated classification of conflict imagery carries inherent ethical weight: misclassification (both false positives and false negatives) can have real consequences for archiving, historical record, and humanitarian decision-making. We caution against deployment without human oversight. Our code and models are released for research purposes only.

7. Acknowledgements

We thank the Nakba-NLP 2026 shared task organizers for their efforts in curating this important dataset and hosting the competition. We also acknowledge the contributions of the broader Palestinian digital archive community in preserving this visual record.

8. Bibliographical References

Alexei Abrahams, Shadi Abudalfa, Mustafa Jarrar, and George Mikros. 2026. The nakbaarchive-classifier shared task on nakba image classification. In *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexey A. Kalinin. 2020. Albumentations: Fast and flexible image augmentations. *Information*, 11(2):125.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Xiaohan Ding, Yiyuan Zhang, Yixiao Ge, Sijie Zhao, Lin Song, Xiangyu Yue, and Ying Shan. 2024. Unireplknet: A universal perception large-kernel convnet for audio, video, point cloud, time-series and image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5513–5524.

Yuxin Fang, Jiaqi Sun, Jianian Wang, Xiaosong Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva-02: A visual representation-level-mimicri transformer for image recognition. *arXiv preprint arXiv:2303.11331*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. volume 17, pages 2096–2030.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742. PMLR.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. 2024. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Vasiljevic, Pauline Sun, Peter Schwartz, Akhil Batra, Vincent Brandao, et al. 2024. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Jian Sun, Laurent Schampers, Ang Li, Christian Szegedy, Caroline Pantofaru, Golnaz Ghiasi, et al. 2022. Maxvit: Multi-axis vision transformer. *arXiv preprint arXiv:2204.01697*.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. In *arXiv preprint arXiv:1807.03748*.

Ross Wightman. 2019. Pytorch image models. https://github.com/rwightman/pytorch-image-models.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. CutMix: Regularization strategy to train strong classifiers with localizable features.

In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8024–8033.

Hongyi Zhang, Moustapha Cai, Yann Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

9. Language Resource References

A. Implementation Details

All experiments were conducted on a single NVIDIA P100 GPU (16 GB) via Kaggle. We used PyTorch 2.0 (Paszke et al., 2019) with `timm` (Wightman, 2019) for backbone loading and `albumentations` (Buslaev et al., 2020) for augmentation. All random seeds are fixed to 42 across `numpy`, `torch`, and `cudnn` (deterministic mode). Code and full configurations are publicly available at `https://github.com/technoob05/hcmus-thefangs-nakba-classifier`.

A.1. Shared Training Protocol

The shared protocol uses AdamW (Loshchilov and Hutter, 2019) ($\beta_1 = 0.9$, $\beta_2 = 0.999$, wd$= 10^{-2}$) with cosine-annealing LR schedule ($\eta_{\min} = 10^{-6}$, linear warm-up over epoch 1), 15 training epochs, and best-checkpoint selection on validation macro F1. Inputs are normalised with ImageNet (Deng et al., 2009) statistics (mean $[0.485, 0.456, 0.406]$, std $[0.229, 0.224, 0.225]$).

A.2. Per-Experiment Hyper-Parameters

Table 5 summarises backbone-specific settings. All other hyper-parameters follow the shared protocol above.

EXP	Backbone	BS	LR	Img	Notes
001	EfficientNet-B4	16	10^{-3}	384	Focal (α=0.5, γ=1.0)
006	DINOv2-ViT-L	8	10^{-4}	518	Frozen 10 ep; +InfoNCE
007	DINOv2-ViT-L	8	10^{-4}	518	Frozen 10 ep; LDAM+DRW
008	DINOv2-ViT-B	8	10^{-4}	518	KAN head (grid=5)
009	MaxViT-Base	8	2×10^{-5}	384	Focal+MixUp (p=0.5, α=0.2)
010	EVA-02-Base	4	10^{-5}	448	MixUp+CutMix (Yun et al., 2019)
015	UniRepLKNet	8	5×10^{-5}	384	Label smoothing 0.1

Table 5: Per-experiment hyper-parameters. BS=batch size, LR=peak learning rate, Img=input resolution.

A.3. Augmentation Cascade (EXP009)

The six-step `albumentations` (Buslaev et al., 2020) cascade: (1) RandomResizedCrop (scale 0.8–1.0) + HFlip + VFlip + Rotate $\pm15°$; (2) OneOf{MotionBlur, GaussianBlur} $p = 0.3$; (3) OneOf{OpticalDistortion, GridDistortion} $p = 0.3$; (4) OneOf{CLAHE, Sharpen, Emboss} $p = 0.3$; (5) HueSaturationValue + RandomBrightnessContrast $p = 0.3$; (6) CoarseDropout (8 holes, 32×32) $p = 0.3$. All other experiments use only flips, rotation, and colour jitter.

A.4. VLM-Regularised Training (EXP006)

EXP006 adds an auxiliary InfoNCE loss (van den Oord et al., 2018) ($\lambda = 0.3$, $\tau = 0.07$) that aligns DINOv2 visual features with text encodings produced by BLIP-2 (Li et al., 2023) and CLIP (Radford et al., 2021). Synthetic captions are generated for each training image; the contrastive objective pulls image representations toward their corresponding captions and pushes away mismatched pairs. The total loss is $\mathcal{L} = \mathcal{L}_{\text{CE}} + 0.3 \cdot \mathcal{L}_{\text{InfoNCE}}$. The intent is to suppress photographer-style memorisation by grounding features in semantic descriptions rather than visual style.

A.5. LDAM Curriculum (EXP007)

EXP007 implements a three-stage loss curriculum for imbalanced learning (Cao et al., 2019). Stage 1 (epochs 1–5): standard cross-entropy to learn initial representations. Stage 2 (epochs 6–10): LDAM loss with class-dependent margins $m_i = C_{\max}/n_i^{1/4}$. Stage 3 (epochs 11–15): LDAM with deferred re-weighting (DRW), where per-class weights are inversely proportional to class frequency. The curriculum transitions aim to first build stable representations before enforcing margin constraints.

A.6. KAN Classification Head (EXP008)

EXP008 replaces the standard MLP head with a two-layer Kolmogorov-Arnold Network (Liu et al., 2024): $768 \to 512 \to 256 \to 2$, using cubic B-spline activations with grid size 5. Unlike standard MLPs which use fixed activation functions (ReLU), KAN layers learn the activation shape itself via spline parameters, enabling non-monotonic decision boundaries. Training proceeds in two phases: backbone frozen (10 epochs), then full fine-tuning (5 epochs).

A.7. Domain Adversarial Training (EXP005)

EXP005 adds a Gradient Reversal Layer (GRL) (Ganin et al., 2016) to the DINOv2 backbone. A photographer-identity discriminator is trained adversarially via the GRL, which negates discriminator gradients during backpropagation, forcing backbone features to become uninformative for author identity whilst remaining informative for damage prediction. The adversarial coefficient is annealed as $\lambda_p = 2/(1 + e^{-10p}) - 1$ where $p \in [0, 1]$

is fractional training progress. This experiment is exploratory and its result is not included in the main comparison table.

Baflah-lamri at NAKBA-NLP 2026: Manual Ground Truth Enrichment

Dr. Abdelouahab Bafleh

Scientific and Technical Research Center for the Development of the Arabic Language (CRSTDLA)
Ouargla Research Unit, Algeria

a.baflah@crstdla.dz

Dr. Lamri Mohamed

Scientific and Technical Research Center for the Development of the Arabic Language (CRSTDLA)
Ouargla Research Unit, Algeria

m.lamri@crstdla.dz

Abstract

This paper presents a detailed description of the team's methodology in participating in Subtask 1 (Transcription Track) of the NAKBA NLP 2026 Shared Task for Arabic Manuscript Understanding. We present a rigorous approach to line-level manual transcription of historical Arabic manuscripts derived from the Omar Al-Saleh memoir collection (1951-1965). Our methodology emphasizes accuracy, consistency, and adherence to diplomatic transcription principles, while addressing the unique palaeographic and physical challenges of Arabic handwriting, such as writing speed, orthographic variation, and the impact of writing tools (e.g., immediate strike-throughs and ink spatter). The work guided by strict transcription guidelines and contextual verification protocols, matching cropped line images with full-page images to resolve ambiguities and automated cropping issues. The team successfully transcribed the entire assigned batch of 500 lines across 368 unique pages, producing reference data comprising 6,719 words and 37,646 characters. This effort contributes to providing highly reliable Ground Truth data, serving as an essential foundation for training and evaluating Handwritten Text Recognition (HTR) models for Arabic manuscripts.

Keywords: Handwritten Text Recognition (HTR), Diplomatic Transcription, Historical Arabic Manuscripts, Natural Language Processing (NLP), Ground Truth, Omar Al-Saleh Memoirs, NAKBA NLP Shared Task.

Introduction

The absence of cognitive content from Arabic manuscripts is often due to private ownership and lack of accessibility, particularly for personal memoirs, historical documents, and other archival materials. Given that computational analysis possesses immense capacity and tremendous speed in processing information, making these documents available for computational analysis would be a significant service, enabling researchers to access and benefit from their content.

Automated understanding of Arabic manuscripts represents a critical challenge in the fields of digital humanities and cultural heritage preservation (Chan et al., 2024). Despite its promising potential for preserving cultural heritage and making it widely accessible, it faces numerous challenges primarily stemming from the complexity of recognizing Arabic handwriting, which varies from one scribe to another.

The NAKBA NLP 2026 Shared Task addresses this challenge by providing a benchmark corpus from the Omar Al-Saleh memoir collection, comprising 16 documents spanning from 1951 to 1965, with an estimated 6,395 pages containing approximately 1,597,025 words (Zaraket et al., 2026). This collection exhibits unique characteristics of mid-twentieth-century Arabic handwriting, offering a crucial resource for developing Arabic HTR technologies.

Since computational models cannot decipher Arabic handwriting without a deep understanding of the palaeographic (calligraphic) and linguistic characteristics of the original text, this paper provides a description of the manual transcription methodology applied to a sample of 500 lines from these memoirs, focusing on analyzing the calligraphic and linguistic challenges in the completed batch.

This paper is not limited to presenting quantitative achievements; rather, it offers an analytical reading of the obstacles encountered during the transcription process. It discusses the issues of ambiguous character shapes caused by writing speed and the necessity of relying on context to resolve them. It clearly differentiates between text truncation resulting from automated image cropping and that originating from the manuscript's intrinsic characteristics. Furthermore, it highlights documentary linguistic

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 128–132
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

phenomena, such as the writing of numerals and the spelling of foreign proper nouns. By presenting these observations on the Ground Truth data, the paper aims to provide qualitative insights that contribute to improving the capacity of deep learning algorithms to process handwritten Arabic texts (with daily memoirs as a case study).

Methodology & Transcription Guidelines

The methodology relied on the principles of Diplomatic Transcription coupled with rigorous contextual verification mechanisms to ensure the construction of highly reliable and historically faithful Ground Truth data. This approach is a fundamental standard in the digitization of historical documents to ensure the text transcribed exactly as written, without linguistic or grammatical corrections that might obscure the original author's stylistic or orthographic characteristics (Zahour et al., 2007).

The transcription process was carried out independently by two native Arabic-speaking annotators to ensure high accuracy and resolve any ambiguities through consensus.

Orthographic Conventions

The transcription process based on a set of strict guidelines to handle recurring orthographic phenomena, which are among the most prominent challenges in Arabic natural language processing (Abandah & Abdel-Karim, 2014):

- **Hamzas:** Strict adherence to the drawing of Hamzas (أ، إ، ؤ، ئ، ء) exactly as they appeared in the original manuscript, regardless of whether they conform to or violate modern standard orthographic rules.
- **Alif Maqsura and Dotted Yaa:** The Alif Maqsura (ى) and the terminal dotted Yaa (ي) transcribed based on their actual visual representation in the text, rather than their linguistic requirements.
- **Ta' Marbuta and Ha':** A precise distinction made between the Ta' Marbuta (ة) and the pronoun Ha' (ه) whenever they were clearly distinguishable. Context used as a deciding factor in cases where dots omitted due to rapid writing.
- **Diacritics:** Transcription was limited exclusively to the diacritical marks clearly visible in the manuscript, ignoring any attempt to infer or add invisible diacritics.

Handling Special Cases

Based on the nature of the images provided in this batch, and guided by the task's instructions (Zaraket et al., 2026), the methodology necessitated the following rules:

- **Unreadable Words and Handwriting Complexity:** Although contextual verification successfully resolved most ambiguities, approximately 20 lines (about 4%) in the completed batch contained entirely incomprehensible words (averaging one to three words per line). This obscurity is not due to physical damage or smudging, but rather to the severe complexity and reduction in character drawing resulting from rapid writing.
- **Marking Missing Words:** Sequential dots (e.g., ...) were used as a placeholder instead of complex tags like [###]. The length of the dot sequence reflects the estimated size of the unreadable word(s) (as seen in reference lines 34, 38, 42, and 46 of the dataset), to accurately locate semantic interruptions.
- **Margins and Insertions:** Neither significant marginal notes nor insertions were recorded in the transcribed batch. In the rare instances of their appearance, the focus remained on the main body of the line (e.g., Line 61).
- **Numerals:** Numbers transcribed in their original format, whether they took the Eastern Arabic form (٠-٩) or the Western form (0-9), reflecting the author's style in recording dates and calculations. In one instance, "962" used, referring to 1962.
- **Foreign Words:** Foreign proper nouns or non-Arabic terms were transcribed exactly as the author wrote them using Arabic letters (phonetic Arabization), preserving Latin characters whenever they were written in the original document.

Dataset Statistics

As a culmination of the adopted "Diplomatic Transcription" methodology, the transcription process resulted in the construction of a reference dataset that faithfully reflects the content of the original manuscript. This section presents a quantitative analysis of the results achieved on the assigned batch, reviewing the coverage scope, statistical metrics of the extracted texts, and readability quality indicators.

Transcription Statistics

The table below reflects the quantitative characteristics of the transcribed data, illustrating the readability level and the volume of extracted texts:

Category	Count / Percentage
Fully readable lines (no missing words)	451 (90.2%)

Category	Count / Percentage
Lines containing unreadable words (marked with ...)	49 (9.8%)
Lines with uncertainty markers or damaged text (e.g., [?])	0 (0.0%)
Lines with foreign language content (Latin characters)	5 (1.0%)
Average characters per line	75.3 ± 16.7
Average words per line	13.4 ± 3.2

Table 1: Transcribed Data Characteristics

Quality Metrics and Linguistic Characteristics

Based on internal verification and statistical analysis of the final transcription file, the following metrics achieved:

Readability Integrity: The percentage of lines that we were able to read and transcribe fully reached 90.2%, while lines containing unreadable words (due to speed and complexity) accounted for 9.8%.

Textual Output: The transcription produced 6,719 words, comprising 37,646 characters.

Numerical Diversity: The memoirs featured a notable presence of numerical data (dates, calculations), with 113 lines (22.6%) containing numbers.

Linguistic Diversity: Foreign names or terms written in Latin characters appeared in five lines (1.0%), reflecting the documentary and political nature of the memoirs.

Challenges & Observations

The transcription process encountered various palaeographic and linguistic challenges. This section presents the most prominent difficult cases observed and how they addressed systematically to ensure data accuracy:

Ambiguous Character Shapes and Handwriting Complexity

The memoirs written in a precise Mashriqi Naskh script, reflecting the author's personality. He adhered to distinctive shapes for certain letters, such as the terminal Nun (written above the line) and the Ha' preceding a final Alif. There was also a special ligature for juxtaposed Ba' and Ta'. Familiarity with these shapes eliminated reading difficulties. However, other challenges arose:

Challenge: Visual confusion of structurally similar letters (especially distinguishing between Ba', Ta', Tha', or terminal Nun and Yaa) when diacritical dots are faded due to rapid writing.

Solution: Examining the surrounding context to verify lexical plausibility and continuously referring to the full-page image to identify the author's usual pattern. In extreme cases where the word's shape was entirely reduced (49 lines), guessing was avoided, and the word was marked with sequential dots (...) to preserve scientific integrity (e.g., Lines 21 and 23).

Foreign Names and Orthographic Variation

Challenge 1 (Foreign Names): Transcribing the names of foreign figures (American and European) written in Arabic letters without certainty of their standard spelling.

Solution: Transcribing the text exactly as written, preserving the original phonetic Arabization used by the author, without attempting to "correct" it (e.g., keeping "فيليب هوف" [Philip Hoff] exactly as the author wrote it).

Challenge 2 (Orthographic Variation): The inconsistent use of certain orthographic rules within the same document, specifically regarding Hamza placement (e.g., مسألة/مسءلة, مبدأ/مبدا).

Solution: Strictly adopting Diplomatic Transcription by transferring the text as is without linguistic intervention, which documents historical variation patterns for future analysis.

Automated Cropping Challenges and Extended Sentences

Challenge 1 (Extended Sentences): Some cropped lines contained a truncated part of a long sentence, making the semantic meaning and syntactic context incomplete.

Solution: Utilizing the full source page image (source_image) to support understanding the overall context of the paragraph, enabling more accurate transcription of the truncated words.

Challenge 2 (Image Cropping): The appearance of horizontal or vertical truncation of letters at the beginnings or ends of lines (e.g., Line 23).

Solution: Methodologically distinguishing between document defects and processing defects. By returning to the original page, it confirmed that this truncation resulted exclusively from the automated cropping process and was not a physical omission.

Physical Constraints & Writing Tools

Challenge 1 (Ink Spatter and Pen Pressure): The use of multiple pens was noted, manifesting in scattered ink dots. Variations in pen pressure produced multiple visual representations for the same letter.

Solution: Overcoming this by zooming in and utilizing context, alongside referencing the original page to verify whether a spot was a diacritic or a transient ink mark.

Challenge 2 (Immediate Strike-throughs): Some lines included rapid strike-throughs (crossing out a word and writing another above or next to it), creating visual noise.

Solution: In adherence to diplomatic transcription, crossed-out words ignored. Focus exclusively placed on transcribing the approved, final text that forms the correct context, to prioritize the extraction of the author's final intended text, avoiding complex tags that might confuse training models.

Challenge 3 (Printed Text): The presence of printed text within the handwritten text, likely because the paper belonged to pre-printed notebooks.

Solution: Adopting only the handwritten text and ignoring the printed text, as it falls outside the context.

Discussion

The work accomplished in this batch offers a qualitative contribution to enriching the Ground Truth data for developing Arabic HTR models. Analyzing the challenges yields several methodological implications:

Palaeographic Characteristics of the Memoir Style

The transcription process demonstrated that "personal memoirs" pose challenges that do not fundamentally differ from carefully copied official documents. The 9.8% of lines containing unreadable words is not a flaw in the document or scanning quality; rather, it reflects an emotional state or rapid flow of thought that led to character reduction. Furthermore, the impact of writing tools (varying pen pressure, ink spatter, and strike-throughs) adds a dynamic character. This provides a crucial indicator: future algorithms must trained on Language Models to predict words based on context, rather than relying merely on the raw visual image.

Impact of Digital Processing (Automated Cropping)

The phenomenon of "truncation" in extended letters—resulting from automated cropping and necessitating continuous reference to full-page images—highlighted a technical weakness in data preparation. This confirms that developing Line Segmentation algorithms tailored to the nature of the Arabic script (characterized by vertical overlap and sub-line extensions) is a fundamental prerequisite for successful automated recognition.

Linguistic Variation and the Importance of Flexible Lexicons

Observations regarding the phonetic Arabization of foreign names and orthographic variation reflect the absence of strict standardization in daily Arabic writing during the mid-twentieth century. This lexical and orthographic diversity requires computational model designers to build Flexible Lexicons capable of accommodating spelling multiplicity without classifying them as "errors".

Conclusion

This paper presented an analytical description of our team's methodology in completing Subtask 1 of the Arabic Manuscript Understanding Shared Task (NAKBA NLP 2026). By applying the principles of Diplomatic Transcription and relying on contextual verification, we successfully transcribed 500 lines assigned rows 100 %. As noted in our challenges, approximately 20 lines contained between one to three unreadable words due to handwriting complexity. To build a highly reliable reference dataset comprising 6,719 words representing historical memoirs spanning between 1951 and 1965.

Our experience demonstrated that the core challenges in dealing with personal memoirs do not lie in physical damage, but rather stem from the characteristics of rapid handwriting (strike-throughs, character reduction, and varying pen pressure), alongside the technical challenges resulting from automated cropping. This batch also provided insights into orthographic variation and the Arabization of foreign names.

Finally, in affirmation of open science principles, all transcribed texts will made available via a dedicated Anonymous GitHub repository under a Creative Commons license (CC-BY-4.0). We hope that this data and the accompanying methodological analyses will contribute to guiding and developing Arabic HTR systems, and to preserving and analyzing Arab documentary heritage more effectively.

Bibliographical References

Abandah, G. A., & Abdel-Karim, S. (2014). Survey and bibliography of Arabic optical text recognition. *Signal Processing*, *98*, 447–451. https://doi.org/10.1016/j.sigpro.2013.11.026

Al-Shatnawi, A., & Al-Zoubi, M. (2024). A systematic literature review of deep learning methods for Arabic handwritten text recognition. *Engineering, Technology & Applied Science Research*, *14*(4), 15234–15245. https://doi.org/10.48084/etasr.6913

Chan, A., Mijar, A., Saeed, M., Wong, C.-W., & Khater, A. (2024). HATFormer: Historic handwritten Arabic text recognition with

Transformers. *arXiv*. https://doi.org/10.48550/arXiv.2410.02179

Zahour, A., Likforman-Sulem, L., Boussellaa, W., & Taconet, B. (2007). Text line segmentation of historical Arabic documents. In *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR 2007)* (pp. 138–142). IEEE. https://doi.org/10.1109/ICDAR.2007.4378688

Zaraket, F., Shalash, B., Hamoud, H., Chamseddine, A., Abid, F. B., Jarrar, M., Chakra, C. A., & Ghanem, B. (2026). AR-MS: Arabic manuscript understanding. In *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026), Palma, Mallorca, Spain, May.*

Free-Gaza at NakbaArchiveClassifier Shared Task: Towards Distinguishing the Destructive Effect of Nakba: NakbaImage Classification using Artificial Intelligence Techniques

Nisreen I. R. Yassin, Enas A. Hakim Khalil

Systems & Information Department, National Research Centre (NRC), Giza, Egypt.
nisreen.yassin20@gmail.com, enaskhalil@gmail.com

Abstract

The accounts of the continuing Palestinian Nakba encompass considerable significance. Over the course of the three years of the conflict, millions of photos from social media have been preserved. The preservation and classification of these data through artificial intelligence tools are essential to guarantee their availability, accessibility, and applicability. This paper presents a highly optimized, resource-constrained machine learning pipeline for binary image classification. The system is designed for the NakbaArchiveClassifier Shared Task 2026, which aims to distinguish between images of destroyed infrastructure and those of intact infrastructure. The system uses two lightweight EfficientNetB0 networks to build a weighted ensemble. With strict hardware limitations of 2GB GPU VRAM, the system achieves an F1-score of 84.16%, ranking 9th on the leaderboard.

Keywords: Nakba-NLP 2026, ensemble system, binary image classification

1. Introduction

Following October 7, 2023, a huge number of infrastructure images were shared on social media, documenting the situation in the Giza strip. Visual testimonies provide undeniable documentary evidence of conflict, displacement, and structural impact. The huge volume of unstructured images presents a substantial bottleneck for historians, journalists, and researchers. Manual classification of these images is not scalable, which motivated the development of automated methods for damage detection and classification.

Automating the classification of these visual archives is a crucial step in rendering historical data machine-understandable. By developing robust computer vision models capable of analysing and indexing archival images, researchers gain scalable retrieval tools. This technological intervention not only ensures the long-term accessibility of these narratives for future generations but also allows researchers to efficiently validate and quantify this visual evidence of historical and ongoing destruction within large-scale digital repositories.

To address the challenges of classifying unstructured images, the NakbaArchiveClassifier Shared Task 2026 shared task is formulated as a binary classification problem (Abrahams et al., 2026). Participants must process raw input images and accurately classify each sample image into one of two classes: destruction or non-destruction.

The task introduces many domain-specific challenges:

- Moderate class imbalance (35% positive class)
- High intra-class variance (varied structural types, perspective)
- Social media compression artifacts and heterogeneous capture devices (visually subtle or partial structural damage)

Our Free-Gaza system uses transfer learning EfficientNetB0, fine-tuned on the provided dataset. It is evaluated through accuracy and F1-score. The proposed system provides an efficient and reproducible baseline for classifying images of infrastructure damage collected from social media.

2. Background

Disaster response monitoring has extensively studied infrastructure damage detection. For post-disaster assessment, early methods primarily relied on remote sensing and satellite imagery. Convolutional neural networks (CNNs) showed strong performance in recognizing structural damage patterns, such as collapsed building roof failure. Recently, the research has expanded to include ground-level imagery shared on social media platforms. The data collected from such platforms is real-time but highly unstructured visual data. Unlike satellite images, social media images exhibit substantial domain variability, including diverse viewpoints, varying lighting conditions, compression artifacts, and heterogeneous capture devices. These factors present significant intra-class variance and make binary classification more challenging, especially when damage is partial or visually subtle.

CNNs have become the dominant method for image classification following the success of deep learning architectures such as AlexNet (Krizhevsky et al., 2012) and ResNet (He et al., 2016). These models learn hierarchical feature representations that capture both low-level

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 133–136
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

textures and high-level semantic structures. Therefore, these models are well-suited for identifying infrastructure damage. In the case of limited data, transfer learning has shown particular effectiveness. By initializing networks with weights pretrained on large-scale datasets such as ImageNet (Deng et al., 2009), models can leverage general visual priors and adapt them to specific-domain tasks through fine-tuning. This method reduces overfitting and improves convergence when labelled data are scarce. The EfficientNet family (Tan and Le, 2019) introduces compound scaling of network depth, width, and resolution to achieve improved trade-offs between accuracy and efficiency. EfficientNetB0, the baseline variant, achieves competitive performance while using significantly fewer parameters than earlier architectures. Such parameter efficiency is advantageous in shared-task environments and moderate-sized datasets, where overfitting is a concern.

While most destruction detection research focuses on contemporary satellite imagery, the methodologies provide a strong foundation for classifying destruction in historical archives. YOLO-based object detection models have been successfully applied to identify destroyed buildings in high-resolution imagery (Zhang et al., 2023). Semantic segmentation models such as hybrid U-Net architectures enable pixel-level classification of damaged structures (Xu et al., 2022). In data-scarce contexts, unsupervised radar-based approaches have demonstrated the feasibility of detecting building destruction without extensive manual labeling (Kaufmann et al., 2024).

3. System Overview

The proposed image classifier was designed to achieve the best performance on complex, unstructured images while overcoming the stringent hardware limitations of 2GB VRAM. The system shifts from an empirical risk-minimization approach to a data-driven ensemble-based approach. ConvNeXt and DenseNet121 are two examples of networks that are deeper than others and achieve better texture recognition. However, empirical evidence during development revealed that these networks' performance was decreased because they had to use smaller batch sizes. To extract as many feature representations as possible from low-resolution 224 x 224-pixel input, the system uses EfficientNetB0 as its global backbone. EfficientNet's compound scaling algorithm will uniformly scale the network's width, depth, and resolution (Tan and Le, 2019).

Transfer learning is used to train EfficientNetB0 using pre-trained weights from ImageNet (Deng et al., 2009). Model A is trained on the original data using standard binary cross-entropy. This caused the model to over-optimize for "Not-destruction" buildings that were easy to identify. At the same time, it struggled with ambiguous landscapes after a disaster, such as partial structural collapse or rubble hidden behind other buildings. Model B is trained using the original training data and a set of external data. The external data includes the AIDER dataset (Kyrkou and Theocharides, 2020; Kyrkou and Theocharides, 2019), pseudo-labels data, which are predictions exhibiting extreme statistical confidence, and the old validation data. For Model B, the learning objective was changed to binary focal cross-entropy (Lin et al., 2017). This algorithm reduces the impact of class imbalance by scaling the loss based on the model's confidence in its predictions and adapting to the varying levels of difficulty across the examples in the dataset. For the enriched dataset, a learning rate is scheduled using a cosine decay learning rate (Loshchilov and Hutter, 2016). The approach entails a smoothly and continuously reduced learning rate over time. The non-linearly decaying toward convergence allowed Model B to converge securely in 12 Epochs.

In the prediction phase, the 8-Way Test-Time Augmentation (TTA) Algorithm (Shorten and Khoshgoftaar, 2019) is implemented. For each test image, the algorithm generates a set of eight augmented views, including the original, a horizontal flip, a vertical flip, rotations of 90, 180, and 270 degrees, and adjustments to brightness and contrast. The model evaluates all eight views and calculates the final prediction probability as the arithmetic mean of the outputs. A weighted soft voting mechanism is used for final classification (Zhou, 2025). Model A is trained purely using original data, while model B is trained using original data and AIDER data; therefore, the ensemble probability is calculated as follows:

$$P_{ensemble} = 0.40P_{model\ A} + 0.60P_{model\ B}$$

4. Experimental Setup

The proposed system depends on two datasets. The first dataset is the original dataset provided by NakbaArchiveClassifier Shared Task 2026. This dataset contains 1599 PNG images, divided into 1400 for training, 199 for validation, and 402 for testing. The dataset demonstrates moderate class imbalance and high visual variability, including disparities in lighting, perspective, resolution, and compression artifacts characteristic of social media platforms. The second dataset is the AIDER (Aerial Image Dataset for Emergency Response Applications). The construction of the AIDER dataset entailed the manual collection of all images about four distinct disaster events. This collection encompassed 521 images of fire and smoke, 526 images of flooding, 511 images of collapsed buildings and rubble, and 485 images of traffic accidents. In addition, 4,863 images were collected for the normal case. The data structure

is summarized in Table 1. Data usage for models A and B is illustrated in Table 2. For replication of the 0.8416 F1-score of NakbaArchiveClassifier Shared Task 2026, the hyperparameter configurations utilized for models A and B are depicted in Table 3. Each of these two models employs an alternative optimization approach to support its distinct functions in the ensemble. While Model A aims to achieve steady, background-level convergence, Model B aggressively selects hard examples from the augmented dataset. The development, training, and inference phases of NakbaArchiveClassifier were built on open-source software. The system relies on standard Python machine learning libraries and the TensorFlow framework.

Classes labels	original training data	Original val. data	AIDER data	Test data
Destruction	494	70	2043	-
Not destruction	906	129	4863	-
total	1400	199	6906	402

Table 1: Dataset structure

	Training	Validation
Model A	Original dataset	Original dataset
Model B	Original (494+906) + AIDER (2043+1600) + Pseudo-labels + Old Val.	Split 10% of the training data.

Table 2: Data usage for model A and model B

Parameter	Model A	Model B
Base architecture	EfficientNetB0	EfficientNetB0
Weights	ImageNet	ImageNet
Resolution	224x224x3	224x224x3
Batch size	32	16
Dropout Rate	0.3	0.4
Classification head	Dense 1 unit sigmoid	Dense 1 unit sigmoid
Voting weight	0.4	0.6

Table 3: Hyperparameter configurations utilized for models A and B

5. Results

Several experiments were performed to evaluate the proposed system. The experiments were conducted on an Intel® Core™ I7 CPU at 2.60 GHz computer using the TensorFlow platform for machine learning. The performance of the proposed system is evaluated on the validation set following each training epoch. The evaluation metrics used include accuracy, macro F1-score, precision, recall, and a confusion matrix.

Model A has been trained on the original data from NakbaArchiveClassifier and evaluated on the validation set of 199 PNG images using standard binary cross-entropy. As shown in Table 4, the validation data contain two classes: destruction (70 images) and not-destruction (129 images), indicating class imbalance. Therefore, the model is biased toward predicting the images as not destroyed. The recall for the destruction class is 0.67, indicating poor classification of destroyed images, with 23 destroyed buildings missing. The confusion matrix of model A is shown in Figure 1.

	Precision	Recall	F1-score	support
destruction	0.80	0.67	0.73	70
Not-destruction	0.84	0.91	0.87	129
accuracy			0.82	199
Macro avg.	0.82	0.79	0.80	199
Weighted avg.	0.82	0.82	0.82	199

Table 4: Classification report of model A

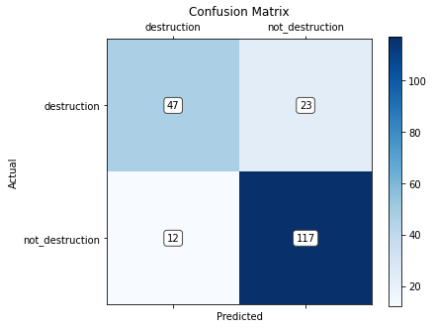


Figure 1: Confusion matrix of model A.

Model B has been trained on a combination of data: original data, AIDER data, pseudo-labels (200 images), and old validation data (199 images). In the validation phase of model B, the validation data has been increased to 1090 images, selected at random from the AIDER data. Table 5 shows the classification report of model B. Its confusion matrix is shown in Figure 2. According to the classification report for model B,

the validation data contains 581 destructive images and 509 non-destructive images. The confusion matrix shows the sensitivity of model B to the destruction images, with a recall of 0.97. This behavior reflects the use of Focal Loss. The overall macro F1-score of model B is 0.85, which is greater than the macro F1-score of model A of 0.80. The ensemble system combines the two models. The system uses a 40/60-weighted soft-voting mechanism and a test-time augmentation mechanism. This system successfully balanced precision and recall across the two classes. The final achieved macro F1-score is 0.8416 on the unlabeled official test set. The ensemble system's classification report is shown in Table 6.

	Precision	Recall	F1-score	support
destruction	0.79	0.97	0.87	581
Not-destruction	0.96	0.71	0.82	509
accuracy			0.85	1090
Macro avg.	0.87	0.84	0.85	1090
Weighted avg.	0.87	0.85	0.85	1090

Table 5: Classification report of model B

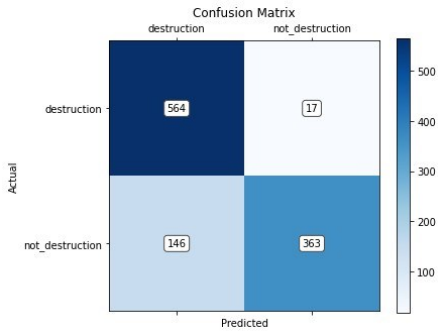


Figure 2: Confusion matrix for model B.

accuracy	0.8557
precision	0.8427
recall	0.8405
F1-score	0.8416
specificity	0.8405
Balanced-accuracy	0.8405

Table 6: Classification report of the ensemble system

6. Conclusion

The development of the NakbaArchiveClassifier pipeline suggests the potential for achieving highly competitive classification metrics. The proposed classification system successfully functioned under significant hardware limitations (a 2-gigabyte virtual memory limit). The system considered both the data and learning objectives to be dynamic variables. Therefore, it was able to overcome severe domain shifts and class imbalance. The evaluation of the proposed ensemble system shows promising results. For future work, our model's capabilities could be further refined with more powerful hardware and additional training data.

7. Bibliographical References

Alexei Abrahams, Shadi Abudalfa, Mustafa Jarrar, and George Mikros. 2026. The nakbaarchiveclassifier shared task on nakba image classification. In Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026), Palma, Mallorca, Spain.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, IEEE, pp. 248-255.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision, pp. 2980-2988.

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983.

Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. Journal of big data, 6, pp. 1-48.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning, PMLR, pp. 6105-6114.

Zhi-Hua Zhou. 2025. Ensemble methods: foundations and algorithms. Chapman and Hall/CRC.

HCMUS_TheFangs at NakbaVirality Shared Task: The Audience is the Message: Escaping the Deep Learning Trap in Conflict-Domain Virality Prediction

Dao Sy Duy Minh*
[1]Faculty of Information Technology, University of Science, Ho Chi Minh, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
23122041@student.hcmus.edu.vn

Abstract

We present HCMUS_TheFangs's system for the Nakba-NLP 2026 Virality Shared Task, which achieves **Rank #1** on the final leaderboard with a test Macro-F1 of **0.7062**, placing first among all competing teams. Our winning system is deliberately simple: a single *Community Target Encoding* feature-the smoothed historical virality rate of the posting subreddit-combined with TF-IDF text features and an XGBoost classifier. This design emerged from a hard-won insight: virality in conflict reporting is determined not by what is posted but by where it is posted. We spend the majority of this paper showing *why* this holds. Through 18 ablation experiments we trace the journey from deep learning failure (cross-attention XLM-R+CLIP scoring 0.2935) to SOTA (0.7062), documenting the "Deep Learning Trap" that ensnares practitioners who apply representation-centric models to domains governed by sociological rather than semantic dynamics. An equally surprising finding emerges along the way: posts with fewer hashtags and cleaner prose consistently outperform heavily tagged, promotional content, a "Less is More" paradox that challenges conventional social media optimization wisdom. All code and experiments are publicly available.

Keywords: Virality Prediction, Multimodal Classification, Community Context, Target Encoding, Conflict Discourse, Feature Engineering, XGBoost, Nakba-NLP

1. Introduction

On October 7, 2023, a conflict erupted that would produce one of the densest information cascades in social media history. Thousands of Gaza-related posts flooded Reddit, Twitter, and Telegram simultaneously. Some reached hundreds of thousands of interactions; others, carrying nearly identical imagery, vanished unread. Understanding *why* is the question at the heart of the Nakba-NLP 2026 Virality Shared Task (Ezzini et al., 2026).

Our first instinct was standard: fine-tune XLM-RoBERTa (Conneau et al., 2020) for text, run CLIP (Radford et al., 2021) and DINOv2 (Oquab et al., 2023) over images, and fuse via cross-attention. The result-test Macro-F1 of 0.2935-was worse than a majority-class baseline. We had fallen into the **Deep Learning Trap**: the assumption that architecture alone will learn sociological structure nowhere encoded in pixels or tokens. The breakthrough came from a different question: not "What does this post look like?" but "Who is reading it?" A single feature-the smoothed historical virality rate of the posting community-paired with TF-IDF and XGBoost, became our winning submission (**Macro-F1 = 0.7062, Rank #1**), more than 27 points above our best deep learning result.

2. Related Work

Early work by Cheng et al. (2014) established that cascade depth depends heavily on structural network properties-follower count, posting time, platform affordances-rather than content semantics. As multimodal encoders became available, studies using CLIP (Radford et al., 2021) showed that image-text alignment correlates with engagement in general-domain news, while work on CrisisMMD (Alam et al., 2018) showed visual features become predictive for humanitarian content only when paired with source metadata.

Recent work reinforces these limitations. Quy et al. (2025) found that metadata extraction and retrieval augmentation substantially outperform standalone visual encoders for event-based analysis. In specialized domains, structured pipelines with explicit grounding frequently surpass end-to-end neural models. Agent behavior is shaped more by environmental framing than by model architecture-a "context over content" principle consistent with the community-dominance pattern we observe.

In political discourse specifically, Del Vicario et al. (2016) showed that ideologically homogeneous communities amplify in-group content regardless of quality, creating an amplification coefficient that dwarfs content-level signals. Work on propaganda detection (Da San Martino et al., 2020) established that rhetorically-charged language provides discriminative signal for classification, though we find it is useful only when conditioned on community context, consistent with frame resonance theory (Benford and Snow, 2000). On the modeling side, gradient boosting consistently outperforms neural models when labeled data is scarce (Chen and Guestrin,

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 137–143
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

2016; Prokhorenkova et al., 2018).

3. Task, Dataset, and the Visual Homogeneity Problem

The NakbaVirality dataset (Ezzini et al., 2026) provides 1,691 labeled training samples and a held-out test set. Each sample consists of a textual caption (a tweet or Reddit post body, roughly a third of which is non-English), an associated image depicting the Gaza conflict, and metadata encoding the source community. Labels follow an ordinal scale-Low, Medium, and High Viral-distributed approximately 40%/50%/10%, a severe imbalance that motivates Macro-F1 as the evaluation metric.

Our first analytical step revealed a sobering truth. We projected CLIP image embeddings into two dimensions via t-SNE and found that the three virality classes overlapped almost completely. Destroyed buildings, crowds of protesters, olive branches-all appear with similar frequency across every virality level. We term this the **Visual Homogeneity Problem**: in a domain where every image depicts some facet of the same conflict, visual variation alone cannot discriminate posts that go viral from those that do not.

Then came the discovery that changed everything. The dataset metadata encodes the originating community for every post, and we found that its predictive power dwarfs every content-based signal. Highly engaged, ideologically cohesive communities show viral rates an order of magnitude higher than general-interest news forums-a tenfold gap that exists *before* a single word of the caption is read. The determining factor is not content quality but audience identity: members of topically focused communities arrive pre-mobilized, whereas general-audience subscribers encounter the same content as one headline among hundreds. No image encoder can recover that sociological fact from pixels alone.

4. System Description

Figure 1 illustrates our winning system alongside the community amplification insight that powers it. The architecture is deliberately minimal: one sociological fact, two text feature sets, and a gradient-boosted classifier. The right panel reveals *why* that single fact is so powerful-a tenfold virality gap between communities that no content encoder can recover from pixels or token sequences alone.

4.1. Community Context Model - Winning System

The winning submission is built around a single question: how viral has content from this community historically been? We extract the source community identifier from the post metadata and encode it as a continuous virality prior through Bayesian smoothed target encoding. Formally, let C_i be the source community for post i and let $\bar{y}_C$ be the mean numeric label across training examples from that community. The encoded value is

$$\hat{\mu}_C = \frac{n_C \cdot \bar{y}_C + m \cdot \bar{y}_{\text{global}}}{n_C + m} \tag{1}$$

where n_C is the community's training sample count, $\bar{y}_{\text{global}}$ is the global label mean, and $m = 10$ is a smoothing hyperparameter. The smoothing term prevents the encoder from memorizing a subreddit that appears only once in training; instead, small communities are pulled toward the global mean, ensuring graceful degradation for out-of-vocabulary communities at test time.

Critically, this encoding is computed *strictly inside* the cross-validation loop. In each of the five stratified folds, the encoder is fit on the training split and applied to the held-out validation split without any leakage of validation labels. For the final submission, the encoder is re-fit on the entire training corpus. Alongside this single community feature we concatenate 1,000-dimensional TF-IDF unigrams (English stopwords removed) and three structural features-character count, word count, and hashtag count-feeding the full 1,003-dimensional vector into an XGBoost classifier with 1,000 estimators, learning rate 0.05, and maximum depth 6. Despite its remarkable simplicity, this system achieves **test Macro-F1 = 0.7062**, the highest score recorded on the shared task leaderboard.

4.2. The Singularity - Stress-Testing Complexity

In parallel with our minimalist system, we developed "The Singularity," a 219-feature Optuna-tuned model designed to test whether a massive assembly of sociological and content features could outperform Bayesian simplicity. It incorporates four feature pillars: Community Identity, Rhetorical Framing, Structural Integrity, and Semiotic Dissonance. While this model provides a rich playground for ablation, it ultimately achieves only 0.6806 on the test set-2.6 points lower than the community-only model. This performance gap serves as the empirical basis for our conclusion that in low-resource conflict domains, complexity introduces more variance than signal. Detailed feature specifications are provided in Appendix A.

5. Experiments and Results

When we submitted our first deep learning model-a cross-attention fusion of XLM-RoBERTa and CLIP-we received a test Macro-F1 of 0.2935 back from

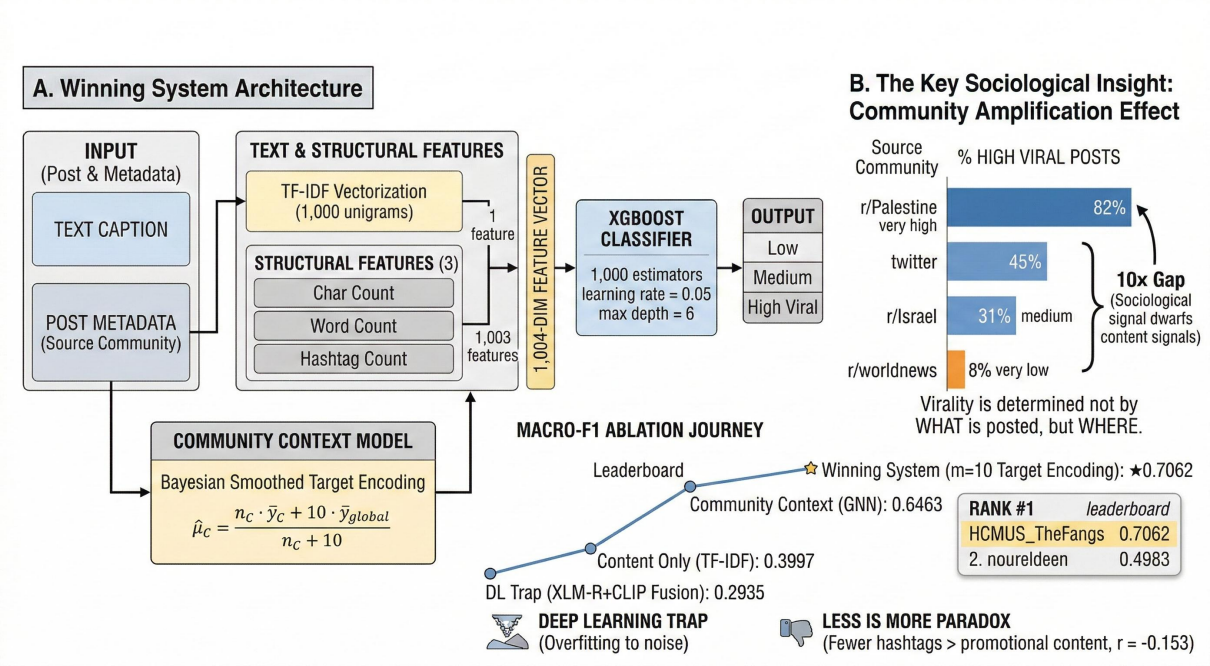


Figure 1: **Left:** Architecture of the winning system (Exp 32). The dominant signal is a single Bayesian-smoothed community target encoding feature derived from post metadata; TF-IDF and structural features contribute supplementary text signal. **Right (top):** The Community Amplification Effect-a tenfold gap in High Viral rate across source communities that no content encoder can recover from pixels or tokens. **Right (bottom):** Ablation progression from deep learning baselines (0.28) to the winning model (0.7062), and our official Rank #1 placement on the competition leaderboard.

the evaluation server. We were, frankly, stunned. Months of GPU time, careful architecture tuning, and multilingual fine-tuning had produced a system that barely outperformed a random guess on a balanced metric. What followed was not a methodological refinement but an intellectual reorientation. Table 1 is the record of that journey: every row is an official test-set score from the evaluation server, and together they tell a clean story about what virality prediction actually requires in conflict-domain social media.

The official competition leaderboard, shown in Table 2, reveals the scale of the advantage our community-context approach delivers. The next best submission reaches only 0.4983-a gap of more than 20 Macro-F1 points. This is not an incremental engineering win; it is the signature of a qualitatively different understanding of what the task is actually measuring.

The results tell a clear story across three phases. Deep learning models failed systematically: our most sophisticated fusion (Exp 01, 0.2935) barely beat random chance, and even the best neural result (Exp 00, 0.5230) fell short of simple TF-IDF plus XGBoost (Exp 02, 0.5328). Content-based feature engineering pushed performance higher-up to 0.5975 with persuasion-technique features-but hit a ceiling still ten points below the winning system.

The decisive leap occurred when community identity entered the model. A GNN-based community embedding alone (Exp 12, 0.6463) already surpassed every content-only system. Richer variants pushed further: the Singularity V2 reached 0.6954, and its Optuna-tuned ensemble successor crossed 0.70. Yet none surpassed the minimal Experiment 32 (**0.7062**), which encodes community context via a single smoothed target encoding feature. We spent weeks engineering 219 features for the Singularity; it scored 0.6806. Stripping almost everything away and keeping the community signal *won*. On a dataset of only 1,691 samples, every additional feature introduced more variance than signal. Simplicity was the correct answer to the bias-variance tradeoff.

6. Analysis and Insights

The most striking pattern in our results is sociological: the same image and caption can go viral or vanish depending entirely on who receives it. Topically focused communities arrive at conflict-related posts already emotionally invested, pre-mobilized by shared identity. Meanwhile, general-interest audiences encounter the identical content as just another headline. This asymmetry produces a tenfold gap in viral rates across communities-a gap our

Exp.	Model / Key Components	Macro-F1 (Test)
Deep Learning Baselines		
04	BLIP-2 Caption Features	0.2866
01	XLM-R + CLIP Cross-Attention Fusion	0.2935
03	DINOv2 + CLIP Dual-Backbone	0.4248
00	CLIP ViT-B/32 + Linear (EDA baseline)	0.5230
Ablation: Content Features Only (no community)		
29	TF-IDF + Structure + CLIP dissonance	0.3280
28	TF-IDF + Structure + Propaganda	0.3935
27	TF-IDF + Structure only	0.3997
02	TF-IDF + 30 Handcrafted (XGBoost)	0.5328
10	+ SNR / Structural Integrity	0.5879
07	+ Persuasion Techniques	0.5975
Community Signal (with content features)		
33	Visual Archetypes (CLIP K-Means only)	0.4152
12	Community Embedding (GNN)	0.6463
16	Community + Propaganda + Arousal	0.6829
20	Singularity V2 (39 features, HistGB)	0.6954
V3-R1	+ Optuna, 3-model ensemble	0.7001
34	Community $\times$ Visual Cluster	0.7028
V3-R3	Singularity: pure XGBoost + 180 interactions	0.6806
32	**Community Target Encoding (submitted)**	**0.7062**

Table 1: Full ablation results sorted by system family. All Macro-F1 scores are official test-set scores from the shared task evaluation server. Exp 32 is our submitted winning system (Rank #1 on the leaderboard).

Rank	Team	Macro-F1	Accuracy
1	**HCMUS_TheFangs (Ours)**	**0.7062**	**0.7305**
2	noureldeen	0.4983	0.6009
3	Digilians	0.4983	0.6009
4	xin1212	0.4559	0.6089
5	ashhadulislam	0.3199	0.4392

Table 2: Official NakbaVirality Shared Task leaderboard.

target encoding explicitly captures. This mirrors recent game-theoretic findings that environmental framing shapes behavior far more than architecture alone.

Equally counterintuitive is the "Less is More" paradox. While conventional wisdom treats hashtags as free discoverability, in the Gaza conflict domain, promotional packaging backfires. Posts written in sparse, direct prose consistently outperform those loaded with hashtags ($r = -0.153$, $p < 0.01$). Rhetorically, promotional language signals distance, whereas unadorned language signals proximity and urgency.

Rhetoric itself is a double-edged sword. Persuasion-technique detectors (flagging loaded language or emotional appeals) improved our models when paired with community context, but *hurt* performance in isolation. As frame resonance theory predicts (Benford and Snow, 2000), a phrase like "ethnic cleansing" acts as shared moral vocabulary in an activist community but as dismissible hyperbole in a general news forum.

The same context-dependence applies to images. Our visual encoders failed as standalone predictors because the visual distribution of viral and non-viral posts in this specific conflict dataset is virtually identical. However, when conditioned on community identity, visual archetypes became highly predictive: an image's value lies not in what it shows, but in how a specific audience interprets it.

7. Ethical Considerations

Predicting virality in conflict domains carries dual-use risks: algorithms that decode engagement can easily be co-opted to optimize disinformation. We publish our methodology because its defensive utility-equipping researchers to audit coordinated amplification-outweighs this danger. Crucially, our target encoding also crystallizes historical bias. By assuming temporal stationarity, the model perpetuates past audience divisions, reminding us that any deployed system must evolve alongside the communities it measures.

8. Conclusion

We set out to predict virality from images and text, only to discover that content matters far less than audience. By encoding a single sociological fact-the historical virality rate of the receiving community-our minimalist XGBoost system achieved **Rank #1** (0.7062 Macro-F1), decisively outperforming complex multimodal ensembles. Our ablation journey from 0.29 to 0.71 affirms that in sociologically structured domains, deep learning overfits, promotional hashtags backfire, and visual features require audience context to become predictive. Future work must explore how these community amplification coefficients transfer across time and platforms. Code is available at github.com/technoob05/NakbaVirality.

Acknowledgments

We thank the Nakba-NLP 2026 shared task organizers for providing the NakbaVirality dataset and the evaluation infrastructure. We also acknowledge the Faculty of Information Technology, University of Science, VNU-HCM, for computational support.

9. Bibliographical References

Takuya Akiba, Shotaro Sano, Toshihiro Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631.

Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. CrisisMMD: Multimodal twitter datasets from natural disasters. In *Proceedings of the 12th International AAAI Conference on Web and Social Media*, volume 12.

Robert D. Benford and David A. Snow. 2000. Framing processes and social movements: An overview and assessment. *Annual Review of Sociology*, 26:611–639.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Justin Cheng, Lada Adamic, P. Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. 2014. Cascades in temporal networks. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 913–924.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, pages 1377–1414.

Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

Saad Ezzini, Salima Lamsiyah, Shadi Abudalfa, Samir El-Amrany, and Walid Alsafadi. 2026. Nakbavirality: Multimodal virality prediction in high-stakes discourse. In *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742. PMLR.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31.

Nguyen Lam Phu Quy, Pham Phu Hoa, Tran Chi Nguyen, Dao Sy Duy Minh, Nguyen Hoang Minh Ngoc, and Huynh Trung Kiet. 2025. Beyond vision: Contextually enriched image captioning with multi-modal retrieval. In *Proceedings of the 3rd Workshop on Event-Centered Information Access (EVENTA), ACM International Conference on Multimedia (MM '25)*, pages 1–9, Dublin, Ireland. ArXiv:2512.20042. Ranked 3rd overall (Track 1).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR.

A. The Singularity: Full Feature Specification

This appendix documents the 219-feature Optuna-tuned XGBoost model (Exp 20-V3) described in Section 4.2. The architecture rests on four core pillars, beginning with *Community Identity* via

Bayesian target encoding ($m = 10$) of the filename-derived community label (shared verbatim with our winning system). This is structurally reinforced by *Rhetorical Framing*, which encodes nine features including seven persuasion-technique scores (e.g., Appeal to Emotion, Loaded Language, Name Calling) that are composited into Emotional and Attack Intensity aggregates. To capture textual hygiene, *Structural Integrity* measures eight features anchored by a Signal-to-Noise Ratio penalizing hashtags, mentions, and URLs—revealing the negative correlation between hashtag density and virality ($r = -0.153$, $p < 0.01$). Meanwhile, *Semiotic Dissonance* quantifies the multimodal gap via three CLIP cosine similarities: between the image and text, and between the image and the anchor phrases "violence" and "peace." A supplementary *Novelty Layer* injects 18 additional features—including GoEmotions BERT arousal scores (Demszky et al., 2020), Moral Foundations metrics, and NMF topic distributions. Ultimately, pairwise products of these features generate roughly 180 polynomial interaction terms, all optimized by Optuna's TPE sampler over 50 trials (Akiba et al., 2019).

B. Full Experiment Chronicle

Table 3 documents the complete experimental trajectory of our research, including experiments not shown in the main ablation table. All Macro-F1 scores are official test-set scores from the shared task evaluation server unless noted otherwise. This chronicle serves both as a reproducibility record and as a catalog of negative results that may inform future work in conflict-domain virality prediction.

C. Deep Learning Failure Analysis

Our deep learning experiments failed systematically rather than randomly.

Cross-Attention Fusion (Exp 01, F1=0.2935). With $\sim$1.3B parameters and only 1,691 training samples, the cross-attention model (XLM-RoBERTa + CLIP ViT-B/32) catastrophically overfit despite focal loss and class weighting. The cross-attention mechanism discovered noise rather than meaningful image-text interactions.

BLIP-2 Captioning (Li et al., 2023) (Exp 04, F1=0.2866). BLIP-2's captions were uniformly generic ("a group of people holding signs," "a destroyed building"), erasing the specificity that might have been useful and producing even less discriminative features than raw CLIP embeddings.

D. Hyperparameter Sensitivity

We report sensitivity analysis for the two most critical hyperparameters.

Smoothing Parameter m. Values tested: $m \in \{1, 5, 10, 20, 50\}$. The optimal $m = 10$ balances memorization (low m overfits small communities) against signal destruction (high m shrinks all encodings toward the global mean).

XGBoost Depth. Depth 6 achieved the best cross-validation Macro-F1 among values 4, 6, and 8. Depth 4 underfits the community-text interaction space; depth 8 overfits on 1,691 samples.

Exp	Name	Architecture / Key Components	Macro-F1	Outcome
Phase 1: Deep Learning Baselines (Feb 6–7)				
00	EDA Baseline	CLIP ViT-B/32 + Focal Loss + EDA features	0.5230	Baseline
01	XLM-R + CLIP	Cross-attention multilingual fusion	0.2935	**Failed**
03	DINOv2 + CLIP	Dual vision backbone with gated fusion	0.4248	Partial
04	BLIP-2 VLM	Image captioning as features	0.2866	**Failed**
05	Nelder-Mead Ensemble	CLIP + XLM-R + GB weight optimization	0.3376	**Failed**
08	Chain-of-Thought	CoT reasoning heads (without full LLM)	0.3659	**Failed**
Phase 2: Feature Engineering Discovery (Feb 7–8)				
02	Text-Only	TF-IDF + 30 handcrafted + XGBoost	0.5328	**GB beats DL**
06	Ordinal Regression	SORD soft ordinal labels on CLIP	0.5375	Marginal
07	Persuasion Techniques	SemEval-style propaganda features	0.5975	Strong
09	Arousal Hypothesis	GoEmotions + VAD arousal features	0.5707	Signal verified
10	Clean Content	SNR + hashtag ablation features	0.5879	"Less is More"
11	Propaganda Features	18 SemEval techniques (merged into 16)	$\sim$0.53	Merged
Phase 3: Community Context Breakthrough (Feb 7–8)				
12	Community GNN	Learnable subreddit embeddings	0.6463	**Breakthrough**
13	Neural Ensemble	CLIP + MLP on all features	0.6384	Below GBM
14	Classical ML	TF-IDF + keywords + SVM/LR	0.4865	Baseline
15	Stacking GBM	Handcrafted + SVD + HistGB	0.4938	Marginal
16	Grand Unification	Community + Arousal + Propaganda	0.6829	Strong
Phase 4: The Singularity and Optimization (Feb 8)				
20	Singularity V2	39 features, 4 pillars, HistGB	0.6954	Near-SOTA
V3-R1	Singularity V3 R1	+ Optuna 50 trials + 3-model ensemble	0.7001	First >0.70
V3-R2	Singularity V3 R2	Removed CatBoost, XGBoost boost	0.6982	Regression
V3-R3	Singularity V3 R3	Pure XGBoost + 180 interactions	0.6806	Regression
V3-R4	Singularity V3 R4	100 Optuna + smart interactions + bagging	0.6931	**Failed**
Phase 5: Hypothesis Testing and Final Model (Feb 13+)				
17	Narrative Framing	Moral Foundations + pronouns only	0.3070	**Failed**
18	Multimodal Dissonance	CLIP text-image similarity gap	0.3880	Weak
19	Topic Frames	NMF topic distributions	0.4230	Partial
21	Causal Discovery	PC Algorithm + DirectLiNGAM	0.6432	Strong
26	Meta-Ensemble	Stack of Exp 14, 15, 16, 21	0.5573	**Failed**
27	Paper Baseline	TF-IDF + structure (no neural)	0.3997	Baseline
28	Propaganda Hyp.	Baseline + propaganda + narrative	0.3935	**Negative**
29	Multimodal Hyp.	Baseline + CLIP dissonance	0.3280	**Negative**
33	Visual Archetypes	CLIP K-Means clustering (50 clusters)	0.4152	Partial
34	Social-Visual	Community $\times$ Visual Cluster	0.7028	Near-SOTA
35	Conflict Intensity	Violence score $\times$ text aggression	0.3680	**Negative**
37	Mixture of Experts	Cluster-gated experts	0.3711	**Mode collapse**
32	**Community Context**	**Target Encoding + TF-IDF + XGBoost**	**0.7062**	**#1 Winner**

Table 3: Complete experiment chronicle across all research phases. Bold "Failed" indicates experiments that performed below practical utility; "Negative" marks controlled hypothesis tests that produced informative null results. The progression reveals a clear narrative arc: deep learning failure $\rightarrow$ feature engineering discovery $\rightarrow$ community context breakthrough $\rightarrow$ simplicity wins.

Xin1212 at NakbaVirality Shared Task: Frozen CLIP with Residual Adapter for Multimodal Virality Classification

Xinyan Zhang
Dalian University of Technology
2737717435@mail.dlut.edu.cn

Bingzhou Yang
Dalian University of Technology
ybz19596612@mail.dlut.edu.cn

April 16, 2026

Abstract

We describe our system for the NakbaVirality shared task on multimodal virality classification. Our final approach uses a frozen LAION CLIP backbone, a lightweight residual adapter over fused text–image embeddings, and a small MLP classification head. On our development split, the best configuration (V9) achieves Macro-F1 of 0.5492 and virality-weighted F1 of 0.5252. On our official test submission, our system obtains F1-score 0.4559 and accuracy 0.6089 according to the platform scorer. We provide implementation details, ablations across multiple versions (Baseline–V9), and practical error analysis for reproducibility.

Keywords: multimodal virality, CLIP, residual adapter, multimodal classification

1. INTRODUCTION

The NakbaVirality task studies multimodal virality prediction in high-stakes geopolitical discourse, where both textual framing and imagery influence reach. The task is a 3-class classification problem (*Low*, *Medium*, *High* virality) evaluated primarily with Macro-F1.

This setting is challenging for three reasons: small dataset, class imbalance, and multilingual noisy content.

Our objective is to improve minority-class sensitivity while preserving stable optimization. Starting from a CLIP-based baseline, we iteratively developed versions V1–V9 and selected V9 as the final system. The key strategy is to keep the CLIP backbone frozen and add a residual adapter.

Main contributions.

- We provide a reproducible versioned study (Baseline to V9).
- We show that frozen-backbone + residual-adapter improves Macro-F1.
- We provide practical error analysis for shared-task settings.

2. BACKGROUND

2.1. Task setup

Input consists of post text and an associated image. Output is one label from {`High viral`, `Medium viral`, `Low viral`}.

2.2. Data description and split protocol

Following the shared task description Ezzini et al. (2026), the dataset is collected from X and Reddit with multilingual content.

We performed a stratified train/dev split ($80/20$, seed 42). This produced $1,691$ training instances in total and a 339-example development split for model selection. The official blind test set used in our submissions contains 872 examples.

2.3. Related modeling direction

Recent multimodal sentiment research reports that naive feature fusion can introduce semantic noise across modalities, reducing classification quality. SECIF addresses this by explicitly combining semantic enhancement and cross-modal interaction fusion Mu et al. (2025).

A complementary direction uses hierarchical cross-modal attention to better capture inter-modal dependencies than text-only pipelines Vamsidhar et al. (2025).

Beyond sentiment analysis, multimodal fusion has also proven useful in social-media fake news detection. A recent BERT-based framework combines text and image information with OCR-derived visual text, showing that image-side textual cues can contribute important evidence Al-alshaqi et al. (2025).

For virality, findings show that community structure and social reinforcement affect diffusion dynamics. More recent multimodal work studies early virality prediction in cross-lingual Reddit settings Dogan et al. (2025).

Finally, conflict-domain discourse analysis across platforms highlights substantial temporal

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 144–146
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

and narrative variation, reinforcing the need for robust multimodal modeling Antonakaki and Ioannidis (2025).

3. SYSTEM OVERVIEW

3.1. Architecture

Our final model (V9) has four stages:

1. CLIP encoder: `laion/CLIP-ViT-B-32-laion2B-s34B-b79K`
2. Late fusion by concatenating text and image embeddings
3. Residual adapter (two-layer MLP) for task-specific correction
4. MLP classifier with dropout for 3-way prediction

Let $z_t, z_i \in R^d$ be CLIP text and image embeddings. We fuse and adapt features as:

$$h = [z_t; z_i] \in R^{2d}, \quad \tilde{h} = h + A(h), \qquad (1)$$

where $A(\cdot)$ is the residual adapter. Final logits are

$$\hat{y} = W_2\sigma\left(W_1\tilde{h}\right). \qquad (2)$$

3.2. Training objective and optimization

We train with weighted cross-entropy and label smoothing ($\epsilon = 0.1$). For a sample with target class y, the smoothed target distribution is $q_k = (1-\epsilon)\mathbf{1}[k=y] + \epsilon/K$, and the loss is $\mathcal{L} = -\sum_{k=1}^{K} w_k q_k \log p_k$, where w_k are class weights and $K = 3$.

4. EXPERIMENTAL SETUP

4.1. Data usage

We trained and selected models on a stratified $80/20$ split of the released training data (seed 42). The development split (339 samples) was used for all local ablations.

4.2. Training configuration (V9)

Key parameters: Batch size 16, Epochs 20, Learning rate $1\mathrm{e}^{-3}$, label smoothing 0.1, and cosine scheduler with warmup.

5. RESULTS

5.1. Version comparison on development split

Table 1 summarizes metrics logged in our study.

Version	Macro-F1	Weighted-F1	Acc.
Baseline	0.4798	0.3924	0.6401
V1	0.5398	0.5072	0.5870
V5	0.5429	0.5052	0.5929
V8	0.5023	0.4633	0.5457
V9	**0.5492**	**0.5252**	0.5959

Table 1: Results on development split.

5.2. Official test submission

The platform reported: F1-score 0.4559, Accuracy 0.6089, Precision 0.6000, Recall 0.6089.

6. CONCLUSION

We presented a frozen-CLIP + residual-adapter approach for NakbaVirality. V9 was the strongest dev configuration and improved minority-class behavior. On our official test submission, the system achieved F1 0.4559.

Mohammed Al-alshaqi, Danda B Rawat, and Chunmei Liu. 2025. A bert-based multimodal framework for enhanced fake news detection using text and image data fusion. *Computers*, 14(6):237.

Despoina Antonakaki and Sotiris Ioannidis. 2025. Cross-platform digital discourse analysis of the israel-hamas conflict: Sentiment, topics, and event dynamics. *arXiv preprint arXiv:2601.02367*.

Sedat Dogan, Nina Dethlefs, and Debarati Chakraborty. 2025. Early multimodal prediction of cross-lingual meme virality on reddit: A time-window analysis. *arXiv preprint arXiv:2510.05761*.

Saad Ezzini, Salima Lamsiyah, Shadi Abudalfa, Samir El-Amrany, and Walid Alsafadi. 2026. The nakbavirality shared task on multimodalvirality prediction in high-stakes discourse. In *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

Guangyu Mu, Ying Chen, Xiurong Li, Li Dai, and Jiaxiu Dai. 2025. Semantic enhancement and cross-modal interaction fusion for sentiment analysis in social media. *PLOS ONE*, 20(4):e0321011.

D Vamsidhar, Parth Desai, Aniket K Shahade, Shruti Patil, and Priyanka V Deshmukh. 2025.

Hierarchical cross-modal attention and dual audio pathways for enhanced multimodal sentiment analysis. *Scientific Reports*, 15(1):25440.

A. REPRODUCIBILITY DETAILS

The final submitted model uses a frozen CLIP backbone with a residual adapter and an MLP classifier. Optimization uses AdamW, cosine schedule with warmup, and label smoothing.

A2NLP at StanceNakba Shared Task: Fine-Tuned AraBERT for Topic-Based Arabic Stance Detection

Alaa Nairat, Aysar Nairat
Birzeit University, Modern University College
Ramallah, Palestine
anairat@birzeit.edu, aysar.nairat@muc.edu.ps

Abstract

This paper describes A2NLP's system for Subtask B of the StanceNakba Shared Task, which addresses cross-topic Arabic stance detection. The goal is to classify sentence–topic pairs into pro, against, or neutral labels. We introduce a topic-conditioned prompting strategy built on `AraBERTv0.2-Twitter`, where each instance is reformulated into a structured prompt that explicitly models the interaction between the sentence and its target topic. The model is trained using 5-fold stratified cross-validation with class-weighted loss to ensure robustness under mild label imbalance. Our final submission achieves a Macro-F1 score of 0.8483 on the official test set, outperforming the `AraBERTv2` baseline (0.810) and ranking fifth overall. Ablation analysis confirms that topic-conditioned prompting substantially improves generalization across topics. The findings demonstrate the importance of structured input design and domain-aligned pretraining for reliable stance detection in dialectal Arabic social media discourse.

Keywords: Arabic NLP, Stance detection, transformer, topic-conditioned prompts, text classification

1. Introduction

Understanding public opinion in online discourse is crucial for social, political, and health-related decision-making, yet manually tracking nuanced stances at scale is infeasible. Stance detection is the automated task of determining whether the author of a text expresses a position *in favor of*, *against*, or *neutral* toward a specific subject of interest, commonly referred to as a *target* (Küçük and Fazli, 2020; AlDayel and Magdy, 2021). Unlike sentiment analysis, stance detection is inherently target-dependent and often requires inferring implicit opinions that are not explicitly stated. This capability is essential for applications such as social media monitoring, political analysis, and public health surveillance, where nuanced public positions are critical and traditional survey-based methods are insufficient (Küçük and Fazli, 2020).

The *StanceNakba Shared Task* 2026 (Aldous et al., 2026), co-located with LREC 2026, provides a challenging benchmark for Arabic stance detection. In particular, **Subtask B** focuses on *Cross-Topic Stance Detection*, requiring a unified model capable of predicting stance across multiple conflict-related topics. This setup reflects realistic deployment scenarios in which systems must generalize beyond a single topic and handle emerging or unseen targets. Arabic, spoken by over 420 million people, remains underrepresented in stance detection research (Guellil et al., 2021; Qu et al., 2024) and poses unique Natural Language Processing (NLP) challenges, including rich morphology, extensive dialectal variation, and informal social media writing with spelling inconsistencies, abbreviations, and emojis (Badaro et al., 2019).

Our Team (**A2NLP**) proposes a prompt-based `AraBERTv0.2-Twitter` model for topic-conditioned Arabic stance detection. Our approach outperforms the official baseline in terms of Macro-F1. Our main contributions are: (1) **Input Reformulation:** Topic-conditioned prompting to model target–text interactions; (2) **Robust Training:** 5-fold stratified cross-validation with class-weighted loss; and (3) **Public Resources:** Public model weights and an interactive *Hugging Face Space* demo. The code implementation can be reached via A2NLP-Code[1], and Interactive demo can be reached via A2NLP-SharedTask-Demo[2]. The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the task and dataset. Section 4 presents our system. Section 5 outlines the experimental setup. Section 6 reports results and error analysis. Section 7 concludes and highlights future directions.

2. Related Work

2.1. Evolution of Arabic Stance Resources

Although Arabic stance detection has progressed in recent years, most early research was dominated by English datasets (Küçük and Fazli, 2020; AlDayel and Magdy, 2021; Guellil et al., 2021). Several Arabic resources have since

[1] https://github.com/eng-aomar/A2NLP-StanceNakba2026-SubtaskB-AraBERTv02-Twitter

[2] https://huggingface.co/spaces/aomar85/A2NLP-StanceNakba-Demo

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 147–159
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

emerged. MAWQIF introduced $4,121$ tweets annotated for stance and related phenomena (Alturayeif et al., 2022). AraStance provided $4,063$ claim–article pairs for fact-checking-oriented stance detection (Alhindi et al., 2021). ArabicStanceX expanded coverage to $14,477$ tweets across 17 topics, enabling multi-topic and zero-/few-shot settings (Alkhathlan et al., 2025). MARASTA further incorporated cross-domain and multi-dialectal variation (Charfi et al., 2024b). Despite these advances, robust cross-topic generalization in dialectal social media remains challenging.

2.2. Computational Methodologies

Early systems relied on traditional classifiers such as SVMs with TF-IDF and n-gram features (Darwish et al., 2017; Baly et al., 2018), later replaced by CNN and LSTM architectures. More recently, Multi-Task Learning (MTL) jointly modeled stance with related tasks (e.g., sentiment, sarcasm), improving shared representations and class balance (Alturayeif et al., 2023). Sequential MTL with hierarchical weighting (SMTL-HW) achieved 88.3% F1, highlighting the benefit of task interaction (Almasoudi et al., 2025).

2.3. Transformer Models

Transformer-based models now dominate Arabic NLP (Antoun et al., 2020). `AraBERTv0.2-Twitter`, pre-trained on 60M Arabic tweets, has shown strong performance in social media stance detection (Antoun et al., 2020; Badran et al., 2024). It underpinned the winning StanceEval 2024 system, enhanced with contrastive learning (Badran et al., 2024), and consistently outperforms MARBERT and CAMeLBERT-da on MAWQIF (Alturayeif et al., 2022, 2024). However, existing work primarily relies on standard fine-tuning, leaving input reformulation and topic-conditioned modeling underexplored.

3. Background

3.1. Task Setup

Subtask B of the StanceNakba 2026 Shared Task (Aldous et al., 2026) addresses cross-topic Arabic stance detection. Each instance consists of a sentence paired with one of two predefined target topics: وجود اللاجئين والمهاجرين في الأردن (من فلسطين، سوريا، العراق) Refugee/Immigrant Presence in Jordan from (Palestine, Syria and Iraq) or التطبيع مع اسرائيل (Normalization with Israel). The task is formulated as sentence pair classification problem with three labels, where the system predicts the stance expressed in the sentence toward the specified topic. The output label belongs to one of three categories: **pro**, **against**, or **neutral**. For example, given the sentence افضل شعب الشعب السوري اختكم من الأردن (The Syrian people are the best people, your sister from Jordan) and the topic وجود اللاجئين والمهاجرين في الأردن (من فلسطين، سوريا، العراق) (Refugee/Immigrant Presence in Jordan from (Palestine, Syria and Iraq)), the appropriate label is **pro**, as the sentence conveys support in relation to the target topic.

3.2. Dataset Details

The dataset used in this subtask is a subset of the MARASTA corpus (Charfi et al., 2024a), consisting of Arabic sentence–topic pairs primarily written in colloquial Arabic and regional dialects such as Palestinian, Jordanian, and Qatari. The genre mainly comprises tweets and short online posts, characterized by informal language, abbreviations, and frequent use of emojis.

The dataset contains $1,024$ instances in total. As shown in Table 1, the label distribution is approximately balanced across the three stance categories: *against* (361 instances, 35.25%), *pro* (348 instances, 33.98%), and *neutral* (315 instances, 30.76%). The prevalence of dialectal variation and informal expressions increases the complexity of the stance detection task, requiring models to generalize across linguistic and stylistic variability.

Topic / Label	Count	Percentage (%)
(التطبيع مع اسرائيل) (Normalization with Israel)		
Against	168	34.22
Neutral	177	36.05
Pro	146	29.73
Sub-total	491	
وجود اللاجئين والمهاجرين (فلسطين، سوريا، العراق) Refugee/Immigrant Presence in Jordan from (Palestine, Syria and Iraq)		
Against	193	36.21
Neutral	138	25.89
Pro	202	37.90
Sub-total	533	
Total (All Topics)	1024	

Table 1: Distribution of Stance Topics for each Label.

4. System Overview

Our system for Subtask B as presented in Figure 1 utilizes a fine-tuned `AraBERTv0.2-Twitter` model, chosen for its robust handling of Modern Standard Arabic (MSA) and the dialectal, emoji-rich nuances of social media text. To optimize performance, we employ dynamic batch-wise padding and class weighting to mitigate label

imbalance.

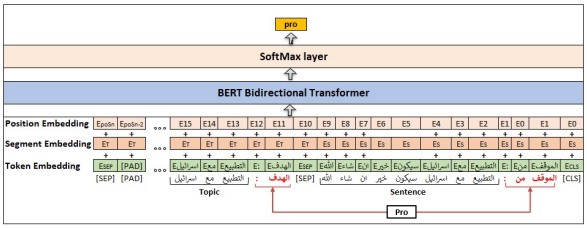


Figure 1: Architecture of the topic-conditioned AraBERT model.

Input Representation: To facilitate cross-topic generalization, we employ a prompt-based approach. Given a sentence x_s and topic x_t, we define static Arabic descriptors $p_1 =$ الهدف: (Target:) and $p_2 =$ الموقف من: (Stance toward:). The input sequence X is constructed as:

$$X = \text{[CLS]}\ p_1\ x_t\ \text{[SEP]}\ p_2\ x_s\ \text{[SEP]}$$

An example of the constructed input sequence is shown below:

```
[CLS] الموقف من: التطبيع مع اسرائيل [SEP] الهدف: التطبيع
مع اسرائيل سيكون خير ان شاء الله [SEP]
Target: Normalization with Israel
[SEP] Stance toward: Normalization
with Israel will be good, God willing
```

Classification Head: The encoder produces contextual embeddings $H \in \mathbb{R}^{L \times d}$. The sequence-level representation, $h_{\text{CLS}} \in \mathbb{R}^d$, is passed through a linear layer f_θ with dropout to compute the predicted probabilities $\hat{y} = \text{softmax}(f_\theta(h_{\text{CLS}}))$.

5. Experimental Setup

5.1. Dataset Details and Preprocessing

Experiments were conducted in Google Colaboratory using 5-fold Stratified Cross-Validation (SCV). For each phase ($N = 824$ and later $N = 1{,}024$), data was split into five stratified folds, with 80% for training and 20% for validation. Cross-validation guided model selection and hyperparameter tuning. Final evaluation was performed on the official Codabench platform. Tables A.5 and 3 report Macro-F1 on the unseen test set.

Preprocessing included Arabic-specific cleaning and normalization. Emojis were converted to text, non-Arabic characters, URLs, mentions, and hashtags were removed. We normalized orthographic variants (Alef forms, Alef Maqsura, Taa Marbuta), removed diacritics and tatweel, reduced extra whitespace, filtered empty samples, and eliminated duplicates.

5.2. Hyperparameter Fine-Tuning

The model was trained for up to 10 epochs with early stopping (patience $= 2$) based on Macro-F1. We applied weight decay (0.01) and dropout (0.2) to mitigate overfitting. Back-translation (Arabic→English→Arabic) generated 842 additional samples for augmentation. Hyperparameters are summarized in Table 2, with extended details in Appendix Table A.2.

Hyperparameter	Value
Max Sequence Length	128 tokens
Batch Size	16
Learning Rate	2×10^{-5}
Weight Decay	0.01
Dropout Rate	0.2
Optimization Metric	Macro-F1
Early Stopping Patience	2 epochs

Table 2: Fine-tuning hyperparameters for stance detection.

6. Results

We evaluated our system on the official test split. Our submission achieved a Macro-F1 score of 0.8483, ranking among the top teams in Subtask B and outperforming the baseline (AraBERT v2) by 0.0383. The full leaderboard results are provided in Appendix A.3 (Table A.5).

We followed a two-phase evaluation: an initial development track for model selection, followed by a final test track for performance verification. Table 3 summarizes our experiments. Prompt-based fine-tuning consistently improved F1 over baseline and augmentation, with the best score obtained using $1{,}024$ samples[3].

Exp	Model	F1	Technique	Size
1	M1[a]	0.777	Baseline	824
2	M1[a]	0.830	Prompt	824
3	M1[a]	0.811	Augmentation Back-translation	$1{,}684^*$
4	M2[a]	0.780	Prompt	824
5	M1[b]	0.836	Prompt	824
6	M1[b]	**0.848**	Prompt	$1{,}024$

Table 3: Subtask B experiments.

M1: AraBERTv02-Twitter; M2: MARBERT
*Augmented samples included.
[a]LR = 3×10^{-5}; [b]LR = 2×10^{-5}

[3]Only representative configurations are shown; additional runs were conducted during model selection.

6.1. Discussion of Key Findings

Our analysis reveals three primary factors that contributed to the model's success. First, the introduction of **topic-conditioned prompts** significantly enhanced the model's ability to generalize to the test set, effectively bridging the gap between training and unseen data. Second, optimization played a crucial role; reducing the learning rate from 3×10^{-5} to 2×10^{-5} improved Macro-F1 from 0.8328 to 0.8363 (absolute gain 0.0035). This consistent cross-validation improvement indicates that a slightly smaller optimization step stabilizes fine-tuning on small stance datasets.

Finally, the impact of data volume was evident in the final phase. Expanding the size of the training to $N = 1,024$ (including validation data released after development) increased Macro-F1 from 0.8364 to 0.8483, highlighting an absolute gain of 0.0119. This demonstrates that while architectural prompts and hyperparameter tuning are important, `AraBERTv0.2-Twitter` is highly data-responsive, benefiting substantially from the greater linguistic variety in the larger dataset.

6.2. Error Analysis

We analyzed misclassifications from out-of-fold (OOF) predictions across the 5-fold stratified cross-validation. Table 4 summarizes the most frequent confusion pairs, and Figure 2 shows the aggregated confusion matrix.

The dominant error is *Neutral* $\rightarrow$ *Against* (32.14%), followed by *Against* $\rightarrow$ *Neutral* (20.24%). This bidirectional confusion suggests difficulty distinguishing implicit opposition from weak or ambiguous sentiment, especially when negative tone is directed toward contextual entities rather than the target itself.

Confusions involving *Pro* are less frequent (*Pro* $\rightarrow$ *Against*: 10.71%, *Pro* $\rightarrow$ *Neutral*: 8.33%), indicating that supportive stances often contain clearer lexical cues. Overall, errors are relatively balanced across classes, suggesting that Macro-F1 gains are not driven by majority-class bias.

The prevalence of Neutral–Against confusions indicates that topic-conditioned prompting does not fully resolve stance ambiguity in cross-topic settings. High-confidence misclassified examples are provided in Appendix A.4(Table A.6).

Confusion Type	Count	Percentage
Neutral $\rightarrow$ Against	27	32.14%
Against $\rightarrow$ Neutral	17	20.24%
Neutral $\rightarrow$ Pro	12	14.29%
Against $\rightarrow$ Pro	12	14.29%
Pro $\rightarrow$ Against	9	10.71%
Pro $\rightarrow$ Neutral	7	8.33%

Table 4: Misclassification counts and percentages for each confusion type from OOF predictions.

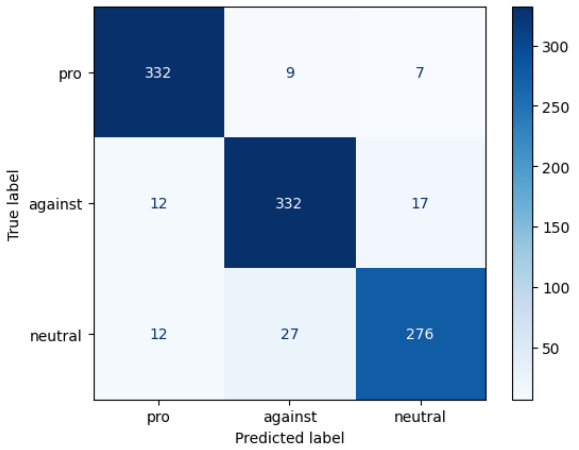


Figure 2: Aggregated confusion matrix for the 5-fold stratified cross-validation on the development set. Results are obtained via OOF inference using our fine-tuned `A2NLP-STANCE` model (based on `AraBERTv0.2-Twitter`) with topic-conditioned prompts.

7. Conclusion

We presented A2NLP's system for Subtask B of the StanceNakba 2026 Shared Task on cross-topic Arabic stance detection. Using topic-conditioned prompts with fine-tuned `AraBERTv0.2-Twitter`, our model reached a Macro-F1 of 0.8483, surpassing the `AraBERTv2` baseline by $+3.83$ points and ranking fifth. Our results show that structured input reformulation and domain-aligned pretraining boost cross-topic generalization, while distinguishing neutral from implicit opposing stances remains challenging. Future work will explore richer contextual modeling to improve stance disambiguation.

8. Limitations

Despite competitive performance, the A2NLP system is limited by a small dataset ($1,024$ instances) restricting generalization across Arabic dialects, difficulty distinguishing neutral from implicit opposition (32.14% misclassified), and evaluation on only two topics, while relying solely on `AraBERTv0.2-Twitter` without external knowledge or multi-task learning.

9. Acknowledgements

We thank the Arab NLP community and the StanceNakba 2026 organizers for providing a valuable benchmark for Arabic stance detection.

10. Bibliographical References

Abeer AlDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Kholoud Khalil Aldous, Md Rafiul Biswas, Mabrouka Bessghaier, Shimaa Ibrahim, Kais Attia, and Wajdi Zaghouani. 2026. StanceNakba shared task: Actor and topic-aware stance detection in public discourse. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

Abeer Almasoudi, Muhammad Arif, Ahlam Hashem, and Esraa Samkari. 2025. Enhanced stance detection for arabic tweets. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 29(6):1262–1272.

Nora Alturayeif, Hamzah Luqman, Zaid Alyafeai, and Asma Yamani. 2024. Stanceeval 2024: The first arabic stance detection shared task. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 774–782.

Nora Saleh Alturayeif, Hamzah Luqman, and Moataz A. Ahmed. 2023. Enhancing stance detection through sequential weighted multi-task learning. *Social Network Analysis and Mining*, 14(1):7.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools*, pages 9–15.

Gilbert Badaro, Ramy Baly, Hazem Hajj, Wassim El-Hajj, Khaled Bashir Shaban, Nizar Habash, Ahmad Al-Sallab, and Ali Hamdi. 2019. A survey of opinion mining in arabic: A comprehensive system perspective covering challenges and advances in tools, resources, models, applications, and visualizations. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(3):1–52.

Mohamed Badran, Mo'men Hamdy, Marwan Torki, and Nagwa M El-Makky. 2024. Alexunlp-bh at stanceeval2024: Multiple contrastive losses ensemble strategy with multi-task learning for stance detection in arabic. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 823–827.

Ramy Baly, Mitra Mohtarami, James Glass, Preslav Nakov, et al. 2018. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 21–27.

Kareem Darwish, Walid Magdy, and Tarek Zanouda. 2017. Improved stance prediction in a user similarity feature space. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 145–148.

Imane Guellil, Houda Saadane, Faiçal Azouaou, Billel Gueni, and Damien Nouvel. 2021. Arabic natural language processing: An overview. *Journal of King Saud University - Computer and Information Sciences*, 33(5):497–507.

Dilek Küçük and Cemal Aykut Nufus Fazli. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Xiaoye Qu, Yingjie Gu, Qingrong Xia, Zechang Li, Zhefeng Wang, and Baoxing Huai. 2024. A survey on arabic named entity recognition: Past, recent advances, and future trends. *IEEE Transactions on Knowledge and Data Engineering*, 36(3):943–959.

11. Language Resource References

Tariq Alhindi, Amal Alabdulkarim, Ali Alshehri, Muhammad Abdul-Mageed, and Preslav Nakov. 2021. Arastance: A multi-country and multi-domain dataset of arabic stance detection for fact checking. *arXiv preprint arXiv:2104.13559*.

Ali Alkhathlan, Faris Alahmadi, Faris Kateb, and Hend Al-Khalifa. 2025. Constructing and evaluating arabicstancex: a social media dataset for arabic stance detection. *Frontiers in Artificial Intelligence*, 8:1615800.

Nora Saleh Alturayeif, Hamzah Abdullah Luqman, and Moataz Aly Kamaleldin Ahmed. 2022. Mawqif: A multi-label arabic dataset for target-specific stance detection. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 174–184.

Anis Charfi, Mabrouka Ben-Sghaier, Andria Samy Raouf Atalla, Raghda Akasheh, Sara Al-Emadi, and Wajdi Zaghouani. 2024a. Marasta: A multi-dialectal arabic cross-domain stance corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11060–11069.

Anis Charfi, Mabrouka Bessghaier, Andria Atalla, Raghda Akasheh, Sara Al-Emadi, and Wajdi Zaghouani. 2024b. Stance detection in arabic with a multi-dialectal cross-domain stance corpus. *Social Network Analysis and Mining*, 14(1):161.

A. Appendix: Supplementary Materials, Software, and Data

This appendix provides implementation details and representative error examples to facilitate reproducibility and qualitative analysis.

A.1. Reproducibility Details

Framework / Library	Version	URL
Transformers	5.0.0	https:https://github.com/huggingface/transformers
PyTorch	2.9.0+cu128	https://pytorch.org
Datasets	4.0.0	https://github.com/huggingface/datasets
Tokenizers	0.22.2	https://github.com/huggingface/tokenizers

Table A.1: Environment setup.

Hyperparameter	Value
Number of epochs	10
Train batch size	16
Evaluation batch size	16
Learning rate	$2e-5$
Weight decay	0.01
Dropout	0.2 (hidden)
Evaluation strategy	Steps
Evaluation steps	50
Save strategy	Steps
Save steps	50
Push to Hugging Face Hub	True
Hub model ID	Dynamic per fold
Load best model at end	True
Metric for best model	Macro-F1
Greater is better	True
FP16	True
Logging strategy	Steps
Logging steps	50
Report to	None
Early stopping patience	2

Table A.2: Extended hyperparameter values for full model replication.

A.2. Additional Backbone Evaluation

To ensure reproducibility, we provide the specific model weights and configurations for each experiment conducted in Subtask B. Table A.3 details the performance metrics, fine-tuning techniques, and dataset sizes used, with corresponding Hugging Face repository links provided in the table notes for each unique model checkpoint.

Table A.4 reports additional pretrained Arabic models evaluated under identical training settings during development. These models were not selected for final submission due to lower Macro-F1 performance.

A.3. Leaderboard Results (Full Ranking)

For completeness, we include the full official leaderboard for Subtask B in the appendix. As shown in Table A.5, our system (A2NLP) ranks 5th overall with a Macro-F1 score of 0.8483, outperforming the AraBERT v2 baseline and placing among the top-performing teams.

Exp.	Model	F1	Technique	Size
1	M1[a]	0.777	Baseline	824
2	M1[a]	0.830	Prompt	824
3	M1[a]	0.811	Augmentation	1,684*
4	M2[a]	0.780	Prompt	824
5	M1[b]	0.836	Prompt	824
6	M1[b]	**0.848**	Prompt	1,024

Model URLs by Experiment:
Exp 1: https://huggingface.co/aomar85/StanceNakba-EXP1
Exp 2: https://huggingface.co/aomar85/StanceNakba-EXP2
Exp 3: https://huggingface.co/aomar85/StanceNakba-EXP3
Exp 4: https://huggingface.co/aomar85/StanceNakba-EXP4
Exp 5: https://huggingface.co/aomar85/StanceNakba-EXP5
Exp 6: https://huggingface.co/aomar85/A2NLP-STANCENAKBA2026-CROSS-TOPIC

M1: AraBERTv02-Twitter; M2: MARBERT
*Augmented samples included; [a]LR = 3×10^{-5}; [b]LR = 2×10^{-5}

Table A.3: Subtask B experiments and associated Hugging Face repositories.

Exp.	Model	Accuracy	Macro-F1
7	CAMeLBERT-Mix-Sentiment†	0.7459	0.7454
8	MARBERT†	0.6409	0.6354

Model URLs by Experiment:
Exp 7: https://huggingface.co/aomar85/StanceNakba-EXP7
Exp 8: https://huggingface.co/aomar85/StanceNakba-EXP8

Table A.4: Additional backbone models evaluated on the validation test set.

Rank	Team / Participant	Macro F1
1	wafaa	0.8724
2	mozalak	0.8607
3	U4RASD	0.8601
4	sarah_yassine	0.8562
5	**A2NLP (ours)**	0.8483
6	asmaa_q	0.8328
7	tokakhaled	0.8110
Baseline (AraBERT v2)		0.8100
8	AyahVerse	0.7904
9	jahidhasan	0.7407
10	walisa_alam	0.7317

Table A.5: Final leaderboard results for Subtask B. Our system (A2NLP) and the AraBERT v2 baseline are highlighted.

A.4. Representative Misclassifications

Sentence	Topic	Ground Truth	Prediction	Confidence
لا يزال اللاجئون يعانون من الآثار الاقتصادية الناجمة عن جائحة كورونا وارتفاع تكلفة المعيشة، والآن أيضا من ارتفاع تعرفة الخدمات Refugees still suffer from the economic impacts of the COVID-19 pandemic, the high cost of living, and now rising service fees.	وجود اللاجئين والمهاجرين في الأردن (من فلسطين ، سوريا ، العراق) Refugee/Immigrant Presence in Jordan from (Palestine, Syria and Iraq)	pro	neutral	97.95%
التطبيع مع الاسد اوسخ من التطبيع مع اسرائيل Normalization with Assad is filthier than normalization with Israel.	التطبيع مع اسرائيل Normalization with Israel	neutral	against	97.63%
أتوقع انو إخوانا السوريين مقدرين الوضع الي بتمر فيه الأردن وهم اخواننا ونسايبنا وقرايبنا لكن الوضع ما بتحمل (الي مش بإيدك بكيدك) الله يكون بعونهم وبعونا I believe our Syrian brothers understand the situation that Jordan is going through. They are our brothers, relatives, and family. However, the situation is becoming difficult to withstand what is beyond your control can still weigh heavily on you. May God help them and help us.	وجود اللاجئين والمهاجرين في الأردن (من فلسطين ، سوريا ، العراق) Refugee/Immigrant Presence in Jordan from (Palestine, Syria and Iraq)	neutral	pro	97.58%
يا سلام على الاخوة الاماراتيين التطبيع حتى في التعاليق ...ربنا معاكم O wow, the Emirati brothers, normalization even in the comments... God be with you.	التطبيع مع اسرائيل Normalization with Israel	neutral	pro	97.00%
الحديث لاجل فلسطين علامة مهمة لصحة الضمير والوعي سلام من ارض العراق لأصحاب الارض المقاومين سادة الرجال في زمن التطبيع المخنث Speaking up for Palestine is a clear sign of a living conscience and true awareness. Greetings from the land of Iraq to the rightful owners of the land the steadfast and resilient true men in a time of normalization.	التطبيع مع اسرائيل Normalization with Israel	against	pro	96.99%
نقدر فعلاً اخواننا السوريين و نتمنالهم الخير و الأمن و أن ترجع سوريا بخير ، بس الاردن مش قادرة تطعمي حالها عشان تطعمي غيرها .. We genuinely value our Syrian brothers and sisters and wish for Syria's safety and recovery, but Jordan is struggling to sustain its own people right now, making it hard to support others.	وجود اللاجئين والمهاجرين في الأردن (من فلسطين ، سوريا ، العراق) Refugee/Immigrant Presence in Jordan from (Palestine, Syria and Iraq)	against	pro	96.34%

Continued on next page

Sentence	Topic	Ground Truth	Prediction	Confidence
الأفضل ما حدا يرجع هناك لأنه مصيبة لو رجعوا عسوريا It's better that no one goes back there because it would be a disaster if they returned to Syria.	وجود اللاجئين والمهاجرين في الأردن (من فلسطين ، سوريا ، العراق) Refugee/Immigrant Presence in Jordan from (Palestine, Syria and Iraq)	neutral	against	96.10%
الأردن يكفيها إخواننا السوريين ومو ناقصة نتمنى أن يتم توزيع الجرحى التي تستدعي حالتهم النقل على مستشفيات دول الخليج Jordan has done its part for our Syrian brothers and can't take on any more. We hope that the wounded whose cases require transfer will be distributed among the hospitals of the Gulf countries	وجود اللاجئين والمهاجرين في الأردن (من فلسطين ، سوريا ، العراق) Refugee/Immigrant Presence in Jordan from (Palestine, Syria and Iraq)	against	pro	95.66%
الوقف إلى جانب إخواننا السوريين واجب قومي وطني و صدقت معاليك المملكة تحملت الكثير الكثير ولكن مصلحة الاردن العليا فوق اي اعتبارات Standing by our Syrian brothers is a national and patriotic duty, and you are right, Your Excellency the Kingdom has endured a great deal. However, Jordan's supreme interest stands above all other considerations.	وجود اللاجئين والمهاجرين في الأردن (من فلسطين ، سوريا ، العراق) Refugee/Immigrant Presence in Jordan from (Palestine, Syria and Iraq)	against	pro	95.64%
اسرائيل لا تحترم المطبعين ولا #التطبيع #فلسطين_ليست_قضيتي Israel respects neither those who normalize nor normalization #Palestine_is_not_my_cause	التطبيع مع اسرائيل Normalization with Israel	neutral	against	95.35%
طيب وين المشكله ؟ قلوب صادقه تبحث عن السلام وانشاء علاقه مع العرب ، وطبعاً التطبيع بعد انهاء الاحتلال وانشاء دوله فلسطينيه So where is the problem? These are sincere hearts seeking peace and building relations with the Arabs—and of course, normalization would follow the end of the occupation and the establishment of a Palestinian state.	التطبيع مع اسرائيل Normalization with Israel	neutral	pro	95.19%
إن شاء الله يثور الشعب على هذا الذل الإرهابيون الصهاينة يقتلون الفلسطينيين الدول العربية تطبيع مع هؤلاء الارهابيين الصهاينة بابتسامة باعوا دينهم ومكانهم المقدس God willing, the people will rise against this humiliation. The Zionist terrorists are killing Palestinians, while Arab countries normalize with these Zionist terrorists with a smile; they have sold their religion and their holy places.	التطبيع مع اسرائيل Normalization with Israel	neutral	against	93.78%

Continued on next page

Sentence	Topic	Ground Truth	Prediction	Confidence
عبئ على الأردن يا خنزير ؟؟ اقتصاد الأردن لم يرى الخير منذ قدومهم؟؟ اولا الأردن منهوبة ومن أفقر دول العالم قبل أزمة سوريا بقرووون وجميعنا ك أردنيين عارفين النهب والسرقة لأموال الشعب من حكومتنا الساقطة والساقط الي يسحج لها . بشو اثروا علينا السوريين؟؟ اخذوا منا بيوتنا ولا اخذوا من اهالينا؟؟؟ السوريين في منهم احسن منا بمراحل وسوريا يلي بتحكي عنها كانت بيوم عاصمتنا عاصمة الشام قبل التقسيمة الوسخة الي صارت . سوريا الي ما بحس فيها وبشعبها ما عنده لا ذمة ولا ضمير وحرام فيه يعيش في بلد عربي وبين العرب. تعرف ادعي دعوه وحده دايما ورح اضل ادعيها على كل شخص ما يقدر السوريين وما يحس فيهم ، الله يطعمك بحرب تشتت اهلك وتجعلك تتشتت انت وعيالك وتدور على مكان يحتويك وما تلاقيه ، هذه الدعوه مش لك شخصيا هذه لكل شخص ما يحس بيلي بيعانيه السوريين بس What 'good' was there since they arrived? First of all, Jordan has been looted and has been one of the poorest countries in the world for centuries, long before the Syrian crisis. All of us as Jordanians know the looting and theft of the people's money by our failed government and the sycophants who cheer for them. How have the Syrians affected us? Did they take our homes or take from our families? Many Syrians are far better than us, and the Syria you speak of was once our capital—the capital of the Levant—before the dirty partition that occurred. Anyone who doesn't feel for Syria and its people has neither integrity nor a conscience, and doesn't deserve to live in an Arab land among Arabs. I have one prayer that I will always repeat for anyone who doesn't appreciate or feel for the Syrians: May God afflict you with a war that scatters your family and forces you and your children to wander, searching for a place to take you in and finding none. This prayer isn't for you personally, but for anyone who doesn't feel the suffering of the Syrians.	وجود اللاجئين والمهاجرين في الأردن (من فلسطين ، سوريا ، العراق) Refugee/Immigrant Presence in Jordan from (Palestine, Syria and Iraq)	pro	against	92.76%

Continued on next page

Sentence	Topic	Ground Truth	Prediction	Confidence
احترام اليهود شي جيد المفروض نحترم جميع الأديان السماوية و المذاهب اما موضوع التطبيع مع اسرائيل مرفوض Respecting Jews is a good thing; we should respect all divinely revealed religions and sects. However, the issue of normalization with Israel is rejected.	التطبيع مع اسرائيل Normalization with Israel	against	pro	89.32%
التطبيع الحقيقي يبدأ بالتطبيع والسلام مع فلسطين والفلسطينيين، بما يمكنهم من العيش بكرامة على أرضهم ويزيل عن أذهاننا مشاهد الذل والإهانة والقمع التي تعرضوا لها على يد الجنود والمستوطنين الإسرائيليين لعقود، عندها سيكون هناك تطبيع مع الشعوب العربية وليس مع سلطات معزولة عن شعوبها True normalization begins with normalization and peace with Palestine and the Palestinians, enabling them to live with dignity on their land and erasing from our minds the scenes of humiliation, insult, and oppression they have endured at the hands of Israeli soldiers and settlers for decades. Only then will there be normalization with the Arab peoples, rather than with authorities isolated from their own populations.	التطبيع مع اسرائيل Normalization with Israel	against	pro	88.94%
@khaberni صاروا السوريين علاقة تعلقوا عليها سرقاتكم ورفاهيتكم ورواتبكم اللي ما بتستحقوها والله عيب الاردن بعاني لانوا في حرامية كثير وسرقات ومحسوبيات Syrians became an excuse to hang your thefts and your unearned salaries on.	وجود اللاجئين والمهاجرين في الأردن (من فلسطين ، سوريا ، العراق) Refugee/Immigrant Presence in Jordan from (Palestine, Syria and Iraq)	pro	against	88.53%
المهم. جاسم متي تزور اسرائيل. ان شاء الله. اخبارك مع التطبيع. امورك ماشيه. والدعم. مستمر Anyway Jassim, when will you visit Israel? How is normalization going for you?	التطبيع مع اسرائيل Normalization with Israel	neutral	pro	87.41%
بخصوص التطبيع هذي اسطوانه الغرب ومرتزقتها اتمنى من المحبين لبلاد الحرمين الشريفين عدم التطرق لها في التقرير Regarding normalization, this is a broken record of the West and its mercenaries.	التطبيع مع اسرائيل Normalization with Israel	neutral	against	87.37%
لست مع اعادتهم ليصبحوا ضحية لبشار الأسد وللحرب I am not for returning them only for them to become victims of Bashar al-Assad and the war.	وجود اللاجئين والمهاجرين في الأردن (من فلسطين ، سوريا ، العراق) Refugee/Immigrant Presence in Jordan from (Palestine, Syria and Iraq)	pro	against	86.22%

Continued on next page

Sentence	Topic	Ground Truth	Prediction	Confidence
خلوكم بس ليش ايمال بتسبونا Just stay, but why do you always insult us?	وجود اللاجئين والمهاجرين في الأردن (من فلسطين ، سوريا ، العراق) Refugee/Immigrant Presence in Jordan from (Palestine, Syria and Iraq)	neutral	against	93.58%

Table A.6: Representative high-confidence misclassified examples from out-of-fold predictions, illustrating systematic stance ambiguities.

Ketaba-OCR at AR-MS NakbaNLP 2026: Efficient Adaptation of Vision-Language Models for Handwritten Recognition

Hassan Barmandah[1,2], **Fatimah Emad Eldin**[1,3], **Khloud Al Jallad**[1,4], **Omer Nacar**[1,5]
[1] NAMAA Community, [2] Umm Al-Qura University, [3] Trouve Labs
[4] Syrian Society for Startups and Research, [5] Tuwaiq Academy
s445001043@uqu.edu.sa, Fatimah@trouve.works,
khloud.aljallad@syssr.org, o.najar@tuwaiq.edu.sa

Abstract

This paper presents Ketaba-OCR-LoRA, a system developed for the NakbaNLP 2026 Shared Task on Arabic Manuscript Understanding (Subtask 2), which targets the transcription of the historically significant Omar Al-Saleh Memoir Collection written in Ruq'ah and Naskh scripts. We propose a parameter-efficient adaptation of a publicly available pretrained Arabic-English Handwritten Text Recognition (HRT) model, originally trained on handwritten corpora including the Muharaf dataset. Instead of adapting general Vision-Language Models from scratch, we fine-tune the HRT backbone using Low-Rank Adaptation (LoRA) and 4-bit quantization (QLoRA), reducing memory requirements from 40GB to approximately 8GB. Our final submission combines multiple model variants through a novel Linear+Boost weighted ensemble strategy. Our approach achieves a CER of 0.0819 and WER of 0.2588 on the blind test set (per-line evaluation), ranking **1st** on per-line evaluation; on the official corpus-wide leaderboard, we rank **3rd** (CER 0.0938, WER 0.2996). This work demonstrates that specialized pretrained HRT models substantially outperform general-purpose Vision-Language Models for Arabic manuscript transcription, and that parameter-efficient fine-tuning provides a practical and reproducible approach for low-resource cultural heritage digitization.

Keywords: Arabic HRT, OCR, LoRA, Fine-tuning, Handwritten Text Recognition, Ensemble

1. Introduction

The NakbaNLP 2026 Shared Task (Subtask 2) addresses a critical challenge in Digital Humanities: automating the transcription of historical Arabic manuscripts. The task evaluates systems on the Omar Al-Saleh Memoir Collection (1951–1965), which comprises over 1.5 million words of culturally significant Palestinian testimonies (Hamoud et al., 2026). The manuscripts are written in historical Ruq'ah and Naskh scripts, for which robust OCR is essential for scholars and cultural preservation.

Rather than training a general-purpose Vision-Language Model (VLM) from scratch, our core strategy relies on parameter-efficient transfer learning. We utilize a specialized Handwritten Text Recognition (HRT) backbone (Sherif, 2025) pretrained on diverse datasets like Muharaf (Aladmani et al., 2024). To adapt this foundation to the specific calligraphy of the memoirs without prohibitive computational costs, we fine-tune the model using Low-Rank Adaptation with 4-bit quantization (QLoRA) (Hu et al., 2022; Dettmers et al., 2023). Our experiments yielded the following key outcomes:

- **Ranking & Performance:** Our system ranks **1st** on per-line evaluation (CER 0.082, WER 0.259) and **3rd** on the official corpus-wide leaderboard (CER 0.0938, WER 0.2996).
- **HRT vs. Generalist VLMs:** Specialized, Ketaba-OCR-LoRA models drastically outperform zero-shot generalist VLMs, which struggle with historical Arabic handwriting.
- **Parameter Efficiency:** QLoRA efficiently bridged the domain gap, reducing CER from 0.58 to 0.08 with minimal computational overhead.
- **Ensemble Innovation:** Our Linear+Boost weighting strategy improved CER by 7.4% over standard inverse-CER weighting.

Our model weights and source code are publicly available.[1]

2. Background and Related Work

Arabic Handwritten Text Recognition (HRT) is inherently complex due to its cursive script, context-dependent allographs, and diacritics (Wasfy et al., 2025). While early encoder-decoder models like TrOCR (Li et al., 2023) struggle with degraded historical manuscripts, recent state-of-the-art methods rely on synthetic data to overcome resource scarcity. For instance, Wasfy et al. (2025) introduced *Qari-OCR* using diverse synthetic fonts, and Hennara et al. (2025) developed *Baseer* (Hennara et al., 2025) to advance low-resource Arabic document understanding. Concurrently, Large Multimodal Models (LMMs) like Qwen-VL (Bai et al., 2023) offer powerful end-to-end vision-language generation. Multilingual models like GLM-OCR (Zai-org, 2026)

[1] Model weights and code: https://huggingface.co/HassanB4/Ketaba-OCR-LoRA

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 160–170
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

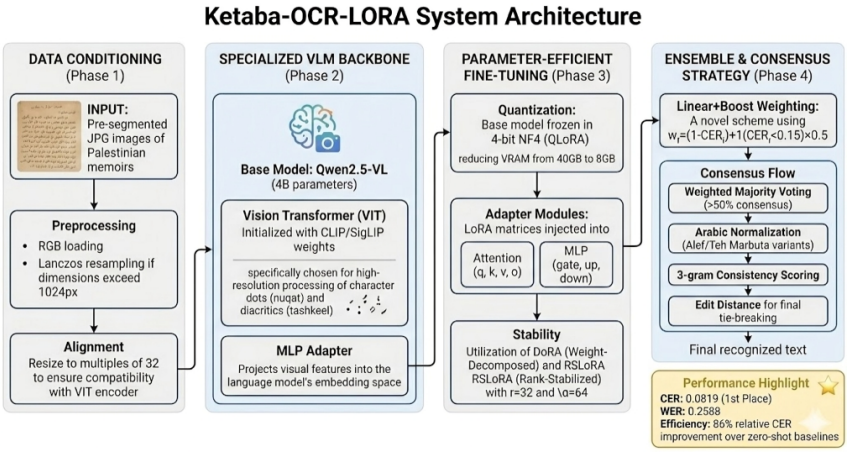


Figure 1: Overview of the Ketaba-OCR-LoRA pipeline.

also exist. However, without domain adaptation, these generalist models often hallucinate modern orthography when processing ambiguous historical texts.

2.1. Task Setup and Data

We participate in the NakbaNLP 2026 Shared Task (Subtask 2: Automatic Manuscript OCR) (Hamoud et al., 2026), which focuses on transcribing the Omar Al-Saleh Memoir Collection (1951–1965) in historical Ruq'ah and Naskh script. The input consists of pre-segmented JPG line images, and the target output is a character-level Unicode transcription that preserves orthography and diacritics (tashkeel). The shared-task dataset includes approximately 15,962 training samples, 1,774 development samples, and 2,095 test samples. In this paper, the 2,095 figure refers to the test split used during development, while the official shared-task ranking was computed on a final blind CodaBench evaluation with 2,671 hidden test images. Evaluation is based on Character Error Rate (CER) and Word Error Rate (WER).

Our submission is fine-tuned only on the official shared-task data. The Muharaf Dataset (Aladmani et al., 2024), the IAM Handwriting Database (Marti and Bunke, 2002), and other handwritten corpora mentioned with the Arabic-English HRT backbone (Sherif, 2025) were used only for backbone pretraining.

3. System Overview

Our proposed system, Ketaba-OCR-LoRA, is built upon Arabic-English HRT model based on Qwen2.5-VL (Sherif, 2025). The overall system design is illustrated in Figure 1. This model was selected for its high-resolution visual encoder, essential for capturing Arabic diacritics (tashkeel) and letter dots (nuqat).

3.1. Data Preprocessing

We preprocess manuscript line images before feeding them to the model. Each image is loaded in RGB; if either dimension exceeds 1024 pixels, we downscale it so that the longer side is capped at 1024 using Lanczos resampling. We then resize the image so that both dimensions are multiples of 32, as required by the vision encoder.

No additional augmentation or normalization is applied; transcriptions are used as provided.

3.2. Base Model: Arabic-English HRT based on Qwen2.5-VL

This 4B-parameter model (Bai et al., 2023) was fine-tuned on handwritten datasets including the Muharaf Dataset (Aladmani et al., 2024) and IAM Handwriting Database (Marti and Bunke, 2002). Its Vision Transformer encoder and MLP adapter provide robust priors for Arabic and Latin handwritten text recognition.

3.3. Low-Rank Adaptation (LoRA)

To efficiently adapt the pretrained HRT model to the Al-Saleh manuscript style without catastrophic forgetting, we employ LoRA (Hu et al., 2022). Instead of updating all model parameters W, we decompose the update ΔW into two low-rank matrices:

$$W' = W + \Delta W = W + BA \quad (1)$$

where W is frozen in 4-bit precision (NF4 format, QLoRA) (Dettmers et al., 2023). We use rank $r = 32$ with scaling factor $\alpha = 64$, along with DoRA (Weight-Decomposed Low-Rank Adaptation) (Liu et al., 2024) and RSLoRA (Rank-Stabilized LoRA) (Kalajdzievski, 2023) for improved training stability.

System / Configuration	Test Set CER ↓	Test Set WER ↓	Blind Test Set CER ↓	Blind Test Set WER ↓
1. Baselines				
Organizer Baseline	0.5840	0.8810	0.5910	0.8850
2. Fine-Tuned Models (Non-Merged LoRA)†				
Ketaba-OCR-LoRA:Qwen2.5-VL†	0.0810	0.1150	0.0884	0.2699
Ketaba-OCR-LoRA + Ensemble (Ours)⋆	–	–	**0.0819**	**0.2588**
3. Fine-Tuned Models (with Optimizations)§				
Ketaba-OCR-LoRA:Qwen2.5-VL Improvement1§	–	–	0.1133	0.3104
Fine-Tuned QARI-3	0.2782	0.5814	0.2635	0.5521
4. Other / External Models				
Arabic OCR 4bit Qwen2.5-VL-3B-v2	–	–	0.3234	0.6203

Table 1: Comparative results for fine-tuned models and external systems. Our row reports **per-line** CER/WER (1st on per-line); official corpus-wide ranking is 3rd (Table 2). Additional baseline experiments are in Table 6 (Appendix G).

LoRA adapters are injected into attention projections (q, k, v, o) and MLP layers (gate, up, down projections). We optimize with AdamW (Loshchilov and Hutter, 2019) using the Hugging Face Transformers library (Wolf et al., 2020).

4. Experimental Setup

4.1. Baseline Models

We evaluated two primary approaches:

1. **Zero-Shot Inference:** We tested inference without fine-tuning using Sherif's pretrained Arabic-English HRT model (Sherif, 2025) (`sherif1313/Arabic-English-handwritten-OCR-v3`[2]) based on Qwen2.5-VL (Bai et al., 2023). For comparison, we also evaluated general-purpose Vision-Language Models including Qwen2.5-VL directly (Bai et al., 2023), QARI-OCR (Wasfy et al., 2025), and GLM-OCR (Zai-org, 2026). A comprehensive list of all baseline models tested is provided in Table 7 (Appendix G.1).
2. **LoRA Fine-Tuning (Ours):** Fine-tuning Sherif's HRT model using the hyperparameters detailed in Appendix D. We used `peft` (Mangrulkar et al., 2022) for adapter injection and `bitsandbytes` (Dettmers et al., 2022) for quantization.

4.2. Evaluation Metrics

The main evaluation metrics are Character Error Rate (CER) and Word Error Rate (WER), detailed in Appendix A.

4.3. Ensemble Strategy

Our final submission employs a weighted ensemble combining predictions from six model configurations. After testing 30 weighting strategies, we found that a **Linear+Boost** scheme outperformed traditional inverse-CER weighting.

Linear+Boost Weighting: Instead of $w_i = \frac{1}{\text{CER}_i}$, we use:

$$w_i = (1 - \text{CER}_i) + \mathbf{1}[\text{CER}_i < 0.15] \times 0.5 \quad (2)$$

This applies a linear decay based on CER, plus a bonus of 0.5 for models with CER below 0.15. This gives weights [0.24, 0.24, 0.14, 0.14, 0.13, 0.12] for our six models, providing a more balanced distribution than inverse-CER while still favoring top performers.

The full ensemble algorithm—including weighted majority voting, Arabic normalization, n-gram consistency, and edit distance consensus—is detailed in Appendix E.

5. Results

5.1. Official Leaderboard and Per-Line Results

The shared task reports two evaluation schemes: **corpus-wide** (entire test set as a single sequence), used for the official ranking, and **per-line** (CER/WER computed per example then averaged). Our team ranks **3rd** on the official (corpus-wide) leaderboard (CER 0.0938, WER 0.2996) and **1st** on per-line evaluation (CER 0.0819, WER 0.2588; Table 1). The full leaderboards (corpus-wide and per-line) are in Appendix B (Table 2 and Table 3).

[2] https://huggingface.co/sherif1313/Arabic-English-handwritten-OCR-v3

As shown in Table 1, our fine-tuned system significantly outperforms all baselines, confirming the superiority of specialized HRT adaptation. Additional evaluations are in Appendix G.

5.2. Error Analysis

Detailed error analysis, including visual comparisons and architectural behavior across varying sequence lengths, is in Appendix F.

5.3. Ablation Study

To evaluate the impact of our parameter-efficient fine-tuning, we compared general-purpose VLMs against specialized Arabic OCR systems (Table 8).

Our zero-shot inference study (Appendix C) highlighted a significant domain gap (0.2032 CER), confirming the necessity of domain-specific adaptation. Base VLMs yielded over 40,000 substitutions, while MSA-centric models (Qari series) suffered catastrophic over-segmentation ($>$200,000 insertions). Our LoRA adaptation reduced these errors by an order of magnitude (Table 8).

5.3.1. The Role of Ensembling

Applying our Linear+Boost ensemble strategy provided a critical final performance boost. It improved the blind test CER from 0.0884 (achieved via standard inverse-CER weighting) to 0.0819, representing a 7.4% relative improvement by effectively bypassing the generative hallucinations of individual models. More details on the discrepancy analysis of the blind test set, and how ensembling actively corrects catastrophic looping and contextual drift, are provided in Appendix H (specifically Table 9 and Table 10).

6. Discussion

Specialized pretrained HRT models substantially outperform general-purpose VLMs for Arabic manuscripts, as they explicitly encode morphological priors and possess invariances to noise, fading, and ink bleeding—properties absent in models trained on clean internet images. QLoRA provides a practical adaptation paradigm, achieving 86% relative CER improvement (0.58 to 0.08) with minimal overhead (8GB vs. 40GB). Our Linear+Boost ensemble strategy, outperforming 29 alternative configurations including inverse-CER and exponential decay, demonstrates the importance of balanced weighting over extreme disparities.

Limitations include system specialization to the Al-Saleh memoirs with unexplored generalization to other historical periods, and disproportionately high WER (0.26) versus CER (0.08) due to Arabic agglutination complexity. Future work should address morphologically-aware postprocessing and constrained decoding against historical dictionaries.

7. Conclusion

We presented Ketaba-OCR, which ranks 1st on per-line evaluation (CER 0.0819, WER 0.2588) and 3rd on the official corpus-wide leaderboard (CER 0.0938, WER 0.2996) at NakbaNLP 2026, by fine-tuning a specialized pretrained HRT model via QLoRA and combining predictions through our novel Linear+Boost ensemble strategy. Our key contributions are: (1) demonstrating that parameter-efficient adaptation of domain-specific HRT models substantially outperforms generalist VLMs for historical Arabic manuscripts; (2) introducing a Linear+Boost weighting scheme that outperforms traditional inverse-CER ensemble methods by 7.4%; and (3) providing a systematic comparison of 30 ensemble configurations.

While our open-source system advances digital humanities—specifically enabling the preservation of historical Palestinian narratives—it remains optimized for the Al-Saleh memoirs. Consequently, it requires further adaptation for other calligraphic styles, and historians should validate transcriptions prior to scholarly use. Ultimately, this work provides a reproducible, computationally efficient blueprint for digitizing low-resource cultural heritage texts.

Ethics Statement

This work supports the digitization of historical Arabic manuscripts and the accessibility of culturally significant Palestinian testimonies (Omar Al-Saleh Memoir Collection), with potential benefit for scholars and cultural preservation. The shared task data consist of historical testimonies; consent and curation are the responsibility of the task organizers, and we use the data only as provided. Automated transcriptions may contain errors; we recommend validation by domain experts before scholarly or public use. Our model weights and code are released to support reproducibility (see footnotes).

Acknowledgements

We thank the NakbaNLP 2026 organizers for access to the Omar Al-Saleh Memoir Collection. We acknowledge the Arabic-English HRT model based on Qwen2.5-VL (Sherif, 2025), and the Hugging Face community for PEFT and bitsandbytes libraries.

References

Mehreen Aladmani, Erin Woertz, David Juen, Aashish Agarwal, Stephen Houdek, Gerik Jager, Razan Abuaita, and Nizar Habash. 2024. Muharaf: Manuscripts of handwritten Arabic dataset for cursive text recognition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6651–6661, Torino, Italy.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. 8-bit optimizers via block-wise quantization. In *International Conference on Learning Representations*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36.

Hadi Hamoud, Ahmad Ali Chamseddine, Bilal Shalash, Firas Ben Abid, Mustafa Jarrar, Chadi Abou Chakra, Bernard Ghanem, and Fadi A. Zaraket. 2026. NAKBA NLP 2026: Shared Task on Arabic Handwritten Manuscript Understanding (Palestine Memory–Omar Al-Saleh Memoir). In *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

Khalil Hennara, Muhammad Hreden, Mohamed Motasim Hamed, Ahmad Bastati, Zeina Aldallal, Sara Chrouf, and Safwan AlModhayan. 2025. Baseer: A vision-language model for arabic document-to-markdown OCR. *arXiv preprint arXiv:2509.18174*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Damjan Kalajdzievski. 2023. A rank stabilization scaling factor for fine-tuning with lora. *arXiv preprint arXiv:2312.03732*.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13094–13102.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. DoRA: Weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 32100–32121. PMLR.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

Urs-Viktor Marti and Horst Bunke. 2002. The IAM-database: An English sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5(1):39–46.

Ahmed Sherif. 2025. Arabic-english handwritten ocr v3. https://huggingface.co/sherif1313/Arabic-English-handwritten-OCR-v3. Vision-Language Transformer model fine-tuned for Arabic and English handwritten text recognition. Apache-2.0 License. Accessed: 2026-01-10.

Ahmed Wasfy, Omer Nacar, Abdelakreem Elkhateb, Mahmoud Reda, Omar Elshehy, Adel Ammar, and Wadii Boulila. 2025. Qari-ocr: High-fidelity arabic text recognition through multimodal large language model adaptation. *arXiv preprint arXiv:2506.02295*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Zai-org. 2026. Glm-ocr: A multimodal large language model for ocr and document understanding. https://github.com/zai-org/GLM-OCR. Accessed: 2026-02-19.

A. Appendix: Evaluation Metrics

$$\text{CER} = \frac{\text{Levenshtein}(\text{Pred}, \text{Ref})}{\text{Length}(\text{Ref})} \tag{3}$$

Word Error Rate (WER) assesses transcription usability:

$$\text{WER} = \frac{S + D + I}{N} \tag{4}$$

where S, D, I are substitutions, deletions, insertions, and N is reference word count.

B. Appendix: Official Leaderboards (Corpus-Wide and Per-Line)

The shared task reports both corpus-wide and per-line CER/WER. Table 2 gives the official (corpus-wide) leaderboard; Table 3 gives the per-line leaderboard (CER/WER computed per example and averaged). Ketaba-OCR (ours) ranks 3rd on corpus-wide and 1st on per-line (Table 1).

C. Appendix: Zero-Shot Ablation Study

D. Appendix: Hyperparameters and Model Details

The hyperparameters and model configuration used for the LoRA fine-tuning experiment are listed in Table 5. These were selected based on extensive hyperparameter search on the validation set, prioritizing Character Error Rate minimization.

D.1. Key Configuration Justifications

Choice of Base Model: Sherif1313's Arabic-English handwritten OCR model was selected because:

1. It is pretrained on Arabic and English handwritten text from multiple sources (Muharaf, IAM, proprietary collections)
2. It exhibits stronger robustness to manuscript-specific degradations compared to general VLMs
3. With 4-bit quantization, the model fits in $\sim$8GB VRAM despite having 4B+ parameters, enabling efficient fine-tuning on consumer hardware

LoRA Rank Selection: Ablation studies indicated that rank 32 with DoRA and RSLoRA provides the optimal trade-off between model capacity and computational efficiency. Ranks below 16 underfit; higher ranks show diminishing returns in CER improvement.

Quantization Strategy: 4-bit NF4 quantization via QLoRA reduces memory usage from $\sim$40GB (full precision) to $\sim$8GB, enabling training on consumer-grade GPUs while preserving model quality.

E. Appendix: Ensemble Method Details

Our final winning submission employed our Linear+Boost weighted ensemble strategy. The ensemble achieved CER 0.0819 and WER 0.2588 on the blind test set.

Models Combined with Linear+Boost Weights:

- Ketaba-OCR-LoRA: CER 0.09, weight 0.24
- Ketaba-OCR-LoRA variant): CER 0.11, weight 0.24
- Zero-shot HRT with LoRA: CER 0.18, weight 0.14
- Zero-shot HRT baseline: CER 0.20, weight 0.14
- Fine-tuned QARI: CER 0.26, weight 0.13
- Arabic OCR 4-bit: CER 0.32, weight 0.12

Weight Computation (Linear+Boost):

$$w_i = \text{normalize}\left((1 - \text{CER}_i) + \mathbf{1}[\text{CER}_i < 0.15] \times 0.5\right) \tag{5}$$

Why Linear+Boost Outperforms Inverse-CER:

- Inverse-CER ($w = 1/\text{CER}$) creates extreme weight disparities (best model gets 3.5$\times$ weight of worst)
- Linear decay provides smoother distribution, allowing weaker models to contribute diversity
- The boost term (+0.5 for CER < 0.15) ensures top 2 performers have sufficient influence
- Combined effect: better balance between trusting top models and leveraging ensemble diversity

Ensemble Algorithm:

1. Apply Linear+Boost weights to all model predictions

Rank	Team	CER ↓	WER ↓
1	Misraj Ai	0.079	0.244
2	Oblevit	0.0925	0.3268
3	**Ketaba-OCR (Ours)**	**0.0938**	**0.2996**
4	Latent Narratives	0.105	0.3106
5	Al-Warraq	0.1142	0.378
6	Not Gemma	0.1217	0.3063
7	NAMAA-Qari	0.195	0.5194
8	Fahras	0.2269	0.5223
9	baseline	0.3683	0.6905

Table 2: Official leaderboard (corpus-wide evaluation; used for ranking). All participating teams and organizer baseline. Ketaba-OCR (ours) ranks 3rd.

Rank	Team	CER ↓	WER ↓
1	**Ketaba-OCR (Ours)**	**0.0819**	**0.2588**
2	Misraj Ai	0.0895	0.2516
3	Latent Narratives	0.1002	0.2845
4	Oblevit	0.1049	0.3316
5	Al-Warraq	0.1059	0.3468
6	Not Gemma	0.11	0.3126
7	Fahras	0.1819	0.4303
8	NAMAA-Qari	0.2032	0.5031
9	baseline	0.2811	0.5884

Table 3: Per-line leaderboard (CER/WER computed per example, then averaged). Ketaba-OCR (ours) ranks 1st (Table 1).

2. If weighted consensus $>$50% for any prediction, select it
3. Normalize Arabic text (alef variants, teh marbuta) and re-vote
4. Score remaining candidates by 3-gram consistency with other models
5. Final tie-breaking: minimum average edit distance to all candidates

Ablation of Weighting Strategies (30 configs tested): Config 18 (Linear+Boost$<$0.15) achieved the best blind test CER of 0.0819, outperforming inverse-CER (0.0884), exponential decay (0.0856), and rank-based methods (0.0871).

F. Appendix: Error Analysis Details

F.1. Qualitative Error Analysis

Figure 2 presents representative failure cases highlighting the differences between the zero-shot baseline, QARI-3, and our proposed Ketaba-OCR-LoRA model. In the first example, the ground truth word is العنوان. The zero shot model produces an incomplete prediction لعنو, omitting both the initial definite article ال and the final character indicating insufficient adaptation to Arabic morphological structure. In contrast, both QARI-3 and our proposed method successfully recover the complete word. The second example illustrates a more challenging scenario, where a horizontal artifact partially occludes the upper region of the text. The zero-shot model again fails to reconstruct the word correctly, while QARI-3 produces a near-correct output. Our proposed model accurately recovers the full word despite visual corruption, indicating improved robustness to noise and structural distortions.

Moreover, we expanded the error analysis for long sentences with punctuation. For instance, Figure 3 presents a challenging long-text example containing multiple words, punctuation, and complex Arabic morphology. Such samples require accurate character recognition, proper word boundary modeling, and contextual coherence. The zero-shot baseline shows character-level distortions and incomplete words, while QARI-3 partially improves structural consistency but remains sensitive to noise. In contrast, our proposed model, Ketaba-OCR-LoRA, accurately reconstructs the full sentence, demonstrating enhanced robustness and superior context-aware modeling for long text sequences.

F.2. Quantitative Error Analysis by Sequence Length

The comparative analysis of Character Error Rate (CER) against sequence length shows distinct ar-

Model	Test Set		Blind Test Set	
	CER ↓	WER ↓	CER ↓	WER ↓
HRT: Zero-Shot Qwen2.5-VL	0.1690	0.4987	0.2032	0.5031
HRT: Zero-Shot Qwen2.5-VL (LoRA)[‡]	–	–	0.1834	0.4644

Table 4: Zero-shot inference ablation study. [‡]Merged adapters. These results demonstrate the baseline performance of the pretrained HRT model without fine-tuning, highlighting the domain gap for unseen historical manuscripts.

Parameter	Value	Parameter	Value
Base Model Configuration			
Base Model Name	Arabic-English-handwritten-OCR-v3	Model Architecture	Vision-Language Transformer
Base Model Size	∼4.07B parameters	Pretraining Data	Muharaf, IAM, Custom
Quantization Configuration			
Quantization Scheme	4-bit NF4 (QLoRA)	Compute Dtype	bfloat16
Double Quant	True	–	–
LoRA Adapter Configuration			
LoRA Rank (r)	32	LoRA Alpha (α)	64
Target Modules	q, k, v, o, gate, up, down	LoRA Dropout	0.05
DoRA	True	RSLoRA	True
Training Hyperparameters			
Learning Rate	2×10^{-5}	Optimizer	AdamW (fused)
LR Scheduler	Cosine	Warmup Steps	200
Batch Size	1 (per GPU)	Gradient Accumulation	4
Effective Batch Size	4	Number of Epochs	1
Max Gradient Norm	1.0	Weight Decay	0.01
Max Sequence Length	2048	Max Image Size	1024
Evaluation Strategy	steps	Eval Steps	500
Save Steps	500	Files per Chunk	1000

Table 5: Comprehensive hyperparameters for QLoRA fine-tuning of Ketaba-OCR-LoRA. The base model (Arabic-English HRT) was pretrained on diverse handwritten datasets.

chitectural behaviors among the evaluated models (see Figure 4). Ketaba-OCR-LoRA demonstrates stability, maintaining a near-zero CER regardless of text length, indicating robust handling of long-range spatial dependencies. In contrast, the Zero-Shot model demonstrates a notable short-text penalty, where error rates are highest for sequences under 20 characters and improve as length increases, indicating heavy reliance on surrounding linguistic context to disambiguate characters, a luxury not afforded by shorter snippets. Conversely, QARI-3 shows a classic performance decay, with CER scaling positively with sequence length before plateauing between 0.15 and 0.20. This degradation in QARI-3 is likely caused by attention saturation or error accumulation in its recurrent components. Overall, while Ketaba-OCR-LoRA is the most versatile, the results highlight a critical trade-off between contextual reliance in Zero-Shot models and the scaling limitations of the QARI-3 architecture.

F.3. Model Comparison Analysis

A comparative performance analysis of various models based on Character Error Rate (CER) and Word Error Rate (WER) is presented in Table 6 and Table 8. Qwen2.5-3B and Qwen2.5-7B models significantly outperform the remaining models, maintaining both CER and WER below the 1.0 threshold. Interestingly, while the Qwen series generally shows lower error rates, the Qwen2.5-VL-4B model shows a notable spike in error metrics compared to its smaller counterparts, indicating that model scaling or architectural changes in that specific configuration may have caused degradation.

Among the Qari models, performance is more volatile; Qari-0.3 serves as the strongest baseline in its group, while Qari-OCR-2B and Qari-0.1 show the highest error rates. A key observation is the relationship between character and word accuracy: in most models, WER tracks closely with CER, but Qari-0.1 and Qwen2.5-VL-4B show wider divergence, potentially indicating issues with linguistic consistency or spacing despite relatively better character recognition. Overall, the Qwen2.5 archi-



Figure 2: Qualitative Samples

tecture remains the most robust for this specific task, providing the highest levels of transcription accuracy.

G. Appendix: Additional Experiments (Test Set Only)

G.1. Baseline Models

Model / Configuration	CER ↓	WER ↓
Qwen2.5-VL-7B	0.6808	0.9198
Qwen2.5-VL-3B	0.6213	0.8628
QARI v0.3	0.5293	0.8772
QARI v0.1	0.7127	0.9296
QARI-OCR 2B	0.6840	0.9126
GLM-OCR	0.9999	0.9999

Table 6: Additional baseline model experiments evaluated on the test set only (blind test set results unavailable). These include general-purpose Vision-Language Models (Qwen2.5-VL), domain-specific Arabic OCR systems (QARI variants), and multilingual models (GLM-OCR).

The high insertion counts for Qari models (>100,000) indicate aggressive over-segmentation, where stylistic ligatures are misinterpreted as word boundaries. Our LoRA-adapted system eliminates these structural hallucinations.

Model / Configuration	Type
Arabic-English HRT (v3)	Specialized HRT
Qwen2.5-VL-7B	General VLM
Qwen2.5-VL-3B	General VLM
QARI v0.3	Arabic OCR
QARI v0.1	Arabic OCR
QARI-OCR 2B	Arabic OCR
GLM-OCR	Multilingual LMM
Arabic OCR 4-bit Qwen2.5-VL-3B-v2	Quantized VLM

Table 7: Comprehensive list of baseline models evaluated. Specialized (HRT) and general-purpose (VLM) models were compared to assess domain adaptation benefits. Detailed performance metrics are provided in Table 1 and Table 4.

H. Appendix: Blind Set Discrepancy and Hallucination Analysis

Since official ground truth annotations were not released for the blind test set, we constructed a *Pseudo-Ground Truth* (Pseudo-GT) from the ensemble's final predictions via weighted majority voting (Linear+Boost, Section 4.3). We consider this a reliable proxy: our ensemble achieved a corpus-wide CER of 0.0938 on the blind test set, meaning the predictions deviate from the true references by less than 10% at the character level. The resulting Pseudo-GT thus closely approximates the actual ground truth, making it suitable for qualitative failure analysis across individual models.

This analysis reveals systematic failure modes in individual models—specifically generative hallucination loops and contextual drift—and demonstrates the corrective power of ensemble consensus.

H.1. Catastrophic Hallucinations (Total Disagreement)

Table 9 illustrates the most severe failure modes observed in individual models, specifically generative repetition loops. When the autoregressive decoder loses attention alignment, it collapses into infinite lexical looping (e.g., repeating the word وجنرالهم) or numerical hallucinations (e.g., looping ١١٨). Our ensemble strategy successfully mitigates these errors, relying on the robust consensus of the Linear+Boost weighting to extract coherent sequences.

H.2. Ensemble Correction in Lower Disagreement

Table 10 demonstrates cases of "Lower Disagreement," where two models agree and one dissents. Here, the standalone HRT_LoRA_Merged model



Figure 3: Qualitative Sample 3

Table 8: Detailed Error Metrics for Baseline Models and Variants

Model	CER	WER	Subs.	Dels.	Ins.
Qwen2.5-3B	0.7868	0.9625	42,215	9,147	35,724
Qwen2.5-7B	0.8052	0.9922	46,953	15,977	21,065
Qwen2.5-VL-4B*	3.7699	3.0734	56,407	30,202	48,983
Qari-0.3	2.6551	1.8641	49,762	3,053	145,225
Qari-0.1	4.5836	2.7653	49,612	15,009	111,262
Qari-OCR-2B	4.2499	4.5788	45,142	7,381	223,041

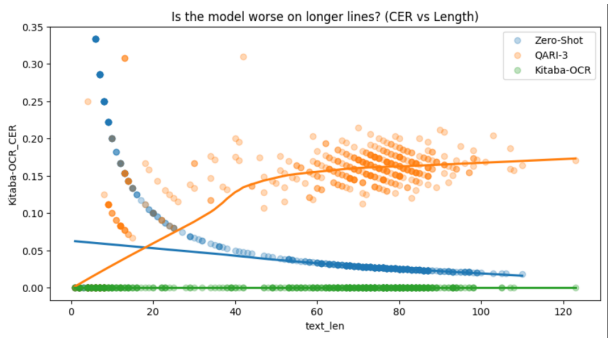


Figure 4: CER vs Sequence Length

occasionally suffers from premature truncation and contextual drift, hallucinating shorter, semantically unrelated phrases. The ensemble leverages consensus to output the structurally accurate transcription.

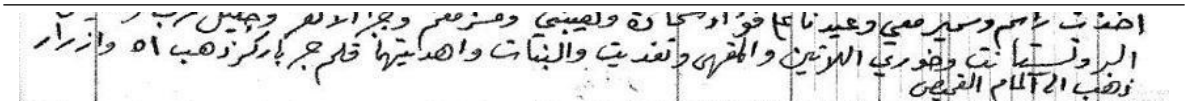
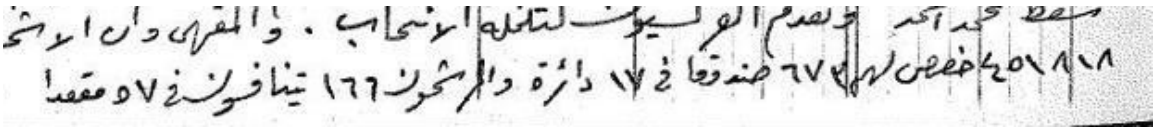
Model	Prediction (Truncated)
	
Image ID: `1962_p064_10035`	
HRT LoRA Merged	أجداد سَرم وسَمير معي وبعدنا عاقو…
Ketaba-OCR-LoRA_1	ذهب إلى إكمال الغربة وسمر معي… وجنرالهم وجنرالهم… *[Infinite Loop]*
Ensemble	اخذت راسم وسمير معني وعيدنا على فؤاد…
	
Image ID: `1962_p176_10069`	
HRT LoRA Merged	٨٨١ ٣٥،٤ عدد واحد وعدم الموسيقى…
Ketaba-OCR-LoRA	٨١ ٨١ ٨١ ٨١ ٨١ ٨١ ٨١ ٨١ ٨١ ٨١… *[Numerical Loop]*
Ensemble	٨١ ١٥٤ خصوص لهم ٣٧٦ صندوقا في…

Table 9: Examples of severe generative hallucinations in single models compared to the stable ensemble output. The autoregressive loops in Submission_1 are completely bypassed by the ensemble consensus.

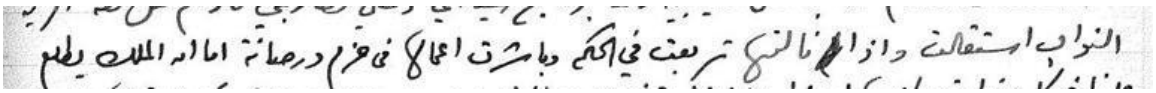
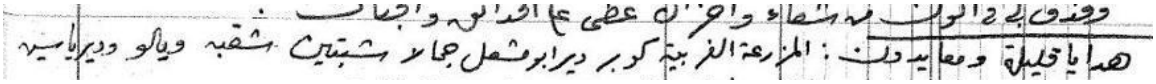
Source	Text Sequence (Truncated)
	
Image ID: `1956_p080_10012`	
Pseudo-GT	النواب استقالت وأوامرنا لنها تربعت…
HRT_LoRA_Merged	النواب استقالت واذا فانها تربعت في… *[Drift]*
Ketaba-OCR-LoRA	النواب استقالت وأوامرنا لنها تربعت…
Ensemble	النواب استقالت وأوامرنا لنها تربعت…
	
Image ID: `1962_p077_10032`	
Pseudo-GT	وقد قبى وألوس من سكاء واحزن…
HRT_LoRA_Merged	وفوق في والونك من سطاء واخزان… *[Truncation]*
Submission_1	وقد قبى وألوس من سكاء واحزن…
Ensemble (Ours)	وقد قبى وألوس من سكاء واحزن…

Table 10: Correction of contextual drift and truncation. When individual models fail to capture the full manuscript line, the ensemble defaults to the more complete, text-anchored sequence.

No Overfit at NakbaArchiveClassifier Shared Task: A Swin Transformer-Based System for Destruction Image Classification

Mohamed F. Mohamed[1,2], Samar M. Abd El-Mageed[1,2], Ensaf H. Mohamed[1,3]
[1]Computer Science Department, Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt
[2]Computer Science Department, Faculty of Computers and Information, Luxor University, Luxor, Egypt
[3]School of Information Technology and Computer Science, Nile University, Giza, Egypt
mfathy@fci.luxor.edu.eg, samar.mahmoud@fci.luxor.edu.eg, enmohamed@nu.edu.eg

Abstract

Automated destruction identification from visual data plays a critical role in large-scale documentation, humanitarian analysis, and digital archiving of conflict-related events. Within this context, the Nakba-NLP 2026 Workshop introduced a shared task aimed at training and evaluating a binary image classification model to distinguish between destroyed or damaged infrastructure and intact infrastructure. However, the limited dataset size and the visual variability of real-world scenes make this task particularly challenging. This work presents a Swin Transformer–based framework tailored for destruction image classification. The proposed model employs a hierarchical Swin Transformer backbone for robust feature extraction, followed by a multi-layer perceptron classifier for decision-making. To address the limited data issue, transfer learning and a customized training strategy are applied to adapt the model effectively without full end-to-end retraining. Furthermore, a semi-supervised data expansion approach is utilized to enlarge the training set from 1,400 to 10,000 images, improving model generalization and robustness. Experimental results on the official blind test set demonstrate strong performance, achieving an F1-score of 86.55% and an accuracy of 87.81%, ranking 5th in the shared task.

Keywords: image classification, deep learning, Swin transformer, multi-layer perceptron network, transfer learning

1. Introduction

Visual content is now an important tool for documenting real-world events and preserving historical memory. Images are essential in humanitarian and conflict-related contexts because they help document infrastructure damage, facilitate large-scale analysis, and support documentation efforts. The NakbaArchiveClassifier (Abrahams et al., 2026) is a shared task that aims to automatically distinguish between destruction and non-destruction images under a realistic evaluation setting. The task presents several challenges. First, it is difficult to train deep neural networks from scratch due to the limited size of the available labeled dataset. Second, robust feature extraction is challenging due to the visual variability of scenes, which includes various viewpoints, lighting conditions, degrees of damage severity, and background clutter. Third, the evaluation is conducted on a blind test set through the CodaBench [1] platform, requiring models that generalize well to unseen data.

To address these challenges, this paper proposes a Swin Transformer–based classification framework with a training strategy tailored to the shared task environment. Instead of training from scratch, pretrained ImageNet-1K weights are utilized to leverage rich visual representations learned from large-scale data. To prevent overfitting and preserve pretrained knowledge, the backbone is frozen, and only a task-specific head block is trained. The head incorporates a Swin stage followed by a multi-layer perceptron (MLP) classifier. Furthermore, to mitigate dataset size limitations, a semi-supervised data expansion strategy is employed to increase the training set using pseudo-labeling and manual verification.

The main contributions of this work can be summarized as follows:

- A Swin Transformer–based model with an MLP classifier head and customized training strategy is proposed for destruction image classification.
- A semi-supervised data expansion approach is employed to increase the training dataset.
- Experimental results on the official test set demonstrate that the proposed classification system achieves strong and balanced performance, reaching an F1-score of 86.55% and an accuracy of 87.81%, and ranking 5th in the shared task [2].

[1] https://www.codabench.org/competitions/12654/#/pages-tab

[2] https://www.codabench.org/competitions/12654/#/results-tab

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 171–176
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

2. Background

Image classification is a fundamental task in computer vision that aims to assign a predefined label to an input image based on its visual content. Over the years, classification techniques have evolved significantly, progressing from traditional handcrafted feature-based approaches to deep learning-based methods that automatically learn feature representations. The emergence of deep learning, particularly convolutional neural networks (CNNs), revolutionized image classification. Architectures such as AlexNet(Krizhevsky et al., 2012), VGG(Simonyan and Zisserman, 2015), ResNet(He et al., 2016), and EfficientNet(Tan and Le, 2019) demonstrated the ability to learn hierarchical spatial features directly from raw pixel data.

Although CNN-based models have a strong capability to capture local patterns, they may struggle to effectively model global context due to their limited receptive fields. To address this limitation, transformer-based architectures have been introduced into vision tasks. Vision Transformers (ViT)(Dosovitskiy et al., 2021) and their variants employ self-attention mechanisms to model long-range dependencies across image patches. Among these variants, the Swin Transformer (Liu et al., 2021) has demonstrated significant improvements over the original ViT by employing a hierarchical structure with shifted window-based self-attention, enabling scalable and computationally efficient modeling of both local and global information.

Consequently, the Swin Transformer is utilized as the core component of the proposed framework. The Swin Transformer study provides multiple architectural variants that differ in size and computational complexity. In this paper, the tiny variant is selected, as it is more suitable for the dataset constraints.

3. Task Description and Dataset

The NakbaArchiveClassifier shared task (Abrahams et al., 2026) is part of the Nakba-NLP 2026 Workshop [3], itself part of LREC 2026 [4]. The task is to develop a binary image classification system that can distinguish between images of damaged or destroyed infrastructure and those of intact infrastructure.

The dataset consists of 2,001 Instagram images posted by Palestinian content creators and journalists in Gaza for documenting events since October 7, 2023. These images are labeled into two classes (destruction or non-destruction) and are divided into training, validation, and testing splits. The details of these distributions are summarized in Table 1.

Table 1: Main Nakba dataset configuration

Split	Label	# Samples	Label Ratio	Split Size	Split Ratio
Train	destruction	494	35.29	1400	69.97
	not_destruction	906	64.71		
Validate	destruction	70	35.18	199	9.95
	not_destruction	129	64.82		
Test	destruction	142	35.32	402	20.09
	not_destruction	260	64.68		
Full Dataset Size		**2001**			

4. System Overview

4.1. Enhanced Dataset

Although the main Nakba dataset provides valuable visual documentation, the size of the training set, which contains 1,400 images, is too small to effectively train a deep learning model. The situation becomes even worse when the classifier employs a transformer-based architecture that typically requires large-scale data. To address this challenge, we employed a semi-supervised training strategy to expand the training set, as illustrated below:

1. A Swin Transformer-based classification model with pretrained weights was trained on the main dataset for three epochs and achieved 86.43% accuracy on the validation set.

2. Several source datasets, which focus on general damage and are partially related to our task, were utilized to increase the number of training images. Table 2 provides an overview of these datasets.

3. The trained model was then used to classify the newly added images into destruction and non-destruction.

4. High-confidence predictions (> 0.9 or < 0.1) were retained as pseudo-labels, while the remaining samples were manually re-annotated to improve data quality.

Using this strategy, a total of 8,600 images were added to the main dataset, bringing the training set to 10,000 images. The final training set configuration is presented in Figure 1.

4.2. Model Architecture

The proposed model adopts a Swin Transformer architecture for destruction image classification and

[3] https://sina.birzeit.edu/nakba-nlp/2026/

[4] https://lrec2026.info/

Table 2: Images data sources

Dataset	Description	Size
(Tuncer, 2023)	It contains images representing different types of damage, including the following classes: (1) debris, (2) damaged building, (3) damaged highway, (4) non-damaged building, and (5) non-damaged highway. These images were collected from five online sources and are stored in JPEG or JPG format with varying dimensions.	4,373 images
(Niloy et al., 2021)	This dataset focuses on disasters and contains images classified into one of the following categories: (1) damaged infrastructure, (2) fire disaster, (3) human damage, (4) land disaster, (5) water disaster, and (6) non-damage.	13,600 images
(sta, 2025)	The dataset contains images of war-related events, where each image is in JPG format and is classified into one of the following categories: (1) combat, (2) destroyed, (3) fire, (4) humanitarian, and (5) military.	1,700 images
(Alam et al., 2020)	It is a consolidated benchmark dataset for social media image classification in disaster response. Its images were collected from sources (Alam et al., 2018) and (Mozannar et al., 2018), and have the purpose of being generally employed for four main tasks. The first task focuses on disaster types, the second on image informativeness, the third on humanitarian categories and actions, and the fourth on damage severity detection.	77,350 images

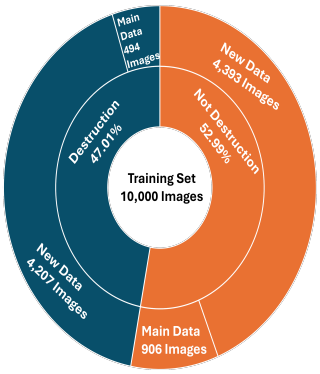


Figure 1: New train dataset configuration

consists of two key components: a backbone block and a head block. Figure 2 presents an overview of the architecture and processing pipeline of the proposed model.

The backbone block is primarily used for feature extraction by employing the first three stages of the Swin Transformer architecture. Specifically, given an input image of size 3×224×224, the backbone first applies patch partitioning followed by linear embedding to convert the image into non-overlapping token representations. The network then processes these tokens through three stages consisting of multiple Swin Transformer blocks (2, 2, and 6 blocks, respectively), with patch merging operations between stages to progressively reduce spatial resolution while increasing feature dimensionality. Each Swin Transformer block consists of two main components: a Window Multi-Head Self-Attention (W-MSA) module and a Shifted Window Multi-Head Self-Attention (SW-MSA) module, both wrapped with layer normalization and residual connections, and followed by a feed-forward MLP module. These window-based attention mechanisms capture local patterns that progressively integrate into a broader contextual representation as the channel depth increases from 96 to 384.

The second component is the head block, whose objective is to classify the input image using the extracted features. Before the final classification process, the head block further refines the representations by employing the last stage of the Swin Transformer architecture, increasing the feature dimension to 768. This strategy customizes the network and allows it to focus more effectively on the target task without requiring full end-to-end retraining, as discussed in Section 4.3. After the final features are condensed via global average pooling and flattening, they are passed to the MLP classification head, which acts as the decision-making module. It consists of two hidden layers with 256 neurons each, followed by a single-neuron output layer. The GELU (Gaussian Error Linear Unit) activation function is employed because it provides smoother nonlinear transformations than ReLU and has demonstrated superior performance in transformer-based architectures. Moreover, dropout layers are incorporated to mitigate overfitting. Finally, the model is trained using the Binary Cross-Entropy (BCE) loss function to optimize the prediction of destruction versus non-destruction labels.

4.3. Training Strategy

The training strategy for this model centers on a transfer learning approach, utilizing Swin Transformer weights pretrained on ImageNet-1K to provide a robust foundation for feature extraction. To prevent disrupting these pretrained features and to avoid overfitting due to the relatively limited dataset size, the backbone is frozen and kept non-trainable throughout the training process. In contrast, only the head block is trained to allow the network to specialize in the target task. The head training is performed in two phases: first, the MLP classification module is trained for 30 epochs to stabilize the decision layers and adapt them to the extracted

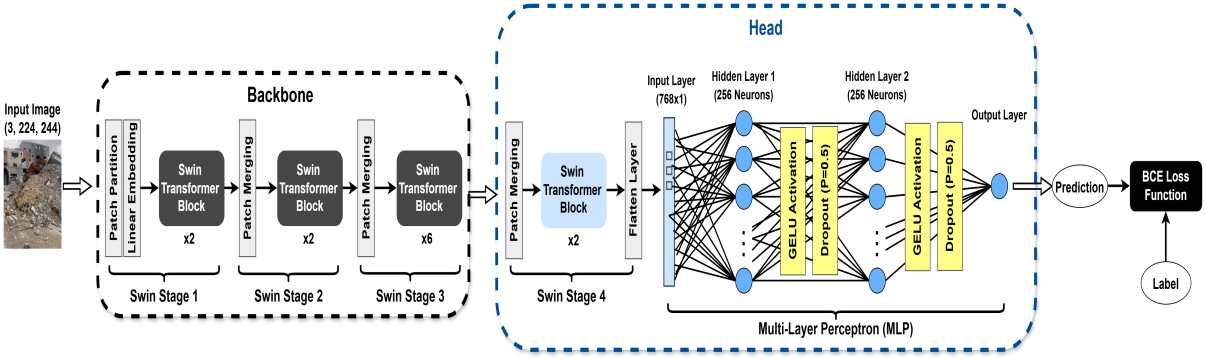


Figure 2: Model architecture

features. Then, the entire head block is trained for an additional 70 epochs to further refine task-specific representations and improve classification performance.

5. Experimental Results

5.1. Implementation Details

The model was implemented in Python using the PyTorch framework. Training and testing were conducted on the Kaggle platform using a Linux-based environment with a 64-bit x86-64 processor architecture. The system was equipped with a single NVIDIA Tesla P100-PCIE-16GB GPU to enable accelerated computation. The input images have dimensions of 224×224×3 (height × width × channels) and are normalized using the mean and standard deviation of ImageNet(Deng et al., 2009). The AdamW(Loshchilov and Hutter, 2019) optimizer is utilized with default parameters except weight_decay set to 1e-4. A batch size of 32 is adopted for training. The initial learning rate is set to 1e-4 for the first train and 1e-5 for the second train. Primary metrics were used to evaluate and compare the model's performance: (1) Accuracy, (2) Precision, (3) Recall, and (4) Macro F1-score.

5.2. Quantitative Results

The proposed model was evaluated on the test set of the NakbaArchiveClassifier Shared Task (Abrahams et al., 2026) by submitting its predictions to the CodaBench platform [5]. It is worth noting that the evaluation was conducted on a blind test set, ensuring strong generalizability validation and fair comparison. Table 3 presents the performance results obtained from the official CodaBench evaluation using multiple metrics.

The model achieved strong classification performance, obtaining an F1-score of 86.55% and an accuracy of 87.81%. The close values between accuracy and F1-score indicate balanced performance across both classes without significant bias. Furthermore, the proposed model ranked 5th among all participating teams in the shared task [6], demonstrating its competitiveness and robustness in comparison with other submitted methods.

Table 3: Model performance on the test set of the Nakba image classification shared task

Evaluation Metric	Score
F1-score	**86.55**
Accuracy	**87.81**
Precision	**86.88**
Recall	**86.26**

5.3. Ablation Analysis

To determine the optimal classification model for destruction identification, suitable for the shared task environment (in terms of dataset characteristics and evaluation metrics), multiple experiments were conducted using different architectural design choices. Table 4 presents the performance results of these experiments on the test dataset using F1-score and accuracy as evaluation metrics.

The reported experiments are those most relevant to the final selected model, where both the backbone and head components were varied in each configuration. Specifically, the backbone options included a transformer-based architecture (Swin), a CNN-based architecture (EfficientNet), and a hybrid CNN–Transformer approach that employs a convolutional block as a preprocessing step followed by Swin as the feature extractor. For the head component, the alternatives included using no additional head (default classifier), an MLP classi-

[5] https://www.codabench.org/competitions/12654/#/pages-tab

[6] https://www.codabench.org/competitions/12654/#/results-tab

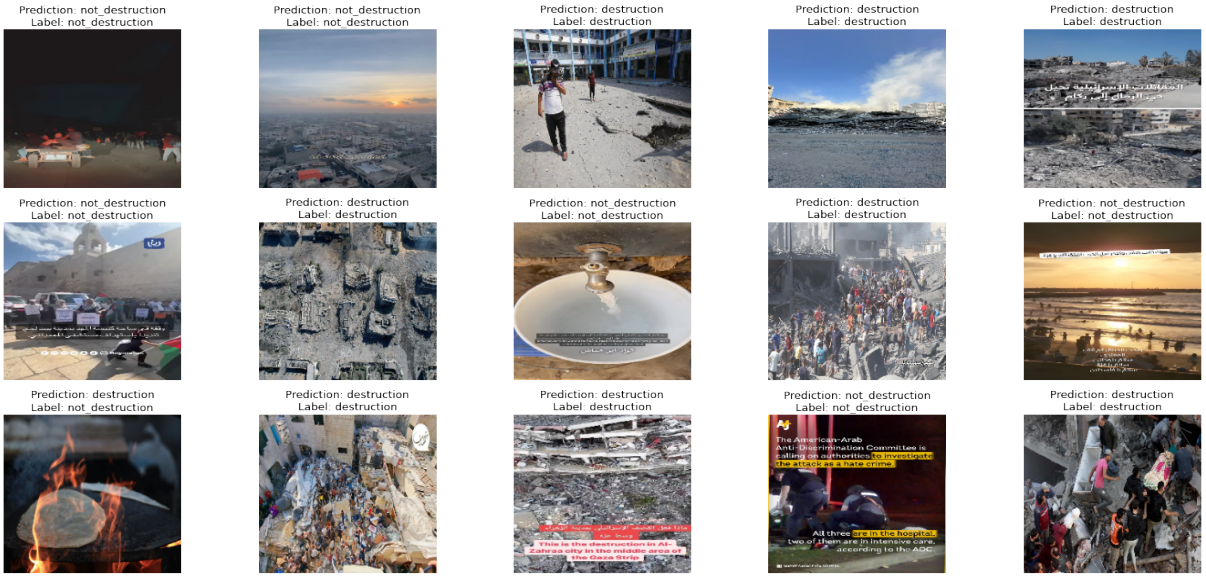


Figure 3: Qualitative results of the proposed framework on the validation set.

Table 4: Ablation analysis of different backbone and head configurations for the Nakba image classification shared task

Backbone	Head	F1-score	Accuracy
Swin	–	84.51	85.82
Swin	MLP	86.55	87.81
Swin	SVM	82.03	83.58
CNN & Swin	MLP	83.56	84.58
EfficientNet	–	84.56	85.82

fier, or a support vector machine (SVM). Overall, the ablation results demonstrate that the Swin Transformer backbone coupled with the proposed MLP head provides the most effective and well-balanced architecture for the target task.

5.4. Error Analysis

To better understand the strengths and limitations of the proposed model, an error analysis was conducted by examining the confusion matrix on the validation set, as illustrated in Figure 4.

The model correctly classified 117 non-destruction images and 54 destruction images, demonstrating strong performance across both categories. However, 12 non-destruction samples were incorrectly predicted as destruction (false positives), while 16 destruction samples were misclassified as non-destruction (false negatives). This number of errors indicates that some destruction cases remain challenging, possibly due to subtle damage patterns or visual similarities between damaged and undamaged structures.

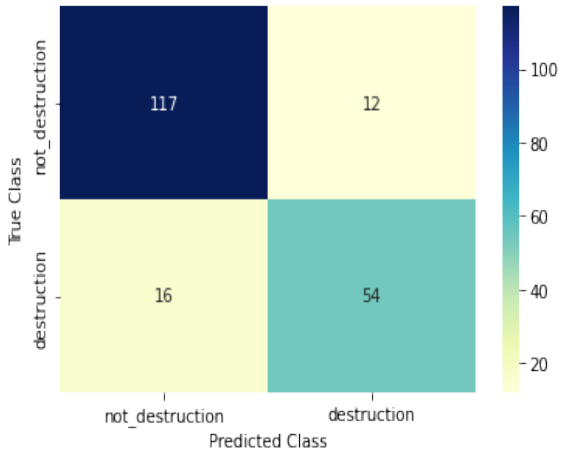


Figure 4: Validation set confusion matrix

Nevertheless, the overall distribution shows that the model maintains a balanced performance between the two classes, without significant bias toward either class.

To qualitatively illustrate the effectiveness of the proposed model, Figure 3 presents representative examples. All these images are randomly sampled from the validation set, which was not used in the training process. Above each image, the true label and the model's prediction are displayed, which highlights the model's ability to capture destruction with a small number of misclassified examples.

6. Conclusion

This paper presented a Swin Transformer-based framework for destruction image classification in the Nakba image classification shared task. The

proposed framework achieves strong performance by integrating a Swin Transformer backbone, an MLP classification head, a customized training strategy with transfer learning, and a semi-supervised data expansion strategy. Moreover, extensive ablation analysis demonstrated the effectiveness of the proposed architectural design, confirming that the Swin backbone coupled with the MLP head provides superior and balanced performance. Despite these achievements, the proposed model still struggles with several issues. These issues include that it has not completely addressed overfitting, classification accuracy requires further improvement, and the model's generalizability is still limited. Future work may explore partial backbone fine-tuning, employing advanced data augmentation techniques, further improvements in the dataset (size and diversity), and adopting multi-head classification.

7. References

2025. stable dataset. https://universe.roboflow.com/nothing-cniqf/stable-d0jqc. Roboflow Dataset.

Alexei Abrahams, Shadi Abudalfa, Mustafa Jarrar, and George Mikros. 2026. The nakbaarchive-classifier shared task on nakba image classification. In *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM)*.

Firoj Alam, Ferda Ofli, Muhammad Imran, Tanvirul Alam, and Umair Qazi. 2020. Deep learning benchmarks and datasets for social media image classification for disaster response. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 151–158.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*. Published as a conference paper at ICLR 2021. Originally submitted to arXiv in October 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, volume 25, pages 1097–1105.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Hussein Mozannar, Yara Rizk, and Mariette Awad. 2018. Damage identification in social media posts using multimodal deep learning.

Fahim Faisal Niloy, Abu Bakar Siddik Nayem, Anis Sarker, Ovi Paul, M Ashraful Amin, Amin Ahsan Ali, Moinul Islam Zaber, AKM Mahbubur Rahman, et al. 2021. A novel disaster image data-set and characteristics analysis using attention model. In *2020 25th International Conference on Pattern Recognition (ICPR)*, page 6116–6122. IEEE.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*. Published as a conference paper at ICLR 2015. Originally submitted in 2014.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, pages 6105–6114.

Türker Tuncer. 2023. Damaged constructions image dataset. https://www.kaggle.com/datasets/turkertuncer/damaged-constructions-image-dataset. Kaggle dataset.

Yafa at StanceNakba: Actor-Level Stance Detection Using Cross-Lingual Approach

Tasnim Zayet, Osama Hamed*, Tasneem Duridi
Department of Computer Science, Birzeit University
Computer Systems Engineering Department, Palestine Technical University - Kadoorie
Department of Computer Science, Palestine Technical University - Kadoorie
{Birzeit, Tulkarm}, West Bank - Palestine
tzayet@birzeit.edu, {osama.hamed, tasneem.duridi}@ptuk.edu.ps
*Corresponding author.

Abstract

This paper addresses the problem of actor-level stance detection in English social media posts concerning the Palestinian issue, a subtask of the StanceNakba-2026 Shared Task. The objective is to classify posts into one of three categories: Pro-Palestine, Pro-Israel, or Neutral, which is more challenging than the traditional favor/against/neutral formulations. This study uses a dataset comprising 1,401 posts, collected from X (formerly Twitter) after October 7, 2023, and annotated with one of the three stance labels. As Yafa's Team, we tried to solve this problem using BERT-based models, which have proven their superiority in similar tasks. Several BERT-based models were fine-tuned and compared, including ARBERT, MARBERT, and PoliBERTweet, among others. Our winning model is the "MARBERT-Y", where the "Y" comes from Yafa, a MARBERT-based model that has achieved a macro-F1 score of 95% on the test set. We argue this to two main factors: the structured and harsh preprocessing steps applied and the fine-tuning process employed. This indicates that domain-adapted transformer models, i.e., those pretrained on large-scale Twitter data are highly effective for politically stance detection tasks.

Keywords: Nakba, stance detection, multi-class classification, BERT-based models, MARBERT, hyperparameters.

1. Introduction

Social media platforms have grown tremendously in recent years. As a result, military conflicts have also moved into the digital space, shaping individuals' opinions around the world in real time (Duridi et al., 2025b). Additionally, people can access the perspectives of all parties involved in a conflict along with supporting evidence. This has led people to take sides. Over the past two years, the war against Gaza has attracted the world's attention. As a consequence, a vast amount of content has spread all over the internet representing various stances. Some support Israel, others stand with the Palestinians, and some maintain a neutral stance (Duridi et al., 2025a). Specifying and analyzing people's stance in an automatic manner is a challenging and important task. It may mobilize and influence public opinions, help in detecting fake news and analyze trends, especially in elections.

This paper outlines the Yafa's Team (yafateam) submission to Subtask A of the StanceNakba-2026 Shared Task (Aldous et al., 2026) at the NakbaNLP 2026 International Workshop (LREC 2026 Conference). Where we tackle the actor-level English stance detection towards the Palestinian-Israeli conflict. The task is to build a multi-class text classification model that can identify an author's political opinion. In other words, the task is to categorize social media post as Pro-Palestine, Pro-Israel, or Neutral. This task is more challenging than other stance detection tasks where the categories are Favor, Against, or Neutral. For example, see the work by Melhem et al. (2024). To approach this task, we employ several BERT-based classification approaches, which investigate the pre-trained transformer model's ability to capture the deep contextual representations of text. We fine-tune BERT-based models using a task-specific dataset consisting of 1,401 English posts. As requested, we had a data-split with 70/15/15 for training, development and testing respectively. Our model learns to distinguish subtle linguistic cues and stance expressions characteristic of this politically sensitive domain, enabling robust classification of author-level political positions.

This paper offers the following three contributions: (i) we present a single-task learning (STL) BERT-based model that utilizes MARBERT to detect actor-level stance on the **Pro-Palestine**, **Pro-Israel**, or **Neutral** opinions, (ii) we describe the list of hyperparameter (HP) that made MARBERT outperform other used BERT-based models, thus ranking Yafa Team 2nd, and (iii) we provide a detailed analysis and discussion of our best-performing model accross thees three opinions.

The paper is organized as follows: Section 2 presents prior and recent research on stance de-

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 177–181
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

tection. Section 3 provides a comprehensive analysis of the social media dataset used. Section 4 describes the proposed systems and experimental setup. Section 5 presents and discusses our experimental results. Finally, Section 6 concludes the paper and suggests ideas for future search.

2. Related Work

Actor-level stance detection focuses on analyzing an individual's opinions towards a specific target and it is an STL that has been widely utilized in elections and conflicts (Mohammad et al., 2016).

Previous studies on actor-level stance (Mohammad et al., 2016) addressed stance detection for specific political targets using English tweets with supervised traditional classifiers, specifically Support Vector Machines (SVMs) and convolutional neural networks (CNN) (Vijayaraghavan et al., 2016). An improvement upon these was done using Bidirectional Long Short-Term Memory (Bi-LSTM) that jointly encodes tweet-target representation (Augenstein et al., 2016). Further advancements were proposed for multi-target stance detection using an ensemble neural approach (Sobhani et al., 2017; Siddiqua et al., 2019).

Recent studies using transformer-based models such as BERT improved actor-level stance detection. Conforti et al. (2020) employed BERT which outperformed earlier architectures in both in-target and cross-target setups, while Kawintiranon and Singh (2021) incorporated political knowledge into BERT's masked language model. Additional work explores zero-shot transfer across targets and topics using BERT-style encoders (Allaway and McKeown, 2020).

Domain-adapted models such as PoliBERTweet (Kawintiranon and Singh, 2022), which was pre-trained on 83 million US election tweets, achieved 3-10% F1 gains over general-purpose models. AraBERT and MARBERT (Abdul-Mageed et al., 2021), pre-trained on large-scale Arabic corpora, outperformed multilingual models like mBERT and XLM-RoBERTa across Arabic NLP benchmarks.

Despite this progress, existing work targets western political contexts with classes favor,against and neutral, leaving cross-cultural conflict scenarios underexplored. Moreover, it remains unclear whether linguistic adaptation outweighs domain adaptation for English tweets related to Arabic politics and cultural context. Our work addresses these gaps by benchmarking multiple BERT-based models, showing that Arabic-based models pre-trained on tweets outperformed both multilingual models and domain-adapted English models, establishing that linguistic and cultural alignment are more critical than topical domain relevance for this type of cross-cultural stance detection.

3. Data and Stats

This section presents an overview of the StanceNakba-2026 dataset and provides descriptive statistics to better understand its structure and distribution. The dataset was collected and annotated by the shared task organizers.

3.1. Dataset Overview

The StanceNakba-2026 dataset is a balanced dataset consisting of 1,401 posts, primarily collected from X (formerly Twitter) social media platform, during the period following October 7, 2023. Each post is annotated with one of three stance labels reflecting the author's position toward the ongoing conflict: **Pro-Israel**, **Pro-Palestine**, or **Neutral**. The task is formulated as a multi-class supervised stance detection problem.

The dataset was released in three separate phases to support the ML/AI model development and evaluation, in line with the shared task design. Table 1 summarizes the composition of the dataset across these phases.

Phase	Total	Pro-Pal.	Pro-Isr.	Neut.
Training	980	326	327	327
Validation	210	55	88	67
Test	211	–	–	–

Table 1: Dataset composition and label distribution across splits.

The training set exhibits a balanced distribution across the three stance categories, which supports effective supervised learning and reduces the risk of strong class bias. The label distribution is illustrated in Figure 1.

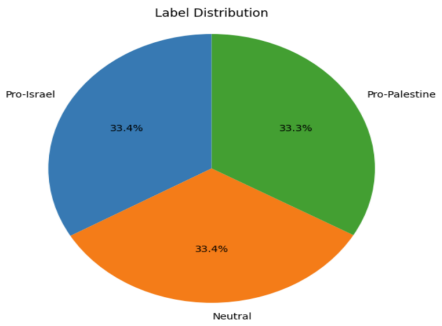


Figure 1: Distribution of stance labels in the training set.

To provide qualitative insight into the nature of the data, Table 2 presents representative examples from each stance category. These examples highlight the diversity in tone, content, and linguistic expression found in the dataset, reflecting the complexity of stance detection in real-world social media discourse.

Label/Pro-	Example
Palestine	Against the genocide that the Zionists commit.
Israel	Finish Hamas. I wish we had a president who would assist Israel to do what needs to be done. Rid the world of these terrorists who murder and violate men, women and children. Enough!
Neutral	The Lord Himself gave that land to the Jews and I will back them till I die.

Table 2: Representative examples from each stance category.

3.2. Data Preprocessing

A systematic preprocessing pipeline was applied to all textual data to improve data quality and maintain consistency throughout the corpus as follows:

- Lowercase normalization: All alphabetic characters were normalized to lowercase to reduce lexical variation and minimize feature space complexity.
- Links normalization: All links were replaced by empty strings.
- User mention Normalization: All user mentions were replaced with uniform text @user.
- Character repetition normalization: character repetitions were removed.
- White space normalization: repeated white spaces were removed; additionally, the tweets were stripped.
- Connecting words normalization: connected words were split using the wordninja library (Anderson, 2019).
- Hashtag normalization: The hashtag notation # was removed, while the hashtag text was kept.

However, punctuation marks and emojis were kept due to their importance in stance detection, as some (e.g., ?,!) can indicate stance. These preprocessing steps produced clean, semantically relevant text optimized for feature extraction and subsequent classification analysis.

4. System Description

This section briefly describes the system we built to take part in the NakbaStance detection Shared Task. Our system is a multi-class classification system that evaluates multiple BERT-based pretrained language models.

4.1. Used Models

4.1.1. XLM-Roberta-base

This is a cross-lingual pretrained language model proposed by (Conneau et al., 2019). It was trained on large-scale multilingual corpora comprising 2.5TB of web-crawled text from 100 languages. The base configuration consists of approximately 125 million parameters, offering a practical trade-off between model capacity and computational cost.

4.1.2. ARBERT and MARBERT

Introduced by (Abdul-Mageed et al., 2021), ARBERT and MARBERT are two Arabic-specific pretrained transformer models that extend the BERT-base architecture, each comprising 12 attention layers, 768 hidden dimensions, and approximately 163 million parameters. The key distinction lies in their pretraining corpora: ARBERT is trained on a 61GB Arabic corpus drawn from Wikipedia, news, and literary sources, whereas MARBERT is pretrained on a large-scale collection of 1 billion tweets (128GB), covering both Modern Standard Arabic and a wide range of dialectal varieties. To accommodate the brevity of tweet content, MARBERT omits the Next Sentence Prediction (NSP) objective during pretraining. Both models surpassed strong multilingual baselines including mBERT and XLM-RoBERTa across different Arabic NLP tasks.

4.1.3. PoliBERTweet

PoliBERTweet extends BERTweet through continued pre-training on 83 million US election tweets (Kawintiranon and Singh, 2022). This specialization improves its handling of political discourse, outperforming BERTweet and other baselines on stance detection and sentiment analysis with 3-10% F1 score gains.

4.1.4. BERT-base-uncased

BERT-base-uncased (Devlin et al., 2019) is the foundational model with 12 layers, 768 hidden units, and 110M parameters. Pre-trained on BookCorpus and Wikipedia using MLM (15% masking) and NSP, it employs WordPiece tokenization (30K vocabulary) and case-insensitive processing.

4.2. Hyperparameter Settings

We fine-tuned UBC-NLP/MARBERT for three-class stance classification using the Hugging Face Trainer framework. The dataset was divided into training (70%), development (15%), and test (15%) splits using stratified sampling to preserve label distribution. A fixed random seed was used to ensure reproducibility.

Optimization was performed using AdamW. The model was trained for a maximum of 8 epochs, with early stopping applied if the macro-averaged F1 score on the development set did not improve for two consecutive epochs. Table 3 describes the list of hyperparameters (HP) applied.

Note that the input sequences were tokenized using the MARBERT tokenizer, i.e., truncated to a maximum length of 128 tokens, and padded to a fixed length. It should be also noticed that no learning rate warm-up or gradient accumulation was applied. Additionally, a customized cross-entropy function with label smoothing was employed to suit the task of multi-class classification and decrease the effect of the confidence problem. Besides label smoothing, the dropout rate in attention layers was increased to 0.2.

Hyperparameter	Value
Learning rate	2×10^{-5}
Weight decay	0.1
Batch size (train / eval)	4 / 4
Maximum sequence length	128
Number of epochs	8
Early stopping patience	2
Optimizer	AdamW

Table 3: Main hyperparameters used for fine-tuning.

4.3. Model Evaluation

Following the organizers' evaluation framework, we use F1-macro metric as the primary one. F1-macro computes the F1 score (i.e. the harmonic mean of precision and recall) for each class independently and averages them equally. This makes it robust against class imbalance.

5. Results & Discussion

As managed by the organizers, the dataset was split into training, development/validation, and testing subsets, allocating 70%, 15%, and 15% of the data, respectively.

During the development phase, our model, i.e., MARBERT-Y was regularly trained on the training split and tested using the development split. It has achieved an F1-macro of 88%, even without applying a harsh pre-processing or changing a lot of HP. However, the validation gave indications that this model is promising.

During the testing phase, our model achieved an F1-macro of 95%. We argue this to the applied harsh pre-processing described in Subsection 3.2, and the adopted set of configurable HP, see subsection 4.2. Table 4 compares the performance of the used BERT-based models according to F1-macro, and showing that MARBERT stood first. Consequently, we call it as MARBERT-Y, where the "Y" stands for Yafa.

Model	Acc.	Prec.	Rec.	F1
ArBert	0.94	0.94	0.94	0.94
MARBERT	0.95	0.95	0.95	**0.95**
PoliBERTweet	0.91	0.91	0.91	0.90
BERT-base-uncased	0.91	0.91	0.91	0.91
XLM-Roberta-base	0.91	0.92	0.91	0.91

Table 4: Performance comparison of BERT-based models.

Initially, we hypothesized that domain-specific political models would achieve superior performance. However, PoliBERTweet underperformed compared to Arabic-specialized models. This indicates that linguistic compatibility is more critical than domain relevance. A translation-based approach to Arabic also proved ineffective due to the semantic distortions introduced during automatic translation. Since the dataset discusses Arabic political topics in English but contains terms whose meaning remains tightly bound to their original linguistic and cultural context, even after translation, we adopted Arabic-specialized models: ArBERT and MARBERT. Although, multilingual models such as XLM-RoBERTa-base achieved competitive results, they still fell short. Finally, ArBERT and MARBERT achieved the best performance with scores of 0.94 and 0.95 respectively. This demonstrates the advantage of large-scale Arabic pre-training, particularly for dialectal expressions, as discussed previously in Section 4.

6. Conclusion

In this paper, the Yafa's team (ranked 2nd on the leaderboard) tried various BERT-based models for the actor-level stance detection, subtask (A) of the NakbaStance-2026 Shared Task. Our best-performing model is the MARBERT-Y. It gets it's name from the MARBERT pretrained model, and the "Y" from Yafa. It has achieved a macro-F1 of 95%. We attribute this model's superiority to three factors: the large-scale collection of tweets used during pretraining, the rigorous preprocessing steps applied, and the set of tuned hyperparameters.

7. Acknowledgements

The authors would like to thank their institutions for their support: Birzeit University and Palestine Technical University - Kadoorie.

8. Bibliographical References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, et al. 2021. Arbert & marbert: Deep bidirectional transformers for arabic. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing, volume 1*, pages 7088–7105.

Kholoud Khalil Aldous, Md Rafiul Biswas, Mabrouka Bessghaier, Shimaa Ibrahim, Kais Attia, and Wajdi Zaghouani. 2026. StanceNakba shared task: Actor and topic-aware stance detection in public discourse. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

Emily Allaway and Kathleen McKeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. *arXiv preprint arXiv:2010.03640*.

Derek Anderson. 2019. wordninja 2.0.0.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. *arXiv preprint arXiv:1606.05464*.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won't-they: A very large dataset for stance detection on twitter. *arXiv preprint arXiv:2005.00388*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1*, pages 4171–4186.

Tasneem Duridi, Lour Atwe, Areej Jaber, Eman Daraghmi, and Paloma Martínez. 2025a. Detection of propaganda and bias in social media: A case study of the israel-gaza war (2023). In *2025 International Conference on New Trends in Computing Sciences (ICTCS)*, pages 204–210. IEEE.

Tasneem Duridi, Areej Jaber, and Paloma Martínez. 2025b. Arabic hate speech detection based on bert models variants. *Egyptian Informatics Journal*, 32:100845.

Kornraphop Kawintiranon and Lisa Singh. 2021. Knowledge enhanced masked language model for stance detection. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 4725–4735.

Kornraphop Kawintiranon and Lisa Singh. 2022. Polibertweet: a pre-trained language model for analyzing political content on twitter. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7360–7367.

Anas Melhem, Osama Hamed, and Thaer Sammar. 2024. Tao at stanceeval2024 shared task: Arabic stance detection using arabert. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 842–846.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.

Umme Aymun Siddiqua, Abu Nowshed Chy, and Masaki Aono. 2019. Tweet stance detection using an attention based neural ensemble model. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1*, pages 1868–1873.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557.

Prashanth Vijayaraghavan, Ivan Sysoev, Soroush Vosoughi, and Deb Roy. 2016. Deepstance at semeval-2016 task 6: Detecting stance in tweets using character and word-level cnns. *arXiv preprint arXiv:1606.05694*.

The Resistant Word at StanceNakba Shared Task: A Topic-Aware Model for Cross-Topic Stance Detection

Mohammed Bahgat[1], **Doaa Salah**[2], **Sarah Yassine**[3]
[1] SUT Egypt, Cairo, Egypt
[2] Cairo University, Cairo, Egypt
[3] Lebanese University, Beirut, Lebanon
mohamed.bahgat@sut.edu.eg
doaa.ibrahim00@eng-st.cu.edu.eg
sarah.yassine.2@st.ul.edu.lb

Abstract

Cross-topic stance detection in Arabic is the task of identifying whether a text expresses a *pro*, *against*, or *neutral* position toward a given issue, and it is particularly challenging under topic shifts and class imbalance. In Subtask B of the StanceNakba 2026 shared task on Arabic cross-topic stance detection, we are given a Levantine Arabic sentence and one of two topics: "Normalization with Israel" or "Refugee/Immigrant Presence in Jordan," and we must classify the expressed stance. A central difficulty is the systematic failure of standard fine-tuning to recognize the minority *neutral* class, driven by majority-class dominance in cross-entropy training and accuracy-based checkpoint selection. To address this, we combine random oversampling with class-weighted cross-entropy loss, and we build an ensemble of four Arabic pre-trained transformers MARBERT, AraBERT Large, XLM-RoBERTa Base, and CAMeL-BERT Mix each trained using Stratified 5-Fold cross-validation. Our final system achieves a macro-F1 of **0.9777** and an accuracy of **97.79%** on the evaluation set.

Keywords: Arabic NLP, stance detection, Levantine dialect, class imbalance, weighted loss, ensemble, transformer

1. Introduction

Stance detection identifies whether a text is *pro*, *against*, or *neutral/none* toward a target. For Arabic social media, this remains difficult because of dialectal variation and implicit or sarcastic language, which often blur stance and sentiment. Recent shared tasks such as StanceEval 2024 reflect growing interest in Arabic stance detection while highlighting remaining challenges for robust modeling (Alturayeif et al., 2024).

We present our submission to Subtask B of the StanceNakba shared task on Arabic cross topic stance detection (Aldous et al., 2026). Given a Levantine Arabic sentence and one of two targets, *Normalization with Israel* or *Refugee/Immigrant Presence in Jordan*, the system predicts *pro*, *against*, or *neutral*. The main challenge is reliable detection of the minority *neutral* class, which is often weakened by standard cross entropy training and accuracy based checkpointing. We address this with class weighted loss, random oversampling, and an ensemble of four Arabic pre trained transformers, MARBERT, AraBERT Large, XLM RoBERTa Base, and CAMeL BERT Mix, trained with Stratified 5 Fold cross validation and combined by probability averaging. Our official submission ranked fourth on the Subtask B leaderboard with 0.8562 macro F1 and 0.8564 accuracy. To support reproducibility, we release our code and experiment configuration in a public GitHub repository.

Our contributions are: (1) a Levantine Arabic cross topic stance system for two politically sensitive targets; (2) an imbalance aware training recipe that improves *neutral* detection; and (3) an ensemble framework combining model diversity, Stratified 5 Fold training, and consistent preprocessing.

2. Background & Related Work

Stance detection predicts whether a text expresses a *pro*, *against*, or *neutral* position toward a target. Recent work has moved from feature based models to neural and transformer approaches, while cross topic stance detection studies whether models trained on seen targets generalize to unseen ones (Khiabani and Zubiaga, 2024; Reuver et al., 2021). For Arabic, progress is still challenged by morphology, dialect variation, and limited annotated resources. Recent efforts such as MARASTA, ArabicStanceX, and StanceEval 2024 have expanded Arabic stance datasets and evaluation settings (Charfi et al., 2024; Alkhathlan et al., 2025; Alturayeif et al., 2024). Our work builds on this line by focusing on Levantine Arabic cross topic stance detection in a political setting.

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 182–186
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

3. System Overview

3.1. Data Description

Subtask B uses a topic specific subset of the MARASTA corpus, a multi dialectal Arabic cross domain stance corpus (Charfi et al., 2024). It contains 1,024 annotated instances split into training, evaluation, and test sets. Each instance is a Levantine Arabic sentence paired with one of two topics: *Normalization with Israel* (*al tatbi' ma' Isra'il*) and *Refugee/Immigrant Presence in Jordan* (*Wujud al Laji'in wal Muhajirin fi al Urdun*). The task is to predict one of three labels: *pro*, *against*, or *neutral*.

Example instances include a *pro* case for *Normalization with Israel* (*al tatbi' ma'a Isra'il sayakun khayran in sha' Allah*), an *against* case for the same topic (*ma fihi tatbi' ma'akum*), and a *neutral* case for *Refugee/Immigrant Presence in Jordan* describing limited work rights for Syrian refugees in Jordan since 2016.

3.1.1. Training Set

The training set contains 843 instances, with a relatively balanced label distribution: 298 *against*, 286 *pro*, and 259 *neutral*. It covers both topics, with 404 instances for *Normalization with Israel* and 439 for *Refugee/Immigrant Presence in Jordan*. For *Normalization with Israel*, the label distribution is 139 *against*, 145 *neutral*, and 120 *pro*; for *Refugee/Immigrant Presence in Jordan*, it is 159 *against*, 114 *neutral*, and 166 *pro*.

3.1.2. Test Set

The test set also contains 181 instances. It is released without labels and used for the final evaluation. The dataset is designed for topic-aware modeling: each sentence is provided with its topic, allowing models to condition predictions on the target.

3.1.3. Evaluation Set

The evaluation set contains 181 instances: 63 *against*, 62 *pro*, and 56 *neutral*. It includes 87 instances for *Normalization with Israel* and 94 for *Refugee/Immigrant Presence in Jordan*. The former is distributed as 29/32/26 (*against*/*neutral*/*pro*), and the latter as 34/24/36.

3.2. Evaluation Metrics

We report Macro-F1 (primary), Accuracy, Precision, and Recall. Macro-F1 averages per-class F1 scores, making it appropriate when performance varies across stance labels and minority classes should be weighted equally (Khiabani and Zubiaga, 2024; Alturayeif et al., 2024). Accuracy measures overall correctness but can be less informative under label imbalance (Alturayeif et al., 2024; Khiabani and Zubiaga, 2024). Precision and Recall quantify, respectively, false-positive control and true-positive coverage per class (Khiabani and Zubiaga, 2024). Here TP, FP, and FN denote true positives, false positives, and false negatives for each class in $\mathcal{C} = \{\text{pro}, \text{against}, \text{neutral}\}$, providing a balanced view of performance on dialectal Arabic stance detection (Alturayeif et al., 2024; Lichouri et al., 2024).

3.3. Methodology

Our system is built on a pipeline of four components applied in sequence: (1) Arabic text preprocessing, (2) class balancing via oversampling, (3) fine-tuning of multiple Arabic transformer models with weighted cross-entropy loss, and (4) probability-averaging ensemble with K-Fold cross-validation. Figure 1 illustrates the complete pipeline.

Given a topic sentence pair, the system first normalises both strings and encodes them as a single input sequence, with the topic as the first segment and the sentence as the second, separated by `[SEP]`. The encoded input is then passed to each of the four transformer models. For every model, Stratified 5-Fold training produces five probability vectors over {*pro*, *against*, *neutral*}. We average these probabilities across folds to obtain one prediction per model, then average again across the four models. The final stance label is the class with the highest averaged probability. For instance, if the final ensemble outputs $(0.08, 0.81, 0.11)$ for (*pro*, *against*, *neutral*), the predicted label is *against*.

3.4. Text Preprocessing

We apply an Arabic-specific normalisation pipeline to both sentences and topic strings prior to tokenisation to reduce social-media noise (inconsistent orthography, diacritics, and platform artefacts) (Charfi et al., 2024). The pipeline removes diacritics and tatweel, standardises common letter variants (alef forms, alef maqsura→ya', ta' marbuta→ha'), removes URLs and mentions, strips the # symbol while keeping hashtag text, and normalises whitespace.

After normalisation, each instance is encoded as a topic–sentence pair, with the topic as the first sequence and the sentence as the second, separated by `[SEP]`. This formatting enables topic-aware attention and supports cross-topic stance classification.

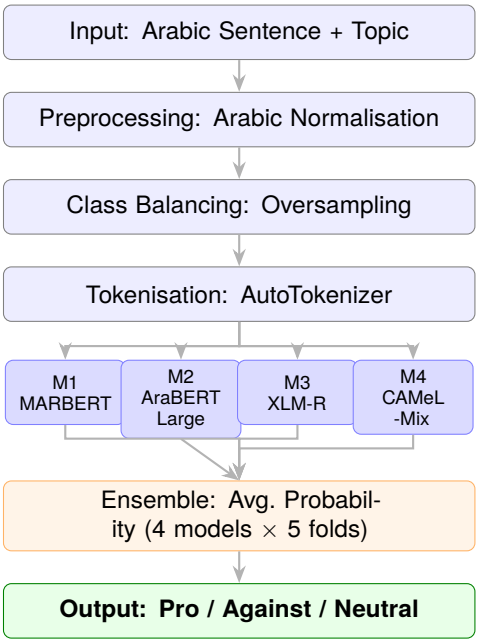


Figure 1: Full system pipeline from raw Arabic input to final stance label. All four models are fine-tuned with K-Fold and Weighted Cross-Entropy Loss; their softmax probabilities are averaged for the final prediction.

3.5. Class Balancing via Oversampling

Although the overall training labels are relatively balanced, we observe topic-level skew (Table **??**) that contributes to under-prediction of the minority *neutral* class. To counter this, we apply random oversampling on the training split using `sklearn.utils.resample`: we upsample each class with replacement (seed = 42) to match the largest class size, then shuffle the resulting balanced set. Oversampling is applied **only** to training; evaluation and test splits remain unchanged to preserve unbiased estimation.

3.6. Pre-Trained Models

We fine-tune four Arabic transformer models chosen for their complementary pre-training data and their demonstrated strength on Arabic NLP tasks (Antoun et al., 2020; Inoue et al., 2021):

- **M1 – MARBERT** (`UBC-NLP/MARBERT`): trained on one billion Arabic tweets spanning multiple dialects; particularly well-suited to dialectal and social media text (Abdul-Mageed et al., 2021).
- **M2 – AraBERT Large** (`aubmindlab/bert-large-arabertv02`): large variant of AraBERT v2, pre-trained on a diverse and large-scale Arabic corpus covering both MSA and web text (Antoun et al., 2020).
- **M3 – XLM-RoBERTa Base** (`FacebookAI/xlm-roberta-base`): a multilingual masked language model pre-trained on 100 languages, offering strong cross-lingual transfer and robustness to out-of-vocabulary Levantine terms.
- **M4 – CAMeL-BERT Mix** (`CAMeL-Lab/bert-base-arabic-camelbert-mix`): trained jointly on MSA and dialectal Arabic text, making it particularly suitable for the mixed-register nature of Levantine social media (Inoue et al., 2021).

The diversity in pre-training corpora of dialectal Twitter data, large-scale Arabic web text, multilingual data, and mixed-register Arabic is intentional. Each model captures different linguistic aspects of the Levantine dialect, and their ensemble corrects individual model weaknesses.

3.7. Weighted Cross-Entropy Loss

Motivation. Standard cross-entropy weights all samples equally, so under class imbalance majority labels dominate training; in our baseline this suppressed learning for the minority *neutral* class, causing systematic under-prediction.

Class-weighted loss. To counter this bias, we apply class-weighted cross-entropy, assigning each class c a weight inversely proportional to its (post-oversampling) frequency in the training set:

$$w_c = \frac{1/|D_c|}{\sum_{c'} 1/|D_{c'}|} \cdot |\mathcal{C}| \tag{1}$$

This normalisation ensures $\sum_c w_c = |\mathcal{C}| = 3$. The weighted loss is:

$$\mathcal{L}_w = -\sum_{c \in \mathcal{C}} w_c \cdot y_c \log p_c \tag{2}$$

where y_c is the ground-truth indicator and p_c is the predicted probability for class c.

Implementation. We implement a custom `WeightedLossTrainer` by subclassing the Hugging Face `Trainer` and overriding `compute_loss`. The weighted-loss implementation is shown in Appendix A.

3.8. K-Fold Cross-Validation Ensemble

To better exploit the limited training data (843 samples) and reduce prediction variance, we train each of the four models using Stratified 5-Fold cross-validation. In each fold, a model is fine-tuned on 80% of the training set and produces softmax probability predictions on the evaluation and test sets. We then average probabilities across folds for each

Hyperparameter	Value
Epochs per fold	5
Batch size	16
Learning rate	2×10^{-5}
Max sequence length	128 tokens
Warmup ratio	0.1
Weight decay	0.01
K-Fold splits	5 (stratified)
Precision	bfloat16
Loss function	Weighted CrossEntropy
Best checkpoint metric	macro-F1
Random seed	42

Table 1: Training configuration, identical for all four models.

model and across the $M = 4$ models to obtain the final prediction with $K = 5$. We use probability averaging rather than majority voting to leverage model confidence information and the full predictive distribution.

3.9. Training Configuration

Table 1 summarises the hyperparameters shared across all four models.

A key departure from the baseline is using **macro-F1** rather than accuracy as the criterion for saving the best checkpoint. Since accuracy rewards majority-class predictions, a model that rarely predicts *neutral* can still achieve high accuracy. Macro-F1 treats all classes equally and directly penalises failure to detect minority classes, producing checkpoints that generalise better across the full label set.

Experiments were run in a Kaggle GPU environment (https://www.kaggle.com/code) using PyTorch v2.8.0+cu126 (https://pytorch.org/), Transformers v4.57.1 (https://huggingface.co/docs/transformers/index), pandas v2.2.2 (https://pandas.pydata.org/), NumPy v2.0.2 (https://numpy.org/), and scikit-learn v1.6.1 (https://scikit-learn.org/stable/). Full scripts and configuration are available in our public repository.

4. Experimental Results

4.1. Model Comparison and Final System

Table 2 compares all evaluated systems on the 181-sample evaluation set. The upper block reports nine individual Arabic transformer models, each trained with Arabic preprocessing, class-weighted loss, random oversampling, and macro-F1 checkpoint selection. The highlighted row is our final ensemble, which combines four models with Stratified 5-Fold cross-validation and probability averaging.

Model / System	Acc.	Prec.	Rec.	F1
Individual fine-tuned models				
bert-base-arabic	0.7403	0.7403	0.7410	0.7401
camelbert-da	0.7624	0.7627	0.7629	0.7627
bert-base-arabertv02	0.7679	0.7733	0.7655	0.7677
MARBERT	0.7734	0.7737	0.7753	0.7735
roberta-eng-ara-128k	0.7790	0.7831	0.7810	0.7765
MARBERTv2	0.7900	0.7899	0.7900	0.7890
ARBERTv2	0.7950	0.7970	0.7950	0.7950
albert-xlarge-arabic	0.7955	0.7979	0.7968	0.7950
camelbert-msa	0.8011	0.8020	0.8034	0.7997
Ensemble system (this work)				
Ens. (4M×5F)	**0.9779**	**0.9778**	**0.9788**	**0.9777**

Table 2: Results on the evaluation set (181 samples), sorted by macro-F1. The highlighted row is the final submitted system.

Among single models, `camelbert-msa` performs best with a macro-F1 of 0.7997. Our ensemble improves substantially over the strongest single model, reaching 0.9777 macro-F1 on the evaluation set.

Official shared-task evaluation (test set). On the hidden official test set, our submission ranked **4th** on Subtask B with macro-F1 = 0.8562, accuracy = 0.8564, macro-precision = 0.8617, and macro-recall = 0.8566.

4.2. Generalization Gap and Error Analysis

The drop from local to official performance (0.9777 vs. 0.8562 macro-F1) suggests optimistic local estimates and some overfitting, likely because the evaluation split comes from the same collection as training, while the hidden test set contains harder lexical and discourse variation.

Detailed diagnostics are available only on the evaluation set. There, *pro* and *neutral* achieve perfect recall, and all four errors come from *against* examples, mostly confused with *neutral*. Thus, the main remaining difficulty is distinguishing implicit or ironic opposition from neutral content under distribution shift.

5. Conclusion

We presented our system for Subtask B of the StanceNakba 2026 shared task on Arabic cross-topic stance detection. To improve minority *neutral* detection, we combined class-weighted cross-entropy, macro-F1 checkpoint selection, random oversampling, and an ensemble of four Arabic transformers trained with Stratified 5-Fold cross-validation and probability averaging. The final system achieved 0.9777 macro-F1 on the evaluation set and 0.8562 macro-F1 on the official hidden test set.

6. Limitations

The gap between local and official performance suggests limited generalization and possible overfitting to the evaluation distribution. In addition, random oversampling does not add new linguistic diversity, the ensemble is computationally expensive, and the dataset covers only two politically sensitive topics in Levantine Arabic, which limits broader generalization. Future work will explore dialect-aware augmentation and instruction-tuned Arabic LLMs for more efficient and robust stance modeling (Alghaslan and Almutairy, 2024).

Acknowledgements

The authors thank the StanceNakba 2026 shared task organizers for providing the dataset and evaluation infrastructure.

7. Bibliographical References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT and MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, pages 7088–7105, Online. Association for Computational Linguistics.

Kholoud Khalil Aldous, Md Rafiul Biswas, Mabrouka Bessghaier, Shimaa Ibrahim, Kais Attia, and Wajdi Zaghouani. 2026. StanceNakba shared task: Actor and topic-aware stance detection in public discourse. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

Mamoun Alghaslan and Khaled Almutairy. 2024. MGKM at StanceEval2024: Fine-tuning large language models for Arabic stance detection. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*, pages 816–822, Bangkok, Thailand. Association for Computational Linguistics.

A. Alkhathlan, F. Alahmadi, F. Kateb, and H. Al-Khalifa. 2025. Constructing and evaluating ArabicStanceX: A social media dataset for arabic stance detection. *Frontiers in Artificial Intelligence*, 8:1615800.

Nora Alturayeif, Hamzah Luqman, Zaid Alyafeai, and Asma Yamani. 2024. StanceEval 2024: The first Arabic stance detection shared task. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*, Bangkok, Thailand. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT4)*, pages 9–15, Marseille, France. European Language Resources Association (ELRA).

Anis Charfi, Marwa Bessghaier, Ali Atalla, Rand Akasheh, Sara Al-Emadi, and Wajdi Zaghouani. 2024. Stance detection in Arabic with a multi-dialectal cross-domain stance corpus. *Social Network Analysis and Mining*, 14(1).

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop (WANLP 2021)*, pages 92–104, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Parisa Jamadi Khiabani and Arkaitz Zubiaga. 2024. Cross-target stance detection: A survey of techniques, datasets, and challenges. *arXiv preprint arXiv:2409.13594*.

Mohamed Lichouri, Khaled Lounnas, Khelil Rafik Ouaras, Mohamed Abi, and Anis Guechtouli. 2024. dzStance at StanceEval2024: Arabic stance detection based on sentence transformers. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*, Bangkok, Thailand. Association for Computational Linguistics.

Myrthe Reuver, Suzan Verberne, Roser Morante, and Antske Fokkens. 2021. Is stance detection topic independent and cross topic generalizable? a reproduction study. In *Proceedings of the 8th Workshop on Argument Mining*, pages 46–56, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A. Weighted-Loss Implementation

```
weights = class_weights_tensor.to(device)
loss_fct = nn.CrossEntropyLoss(weight=weights)
loss = loss_fct(logits.view(-1,
num_labels), labels.view(-1))
```

“Hope” at NakbaArchiveClassifier Shared Task: Transfer Learning-Based CNN Models for Infrastructure Damage Detection

Lojien AlKhidir, Heba Abdelhady

Hamad Bin Khalifa University
Doha, Qatar
loal89338@hbku.edu.qa, heab89248@hbku.edu.qa

Abstract

This paper describes Team Hope’s system for the NakbaArchiveClassifier Shared Task at Nakba-NLP 2026. The task focuses on binary classification of social media images into two categories: destruction and not_destruction. We evaluated multiple convolutional neural network architectures using transfer learning, including ResNet34, ResNet50, EfficientNet-B0, and a fine-tuned ResNet34 variant with staged training. All models were initialized with ImageNet pretrained weights and fine-tuned on the provided dataset of 2,001 images. The dataset is moderately imbalanced and contains visually diverse Instagram images depicting intact and damaged infrastructure. Our best-performing model, ResNet34 trained for 25 epochs with Adam optimizer and a learning rate of 1e-4, achieved 81% accuracy on the evaluation platform. We provide a comparative analysis of the tested architectures and discuss the impact of model depth, training duration, and class imbalance. Given the political and ethical sensitivity of the dataset, we also include a discussion of responsible AI considerations and potential limitations. Our findings suggest that moderate-depth architectures can generalize effectively in low-resource, contextually complex visual classification tasks (Abraham, et al. May 2026).

Keywords: image classification, infrastructure damage detection, transfer learning, convolutional neural networks, Nakba-NLP

1. Introduction

The NakbaArchiveClassifier Shared Task (Abrahams et al., 2026) aims to classify social media images into two categories: destruction and not_destruction. The task contributes to organizing large-scale visual archives related to infrastructure conditions in Gaza.

Automated image classification can support large-scale documentation and indexing of visual material. However, the dataset presents several challenges, including limited size, class imbalance, compression artifacts from social media platforms, and contextual ambiguity in damage representation.

As a multidisciplinary team newly entering computer vision research, we focused on evaluating established convolutional neural network (CNN) architectures using transfer learning (Shin, et al. 2016). Our goal was to compare different pretrained models and analyse their behaviour on a relatively small and context-sensitive dataset.

2. Related Work

Infrastructure damage detection has been widely studied in disaster response contexts such as earthquakes and floods. CNN-based models have demonstrated strong performance in visual damage classification tasks due to their ability to capture hierarchical image features (He, et al. 2016)

Transfer learning from ImageNet-pretrained models is common practice for small and medium-sized datasets, as it allows models to leverage generalized visual representations (Deng et al., 2009; Shin et al., 2016). Prior research suggests that moderate-depth architectures often generalize better than very deep models in low-resource settings.

3. Dataset and Task Description

The dataset provided for the NakbaArchiveClassifier Shared Task consists of 2,001 Instagram images manually labeled into two binary categories: *destruction* and *not_destruction*. The images depict infrastructure scenes collected from social media and include buildings, streets, residential areas, and urban environments under varying visual conditions.

During the development phase, the organizers provided predefined splits:

- **Training set:** 1,400 images
- **Validation set:** 199 images
- **Test set:** 402 images

We strictly adhered to the official splits and did not perform any reshuffling or cross-validation to ensure consistency with the evaluation protocol.

The dataset exhibits a noticeable class imbalance, with approximately 35% of images labeled as *destruction* and 65% labeled as *not_destruction*. This imbalance increases the risk of biased predictions toward the majority class and may influence model behavior during training. In particular, models trained without class weighting may tend to overpredict the *not_destruction* class.

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 187–190
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

The images vary significantly in resolution, camera angle, lighting conditions, and compression artifacts due to their origin on Instagram. Some images depict clear structural collapse, while others contain partial or subtle damage. In certain cases, both intact and damaged structures appear within the same frame, introducing contextual ambiguity.

The evaluation metric used by the shared task platform was accuracy, calculated as the proportion of correctly classified test images. While accuracy provides a straightforward measure of overall performance, it does not fully reflect per-class performance in imbalanced datasets, where alternative metrics such as macro-F1 could provide additional insight (Sokolova and Lapalme 2009).

4. Methodology

4.1 Preprocessing

All images were resized and normalized using standard ImageNet preprocessing statistics (mean and standard deviation) (Den, et al. 2009). We replaced the final fully connected layer of each pretrained network with a new linear layer containing two output neurons corresponding to the binary labels (*destruction* and *not_destruction*).

We used the official train, validation, and test splits provided by the organizers without modification to ensure comparability with other participating systems. Data loading was implemented using PyTorch's DataLoader, with batch sizes of 32 for ResNet models and 16 for EfficientNet-B0.

No additional manual data cleaning or resampling was performed. For EfficientNet-B0, class imbalance was addressed through weighted loss.

4.2 Model Architectures

We evaluated four convolutional neural network architectures initialized with ImageNet pretrained weights. In all cases, the final classification layer was replaced to adapt the models to binary classification.

4.2.1.1 ResNet34

- Pretrained on ImageNet
- Final fully connected layer replaced (2 output units)
- Batch size: 32
- Optimizer: Adam
- Learning rate: 1e-4
- Epochs: 25
- Loss: CrossEntropyLoss

4.2.1.2 ResNet50

- Pretrained on ImageNet
- Final fully connected layer replaced (2 output units)
- Batch size: 32
- Optimizer: Adam
- Learning rate: 1e-4
- Epochs: 10
- Loss: CrossEntropyLoss

4.2.1.3 EfficientNet-B0

- Pretrained on ImageNet
- Final classification layer replaced (2 output units)
- Batch size: 16
- Optimizer: Adam
- Learning rate: 1e-3
- Epochs: 25
- Loss: Weighted CrossEntropyLoss (to mitigate class imbalance)

4.2.1.4 ResNet34 Advanced

- Pretrained on ImageNet
- Final fully connected layer replaced (2 output units)
- Two-stage training strategy:
    - First 5 epochs: backbone frozen, training only final layer
    - Next 5 epochs: full fine-tuning
- Initial learning rate: 1e-3 (final layer)
- Fine-tuning learning rate: 1e-4
- Scheduler: ReduceLROnPlateau
- Batch size: 32

All models were trained using the Adam optimizer (Adaptive Moment Estimation) and standard backpropagation (Kingma and Ba, 2015). The best-performing model on the validation set was selected for final test submission.

5. Experimental Results

We evaluated four pretrained convolutional neural network architectures under consistent training conditions. Model selection was based on validation performance, and final predictions were submitted to the shared task evaluation platform.

Overall, the results indicate relatively close performance across architectures, with accuracy ranging between 79% and 81%. Despite architectural differences in depth and training strategy, performance variation remained within a narrow margin, suggesting that dataset size and class imbalance may constrain achievable gains.

ResNet34 achieved the highest performance, indicating that moderate-depth architectures may generalize better in small-scale datasets compared to deeper networks such as ResNet50. While ResNet50 contains more parameters and higher representational capacity, it did not provide measurable improvement under the current training configuration.

EfficientNet-B0, trained with weighted cross-entropy loss to address class imbalance, achieved comparable but slightly lower performance than ResNet34. This suggests that class weighting alone may not substantially alter model behavior when the imbalance is moderate (Tan and Le 2019).

The advanced ResNet34 configuration incorporating staged freezing and learning rate scheduling did not improve results. The shorter training duration and higher initial learning rate may have limited its convergence potential.

5.1 Accuracy scores of evaluated architectures

The results demonstrate that stable fine-tuning with moderate learning rates and sufficient training epochs contributed more to performance gains than architectural depth or complexity.

Model	Accuracy
ResNet34	81%
ResNet50	80%
EfficientNet-B0	80%
ResNet34 Advanced	79%

Table 1: Accuracy scores of evaluated architectures

6. Error Analysis

Analysis of misclassified samples revealed several recurring patterns. The most common errors included:

- Partial structural damage misclassified as intact when destruction affected only a limited region of the image.
- Distant or small-scale destruction embedded within otherwise intact urban scenes.
- Visual noise and reduced clarity due to Instagram compression artifacts.
- Ambiguous scenes containing both intact and damaged structures, leading to contextual confusion.

A significant proportion of false negatives occurred when damage did not visually dominate the frame. In such cases, the model appeared to rely on global scene characteristics rather than localized damage cues. This suggests that the architectures may prioritize overall texture and structure patterns rather than fine-grained spatial irregularities.

Additionally, class imbalance may have influenced prediction behavior, as models tended to favor the majority *not_destruction* class in borderline cases. This tendency is consistent with the higher prevalence of intact scenes in the training data.

These findings indicate that binary classification of infrastructure damage is not purely a visual clarity problem but also a contextual interpretation challenge. Future improvements may benefit from incorporating attention mechanisms or region-based approaches that focus on localized structural anomalies rather than entire-image representations.

7. Lessons Learned

As a multidisciplinary team newly entering the field of computer vision, this project provided several practical and methodological insights.

First, we learned that architectural complexity does not necessarily guarantee superior performance. Despite evaluating deeper and more sophisticated models, a moderately sized architecture (ResNet34) trained with stable hyperparameters achieved the best results. This highlighted the importance of controlled experimentation over architectural novelty.

Second, we observed that hyperparameter choices such as learning rate, number of training epochs, and class weighting significantly influence performance, particularly in small and moderately imbalanced datasets. Careful adjustment of these parameters proved more impactful than increasing model depth.

Third, we gained a deeper understanding of the challenges posed by real-world social media imagery. Unlike curated benchmark datasets, Instagram images exhibit compression artifacts, varying resolutions, and contextual ambiguity, which complicate binary classification.

Finally, we recognized the importance of critical reflection in machine learning research. Performance metrics alone do not fully capture system behavior, particularly in sensitive domains. Understanding failure patterns and ethical implications is essential for responsible deployment.

This experience reinforced the value of systematic experimentation, transparent reporting, and cautious interpretation of model results.

8. Conclusion

We presented Team Hope's system for the NakbaArchiveClassifier Shared Task, evaluating four pretrained convolutional neural network architectures under consistent experimental conditions. Among the tested models, ResNet34 achieved the highest performance with 81% accuracy on the evaluation platform.

Our results demonstrate that in small and moderately imbalanced datasets, stable training configurations and appropriate hyperparameter selection can have a greater impact on performance than architectural depth alone. The relatively narrow performance gap across models also suggests that dataset characteristics, including visual variability and contextual ambiguity, impose inherent limitations on achievable accuracy.

Beyond quantitative performance, this project emphasized the importance of systematic experimentation, careful interpretation of results, and ethical awareness when working with sensitive visual data. Automated classification systems in such domains should be viewed as supportive analytical tools rather than definitive decision-making systems.

9. Ethical Considerations and Limitations

The dataset used in this shared task is politically and contextually sensitive. Automated classification systems in such domains raise concerns regarding potential misuse, narrative framing, and bias amplification. The dataset originates from specific social media sources, which may introduce representation bias in terms of geography, perspective, and visual framing.

Additionally, the labeling of destruction versus non-destruction may involve subjective interpretation, particularly in cases of partial or ambiguous damage. Such subjectivity can influence model learning and evaluation outcomes.

The relatively small dataset size and class imbalance further limit generalizability. Models trained on this dataset may not transfer reliably to other geographic regions, conflict contexts, or image sources with different visual characteristics.

There is also a risk of automation bias, where outputs from machine learning systems may be interpreted as objective or authoritative despite their inherent limitations. For this reason, the model should not be interpreted as a definitive tool for real-world humanitarian assessment but rather as an experimental research system intended for controlled academic evaluation.

We emphasize the importance of transparency, responsible reporting, and cautious deployment when working with sensitive visual data.

10. Bibliographical References

Alexei Abrahams, Shadi Abudalfa, Mustafa Jarrar, and George Mikros. 2026. The nakbaarchiveclassifier shared task on nakba image classification. In Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026), Palma, Mallorca, Spain.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778.

Mingxing Tan, and Quoc Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning, 6105–6114.

Marina Sokolova, and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. In Information Processing & Management 45 (4), 427–437.

Hoo-Chang Shin, Holger R. Roth, Mingchen Gao, Le Lu, Ziyue Xu, I. Nogues, Jianhua Yao, Daniel Mollura, and Ronald M. Summers. 2016. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. In *IEEE Transactions on Medical Imaging* 35 (5), 1285–1298.

Al-Warraq at AR-MS NAKBA-NLP 2026: Adapting Vision-Language and Transformer Models for Automatic Manuscript OCR/HTR

Sarah Al Hamed, Ahmad Edris Youssef, Aya Hafiz Faris, Alhasan Hamood,

Zainab Kamil, Jana Mukhles Alqasem, Sarah Ayad
Arab Open University, Riyadh, Saudi Arabia
24414598ksa@arabou.edu.sa, 23418335ksa@arabou.edu.sa, 22410231ksa@arabou.edu.sa,
23413390ksa@arabou.edu.sa, 24414598KSA@aou.edu.sa, 24414015KSA@aou.edu.sa,
s.ayad@arabou.edu.sa

Abstract

We present our submission to the NAKBA NLP 2026 Automatic Manuscript OCR/HTR shared task on Arabic manuscripts. The task aims to transcribe manuscript line images into machine-readable Arabic text. Our approach followed an iterative pipeline including model selection, training, error analysis, test-time augmentation, and postprocessing. After evaluating several OCR/HTR models, we selected and trained the most suitable model on the provided manuscript line images and transcriptions. Error analysis showed better character-level performance than word-level performance, which motivated the use of test-time augmentation and text cleaning to improve robustness. The final system achieved a CER of 0.1142 and a WER of 0.378, placing fifth in the shared task. These results show that simple but targeted improvements can support effective Arabic manuscript transcription.

Keywords: OCR, Transformer Model, Vision-Language Model

1. Introduction

Arabic manuscripts are an important part of cultural and scientific heritage. They preserve history, religion, medicine, literature, and many forms of knowledge written over long periods of time. However, much of this material still exists only as handwritten documents or scanned images, which makes access, search, and analysis difficult. Automatic transcription can help convert these manuscript images into machine-readable text and make them easier to preserve, index, and study. Recent work on historical Arabic manuscript datasets also highlights the importance of accurate OCR systems for improving access to these documents.

The goal of this subtask is to develop automatic systems that transcribe Arabic manuscript page images into text. This task is important, but it is also difficult. Arabic script is naturally cursive, and letter shapes change depending on their position in the word. Many letters differ only by dots, and handwritten manuscripts often contain irregular spacing, faded ink, noise, deformation, and inconsistent writing styles. These issues make Arabic manuscript transcription more difficult than modern printed OCR and also more difficult than many Latin-script settings. Recent Arabic HTR studies continue to describe segmentation, overlapping characters, and context dependency as core challenges.

Another challenge is that success at the character level does not always mean success at the word level. A system may correctly recognize many letters but still produce incorrect words because of missing spaces, merged words, or decoding mistakes. For this reason, evaluation in OCR and HTR commonly uses both Character Error Rate (CER) and Word Error Rate (WER), since they reflect different aspects of recognition quality. In manuscript settings, where handwriting is irregular and word boundaries are not always clear, this distinction becomes especially important.

In recent years, deep learning has changed the field of text recognition. Traditional pipelines often relied on hand-crafted features, segmentation rules, and multi-stage processing. In contrast, modern systems can learn directly from image-text pairs and perform end-to-end transcription. This progress is especially relevant for historical manuscripts, where visual variation is high and manual rule design is difficult. Transformer-based OCR and HTR models have shown strong results by learning visual and textual patterns jointly, and large vision-language models are now being explored for OCR and document understanding in multilingual settings.

In this paper, we focus on Arabic manuscript OCR/HTR under the shared task setting. Our work is motivated by the need for practical and robust transcription systems that can handle difficult handwritten line images. The approach combines model selection, training, recovery from technical issues, and simple inference-time improvement strategies. The broader aim is not only to improve recognition accuracy, but also to contribute to the digital preservation and accessibility of Arabic manuscript heritage.

2. Literature Review

Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) are closely

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 191–195
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

related, but they are not the same. OCR is usually used for printed or well-structured text, where character shapes are more regular. In contrast, HTR focuses on handwritten text, where writing styles, letter shapes, and spacing can vary greatly. For manuscript collections, HTR is the more suitable setting because the input is handwritten and of- ten historical. Recent surveys show that HTR has moved from rule-based and segmentation- dependent methods to end-to-end neural models that can work at word, line, and document level (Garrido-Munoz et al., 2025).

Arabic OCR and Arabic HTR are more difficult than many other text recognition tasks. Arabic script is cursive, character forms change according to their position in the word, and many letters differ only by dots or small marks. These properties make segmentation harder and increase confusion between visually similar characters. The task becomes even more difficult in historical manuscripts because of image degradation, faded ink, irregular spacing, writer variation, and non- standard orthography. For these reasons, recent work in Arabic handwritten recognition of- ten prefers line-based and end-to-end approaches instead of fragile character-level segmentation (Garrido-Munoz et al., 2025).

Deep learning has become the main direction in OCR and HTR research. Earlier systems often relied on separate stages such as pre-processing, segmentation, hand-crafted feature extraction, and decoding. More recent approaches replace these stages with neural architectures that learn directly from image-text pairs. A representative example is TrOCR, which uses a Transformer-based visual encoder and a Transformer-based text decoder in an end-to- end framework. TrOCR showed that pre-trained Transformer models can achieve strong results on both printed and handwritten text recognition tasks, while also reducing the need for manually designed OCR pipelines (Li et al., 2021).

This move toward pre-training has also connected OCR and HTR with multimodal vision-language modeling. Recent vision-language models can take an image as input and generate text directly, which makes them attractive for OCR in zero-shot and few-shot settings. Qwen2-VL is one example of this trend. It introduced dynamic resolution handling to better process images of different sizes and reported strong performance on several multimodal understanding tasks, including OCR-related benchmarks (Wang et al., 2024). Such models suggest that large pre-trained multimodal architectures may support multilingual OCR without requiring a fully task-specific design.

Another example is DeepSeek-VL, which was introduced as an open-source vision-language model for real-world multimodal understanding. It was designed to handle diverse visual inputs, including OCR-style content, charts, PDFs, and screenshots (Lu et al., 2024). This is relevant to OCR research because it shows that general-purpose multimodal models are increasingly capable of reading and reasoning over text in images. However, general vision-language models are not automatically sufficient for difficult handwrit- ten Arabic manuscripts. In such settings, image quality, handwriting variation, and rare word forms still require careful adaptation, decoding, and post- processing.

Overall, the literature highlights four main points. First, manuscript transcription should be treated mainly as an HTR task rather than a simple printed OCR task (Garrido-Munoz et al., 2025). Second, Arabic manuscripts remain challenging because of the script structure and the degraded nature of historical documents.

Third, deep learning, especially Transformer-based models such as TrOCR, has significantly improved OCR and HTR performance (Li et al., 2021). Fourth, multimodal pre-trained vision-language models such as Qwen2-VL and DeepSeek-VL are promising for OCR-related tasks, but they still need careful evaluation when applied to handwritten Arabic manuscript data (Wang et al., 2024; Lu et al., 2024).

3. Approach

Our approach focused on building an automatic Arabic manuscript transcription pipeline for the Systems Track. The main goal was to convert cropped manuscript line images into machine-readable Arabic text. Because Arabic manuscripts are visually complex and many existing OCR mod- els are not designed for this type of data, we followed an iterative approach based on model selection, training, error analysis, and inference-time improvement.

We started by testing several OCR and handwritten text recognition models. At the beginning, some of the selected models were not suitable for Arabic script. In particular, one early model was mainly designed for English text, so its performance on Arabic manuscript images was weak. The generated text contained many recognition errors, and the image understanding quality was limited. For this reason, the team compared multiple models and configurations before selecting a better model for Arabic transcription. In total, several candidate models were explored by different team members until a more suitable solution was found. After model selection, we trained the system on the provided manuscript line images and their gold transcriptions. The training process required a long time because the dataset contained around 16,000 training images, and the model needed repeated tuning to improve its ability to recognize Arabic characters and words. During the early stage, the system achieved only about 30% accuracy. However, after repeated experiments, continued training, and changing unsuitable

models, the performance improved significantly and reached about 80% recognition accuracy during development.

The training stage also involved several technical difficulties. Some team members had limited hardware resources, and training on large numbers of images was computationally expensive. In some cases, the training process lasted for many hours or even days. To reduce the risk of losing progress, we adapted the code so that it could save intermediate checkpoints and continue from previous training stages. This was important because interruptions such as power failure could stop the process and lead to loss of many hours of work. By using checkpoint-based recovery, we were able to continue training without restarting from the beginning each time.

Another important challenge appeared after training, when we tried to reload the trained model. A KeyError occurred because the checkpoint file contained only the trained weights, while the processor files were missing. These processor files included essential resources such as the tokenizer and dictionary. To solve this problem, we reused the original processor files from the base model and combined them with the trained weights. This recovery step allowed us to restore the model and continue the experiments successfully.

After obtaining a working model, we evaluated its outputs and analyzed the error types. The results showed an important difference between character-level and word-level performance. The Character Error Rate (CER) was relatively strong, while the Word Error Rate (WER) remained higher. This indicated that the model was often able to rec- ognize many characters correctly, but it still had problems producing correct full words. In particular, the model sometimes merged adjacent words, inserted extra spaces, or produced incorrect spacing before punctuation marks. Therefore, improving word-level quality became an important part of our approach.

To reduce prediction variance and improve robustness at inference time, we applied test-time augmentation (TTA). Each input line image was processed three times using small rotation variations: 0◦, +2◦, and −2◦. These small transformations helped the model handle slight differences in alignment and orientation. After that, a voting strategy was used to select the most frequent predicted text among the three outputs. This simple ensemble-style step helped reduce random recognition variation without the need for additional re- training.

In addition to TTA, we designed a post-processing step to clean the generated Arabic text. This step aimed to correct common formatting and orthographic issues that appeared in the predictions. The cleaning function removed unnecessary spaces before punctuation marks, normalized some Arabic letter forms, and corrected frequent output inconsistencies. It also attempted to separate merged words when possible by using simple Arabic text patterns. This post-processing stage was useful because even when the model recognized characters correctly, the final text still re- quired small corrections to become more readable and closer to the gold transcription.

Time efficiency was another major concern during the final submission stage. Running TTA on all test images increased inference time significantly. At one point, the estimated runtime was longer than the remaining time before the submission deadline. To solve this problem, we reduced the beam search width from 4 to 2 during decoding. This decision decreased the processing time substantially while still preserving acceptable output quality. As a result, the full inference process finished in time and the final submission was uploaded only a few minutes before the competition deadline.

Overall, our approach combined four main elements: careful model selection, repeated training and recovery under limited hardware conditions, inference-time augmentation, and Arabic text post- processing. This pipeline allowed us to improve the system step by step and handle both technical and linguistic challenges. Using this approach, the team achieved competitive results in the shared task, including a CER of 0.1142 and a WER of 0.378, which placed the team among the top-ranked submissions.

4. Discussion

The results show that our system was able to transcribe Arabic manuscript lines with good character-level performance. The final system achieved a CER of 0.1142 and a WER of 0.378, which placed our team in fifth position in the competition. These results show that the selected model was able to learn important visual patterns from the manuscript images and produce useful transcriptions as shown in Table 1.

However, the results also show a clear gap between character recognition and word recognition. The CER was much better than the WER. This means that the model was often able to recognize many characters correctly, but it still had difficulty producing correct full words. This happened because Arabic manuscripts contain connected writing, unclear spacing, and visually similar letters. In some cases, the model merged words, inserted wrong spaces, or produced small orthographic errors. Therefore, word-level prediction remained more difficult than character-level prediction.

To improve robustness, we applied test-time augmentation (TTA). Each image was processed three times using small rotations: 0°, $+2^{\circ}$, and -2°. Then, a voting mechanism selected the most

frequent output. This step helped reduce random variation in the predictions without retraining the model. It provided a simple way to improve stability during inference.

We also applied algorithmic post-processing to clean the generated text. This step normalized some Arabic forms, removed unnecessary spaces before punctuation, and corrected common output inconsistencies. It also helped separate merged words in some cases. This was important because the main weakness of the system was not always at the character level, but often at the word and spacing level.

A practical challenge appeared during the final inference stage. Running TTA on the full test set increased the total processing time, and the estimated runtime became longer than the remaining time before the submission deadline. To solve this problem, we reduced the beam search width from 4 to 2. This decision reduced the runtime from about 15 hours to about 7 hours, which allowed us to finish the full inference process and submit successfully only a few minutes before the deadline. This step shows that system design in shared tasks is not only about accuracy, but also about efficiency and completion under time constraints.

Overall, our work shows that Arabic manuscript OCR/HTR can benefit from a combination of model selection, repeated training, test-time augmentation, and text post-processing. The results are promising, but they also confirm that word-level recognition remains a major challenge in Arabic handwritten manuscripts. In future work, we plan to improve decoding, use stronger language-aware correction methods, and test more specialized Arabic manuscript models in order to reduce WER further.

Table 1: Final system performance and ranking.

Item	Value
Team Name	Al-Warraq
Final Rank	5
Character Error Rate (CER)	0.1142
Word Error Rate (WER)	0.3780
Line Error Rate (LER)	0.3468
Training Effort	100+ hours

5. Conclusion

In this work, we developed an Arabic manuscript OCR/HTR system for the shared task and explored different ways to improve its performance. During this process, we faced several challenges, including unsuitable models, long training time, hardware limitations, and the difficulty of recognizing Arabic handwritten text from unclear manuscript images. Despite these difficulties, we were able to build a working system through repeated model testing, training, inference-time improvement, and text post-processing.

Our results show that automatic transcription of Arabic manuscripts is possible, but it remains a difficult task, especially at the word level. The system achieved better character-level recognition than word-level recognition, which shows that Arabic manuscript OCR still needs stronger handling of spacing, word formation, and noisy handwriting. Even so, the final system achieved competitive results in the task and demonstrated that practical improvements can be achieved through careful experimentation.

This competition also highlights the importance of digital tools for preserving handwritten heritage. With the help of artificial intelligence, historical Arabic manuscripts can become more accessible, searchable, and usable for future research. In future work, we plan to test more specialized Arabic models and improve the text correction stage in order to further reduce recognition errors and improve transcription quality.

6. Acknowledgments

The authors would like to express their sincere gratitude to the organizers of the NAKBA NLP 2026 for providing this valuable and enriching experience. We also would like to thank Dr. Sarah Ayad for her continuous supervision, support, and guidance throughout this work. In addition, we express our appreciation to the Arab Open University for providing an encouraging academic environment that supported learning, research, and student participation in this competition.

7. Bibliographical References

Carlos Garrido-Munoz, Antonio Rios-Vila, and Jorge Calvo-Zaragoza. 2025. Handwritten text recognition: A survey. arXiv preprint arXiv:2502.08417.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yi- juan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2021. Trocr: Transformer-based optical character recognition with pre-trained models. arXiv preprint arXiv:2109.10282.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. 2024. Deepseek-vl: Towards real-world vision-language understanding. arXiv preprint arXiv:2403.05525.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2- vl: Enhancing vision-language model's percep-

tion of the world at any resolution. arXiv preprint arXiv:2409.12191

Sarah Al Hamed, Ahmad Edris Youssef, Aya Hafiz Faris, Alhasan Hamood, Zainab Kamil, Jana Mukhles Alqasem, Adapting Vision-Language and Transformer Models for Automatic Manuscript OCR/HTR of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC), 2026.

Misraj AI at AR-MS NAKBA-NLP 2026: A State-of-the-Art VLM in Arabic Handwritten Text Recognition

Khalil Hennara, Muhammad Hreden, Zeina Aldallal, Sara Chrouf, Safwan AlModhayan
Misraj AI
Khobar, Saudi Arabia
{hennara, hreden, aldallal, sara.chrouf, safwan}@misraj.ai

Abstract

Handwritten Text Recognition (HTR) for Arabic presents unique challenges due to the script's cursive nature, varying writer styles, and morphological complexity. While modern Vision-Language Models (VLMs) have significantly advanced document parsing, their direct application to highly specific cursive domains requires strategic adaptation. This paper details our submission to the Nakba OCR competition, which adapts a 3B-parameter VLM to recognize historical Arabic manuscripts. We employ a progressive training pipeline that utilizes domain-matched data augmentation to bridge the gap between standard printed Arabic OCR and historical handwritten manuscripts. Moving beyond standard decoder-only Supervised Fine-Tuning (SFT), we fine-tune the entire encoder-decoder architecture using differential learning rates. This approach, followed by a final checkpoint merge, allows the model to better resolve the fine visual details of cursive Arabic script. Our final unified model (submitted under the team name **Misraj AI**) establishes a new state-of-the-art (SOTA) on the Nakba dataset, achieving a Word Error Rate (WER) of 0.24 and a Character Error Rate (CER) of 0.08, and officially securing first place on the leaderboard.

Keywords: Arabic OCR, Handwritten Text Recognition, Vision-Language Models, Model Merging, Supervised Fine-Tuning

1. Introduction

Optical Character Recognition (OCR) for handwritten Arabic text presents unique challenges compared to printed text or Latin-based scripts. The cursive nature of Arabic, varying writer styles, and the presence of diacritics demand robust models capable of both precise visual feature extraction and deep linguistic understanding (Salaheldin Kasem et al., 2025; Lorigo and Govindaraju, 2006). Recently, Vision-Language Models (VLMs) have demonstrated state-of-the-art capabilities across a variety of multimodal tasks, moving beyond traditional machine learning architectures for OCR (Hennara et al., 2025; Team et al., 2025; Wu et al., 2025; Li et al., 2025; Comanici et al., 2025). In this work, we detail our submission to the Nakba OCR competition for Arabic handwritten text. We hypothesize that a pre-trained VLM, specifically tailored for Arabic text processing, can be effectively adapted for challenging handwritten domains through strategic data augmentation and progressive fine-tuning. Our contributions are threefold:

1. We demonstrate the efficacy of a two-stage Supervised Fine-Tuning (SFT) pipeline utilizing an Arabic handwritten dataset as an intermediary domain-adaptation step before fine-tuning on the Nakba dataset.
2. We show that unfreezing the vision encoder and training the full encoder-decoder architecture with differential learning rates yields a significant $> 5\%$ improvement over standard decoder-only tuning.
3. We apply checkpoint merging techniques to smooth the loss landscape, yielding our final state-of-the-art results (WER:0.24, CER:0.08) as shown in Table 1.
4. We open-source our adapted model weights [1] and inference pipeline[2] to support the broader research community and facilitate further advancements in handwritten Arabic OCR.

Team	CER	WER
Misraj AI	**0.079**	**0.244**
Oblevit	0.0925	0.3268
3reeq	0.0938	0.2996
Latent Narratives	0.105	0.3106
Al-Warraq	0.1142	0.378
Not Gemma	0.1217	0.3063
NAMAA-Qari	0.195	0.5194
Fahras	0.2269	0.5223
baseline	0.3683	0.6905

Table 1: Official Nakba Competition Leaderboard Results

2. Literature Review

The landscape of Optical Character Recognition (OCR) has undergone a paradigm shift in re-

[1] https://huggingface.co/Misraj/Baseer_Nakba

[2] https://github.com/misraj-ai/Nakba-pipeline

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 196–200
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

cent years, transitioning from traditional cascaded pipelines that rely on separate text detection, segmentation, and recognition modules to end-to-end multimodal architectures (Wei et al., 2024).

2.1. Vision-Language Models for Document AI

Vision-Language Models (VLMs) have recently set new benchmarks in document understanding by directly aligning visual features with autoregressive text generation (Wei et al., 2024). Frontier closed-source models, such as Gemini 2.5, have demonstrated remarkable zero-shot accuracy in reading dense text, interpreting complex spatial layouts, and extracting structured data without the need for traditional OCR engines (Comanici et al., 2025).

In the open-source domain, models like Qwen3-VL have significantly advanced the field by introducing dynamic resolution processing and robust spatial understanding, enabling the model to handle multi-lingual OCR tasks (Wu et al., 2025). Furthermore, specialized frameworks have been developed to handle the strict structural demands of document parsing. For instance, HunyuanOcr offers layout-preserving capabilities that faithfully convert complex PDFs into structured Markdown (Team et al., 2025). Similarly, models in the Monkey-OCR family employ advanced cross-modal alignment techniques specifically optimized for text-rich images, significantly reducing hallucination rates in document-based Question Answering (QA) (Li et al., 2025).

2.2. Advancements in Arabic OCR

Despite the massive leap in general VLM capabilities, handwritten Arabic text remains a notoriously difficult domain. (Salaheldin Kasem et al., 2025).

To address these challenges, a prominent line of work has converged on fine-tuning general-purpose vision-language models on targeted Arabic datasets, with each effort emphasizing a different facet of the problem. Some approaches prioritize diacritized text recognition and font diversity, as seen in Qari (Wasfy et al., 2025), while others focus on dialect-specific orthographic conventions, such as AtlasOCR (Imane Momayiz, 2025), which targets Moroccan Darija. Further along this spectrum, Baseer (Hennara et al., 2025) extends the scope beyond character-level transcription toward holistic Arabic Document Information Extraction, jointly addressing layout, context, and linguistic structure.

2.3. Positioning Our Work

Although models like Gemini 2.5 and Qwen3-VL provide immense general capabilities, there is a clear gap in applying modern VLM architectures to the highly specific domain of cursive Arabic Handwritten Text Recognition (HTR). Our work directly addresses this by adopting the Arabic-centric model as our foundation. Rather than relying on zero-shot capabilities, we demonstrate that a progressive, multi-stage fine-tuning curriculum incorporating domain-matched augmentation and advanced encoder-decoder tuning is essential to unlock SOTA performance on the Nakba dataset.

Metric	NTD	NDD	MD
Total Samples	15,962	2,095	21,196
Unique Characters	137	120	176
Avg. Text Length (Chars)	56.79	57.43	43.71
Mean Aspect Ratio	11.75	11.84	8.50

Table 2: Statistical comparison of datasets used for training.

3. Methodology

3.1. Data Acquisition

The development of our Arabic HTR system leveraged a combination of archival competition data and a large-scale enhancement dataset consisting of historical manuscripts.

3.1.1. Nakba Dataset (Competition Data)

The primary source of data is the Nakba dataset (**ND**), provided as part of the AR-MS NAKBA NLP 2026 shared task (Zaraket et al., 2026). See the shared task overview paper (Hamoud et al., 2026). This dataset consists of high-resolution line images extracted from the *Omar Al-Saleh* memoir collection. As an archival collection, it presents challenges typical of 20th-century historical documents, including varying ink density and age-related paper degradation.

The Nakba training dataset (**NTD**) contains 15,962 valid text samples, while the Nakba development set (**NDD**) includes 2,095 samples. The images are characterized by an extreme mean aspect ratio of approximately 11.75 (614×52 pixels).

3.1.2. Muharaf Dataset

To further enhance the model's ability to generalize across different cursive styles, we incorporated the Muharaf dataset (**MD**) (Saeed et al., 2025), a collection of 21,196 valid samples from historical manuscripts. Muharaf includes a wider variety of calligraphic styles (predominantly Ruq'ah). This external data provides critical exposure to older and more varied handwriting compared to the relatively consistent style of the **ND**. The statistical breakdown of *ND and MD* datasets is summarized in Table 2.

3.2. Data Pre-processing

A selective pre-processing strategy was applied to maintain computational efficiency while preserving the integrity of the competition data. While the original **ND** images were utilized in their provided format, the **MD was converted to grayscale**. This ensures that the supplementary data matches the tonal distribution and visual complexity of the competition set as illustrated in Figure 1.

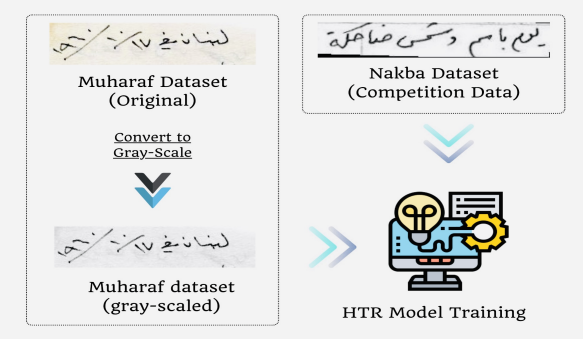


Figure 1: Visual comparison of the Nakba and Muharaf datasets, illustrating the grayscale preprocessing applied to the Muharaf enhancement data.

3.3. Proposed Method

Our methodology adapts a pre-trained Vision-Language Model (VLM) for cursive Arabic handwritten text recognition using a progressive pipeline.

3.3.1. Supervised Fine-Tuning (SFT) Approaches

Our core approach relies on adapting the VLM via standard autoregressive next-token prediction. Given an input image and a text prompt, we apply prompt masking so the cross-entropy loss is calculated exclusively on the generated transcription tokens. This guides the LLM decoder to map visual features to linguistic representations. We investigate two distinct SFT paradigms:

- **Decoder-Only SFT:** The vision tower is frozen, and only the text decoder weights are updated. This provides a computationally efficient mechanism for domain adaptation.
- **Full Encoder-Decoder Tuning:** To better align the vision encoder with the fluid nature of handwritten characters, we unfreeze the vision encoder. To prevent catastrophic forgetting of pre-trained visual representations, we employ *differential learning rates*, applying a significantly smaller learning rate to the vision encoder compared to the text decoder.

3.3.2. Checkpoint Merging

Drawing upon extensive prior research that demonstrates the efficacy of weight averaging in neural network optimization (Yang et al., 2026), we utilize checkpoint merging to maximize generalization. Studies have consistently shown that averaging the weights of multiple checkpoints smooths the model's functional loss landscape. By merging the weights of the top-performing epochs from our training runs, we construct a unified model that effectively mitigates local overfitting to the idiosyncrasies of the training data.

4. Experiments and Results

4.1. Implementation Details and Hyperparameters

All supervised experiments were conducted using the 3B-parameter Baseer (Hennara et al., 2025) across 2 NVIDIA H100 GPUs. To ensure fair comparison, training hyperparameters were standardized across all configurations. Models were trained for 5 epochs using the AdamW optimizer with a weight decay of 0.01 and a cosine learning rate schedule. We utilized a batch size of 128, a maximum sequence length of 1200 tokens, and an input image resolution of 644$\times$644 pixels. For decoder-only training, the text decoder learning rate was set to 1e-4. During full encoder-decoder tuning, the text decoder learning rate remained 1e-4, while the vision encoder utilized a reduced learning rate of 9e-6 to facilitate stable visual adaptation.

4.2. Zero-Shot Evaluation of General-Purpose Models

To establish a performance baseline and assess the out-of-the-box capabilities of frontier Vision-Language Models on this specialized task, we conducted a zero-shot trial using Gemini 3.1 Pro (Google, 2026). We evaluated the model's transcription accuracy directly on the *official development set*. The zero-shot approach yielded a Character Error Rate (CER) of 0.22 and a Word Error Rate (WER) of 0.47.

For comparison, our most basic supervised configuration training the 3B-parameter Baseer model for only three epochs exclusively on the Nakba Dataset (**ND**) achieved a CER of 0.19 and a WER of 0.42. Given that our simple, domain-specific SFT baseline comfortably outperformed a massive state-of-the-art general-purpose VLM, we concluded that the morphological complexity and cursive nature of Arabic handwritten text strictly necessitate dedicated fine-tuning. Consequently, we discontinued further zero-shot explorations with other proprietary models.

Architecture	Dataset / Curriculum	CER	WER
Decoder-Only	**ND**	0.19	0.42
	ND + **MD** (Joint)	0.18	0.42
	MD → **ND** (Sequential)	0.16	0.40
Encoder-Decoder	**MD** (Dec) → **ND** (Enc-Dec)	0.12	0.32
	MD (Enc-Dec) → **ND** (Enc-Dec)	0.10	0.29
Checkpoint Merging	**Encoder-decoder (Full) ep1 + ep5**	**0.08**	**0.24**

Table 3: Comprehensive Results of Training Strategies. *Note: **ND** refers to the Nakba Dataset, and **MD** refers to the Muhraf Dataset. All results reflect the performance on the official hidden test set.*

4.3. Decoder-Only SFT Evaluation

In our initial supervised experiments, we froze the vision encoder and tested three data curricula to evaluate the impact of domain-matched augmentation: (a) **ND** only, (b) **ND** and **MD** combined jointly, and (c) a sequential approach training first on **MD**, followed by **ND**.

As shown in Table 3, the data curriculum significantly impacts downstream performance. Training exclusively on **ND** established a baseline CER of 0.19 and WER of 0.42. Joint training yielded only a marginal improvement (CER: 0.18). However, adopting a sequential domain-adaptation approach (**MD** → **ND**) produced our strongest decoder-only results (CER: 0.16, WER: 0.40), confirming that pre-conditioning the language modeling head on structurally similar handwriting is highly beneficial.

4.4. Full Encoder-Decoder Tuning Evaluation

To better account for the structural variations of the script, we transitioned to full-parameter tuning. We initialized the model using the intermediate decoder-only checkpoint trained *exclusively* on **MD**. From this foundation, we unfroze the vision encoder and applied our differential learning rate strategy to train the full architecture on **ND**. We also tested applying this full encoder-decoder recipe sequentially from scratch (**MD** full-tuning → **ND** full-tuning).

This transition provided dramatic performance gains, proving that the vision tower requires domain-specific visual adaptation for cursive Arabic. Initializing with the **MD**-trained decoder and unfreezing the vision tower for **ND** training dropped the WER substantially to 0.32 and CER to 0.12. Applying the recipe sequentially from scratch further accelerated performance, pushing the metrics down to a CER of 0.10 and WER of 0.29 as show in Table 3.

4.5. Final Submission via Checkpoint Merging

During the evaluation of our best full encoder-decoder run on the hidden test set, we observed slight performance variance across different epochs. To stabilize and optimize our final predictions, we applied weight averaging to the Epoch 1 and Epoch 5 checkpoints, which individually exhibited the strongest blind test results.

As shown in Table 1, the results highlight the exceptional robustness of our progressive fine-tuning approach. Our model demonstrated a massive improvement over the competition baseline, reducing the CER by an absolute 28.9% (from 0.3683 down to 0.079) and the WER by 44.6% (from 0.6905 down to 0.244). Furthermore, our system significantly outperformed all other participating teams. We maintained a comfortable lead of over 5.5% in absolute WER compared to the next best submission (3reeq at 0.2996) and over 1.3% in absolute CER compared to the second-place CER score (Oblevit at 0.0925). This substantial performance gap underscores the necessity and effectiveness of combining domain-matched data augmentation with full encoder-decoder VLM tuning for cursive Arabic HTR.

5. Conclusion

In this paper, we presented a highly effective methodology for Arabic Handwritten Text Recognition (HTR) using Vision-Language Models. By leveraging the Baseer 3B model and implementing a multi-staged training curriculum incorporating domain-matched data from the Muharaf dataset, progressive full-parameter tuning with differential learning rates, and checkpoint merging, we achieved state-of-the-art results, securing top rankings in the Nakba OCR competition.

Our findings demonstrate that while frontier general-purpose models provide a strong baseline, the morphological complexities of cursive Arabic necessitate specialized fine-tuning for high-precision tasks. By meticulously engineering the supervised fine-tuning pipeline, we showed that VLMs can be robustly adapted to challenging historical document transcription, offering a stable solution for complex OCR applications.

6. Ethics Statement

All datasets utilized in this study were obtained through the official NAKBA NLP 2026 shared task and public research repositories, and were used strictly in accordance with their respective licensing agreements and ethical guidelines. Furthermore, we acknowledge that our model's capabilities are intrinsically tied to the specific calligraphic styles and historical periods present in the training distribution. As such, the system may exhibit representation biases if applied to documents from different demographic groups, regions, or eras. We strongly advocate for responsible deployment and emphasize the necessity of human-in-the-loop verification when applying this technology to broader historical archiving efforts.

Bibliographical References

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Google. 2026. Gemini (3.1 pro). Large language model.

Hadi Hamoud, Ahmad Ali Chamseddine, Bilal Shalash, Firas Ben Abid, Mustafa Jarrar, Chadi Abou Chakra, Bernard Ghanem, and Fadi A. Zaraket. 2026. Nakba nlp 2026: Shared task on arabic handwritten manuscript understanding (palestine memoryâ€“omar al-saleh memoir). In *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

Khalil Hennara, Muhammad Hreden, Mohamed Motasim Hamed, Ahmad Bastati, Zeina Aldallal, Sara Chrouf, and Safwan AlModhayan. 2025. Baseer: A vision-language model for arabic document-to-markdown ocr. *arXiv preprint arXiv:2509.18174*.

Abdeljalil Elmajjodi Haitame Bouanane Imane Momayiz, Soufiane Ait Elaouad. 2025. Atlasocr: Open-source ocr for moroccan darija with vision–language models. https://huggingface.co/atlasia/AtlasOCR.

Zhang Li, Yuliang Liu, Qiang Liu, Zhiyin Ma, Ziyang Zhang, Shuo Zhang, Zidun Guo, Jiarui Zhang, Xinyu Wang, and Xiang Bai. 2025. Monkeyocr: Document parsing with a structure-recognition-relation triplet paradigm. *arXiv preprint arXiv:2506.05218*.

Liana Lorigo and Venu Govindaraju. 2006. Offline arabic handwriting recognition: A survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28:712 – 724.

Mehreen Saeed, Adrian Chan, Anupam Mijar, Joseph Moukarzel, Georges Habchi, Carlos Younes, Amin Elias, Chau-Wai Wong, and Akram Khater. 2025. Muharaf: Manuscripts of handwritten arabic dataset for cursive text recognition.

Mahmoud Salaheldin Kasem, Mohamed Mahmoud, and Hyun-Soo Kang. 2025. Advancements and challenges in arabic optical character recognition: A comprehensive survey. *ACM Computing Surveys*, 58(4):1–37.

Hunyuan Vision Team, Pengyuan Lyu, Xingyu Wan, Gengluo Li, Shangpin Peng, Weinong Wang, Liang Wu, Huawen Shen, Yu Zhou, Canhui Tang, et al. 2025. Hunyuanocr technical report. *arXiv preprint arXiv:2511.19575*.

Ahmed Wasfy, Omer Nacar, Abdelakreem Elkhateb, Mahmoud Reda, Omar Elshehy, Adel Ammar, and Wadii Boulila. 2025. Qari-ocr: High-fidelity arabic text recognition through multimodal large language model adaptation.

Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, et al. 2024. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. 2025. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2026. Model merging in llms, mllms, and beyond: Methods, theories, applications, and opportunities. *ACM Computing Surveys*, 58(8):1–41.

Fadi Zaraket, Bilal Shalash, Hadi Hamoud, Ahmad Chamseddine, Firas Ben Abid, Mustafa Jarrar, Chadi Abou Chakra, and Bernard Ghanem. 2026. Ar-ms: Arabic manuscript understanding. In *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

PushingBoundaries at NakbaVirality Shared Task: Recursive Prompt Improvement for Multimodal Virality Classification

Ashhadul Islam[1], **Md Rafiul Biswas**[2], **Samir Brahim Belhaouari**[2], **Wajdi Zaghouani**[3]
[1]KTH Royal Institute of Technology, Stockholm, Sweden
[2]Hamad Bin Khalifa University, Doha, Qatar
[3]Northwestern University in Qatar, Doha, Qatar
aisla@kth.se, {mbiswas,sbelhaouari}@hbku.edu.qa, wajdi.zaghouani@northwestern.edu

Abstract

This paper describes our participation in the NakbaVirality shared task at the NakbaNLP Workshop (LREC–COLING 2026). We investigate Recursive Prompt Improvement (RPI), an instruction-level optimization strategy for virality classification in high-stakes geopolitical discourse. In this work, we propose a self-supervised approach to iteratively improve the classification prompt without human intervention. We begin with a basic prompt that guides the LLM to perform multi-class classification, incorporating contextual information about the tweets. After obtaining predictions, we identify misclassified tweets and feed them back to the model with an instruction to refine and improve the original classification prompt. This process is repeated over multiple iterations to assess whether performance improves over time. Our results show a remarkable improvement in F1 score from the first iteration to the final one. Although the proposed method does not reach the accuracy of models fine-tuned directly on task-specific data, it demonstrates that iterative, self-supervised prompt refinement can serve as a viable proxy for fine-tuning. By leveraging the model's own errors as feedback, this approach reduces reliance on computationally expensive training procedures and heavy GPU usage, while preserving much of the adaptability typically associated with fine-tuned models. This paradigm opens promising avenues for resource-efficient model adaptation and suggests new directions for scalable, low-cost performance improvement without traditional fine-tuning. The code has been made publicly available at GitHub repository.

Keywords: virality, social media, multimodal, LLM

1. Introduction

Social media platforms such as X, Facebook, and Instagram have become central arenas for information diffusion and narrative contestation during geopolitical conflicts (Hussain and Howard, 2013; Castells, 2015). In these contexts, virality is shaped by emotional intensity, polarization, and algorithmic amplification, and can directly influence public opinion and mobilization (Vosoughi et al., 2018; Stray et al., 2023; Kamin, 2019; Tufekci, 2015).

The Nakba and post–October 7th Gaza war discourse exemplify highly polarized, emotionally charged online communication, where moral-emotional language and identity signaling strongly drive engagement (Garimella et al., 2018; Barberá, 2015; Brady et al., 2017). Predicting virality in such settings therefore requires models that go beyond surface lexical features to capture emotional framing and stance.

Conflict discourse is also inherently multimodal, with images, symbols, and memes playing a key role in shaping interpretation and engagement (Highfield and Leaver, 2016; Alam et al., 2024). Prior work shows that integrating textual and visual signals improves social media understanding tasks (Zaghouani et al., 2025; Hasanain et al., 2024), yet multimodal virality prediction in sensitive geopolitical domains remains underexplored.

The NakbaVirality shared task (Ezzini et al., 2026) addresses this gap by challenging participants to predict reach and engagement for both textual and multimodal posts related to the Nakba and the post–October 7th war, fostering research on context-rich and polarizing data.

We participated in the NakbaVirality Shared Task, which focuses on predicting virality levels (low, medium, high) for posts in a high-stakes geopolitical discourse setting. We have seen in previous works LMMs and LLMs can be converted into classifiers by suitable prompting (Islam et al., 2023). However, to improve performance the prompt needs to be changed and tested multiple times manually. Our primary objective in this task was to examine whether *Recursive Prompt Improvement (RPI)*—an iterative system prompt refinement strategy—can serve as a viable alternative to conventional model fine-tuning. Rather than updating model parameters, we explored whether structured prompt evolution driven by misclassified examples could achieve competitive performance in a data-efficient and cost-effective manner.

- **Recursive Prompt Improvement (RPI):** We introduce an iterative prompt refinement framework that leverages misclassified examples to progressively improve task-specific system prompts without parameter updates.
- **Alternative to Fine-Tuning:** We empirically demonstrate that recursive prompt optimiza-

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 201–205
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

tion can substantially improve virality classification performance, highlighting its effectiveness as a lightweight alternative to traditional fine-tuning.

- **Cross-Model Prompt Transferability:** We show that prompts optimized using a smaller open-source model generalize effectively to larger and closed-source LLMs, suggesting that prompts can act as portable task adapters.
- **Cost-Efficient Optimization Pipeline:** We propose a practical workflow where a cheaper model is used for iterative prompt refinement, and the resulting optimized prompt is deployed on more powerful models, reducing computational and financial overhead.

2. Background

The NakbaVirality shared task (Ezzini et al., 2026) formulates virality prediction as a three-class classification problem, requiring systems to label posts as *Low*, *Medium*, or *High* based on a normalized engagement score derived from likes, shares/retweets, and comments. The dataset contains approximately 2,600 curated post–image pairs collected from X and Reddit after October 7, 2023, filtered using conflict-related keywords (e.g., "Gaza," "Nakba," "Palestine," "Israel"). The content reflects highly polarized, multilingual discourse, with the majority in English and Arabic.

Although the task emphasizes multimodal modeling, our work focuses primarily on a text-only setting to isolate the effect of **Recursive Prompt Improvement (RPI)**. We conducted experiments on roughly 1,600 tweet–image pairs, using only 100 samples for iterative prompt optimization and the remainder for validation and evaluation in a data-efficient setup.

Unlike prior approaches centered on parameter fine-tuning or supervised multimodal architectures (Ezzini et al., 2026), we propose RPI as a lightweight, model-agnostic alternative. By iteratively refining the system prompt using misclassified examples, we improve Macro-F1 without updating model weights and demonstrate cross-model transferability. This positions RPI as a computationally efficient strategy for adapting large language models to sensitive, high-stakes virality classification tasks.

3. System Overview

We are given 1,600 tweet–image pairs annotated with three virality classes: *Low (0)*, *Medium (1)*, and *High (2)*. Although the shared task supports multimodal inputs, our primary system focuses on the **text-only setting** in order to evaluate prompt optimization independently from visual reasoning.

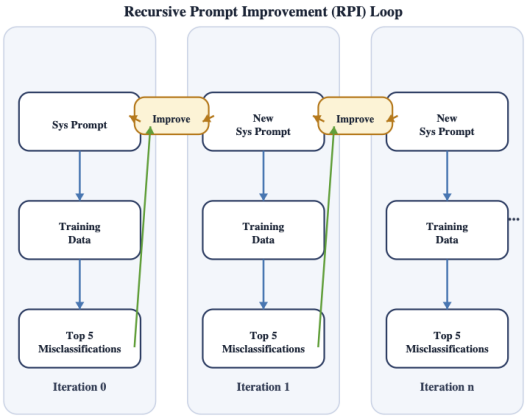


Figure 1: Recursive prompt flow Diagram

3.1. System Configurations

We evaluate three configurations:

- **System A (Baseline Prompt, Epoch 0):** A naïve fixed system prompt defining virality rules.
- **System B (Recursive Prompt Improvement):** Iteratively refined prompt using misclassification feedback.
- **System C (Cross-Model Transfer):** Applying the optimized prompt to different LLM architectures.

3.2. Recursive Prompt Improvement (RPI)

Instead of optimizing model parameters θ, we optimize the **system prompt** P.

Let:

$$D = \{(x_i, y_i)\}_{i=1}^{N}$$

where x_i is tweet text and $y_i \in \{0, 1, 2\}$ is the virality label.

Given model M and prompt P_t at iteration t:

$$\hat{y}_i = M(x_i; P_t)$$

We identify misclassified examples:

$$E_t = \{x_i \mid \hat{y}_i \neq y_i\}$$

The prompt update rule is:

$$P_{t+1} = \text{Improve}(P_t, \text{TopK}(E_t))$$

where the `Improve` function is implemented via a meta-prompt asking the model to revise classification instructions based on observed errors.

Figure 1 illustrates the Recursive Prompt Improvement (RPI) workflow as an iterative refinement process. In each iteration, the current system prompt generates virality predictions on the training subset, which are then compared with ground-truth labels to identify classification errors. Rather than updating model parameters, refinement occurs at the instruction level.

We select the top-5 misclassified tweets as diagnostic signals, highlighting weaknesses in the prompt formulation, such as ambiguous boundaries between Medium and High virality or insufficient emphasis on urgency cues. These examples are incorporated into a meta-prompt that revises the classification guidelines. The resulting updated system prompt is used in the next iteration.

This closed-loop process can continue for multiple rounds, progressively clarifying decision rules and improving separation between Low, Medium, and High virality classes without gradient-based fine-tuning.

Each iteration therefore consists of three structured stages:

1. **Prediction Stage:** Apply the current system prompt to generate virality labels.
2. **Error Analysis Stage:** Compare predictions with ground truth and extract the top-5 most informative misclassifications.
3. **Prompt Refinement Stage:** Update the system prompt using a meta-instruction informed by these errors.

Through this structured cycle, prompt engineering becomes a systematic optimization procedure rather than a one-shot manual design process.

4. Experimental Setup

We did not use any additional labeled datasets, external annotations, or handcrafted features beyond the data provided by the shared task. Our methodology is intentionally lightweight and relies primarily on large pre-trained language models and structured prompt design rather than parameter fine-tuning.

Our approach leverages:

- **Pre-trained LLMs:** DeepSeek-chat, OpenAI `gpt5.2`, and Qwen2.5-VL-7B-Instruct, accessed via the OpenRouter API. We rely on the models' pre-trained world knowledge and contextual reasoning capabilities for interpreting geopolitical and emotionally charged content.
- **Data partitioning strategy:** The dataset is divided into train/dev/test splits, with 100 samples used for iterative prompt optimization and the remaining data used for validation and testing.
- **Iterative refinement setup:** The recursive loop was executed for up to 30 epochs, with the best-performing prompt observed at epoch 7.
- **Computational environment:** Experiments were conducted using Google Colab Pro with an NVIDIA A100 GPU (40GB GPU RAM) and 80GB system RAM.
- **API usage and cost:** Running the closed-source model (`openai/gpt5.2`) for approximately 3,000 prompts across two epochs incurred an estimated cost of $15 USD.

5. Results

Figure 2 shows the Macro-F1 scores across iterations for both training and validation sets. The validation trend closely follows the training curve, indicating stable generalization during prompt refinement. Performance improves in the early iterations and reaches its maximum validation score at Epoch 7, after which gains plateau. This suggests that the recursive prompt optimization converges within a small number of iterations without evident overfitting.

It is instructive to compare the initial prompt with the final prompt obtained after iterative self-supervised refinement.

Initial Prompt

```
You are a tweet virality scor-
ing model.
Your job:  read the tweet
text and return a virality
score.
Virality score rules:
0 = low viral (ordinary,
niche, no hook, low engage-
ment potential)
1 = medium viral (interesting,
relatable, decent hook, moder-
ate engagement potential)
2 = high viral (strong hook,
highly shareable, controver-
sial, emotional, breaking
news, meme-worthy, strong
CTA)
```

Final Self-Supervised Prompt

```
You are a tweet virality scor-
ing model.  Your task:  read
a tweet and return a viral-
ity score (0=low, 1=medium,
```

```
2=high). Consider emotional
impact, public interest, and
potential for shares or dis-
cussions. Assign low scores
(0) to tweets that are short,
lack details, report routine
developments, or contain ex-
treme claims without evidence
of broad impact. Reserve
high scores (2) only for
tweets with urgent, highly
contentious elements likely
to spark widespread debate or
action. Be cautious with
tweets about ongoing con-
flicts or political issues;
default to lower scores un-
less they contain urgent
breaking news, major policy
shifts, or highly provocative
accusations with clear poten-
tial for mass discussion.
```

The refined prompt demonstrates a clearer operationalization of the scoring criteria. While the initial prompt provides broad qualitative descriptors (e.g., "strong hook" or "highly shareable"), the final prompt introduces more structured decision boundaries and explicit cautionary rules. In particular, it distinguishes between emotional intensity and verifiable public impact, discourages over-scoring routine or unsubstantiated claims, and provides nuanced handling of political or conflict-related content. Such refinements enable the model to better capture subtle contextual cues that would otherwise require extensive manual inspection. By encoding these nuanced distinctions directly into the prompt, the self-supervised process effectively internalizes criteria that a human annotator might take considerable time to apply consistently across large volumes of tweets.

5.1. Cross-Model Settings

The recursive prompt refinement procedure was first conducted using text-only LLMs: DeepSeek-chat (open-source) and OpenAI `gpt5.2` (closed-source) to assess cross-model generalization. This setup allows us to examine whether improvements achieved through recursive refinement on one model transfer effectively to other architectures. To further investigate multimodal adaptability, we conducted additional experiments using Qwen2.5-VL-7B-Instruct, a vision-language model. For the multimodal model, the prompt was minimally adapted to explicitly reference both textual and visual inputs (e.g., "You are a tweet and image virality scoring model"), while preserving the refined classification instructions learned during recursive optimization.

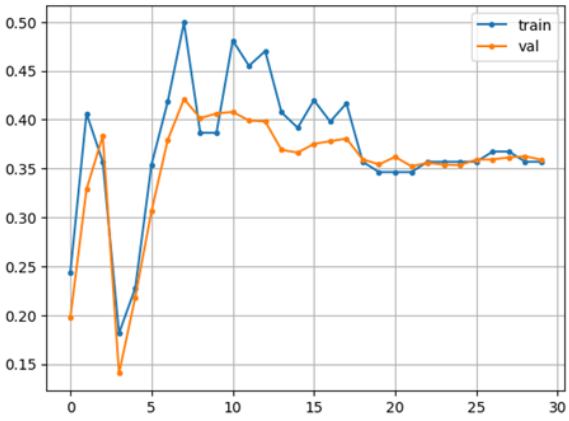


Figure 2: Measurement of F1-score over number of epoch

Model	Baseline	RPI	Gain
Deepseek-chat	0.20	**0.44**	**+120%**
Openai/gpt5.2	0.25	**0.42**	+68%
Qwen2.5-VL-7B-Instruct	0.20	**0.30**	+50%

Table 1: Relative improvement after Recursive Prompt Improvement (RPI).

Table 1 shows that the optimized prompt (Epoch 7) consistently outperforms the baseline configuration. DeepSeek exhibits the largest improvement (+0.24 absolute F1), indicating strong alignment between prompt refinement and model behavior. OpenAI `gpt5.2` also benefits substantially (+0.17), demonstrating that prompt-level optimization is transferable even to larger, closed-source systems.

5.2. Multimodal Settings

In the multimodal setting, Qwen2.5-VL-7B-Instruct shows a measurable improvement (+0.10), suggesting that the refined decision boundaries encoded in the optimized prompt remain beneficial when image information is introduced. Although the gain is smaller compared to text-only LLMs, the improvement confirms that Recursive Prompt Improvement produces architecture-agnostic refinements that generalize beyond the model used during optimization.

6. Conclusion

We investigated Recursive Prompt Improvement (RPI) as a parameter-free alternative to fine-tuning for virality classification. Iterative, error-driven prompt refinement consistently improved performance across multiple language models and demonstrated cross-model transferability. Several multimodal models behaved unstably: `mbzuai/ain` achieved low performance (F1 =

0.14), while Molmo and `idefics2` collapsed to single-class predictions. In contrast, text-focused LLMs were more stable and achieved stronger results, suggesting the task relies heavily on textual reasoning. Prompt optimization performed with a smaller open-source model transferred effectively to others. Notably, DeepSeek outperformed the larger closed-source `gpt5.2`, showing that scale alone does not ensure better performance under prompt-based refinement.

Acknowledgments

This work was funded by the NPRP Grant No. 14C-0916-210015 from the Qatar National Research Fund, part of the Qatar Research Development and Innovation Council (QRDI). The findings and conclusions expressed in this paper are solely those of the authors.

7. References

Firoj Alam, Md Rafiul Biswas, Uzair Shah, Wajdi Zaghouani, and Georgios Mikros. 2024. Propaganda to hate: A multimodal analysis of arabic memes with multi-agent llms. In *International Conference on Web Information Systems Engineering*, pages 380–390. Springer.

Pablo Barberá. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political analysis*, 23(1):76–91.

William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318.

M Castells. 2015. Networks of outrage and hope: Social movements in the internet age. john wiley & sons.

Saad Ezzini, Salima Lamsiyah, Shadi Abudalfa, Samir El-Amrany, and Walid Alsafadi. 2026. The nakbavirality shared task on multimodalvirality prediction in high-stakes discourse. In *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 world wide web conference*, pages 913–922.

Maram Hasanain, Md Arid Hasan, Fatema Ahmad, Reem Suwaileh, Md Rafiul Biswas, Wajdi Zaghouani, and Firoj Alam. 2024. Araieval shared task: propagandistic techniques detection in unimodal and multimodal arabic content. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 456–466.

Tim Highfield and Tama Leaver. 2016. Instagrammatics and digital methods: Studying visual social media, from selfies and gifs to memes and emoji. *Communication research and practice*, 2(1):47–62.

Muzammil M Hussain and Philip N Howard. 2013. What best explains successful protest cascades? icts and the fuzzy causes of the arab spring. *International studies review*, 15(1):48–66.

Ashhadul Islam, Md Rafiul Biswas, Wajdi Zaghouani, Samir Brahim Belhaouari, and Zubair Shah. 2023. Pushing boundaries: Exploring zero shot object classification with large multimodal models. In *2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 1–5. IEEE.

Julia Kamin. 2019. *Social Media and Information Polarization: Amplifying Echoes or Extremes?* Ph.D. thesis, University of Michigan.

Jonathan Stray, Ravi Iyer, and Helena Puig Larrauri. 2023. The algorithmic management of polarization and violence on social media. Technical report, Knight First Amendment Institute, Columbia University. KnightColumbia.org, forthcoming.

Zeynep Tufekci. 2015. Algorithmic harms beyond facebook and google: Emergent challenges of computational agency. *Colo. Tech. LJ*, 13:203.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

Wajdi Zaghouani, Md Rafiul Biswas, Mabrouka Bessghaier, Shimaa Ibrahim, George Mikros, Abul Hasnat, and Firoj Alam. 2025. Mahed shared task: Multimodal detection of hope and hate emotions in arabic content. In *Proceedings of the third arabic natural language processing conference: shared tasks*, pages 560–574.

U4RASD at StanceNakba Shared Task: Data Augmentation and Auxiliary Objectives for Arabic Stance Detection

Nancy Hamdan[1], **Aya Jouni**[1], **Aya Saïd**[2], **Fadi A. Zaraket**[1,2]
[1]Arab Center for Research and Policy Studies, [2]American University of Beirut
{nhamdan, ajouni, fzaraket}@dohainstitute.edu.qa; aas165@mail.aub.edu

Abstract

This paper describes a submission to Track B of the StanceNakba Shared Task on Arabic cross-topic stance detection in the political domain. We investigate LLM-based data augmentation, auxiliary training objectives including contrastive and multi-task learning, zero-shot prompting, and a preliminary terminology-based clustering approach. Our final system, based on MARBERTv2 with dialect-aware LLM-based augmentation, achieved 86% macro-F1 on the blind test set and ranked 3rd out of 10 teams. Our results show that dialect-aware augmentation substantially improved performance in a low-resource Arabic stance detection setting, while not all auxiliary objectives or clustering-based strategies yielded consistent gains. We release our code at https://acr.ps/1L9B9Tw.

Keywords: Arabic Stance Detection, Data Augmentation, Auxiliary Objectives

1. Introduction

Stance detection concerns the automated identification of the position expressed in an input text towards a specified target idea, object, or proposition (Somasundaran and Wiebe, 2010; Mohammad et al., 2016). Stance *favor*, *against*, and *neutral* labeling differs from sentiment analysis as it also concerns the relationship between the text and the specified target (Mohammad et al., 2016; Alturayeif et al., 2024). This makes stance nuanced and sensitive to *targets* where opinions may be implicit or sarcastic.

Stance detection plays an important role in analyzing public opinion and political polarization and can complement traditional surveys and polls (Küçük and Can, 2020). It can be formulated as either single-target, where one separate model is trained per target, or multi-target, where a unified model takes the target as part of its input and learns jointly across several targets (Sobhani et al., 2017; Xu et al., 2018).

This paper describes our submission to Track B of the StanceNakba Shared Task (Aldous et al., 2026), which focuses on Arabic cross-topic stance detection for the normalization with Israel (Topic 1) and refugee presence in Jordan (Topic 2) topics. We build upon the shared task baseline system and make the following contributions.

- We explore data augmentation strategies and identify an effective approach.
- We investigate auxiliary training objectives, including contrastive and multi-task learning.
- We conduct a preliminary study of terminology-based clustering using stance-specific embedding centroids.
- We evaluate zero-shot prompting as a training-free alternative.

Our best performing system, based on MARBERTv2 (Abdul-Mageed et al., 2021) with dialect-aware LLM-based data augmentation, achieved 88.7% macro-F1 on the validation set and 86% on the blind test set, ranking 3^{rd} out of 10 teams. Our results show that dialect-aware augmentation substantially improved performance, while not all augmentation, auxiliary objectives, or clustering-based strategies yielded consistent gains. We further analyze the interaction between contrastive learning and data augmentation using effective rank analysis.

2. Background: Setup and Dataset

Track B focuses on Arabic cross-topic stance detection. Given a social media post and an associated topic, the model predicts a *pro*, *against*, or *neutral* stance label.

The dataset, subset of MARASTA (Charfi et al., 2024), provides 1,205 Arabic posts covering **Topic 1:** (*Normalization with Israel*) and **Topic 2:** (*Refugee/Immigrant Presence in Jordan*). Table 1 presents example samples. Table 2 summarizes the distribution across the training and development subsets. The relatively small size of the dataset, combined with Arabic dialectal variation and rich morphology, increases model generalization difficulty.

3. Related Work

Stance detection was formalized in the target-specific setting in SemEval-2016 (Mohammad et al., 2016), and later extended to multi-target and multi-task settings (Sobhani et al., 2017; Xu et al., 2018; Li and Caragea, 2021).

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 206–211
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

ID	Sentence	Topic	Label
1	انا من فلسطين واتمنى أنهم العراقيين يجو لا عنا *I am from Palestine, and I hope Iraqis come to us.*	Topic 2	Pro
6	والله لو تصيحون الى ان تنشف حلوقكم مافيه تطبيع معكم *Even if you shout until your throats dry, there will be no normalization with you.*	Topic 1	Against
26	سفير الدولة في الأردن يزور المخيم الأردني للاجئين السوريين *The country's ambassador to Jordan visits the Jordanian camp for Syrian refugees.*	Topic 2	Neutral

Table 1: Sample excerpts of instances from the Track B training dataset.

Topic	Label	Train	Dev	Topic total
Norm. w/ Israel	Pro	120	26	491
	Against	139	29	
	Neutral	145	32	
Refugees in Jordan	Pro	166	36	533
	Against	159	34	
	Neutral	114	24	
Total		**843**	**181**	**1,024**

Table 2: Track B label distribution by topic for the training and development datasets.

Arabic stance detection remains relatively under-resourced, though recent efforts have expanded datasets and modeling approaches. The MAWQIF dataset (Alturayeif et al., 2022), used in StanceEval 2024 (Alturayeif et al., 2024), includes stance along with sentiment and sarcasm labels. The winning system incorporated multi-task learning and contrastive loss objectives (Badran et al., 2024). ArabicStanceX later introduced a larger multi-topic dataset for Arabic stance detection (Alkhathlan et al., 2025).

4. System Overview

Our system is based on fine-tuning an encoder LLM for stance detection. We explored four directions to improve performance: (1) data augmentation, (2) auxiliary training objectives, (3) terminology-based clustering, and (4) zero-shot prompting. We provide a concrete algorithmic walkthrough for each component.

4.1. Data Augmentation

We explored LLM-based augmentation strategies to increase data diversity, improve generalization and compensate for the limited size of the data.

Counterfactual Stance Augmentation We used Gemini 2.5 Flash Lite (Google, 2025b) to generate counterfactual variants of each training sample. For every instance labeled as *pro*, *against*, or *neutral*, the model generated two additional variants corresponding to the remaining stance labels while preserving the topic and applying minimal edits.

External Data Augmentation We incorporated data from an internal benchmark on politically sensitive topics.[1] For Topic 1, we used Palestine/Israel-related data, and for Topic 2, refugee/immigration-related data, then expanded them with Gemini 2.5 Flash Lite while preserving topic and stance labels.

Dialect-Aware Augmentation We used Gemini 3 Flash Preview (Google, 2025a) to generate paraphrased versions of each training sample. For each training instance, the model generated three paraphrases while strictly preserving both stance and topic. The prompt enforced dialectal consistency: dialectal samples remain dialectal, MSA samples remain MSA, and mixed samples remain mixed. This increased lexical diversity while preserving the original data distribution.

4.2. Auxiliary Training Objectives

In addition to data augmentation, we experimented with modifying the training objective.

Contrastive Learning We combined supervised contrastive loss using L2 distance (Hadsell et al., 2006) with cross-entropy loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{contrastive}}, \quad \lambda = 0.1.$$

The contrastive term encourages samples with the same stance to have closer embeddings while separating different stance classes.

Multi-Task Learning We explored multi-task learning (Ruder, 2017) using LLM-generated sarcasm and sentiment labels. We used Gemini 2.5 Flash Lite to produce the auxiliary annotations, and the model was trained with three classification heads using uncertainty-based loss weighting (Cipolla et al., 2018).

[1] https://acr.ps/1L9F2Dn

Hyper-parameter	Value
Maximum sequence length	512 tokens
Train batch size	32
Learning rate	2.0×10^{-5}
Number of epochs	10 (30 with augmented data)
Warm-up ratio	10%
Weight decay	0.01
Optimizer	AdamW

Table 3: Training hyper-parameters used for model fine-tuning.

Preliminary Terminology-Based Clustering aims to characterize the stance towards a topic with specific terms and entities. A domain expert could define stance-indicative terminology and use it to infer stance. We investigated an automatic construction alternative.

For each topic and stance label, we collected 10 representative articles and embedded them using Qwen3 8B embeddings (Zhang et al., 2025). We computed a centroid embedding for each stance group. Due to limited experimentation we excluded this promising method from our submission, and we plan to further explore it in future work.

4.3. Zero-Shot Prompting

We evaluated zero-shot prompting with Gemini 2.5 Flash Lite (Google, 2025b), prompting it to predict *pro*, *against*, or *neutral* given the topic while accounting for sarcasm and dialect.

5. Experimental Setup

In this section, we describe the data preprocessing, training configuration, and evaluation protocol we followed.

Data Preprocessing We applied standard Arabic normalization: removing diacritics, tatweel (ـ), URLs, mentions, and hashtag symbols; normalizing whitespace; and unifying common character variants such as alef (ا آ إ أ), alef maqsura (ى), and taa marbuta (ة).

Training Configuration All models were trained for 10 epochs, or 30 with augmented data, using a maximum sequence length of 512 and batch size 32. Full hyperparameters are listed in Table 3.

Models We trained *AraBERT* v0.2-base (Antoun et al., 2020) as the baseline provided by the organizers, and evaluated *MARBERT* v2 (Abdul-Mageed et al., 2021). We adopted *MARBERT*, as it outperformed *AraBERT*, in our subsequent experiments.

Models were trained using HuggingFace Transformers (Wolf et al., 2020) and we used PyTorch Metric Learning (Musgrave et al., 2020) for the contrastive loss function. For data augmentation and zero-shot prompting, we used the OpenRouter[2] platform. Appendix A shows the version numbers and URLs for all tools and libraries used.

Evaluation Models were trained on the official training set, with augmented samples when applicable, and evaluated using stratified 5-fold cross-validation on the training data and the full development set. We report macro-F1 as the primary metric, along with macro precision, macro recall, and accuracy. The model with the best development set performance was selected for our final submission.

6. Results

Table 4 reports performance on the training set using 5-fold cross-validation and on the development set. We report macro precision, recall, F1, and accuracy. Our submitted system achieved a macro-F1 score of 86.0 on the blind test set as shown in Table 5.

Backbone Model Comparison Replacing *AraBERT* with *MARBERT* improved macro-F1 by approximately 5.5 points on the development set. This is likely due to better domain alignment, as *MARBERT* was exclusively pretrained on Arabic tweets, similar to the shared task data.

Effect of Contrastive and Multi-Task Learning Adding contrastive loss improved *MARBERT* by about 1.1 macro-F1 points, while multi-task learning with LLM-generated sarcasm and sentiment labels reduced performance by about 1 point, likely due to noise in the auxiliary annotations.

Effect of Data Augmentation Counterfactual and external augmentation reduced performance by about 1 macro-F1 point relative to *MARBERT*. In contrast, dialect-aware augmentation yielded the best results, improving macro-F1 by about 5.4 points and corresponding to our submitted system. Combining it with contrastive loss reduced performance by 2.8 points.

Zero Shot Prompting Gemini 2.5 Flash Lite achieved a macro-F1 of 77.4 in the zero-shot setting, approaching the baseline fine-tuned model without task-specific training.

[2]https://openrouter.ai/

Table 4: Performance comparison across different stance detection configurations on the training set using 5-fold stratified cross-validation (CV), and on the development set (Dev). All metrics are Macro-averaged except for Accuracy. **Bold** indicates best performance; underline indicates second-best performance per metric and category. The final submitted system is highlighted.

ID	Description	Precision		Recall		Macro F1		Accuracy	
		CV	Dev	CV	Dev	CV	Dev	CV	Dev
Overall Performance									
1	Baseline (AraBERTv0.2)	.795	.783	.791	.779	.791	.780	.791	.779
2	MARBERTv2	.819	.833	.812	.834	.812	.833	.814	.834
3	MARBERTv2 + Contrastive Loss	.809	.848	.804	.843	.804	.844	.805	.845
4	MARBERTv2 + Multitask Learning	.781	.830	.770	.821	.772	.823	.773	.823
5	MARBERTv2 + Counterfactual + External Aug.	.977	.822	.976	.823	.977	.822	.976	.823
6	MARBERTv2 + Dialect-Aware Aug.	.949	.888	.948	.887	.948	**.887**	.948	.889
7	MARBERTv2 + Dialect-Aware Aug. + Contrastive Loss	.943	.861	.943	.859	.943	.859	.943	.861
8	Gemeni 2.5 Flash Lite + Zero Shot Prompting	-	.819	-	.769	-	.774	-	.773
9	Qwen 3 8B + Clustering	-	.340	-	.357	-	.307	-	.364
Topic 1: Normalization with Israel									
6	MARBERTv2 + Dialect-Aware Aug.	.927	.839	.926	.846	.926	**.841**	.925	.839
7	MARBERTv2 + Dialect-Aware Aug. + Contrastive Loss	.917	.795	.920	.798	.918	.797	.917	.793
Topic 2: Refugees and Immigrants in Jordan									
6	MARBERTv2 + Dialect-Aware Aug.	.970	.933	.968	.933	.969	**.933**	.969	.936
7	MARBERTv2 + Dialect-Aware Aug. + Contrastive Loss	.967	.930	.967	.920	.967	.924	.967	.925

System	Macro-F1 (Blind Test)
MARBERTv2 + Dialect-Aware Augmentation	.860

Table 5: Official blind test performance of our submitted system to track B in the StanceNakba shared task.

Terminology-Based Clustering The preliminary clustering approach performed substantially worse than supervised models, achieving 30.7 macro-F1 on the development set. This indicates that simple centroid-based clustering with limited seed documents is insufficient for reliable stance classification.

Topic Level Analysis Across configurations, performance was consistently lower on Topic 1 than Topic 2. The highest macro-F1 on Topic 1 was 84.1, compared to 93.3 on Topic 2.

Error Analysis Most errors for the submitted system on the development set occur between the *neutral* and *against* classes (7 neutral samples were predicted as against and 6 against samples as neutral), while *pro* is more reliably detected. For example, sarcastic or implicit posts such as "الاردن حيث البطاله المستفحله للاردنيين الحكومه تعد بتوفير 200الف فرصه عمل للاجئين السوريين؟!" ("Jordan, where unemployment is rampant among Jordanians, while the government promises 200,000 job opportunities for Syrian refugees?!"; neutral classified as against) are misclassified due to implicit stance expression. These cases are inherently ambiguous, and some *neutral*/*against* confusions may reflect plausible alternative interpretations of the stance.

Combining dialect-aware augmentation with contrastive loss degraded performance compared to augmentation alone. We hypothesize that this is related to dimensional collapse in contrastive learning (Jing et al., 2022). To examine this, we computed the effective rank of the learned representations using RankMe (Garrido et al., 2023). The submitted system achieved a RankMe score of 3.18, while the model trained with augmentation and contrastive loss scored 1.70, a 46.5% reduction. This suggests reduced diversity of learned representations when contrastive learning was combined with data augmentation.

7. Conclusion

We presented our submission to Track B of the StanceNakba Shared Task (Aldous et al., 2026). Our system based on MARBERTv2 with dialect-aware LLM-based augmentation ranked 3rd on the blind test set, showing that controlled dialect-aware augmentation improves performance, while auxiliary objectives did not consistently help. This work is limited by the small dataset size and its restriction to two topics, which may affect generalizability. In future work, we will evaluate the approach on broader datasets and further explore clustering-based methods.

8. Bibliographical References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Kholoud Khalil Aldous, Md Rafiul Biswas, Mabrouka Bessghaier, Shimaa Ibrahim, Kais Attia, and Wajdi Zaghouani. 2026. StanceNakba shared task: Actor and topic-aware stance detection in public discourse. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

Ali Alkhathlan, Faris Alahmadi, Faris Kateb, and Hend Al-Khalifa. 2025. Constructing and evaluating arabicstancex: a social media dataset for arabic stance detection. *Frontiers in Artificial Intelligence*, Volume 8 - 2025.

Nora Alturayeif, Hamzah Luqman, Zaid Alyafeai, and Asma Yamani. 2024. Stanceeval 2024: The first arabic stance detection shared task. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 774–782, Bangkok, Thailand. Association for Computational Linguistics.

Nora Saleh Alturayeif, Hamzah Abdullah Luqman, and Moataz Aly Kamaleldin Ahmed. 2022. Mawqif: A multi-label Arabic dataset for target-specific stance detection. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 174–184, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resources Association.

Mohamed Badran, Mo'men Hamdy, Marwan Torki, and Nagwa El-Makky. 2024. AlexUNLP-BH at StanceEval2024: Multiple contrastive losses ensemble strategy with multi-task learning for stance detection in Arabic. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 823–827, Bangkok, Thailand. Association for Computational Linguistics.

Anis Charfi, Mabrouka Ben-Sghaier, Andria Samy Raouf Atalla, Raghda Akasheh, Sara Al-Emadi, and Wajdi Zaghouani. 2024. Marasta: A multi-dialectal arabic cross-domain stance corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11060–11069.

Roberto Cipolla, Yarin Gal, and Alex Kendall. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7482–7491.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. 2023. Rankme: assessing the downstream performance of pretrained self-supervised representations by their rank. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Google. 2025a. Gemini 3 flash: frontier intelligence built for speed. Google Blog.

Google. 2025b. We're expanding our gemini 2.5 family of models. Google Blog.

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. 2022. Understanding dimensional collapse in contrastive self-supervised learning.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Comput. Surv.*, 53(1).

Yingjie Li and Cornelia Caragea. 2021. A multi-task learning framework for multi-target stance detection. In *Findings of ACL-IJCNLP*, pages 2320–2326.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Kevin Musgrave, Serge J. Belongie, and Ser-Nam Lim. 2020. Pytorch metric learning. *ArXiv*, abs/2008.09164.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557, Valencia, Spain. Association for Computational Linguistics.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124, Los Angeles, CA. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chang Xu, Cécile Paris, Surya Nepal, and Ross Sparks. 2018. Cross-target stance classification with self-attention networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 778–783, Melbourne, Australia. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models.

A. Tools and Libraries

Table 6 summarizes the tools and libraries we used in our experiments and their corresponding versions and URLs.

Tool / Library	Version	Usage
PyTorch	`2.10.0`	Deep learning framework
HuggingFace Transformers	`4.49.0`	Model training and fine-tuning
PyTorch Metric Learning	`2.9.0`	Contrastive loss implementation
OpenAI Python SDK	`2.16.0`	Client library used to access OpenRouter-hosted LLMs
OpenRouter API	`v1`	LLM provider endpoint used for data augmentation and zero-shot prompting

Table 6: Tools, libraries, and APIs used in the experiments, with versions and URLs.

AyahVerse at NakbaArchiveClassifier Shared Task: Architectural Trade-offs and Decision Calibration for Humanitarian Image Classification

Ibad-ur-Rehman Rashid, Akhtar Ali
Government Post Graduate College, Mansehra, Affiliated with Hazara University, Pakistan;
Northeastern University, China
ibad@gcm.edu.pk, ata@mails.neu.edu.cn

Abstract

This paper presents our submission to the Nakba-NLP 2026 Shared Task on binary image classification, where the goal is to categorize images of Gaza infrastructure as destroyed or intact. To address the challenges of class imbalance and resource-constrained deployment, we evaluated three convolutional architectures: ResNet50, MobileNetV2, and EfficientNet-B0, combined with a post-hoc threshold optimization step. Our results show that lightweight architectures are competitive with heavier models for this task, with EfficientNet-B0 achieving the highest Test F1-score of 0.85 despite having significantly fewer parameters than ResNet50. We further investigated the effect of input resolution, finding that increasing resolution improved ResNet50's performance, though it remained below lightweight alternatives. Finally, we demonstrate that shifting the binary decision threshold from the default 0.50 to an optimized 0.45 improved ResNet50's Test F1 from 0.79 to 0.81 by recovering recall for the minority destroyed class. Notably, this adjustment was only needed for ResNet50, while EfficientNet-B0 and MobileNetV2 performed best at the default 0.50, suggesting that larger models are more prone to majority-class bias. Overall, these results provide a systematic analysis of architectural efficiency and threshold behavior under class imbalance, offering practical insights for damage classification in resource-constrained crisis settings.

Keywords: Image Classification, EfficientNet, MobileNet, Edge Computing, Threshold Optimization, Humanitarian AI

1. Introduction

The Nakba Image Classification task(Abrahams et al., 2026) highlights the urgent need to categorize millions of images archived from social media by distinguishing destroyed infrastructure from intact environments. While academic competitions typically prioritize absolute accuracy through massive ensembles, practical deployment in humanitarian contexts, such as field hospitals or local NGO servers, often requires models that can run without high-end GPUs.

This paper presents the AyahVerse team's binary classification system. Instead of solely maximizing predictive metrics via scaling, we evaluate the trade-offs necessary for field-ready deployment. We structure our experiments around three simple questions:

1. How do lighter, mobile-friendly models (like MobileNetV2 and EfficientNet-B0) compare to standard, heavy models (like ResNet50) for this specific task?
2. Does using larger, higher-resolution images actually help the model spot damage, or does it just confuse the model with noisy details?
3. How much does adjusting the standard decision threshold help the model handle datasets where one class (intact buildings) heavily outnumbers the other (destroyed buildings)?

Our code is available at Github.Our code is available at Github.[1]

2. Background

2.1. Related Work

Our system design is informed by three intersecting areas of research: humanitarian image classification, efficient deep learning architectures, and decision calibration for imbalanced data.

Crisis Informatics and Image Classification

The automated analysis of social media imagery during crises has gained significant traction. Datasets like CrisisMMD (Alam et al., 2018) have supported research into categorizing disaster-related images to aid humanitarian response. However, much of the existing literature focuses on maximizing absolute accuracy using massive ensemble models (Nguyen et al., 2017). Our work pivots from this trend by prioritizing deployability in resource-constrained humanitarian contexts, specifically for the Nakba-NLP damage assessment task.

[1] https://github.com/Ebad-urRehman/nakba_archive_efficient_classification

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 212–216
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Efficient Architectures for Edge Deployment

While deep residual networks like ResNet (He et al., 2015) have historically dominated image classification, their computational overhead makes them unsuitable for edge deployment (e.g., field laptops or localized NGO servers). Recent advancements have focused on mobile-optimized architectures. MobileNetV2 (Sandler et al., 2019) introduced inverted residuals and linear bottlenecks to significantly reduce parameter counts. Similarly, EfficientNet (Tan and Le, 2020) proposed a compound scaling method that balances network depth, width, and resolution. We directly compare these architectures to determine the optimal trade-off between parameter count and damage-recognition capability.

Mitigating Class Imbalance

This dataset is highly imbalanced, as intact infrastructure heavily outnumbers destroyed infrastructure. Standard approaches to class imbalance include resampling techniques or cost-sensitive learning, such as Focal Loss (Lin et al., 2018), which reduces the loss contribution of easy, well-classified examples to focus training on hard cases. Beyond the loss function, shifting the binary decision threshold (threshold moving) is a proven technique to calibrate classifiers against asymmetric misclassification costs (Provost et al., 1998). We extend this by searching for the optimal threshold during the validation phase to improve minority-class recall without severely degrading precision.

2.2. Dataset

The shared task requires binary classification on a dataset of 2,001 Instagram images collected between October 2023 and December 2025. The data features a significant class imbalance: 706 images (35.3%) of direct conflict and destruction versus 1,295 (64.7%) intact scenes.

Class	Number of Images	Percentage
Destroyed Infrastructure	706	35.3%
Intact Infrastructure	1,295	64.7%
Total	**2,001**	**100%**

Table 1: Class distribution in the Nakba Image Classification dataset showing significant imbalance (35.3% destroyed vs. 64.7% intact).

3. System Overview

To mitigate the class imbalance inherent to the dataset, we adopt two complementary strategies: replacing standard binary cross-entropy with a customized Focal Loss ($\alpha = 1$, $\gamma = 2$) at training time, and applying post-hoc threshold optimization at inference time. Together, these address both the learning and prediction stages of the classification pipeline.

3.1. Data Preprocessing and Augmentation

All images were normalized using ImageNet mean and standard deviation values. During training, we applied random resized crops, horizontal flips, random rotation, and color jittering to improve generalization. All models were evaluated at a standard resolution of 224×224, with an additional variant ResNet50 tested at 334×334 to investigate the effect of higher resolution.

3.2. Architectural Pipelines

We established three distinct pre-trained pipelines to evaluate architectural trade-offs:

- **Heavyweight Baseline (ResNet50):** A 50-layer deep residual network ($\sim$25.5M parameters).
- **Lightweight Competitor (MobileNetV2):** A mobile-optimized architecture ($\sim$3.4M parameters).
- **Balanced Competitor (EfficientNet-B0):** A compound-scaling architecture ($\sim$5.3M parameters).

3.3. Decision Calibration

Standard binary classifiers apply a default decision boundary of $t = 0.5$ to the sigmoid probability output $p_i = P(\hat{y}_i = 1|x_i)$. However, in highly imbalanced disaster datasets, this static boundary often suppresses minority-class predictions. To account for the asymmetric costs of misclassifying destroyed infrastructure, we parameterize the predicted class label $\hat{y}_i(t)$ as a function of an adjustable threshold t:

$$\hat{y}_i(t) = \begin{cases} 1, & \text{if } p_i \geq t \\ 0, & \text{otherwise} \end{cases}$$

Rather than relying on the default, we implement a post-hoc calibration step. During the validation phase, we perform a grid search over a bounded interval $t \in [0.3, 0.75]$ to find the optimal threshold t^* that maximizes the Macro F1 score on the validation set (V):

$$t^* = \underset{t \in [0.3, 0.75]}{\arg\max} \left(\text{Macro F1}(y_V, \hat{y}(t)_V)\right)$$

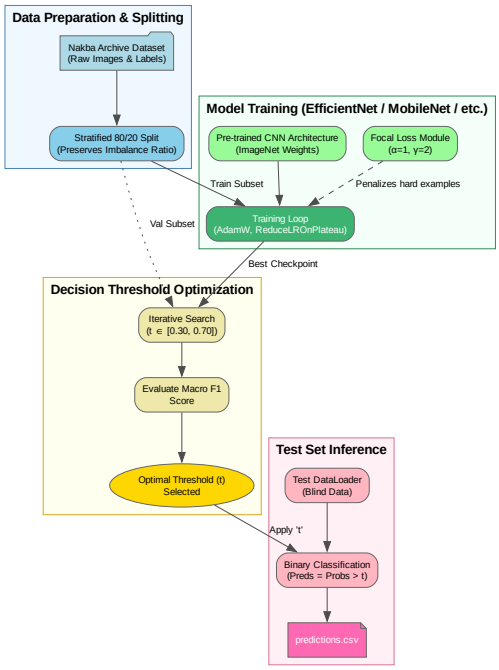


Figure 1: System architecture highlighting data splitting, EfficientNet-B0 training, and post-hoc threshold optimization.

While implemented as a lightweight grid search over $t \in [0.3, 0.75]$, this step is modular and can be applied to any binary classifier independently of the training procedure. The threshold yielding the highest validation metric is then frozen as the operational boundary for all final test set inferences as shown in Figure 3.

4. Experimental Setup

We combined the official training and validation sets provided by the shared task, which makes a total of 1,599 images, and performed an 80/20 stratified split to create our own training (1,279 images) and validation (320 images) sets for all experiments. All models were trained on a single T4 GPU using the AdamW optimizer (initial lr=1×10^{-4}), a ReduceLROnPlateau scheduler (patience=3), and the best checkpoint was saved based on validation Macro F1. All three architectures were evaluated under identical hyperparameter configurations, with only the model swapped between runs. Standard models processed 224×224 RGB imagery, while a secondary ResNet50 variation was scaled to 334×334 to test high-resolution sensitivity.

Dataset Split	Total Images
Combined (Train + Val)	1,599
Training Split (80%)	1,279
Validation Split (20%)	320

Table 2: Combined official Train+Val sets (1,599 images) with 80/20 stratified train-validation split for experiments.

5. Results and Analysis

Our quantitative evaluation revealed distinct performance dynamics across architectures and thresholds. Table 3 outlines the metric performance on the blind test set, where our best official submission achieved a Test Macro F1 of 0.85 using EfficientNet-B0.

Model Architecture	Resolution	Threshold (t)	Val Macro F1	Test Macro F1
ResNet50 (Baseline)	224x224	0.50	0.81	0.79
ResNet50 (High Res)	334x334	0.50	0.87	0.83
MobileNetV2	224x224	0.50	0.83	0.84
EfficientNet-B0	**224x224**	**0.50**	**0.89**	**0.85**
ResNet50 (Optimized)	224x224	0.45	0.82	0.81

Table 3: Test set evaluation across architectural, resolution, and threshold variations.

5.1. Architectural and Resolution Trade-offs

EfficientNet-B0 achieved the strongest performance in this task, achieving an F1-score of 0.85 while remaining significantly lighter than ResNet50. Surprisingly, MobileNetV2 (0.84) also outperformed the heavier ResNet50 baseline (0.79). This indicates that massive parameter depth may not be strictly necessary for identifying general disaster markers, as depthwise separable convolutions are capable of capturing structural damage patterns effectively.

Furthermore, increasing spatial resolution to 334×334 improved ResNet50's test performance to 0.83, though it still fell below MobileNetV2 (0.84) and EfficientNet-B0 (0.85) as shown in Table 3.

5.2. Threshold Optimization Impact

During validation, the standard $t = 0.5$ yielded an F1 of 0.81. Shifting the threshold to $t = 0.45$ yielded a marginal validation improvement (0.81 $\rightarrow$ 0.82), which nonetheless translated to a more meaningful gain on the blind test set (0.79 $\rightarrow$ 0.81). Notably, this adjustment was only necessary for ResNet50, as EfficientNet-B0 and MobileNetV2 achieved their best results at the default $t = 0.50$ despite the same class imbalance, suggesting that larger models are more susceptible to majority-class bias.

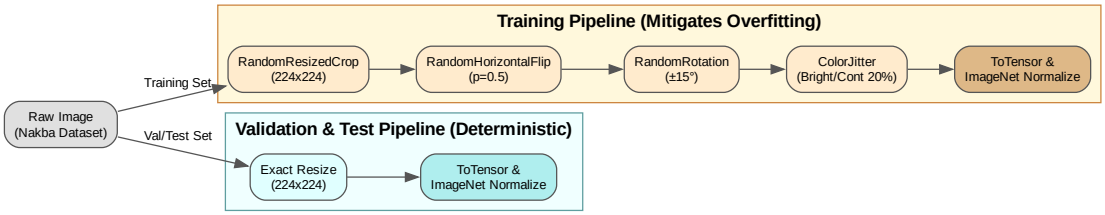


Figure 2: Data preparation and augmentation pipelines. The training set undergoes heavy spatial and color transformations to mitigate overfitting, while the validation and test sets follow a deterministic resizing and normalization path.

Threshold Optimization on Validation Set

Threshold Sweep Results	
Threshold (t)	**Macro F1**
0.35	0.7977
0.40	0.8039
0.45	**0.8214**
0.50	0.8197
0.55	0.7898
...	...

Default (0.5): 0.8197 F1

Optimal (0.45): 0.8214 F1

Improvement: +0.0017

Figure 3: Threshold optimized on validation set (t=0.45) and applied to test set, improving ResNet50 F1 from 0.79 to 0.81.

6. Conclusion

Our experiments demonstrate that lightweight architectures are not only computationally efficient but also more robust to class imbalance for this task. EfficientNet-B0 achieved the highest Test F1 of 0.85 while requiring far fewer parameters than ResNet50. Increasing ResNet50's input resolution improved test performance (0.79 $\rightarrow$ 0.83), though it remained below lightweight alternatives. Most notably, threshold calibration was only necessary for ResNet50, while EfficientNet-B0 and MobileNetV2 performed well at the default threshold, suggesting that larger models may require additional post-hoc calibration in imbalanced settings. These findings contribute a systematic comparison of architectural efficiency and threshold behavior under class imbalance, offering practical insights for future crisis classification systems deployed in resource-constrained environments.

7. Bibliographical References

Alexei Abrahams, Shadi Abudalfa, Mustafa Jarrar, and George Mikros. 2026. The nakbaarchiveclassifier shared task on nakba image classification. In Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026), Palma, Mallorca, Spain.

Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. Crisismmd: Multimodal twitter datasets from natural disasters.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. Focal loss for dense object detection.

Dat Nguyen, Kamela Ali Al Mannai, Shafiq Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. 2017. Robust classification of crisis-related data on social networks using convolutional neural networks. Proceedings of the International AAAI Conference on Web and Social Media, 11(1):632–635.

Foster J. Provost, Tom Fawcett, and Ron Kohavi. 1998. The case against accuracy estimation for comparing induction algorithms. In Proceedings of the Fifteenth International Conference on Machine Learning, ICML '98, page 445–453, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2019.

Mobilenetv2: Inverted residuals and linear bottlenecks.

Mingxing Tan and Quoc V. Le. 2020. Efficientnet: Rethinking model scaling for convolutional neural networks.

mlenthusiast at NakbaArchiveClassifier Shared Task: A Lightweight SVM-Gated Ensemble of EfficientNets for Image Classification

Md. Ajwad Hossain
Department of Electronics and Telecommunication Engineering
Chittagong University of Engineering & Technology (CUET)
Chattogram, Bangladesh
md.ajwadhossain@gmail.com

Abstract

Image classification under strict time constraints requires a delicate balance between feature complexity and computational overhead. This paper presents an optimized ensemble methodology developed for the NAKABA competition, focusing on identifying structural destruction. We propose a hybrid architecture that leverages two distinct Convolutional Neural Networks (EfficientNetB0 and EfficientNetB3) as base feature extractors, coupled with a Support Vector Machine (SVM) functioning as a meta-classifier. Instead of standard probability averaging or processing high-dimensional embeddings directly, the Meta-SVM acts as a learned gating mechanism to optimally combine the low-dimensional probability predictions of the base models. This ensures robust performance without the latency of heavier deep learning architectures. Empirical results demonstrate the efficacy of this approach. The model achieved a validation accuracy of 0.884 and a weighted F1-score of 0.885, with a notable F1-score of 0.839 on the challenging 'destruction' class. On the official NAKABA leaderboard test set, the ensemble maintained strong generalization, achieving an F1-score of 0.831 and an accuracy of 0.845, which secured the 12th position overall and proved the model's high effectiveness within the competition's strict operational constraints.

Keywords: Ensemble Learning, EfficientNet, Support Vector Machine, Stacking, Time-Constrained Inference

1. Introduction

Identifying critical states (e.g., complete structural "destruction") from photographs is a complex visual task. Importantly, predictions must be made under strict latency constraints. In real-world scenarios like post-disaster assessment or real-time structural health monitoring, models must deliver near-real-time inference on constrained hardware (Fan, 2024). In such settings, lightweight network architectures are essential, since they must execute rapidly on remote devices, such as drones, with limited computational capacity. This creates a critical trade-off: we require models with very high accuracy for safety-critical damage detection, alongside very low inference latency.

Standard deep neural networks (e.g., large ResNets or Transformers) can achieve high accuracy but typically incur prohibitive latency. Recent work has introduced model families that are far more compact: for example, EfficientNet models use compound scaling to balance depth, width, and resolution, yielding state-of-the-art accuracy with far fewer parameters. In particular, EfficientNet architectures achieve top-tier ImageNet accuracy while being significantly smaller and faster at inference than previous benchmarks (Tan and Le, 2019). Successive iterations further focus on parameter efficiency and training speed, producing models that execute even faster (Tan and Le, 2021).

Meanwhile, ensemble methods are known to improve robustness by combining multiple classifiers. As demonstrated by Dietterich (2000), aggregating diverse models frequently outperforms any single constituent model. In theory, larger ensembles yield better results, but naively ensembling multiple large Convolutional Neural Networks (CNNs) violates strict time budgets. Thus, we leverage ensembles strategically: we utilize multiple EfficientNet-based feature extractors (each relatively small and fast) and merge their outputs via a highly optimized, lightweight decision rule.

Recent advancements in ensemble fusion have explored geometric stacking methods, such as solving maximum-weight rectangle problems in high-dimensional spaces to separate classes (Wu et al., 2023, 2024). While effective, these can introduce unnecessary computational overhead. Instead, we propose an architecture that decouples heavy feature extraction from the final decision using a streamlined geometric approach. Multiple EfficientNet variants (B0 and B3) run in parallel to extract features, and a Support Vector Machine (SVM) (Cortes and Vapnik, 1995) meta-classifier combines their probabilistic outputs. By embedding the base-model predictions into a highly compressed, 4-dimensional space and solving for an optimal linear hyperplane, the meta-classifier effectively learns a robust gating mechanism. In summary, by combining efficient CNN feature extractors with a fast SVM meta-learner, our pipeline achieves high accuracy on the structural damage classification

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 217–220
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

task while strictly satisfying the competition's latency requirements.

2. Dataset and Task Description

The proposed ensemble architecture was developed and evaluated using the official dataset provided by the NAKABA competition (Abrahams et al., 2026). The primary objective of this task is the binary classification of structural images to rapidly identify critical damage.

2.1. Data Composition

The development dataset consists of a total of 2,001 labeled photographs categorized into two mutually exclusive classes: `not_destruction` (1,295 instances) and `destruction` (706 instances). This natural class imbalance reflects the realities of post-disaster structural assessments. To ensure robust model evaluation and tune our meta-classifier without overfitting, this available data was partitioned into:

- **Training Set:** 1,400 images utilized to train the base EfficientNet feature extractors and the Meta-SVM.
- **Validation Set:** 199 images used to evaluate local generalization and tune hyperparameters (e.g., the SVM regularization parameter C).

The final evaluation was conducted on an unseen hidden test set maintained by the competition organizers on the official leaderboard.

2.2. Preprocessing and Optimization

Given the variance in real-world structural photographs, a strict preprocessing pipeline was applied prior to feature extraction. To align with the compound scaling principles of the EfficientNet family (Tan and Le, 2019), architecture-specific input resolutions were utilized to maximize feature extraction efficiency and preserve pre-trained weight integrity.

Images processed by the EfficientNetB0 baseline were resized to 224×224 pixels, providing a lightweight input footprint for rapid foundational pattern recognition. Conversely, images processed by the EfficientNetB3 baseline were resized to 300×300 pixels, allowing the deeper network to capture more complex, fine-grained spatial features without distortion. Furthermore, all pixel values were normalized using standard ImageNet mean and standard deviation metrics to ensure stable convergence and consistent feature mappings across both parallel extraction branches.

3. Methodology

The core of the proposed architecture is a meta-learning ensemble utilizing a stacking mechanism to combine predictive signals. The pipeline is divided into feature extraction, base probability generation, and meta-classification.

3.1. Base Feature Extractors

We utilize two pre-trained models from the EfficientNet family (Tan and Le, 2019), chosen for their optimal scaling of depth, width, and resolution:

- **EfficientNetB0:** Serves as a highly lightweight, rapid feature extractor capable of capturing foundational visual patterns with minimal computational overhead.
- **EfficientNetB3:** A moderately larger variant that captures more complex, fine-grained spatial features without suffering from the extreme latency associated with heavier architectures.

In our pipeline, the final classification heads of both networks are bypassed. Instead, the models are used to output dense spatial feature embeddings (1280 and 1536 dimensions, respectively) for each input image.

3.2. SVM Meta-Classifier via Stacking

Instead of employing simple soft-voting or training an additional dense neural network layer on the high-dimensional embeddings—which would significantly increase inference latency and risk overfitting—we utilize a Support Vector Machine (SVM) (Cortes and Vapnik, 1995) in a stacked generalization architecture.

Figure 1 outlines this approach. To minimize latency, we do not concatenate the massive EfficientNet feature vectors directly. Instead, we train two independent Base SVMs that reduce these dense features into 2-dimensional probability vectors representing the likelihood of the 'destruction' and 'not_destruction' classes.

These outputs are then concatenated into a highly compact, 4-dimensional vector. Finally, a linear Meta-SVM ($C = 0.1$) evaluates this combined input. In this stacked configuration, the Meta-SVM effectively acts as a learned gating mechanism. By finding the optimal hyperplane in this highly compressed probability space, the meta-classifier learns exactly which base model's predictions to trust under varying conditions, ensuring the final classification is both highly accurate and exceptionally fast.

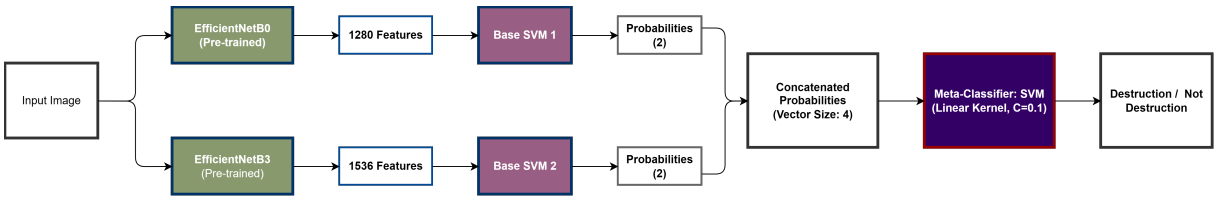


Figure 1: Proposed Hybrid Architecture: Input images are processed in parallel by EfficientNet variants to extract dense feature embeddings. These features are reduced to probability vectors by Base SVMs, which are then concatenated and classified by the final SVM meta-classifier.

3.3. Implementation Details and Hardware

To ensure complete reproducibility and optimal computational efficiency, all experiments were conducted within a Google Colaboratory (Colab) environment, accelerated by an NVIDIA T4 GPU.

A global random seed of 42 was strictly enforced across all computational libraries (including data partitioning, feature extraction, and SVM initialization). During the preliminary processing and training phases, the dataset was ingested using a batch size of 32.

4. Experiments and Results

4.1. Validation Performance

During local evaluation, the ensemble demonstrated marked improvements over the standalone models. The meta-classifier successfully filtered out individual model errors.

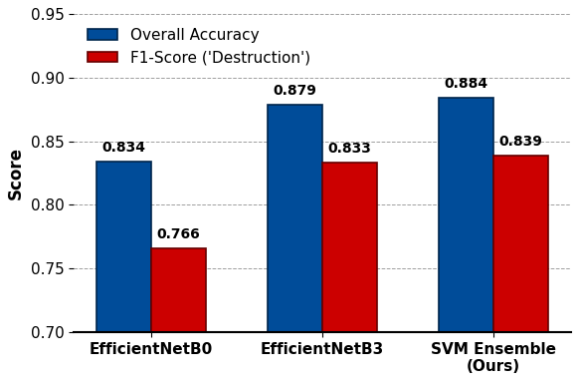


Figure 2: Performance Comparison: The ensemble model shows a distinct improvement in identifying the 'destruction' class compared to the standalone baselines.

The F1-score for the critical 'destruction' class improved to 0.839 compared to the best individual base model (EfficientNetB3), which achieved 0.833.

4.2. Leaderboard Test Performance

The ultimate validation of the time-constrained ensemble was its performance on the official hidden test set of the NAKABA competition. The model generated predictions efficiently and yielded highly competitive results on the leaderboard:

Metric	Score
Test Accuracy	0.84577
Test F1-Score	0.83123

Table 1: Official NAKABA Leaderboard Test Results

4.3. Discussion of Ensemble Effectiveness

The proposed ensemble demonstrated a distinct advantage over individual feature extractors by successfully leveraging their complementary strengths. While the EfficientNetB3 base model provided a strong foundation (validation accuracy of 0.879 and weighted F1-score of 0.880), it struggled slightly more with the critical 'destruction' class.

By extracting the output probabilities of both EfficientNetB0 and EfficientNetB3 and feeding them into an SVM meta-classifier—tuned with a linear kernel and a regularization parameter of $C = 0.1$—the architecture effectively learned to correct individual model errors. This meta-learning approach yielded a cross-validation weighted F1-score of 0.9900 during training, indicating a highly effective integration of the base models' probabilistic signals.

The final validation performance confirmed this advantage. The ensemble pushed the overall accuracy to 0.884 and achieved a highly balanced predictive capability across classes. Specifically, the F1-score for 'not_destruction' reached 0.910, while the performance on the more challenging 'destruction' class saw a meaningful improvement to 0.839. This confirms that simple linear combinations of probabilities via a meta-classifier can be exceptionally powerful when the base models capture distinct yet valuable spatial hierarchies.

5. Conclusion and Future Work

This study details an efficient, SVM-gated ensemble methodology tailored for the time-constrained NAKABA image classification challenge. By decoupling heavy feature extraction from the final decision-making layer, the architecture maximizes the utility of EfficientNet variants without incurring prohibitive latency.

The empirical results validate the robustness of this approach. The ensemble model not only surpassed the standalone base models—achieving an overall accuracy of 0.884 and a weighted F1-score of 0.885—but also provided a critical performance uplift in identifying structural destruction. The final system successfully generated rapid predictions for the unseen test dataset, yielding highly competitive leaderboard scores that secured the 12th position overall.

Future work will explore the integration of additional lightweight base extractors and the potential of non-linear gating mechanisms to further refine decision boundaries under strict operational constraints.

6. Limitations

While the proposed SVM-gated ensemble achieves high accuracy, its performance is fundamentally constrained by the diversity of the training data. The model was optimized specifically for the structural damage profiles present in the NAKABA dataset. Deploying this system to novel geographic regions or vastly different architectural styles may require domain adaptation or fine-tuning of the base EfficientNet feature extractors.

7. Acknowledgements

The author would like to express gratitude to the organizers of the NAKABA challenge for providing a rigorous and stimulating platform for applied machine learning.

8. Bibliographical References

Alexei Abrahams, Shadi Abudalfa, Mustafa Jarrar, and George Mikros. 2026. The nakbaarchive-classifier shared task on nakba image classification. In *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Thomas G. Dietterich. 2000. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ching-Lung Fan. 2024. Deep neural networks for automated damage classification in image-based visual data of reinforced concrete structures. *Heliyon*, 10(19):e38104.

Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.

Mingxing Tan and Quoc V. Le. 2021. Efficientnetv2: Smaller models and faster training. *CoRR*, abs/2104.00298.

Shengli Wu, Jinlong Li, and Weimin Ding. 2023. A geometric framework for multiclass ensemble classifiers. *Machine Learning*, 112:4929–4958.

Wei Wu, Liang Tang, Zhongjie Zhao, and Chung-Piaw Teo. 2024. Enhancing binary classification: A new stacking method via leveraging computational geometry.

Pixel at NakbaArchiveClassifier Shared Task: ConvNeXt-Based Ensemble for Destruction Detection

Rahaf Jaber
Independent Researcher
Al-Ahsa, Saudi Arabia
rahaf.fathi.jaber@gmail.com

Abstract

This paper describes our submission to the Nakba Image Classification Shared Task at the Nakba-NLP 2026 workshop. The task requires binary classification of social media images into two categories: *destruction* and *not_destruction*. The dataset includes approximately 1,600 annotated development images and 400 held-out test images, collected from Instagram posts published in Gaza between October 2023 and December 2025. High variability in viewpoint, lighting, and image quality, coupled with the inherent complexities of identifying structural damage in dense urban environments, makes this task particularly challenging. Our system utilizes a pretrained ConvNeXt-Tiny backbone fine-tuned through a stratified 5-fold cross-validation framework. To mitigate class imbalance, we implement a weighted cross-entropy loss function. During the inference phase, we employ an ensemble strategy that averages predictions across all five fold-specific models, and test-time augmentation (TTA) is applied to enhance robustness. The final ensemble achieved a Macro F1-score of 0.8952 and an accuracy of 0.9055 on the official test set. Our results suggest that the integration of modern convolutional architectures with robust ensembling and augmentation strategies provides a reliable baseline for automated destruction detection.

Keywords: image classification, destruction detection, ensemble learning, transfer learning, cross-validation

1. Introduction

The Nakba Image Classification Shared Task, organized as part of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), addresses the critical problem of automatically distinguishing images depicting destroyed buildings from those showing intact structures (Abrahams et al., 2026). As large-scale digital archives of conflict-zone imagery grow, manual categorization becomes increasingly infeasible. Automatic classification is therefore essential for improving the accessibility of these archives, supporting humanitarian documentation, and enabling downstream longitudinal analysis of structural damage, similar to efforts seen in remote sensing for disaster relief (Wang et al., 2024).

The dataset provided for this task is uniquely challenging, consisting of images collected from Instagram and published by Palestinian content creators and journalists in Gaza between October 2023 and December 2025. Unlike standardized benchmark datasets, these images are sourced from real-world social media environments characterized by extreme variability in lighting conditions, sensor quality, and camera viewpoints.

The task is formulated as a binary classification problem with two labels: *destruction* and *not_destruction*. Several factors complicate this distinction. First, background clutter, such as rubble-filled streets adjacent to intact buildings can confuse standard classifiers. Second, partial structural damage and occlusions caused by smoke or dust require the model to capture fine-grained visual features rather than relying on coarse global textures. Finally, the inherent class imbalance in archival data necessitates a robust training strategy to avoid bias toward the majority class (Johnson and Khoshgoftaar, 2019).

In this paper, we present a system based on a pretrained ConvNeXt-Tiny backbone, a modernized convolutional architecture that incorporates design principles from vision transformers (Liu et al., 2022). Our methodology emphasizes robustness through four key pillars:

- **Transfer Learning:** Leveraging high-quality feature hierarchies from ImageNet-1k pretraining (Deng et al., 2009).
- **Stratified Ensembling:** Using 5-fold cross-validation to ensure the model generalizes across different data distributions.
- **Class Balancing:** Implementing a weighted cross-entropy loss to penalize the misclassification of the destruction class more heavily.
- **Test-Time Augmentation (TTA):** Enhancing inference-time stability by averaging predictions across multiple image orientations, a technique proven to improve performance in noisy visual domains (Shanmugam et al., 2021).

Our final system achieves a Macro F1-score of 0.8952 and an accuracy of 0.9055 on the official test set. These results demonstrate that modern

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 221–225
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

convolutional architectures, when combined with careful ensembling and augmentation strategies, provide strong performance for destruction detection in complex social media imagery.

2. Related Work

Image classification has evolved substantially over the past decade, progressing from early deep convolutional neural networks (CNNs) to architectures that integrate design principles inspired by transformers. Residual connections introduced in ResNet enabled the effective training of very deep networks, establishing a strong foundation for large-scale visual recognition. More recently, research has explored incorporating transformer-style components, such as large receptive fields and modern normalization strategies, into convolutional frameworks.

ConvNeXt (Liu et al., 2022) represents a practical extension of this line of work. By modernizing standard convolutional blocks with larger kernels, inverted bottlenecks, and GELU activations, ConvNeXt narrows the performance gap between CNNs and Vision Transformers while preserving the spatial inductive biases inherent to convolutions. Pretraining on large-scale datasets such as ImageNet (Deng et al., 2009) further enhances its suitability for transfer learning on domain-specific tasks with limited annotated data.

2.1. Optimization, Regularization, and Ensembling

Fine-tuning pretrained models on moderate-sized datasets requires careful optimization. AdamW (Loshchilov and Hutter, 2017a), with its decoupled weight decay, has become a standard choice for stabilizing training and improving generalization. Learning rate scheduling strategies such as cosine annealing with warm restarts (Loshchilov and Hutter, 2017b) provide smooth decay and periodic exploration, which can be beneficial in short training cycles typical of shared tasks.

Ensemble learning has long been recognized as an effective technique for reducing model variance and improving generalization (Dietterich, 2000). Aggregating predictions from models trained on different data splits is particularly advantageous when the dataset size is limited or when input data exhibits high variability, as is common in social media imagery.

2.2. Destruction Detection in Crisis Contexts

Automated building damage classification has been widely studied in disaster response and post-conflict assessment, particularly using aerial and satellite imagery (Wang et al., 2024; Al Shafian and Hu, 2024). These approaches typically focus on distinguishing between intact and damaged structures using CNN-based models trained on relatively structured visual inputs.

In contrast, social media imagery presents additional challenges, including extreme viewpoint variation, motion blur, compression artifacts, and inconsistent framing. Research in crisis informatics suggests that robust training strategies and inference-time augmentation can help mitigate noise in user-generated content (Shanmugam et al., 2021). Furthermore, class imbalance is a recurring issue in archival datasets, where intact structures often outnumber damaged ones. Weighted loss functions and related strategies have been shown to improve balanced performance under such conditions (Johnson and Khoshgoftaar, 2019).

Our work builds upon these lines of research by combining a modernized convolutional backbone with stratified cross-validation and probability-level ensembling, tailored specifically to the variability of social media imagery in conflict-related archives.

3. System Description

This section describes our data preprocessing pipeline, model architecture, training strategy, and inference procedure. The architecture of our proposed system follows a robust pipeline consisting of high-resolution preprocessing, a modernized convolutional backbone, and a multi-stage ensemble strategy. A high-level overview of the inference architecture is illustrated in Figure 1.

3.1. Data Preparation and Augmentation

The dataset provided for this task consists of 2,001 high-resolution images sourced from social media. The distribution of classes is inherently imbalanced, reflecting the real-world nature of archival documentation: 706 images (35.3%) are labeled as *destruction*, describing various levels of structural damage, while 1,295 images (64.7%) are labeled as *not_destruction*, showing intact buildings or non-structural scenes.

We combine the provided training and validation splits into a single development pool to maximize the data available for learning while maintaining rigorous evaluation through stratified 5-fold cross-validation. Stratification is critical in this context to preserve the class ratios (*destruction* vs. *not_destruction*) across all folds, ensuring that the model's performance metrics are not skewed by distribution shifts.

All images are resized to 384×384 pixels to preserve fine-grained structural details necessary for

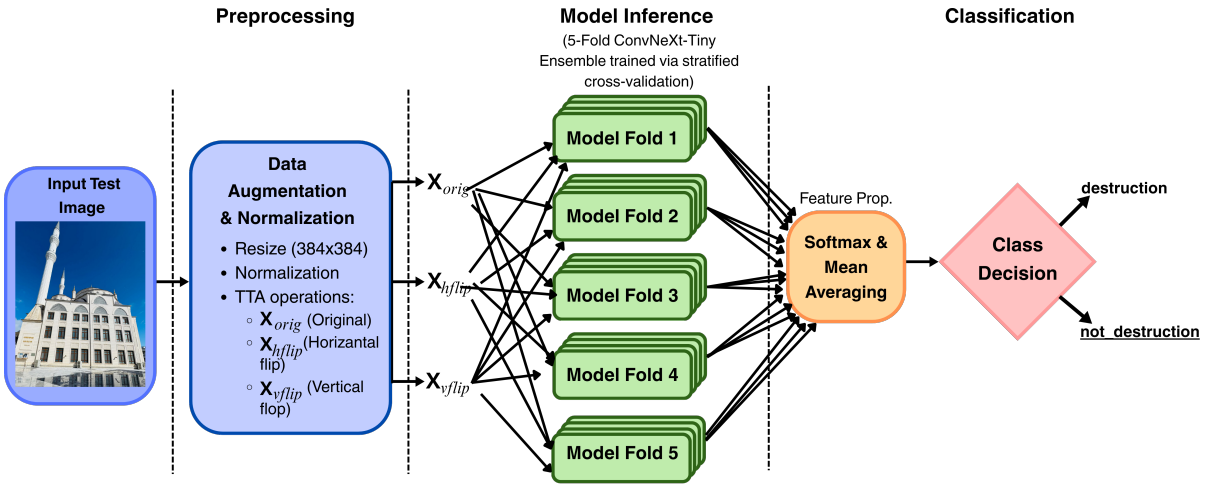


Figure 1: Inference framework of the proposed 5-fold ConvNeXt-Tiny ensemble with test-time augmentation (TTA). Augmented inputs are processed by fold-specific models, and predictions are combined through softmax-based mean averaging to produce the final class label.

damage detection. During the training phase, we apply a stochastic augmentation pipeline to improve generalization, including random horizontal flipping and random rotations. Images are normalized using the ImageNet-1k mean and standard deviation (Deng et al., 2009).

3.2. Backbone Architecture

We employ the ConvNeXt-Tiny architecture (Liu et al., 2022) as our primary feature extractor. ConvNeXt was selected due to its modernized design, which incorporates elements from Vision Transformers (ViTs), such as larger kernel sizes, inverted bottlenecks, and Layer Normalization, while retaining the inductive bias and computational efficiency of fully convolutional networks. We initialize the backbone with weights pretrained on ImageNet-22k and fine-tuned on ImageNet-1k. The final 1000 class linear head is replaced with a custom binary classification head consisting of a single linear layer producing two logits corresponding to the task labels.

3.3. Training Strategy

We train the model using stratified 5-fold cross-validation. For each fold, a separate model is trained and the best checkpoint is selected based on Macro F1-score on the validation split. Furthermore, the model is optimized using the AdamW optimizer (Loshchilov and Hutter, 2017a) with an initial learning rate of 3×10^{-5} and a weight decay coefficient of 0.05 to prevent overfitting on the relatively small dataset.

To mitigate the inherent class imbalance, we utilize a weighted cross-entropy loss function. Based on the class distribution in the training set, we assign a weight of 1.84 to the *destruction* class and 1.0 to *not_destruction*. This weighting scheme forces the model to prioritize sensitivity toward structural damage, which is the minority class in the provided archive.

3.4. Ensemble and Test-Time Augmentation

After training, we obtain five models corresponding to the five folds. During the inference phase, we adopt a dual-level ensemble strategy to maximize stability. First, we apply Test-Time Augmentation (TTA), generating three distinct views for every input image: the original view, a horizontal flip, and a vertical flip. Second, these three views are passed through all five fold-specific models (the best checkpoint from each fold). Let $P_{f,v}$ be the softmax probability vector for fold $f \in \{1 \dots 5\}$ and view $v \in \{1 \dots 3\}$. The final prediction $\hat{y}$ is computed by averaging the softmax probabilities across all 15 permutations (5 folds $\times$ 3 TTA views), with the final label determined by selecting the class with the highest mean probability:

$$\hat{y} = argmax \left(\frac{1}{15} \sum_{f=1}^{5} \sum_{v=1}^{3} P_{f,v} \right) \quad (1)$$

This strategy effectively reduces individual model variance and accounts for orientation differences in archival social media photography.

3.5. Implementation Details

The system is implemented using the PyTorch framework and the `timm` (PyTorch Image Models)

library. All experiments were conducted on a single NVIDIA T4 GPU using Google Colab. The code and pretrained weights will be made available upon acceptance.

4. Results and Discussion

The performance of the proposed ensemble system was evaluated using the official shared task metrics: Macro F1-score and Accuracy. Although Accuracy reflects overall correctness, model selection during cross-validation was based primarily on Macro F1-score, ensuring balanced consideration of both the *destruction* and *not_destruction* classes under potential class imbalance.

Table 1 reports the final performance on the official held-out test set.

Metric	Score
Macro F1-score	0.8952
Accuracy	0.9055

Table 1: Performance of the proposed system on the official test set.

The system achieves strong performance across both metrics, with a Macro F1-score of 0.8952 and an Accuracy of 0.9055. Performance remained consistent across folds, suggesting that the stratified 5-fold setup helped reduce variance due to data partitioning. Averaging predictions at the probability level further improved robustness on unseen data.

For completeness, we also evaluated two alternative model configurations. A single ConvNeXt-Tiny model trained on the original training/validation split achieved a Macro F1-score of 0.8864, highlighting the benefit of the stratified 5-fold ensemble in reducing variance and improving robustness. Another configuration incorporating AutoAugment and label smoothing yielded a Macro F1-score of 0.8904, slightly lower than the proposed approach. These comparisons emphasize that careful class weighting, cross-validation, and test-time augmentation are more effective than simply adding advanced augmentations or smoothing strategies.

Several factors likely contributed to the superior performance of the proposed system. First, the ConvNeXt-Tiny backbone effectively captures multi-scale structural patterns typical of damaged buildings. Its convolutional design preserves spatial inductive biases, which are beneficial for recognizing localized destruction cues such as fractured walls, exposed interiors, and scattered debris in 384×384 images.

Second, Test-Time Augmentation (TTA) improved prediction stability for images collected from mobile devices. Social media imagery often exhibits inconsistent framing and orientation; applying horizontal and vertical flips reduces sensitivity to these variations and encourages more consistent predictions.

Overall, the combination of transfer learning, stratified cross-validation, probability-level ensembling, and inference-time augmentation provides a reliable and reproducible baseline for destruction detection in social media archives, and clearly outperforms simpler or alternative training strategies.

5. Conclusion

This paper presented our submission to the Nakba Image Classification Shared Task at Nakba-NLP 2026. We developed a binary image classification system based on a fine-tuned ConvNeXt-Tiny backbone, trained using stratified 5-fold cross-validation and combined through probability-level ensembling. Test-time augmentation was applied during inference to improve robustness to orientation and framing variations commonly observed in social media imagery.

The final system achieved a Macro F1-score of 0.8952 and an Accuracy of 0.9055 on the official held-out test set. These results indicate that modern convolutional architectures, when paired with careful training and evaluation strategies, remain effective for destruction detection in noisy, real-world settings. In particular, class-weighted optimization and cross-validated ensembling played an important role in maintaining balanced performance across both classes.

Overall, the proposed approach provides a reproducible and competitive baseline for automated destruction classification in archival social media collections.

6. Limitations and Future Work

Despite strong performance, several limitations remain. First, the dataset size is relatively modest, which may limit generalization to rare or atypical visual patterns of structural damage. The model may struggle with subtle cases where damage is partially visible or visually ambiguous. Second, images collected from social media vary substantially in resolution, compression artifacts, and viewpoint, which can introduce unpredictable failure cases.

In addition, the current system is purely visual. It does not leverage accompanying textual captions, hashtags, or temporal metadata, which may provide complementary contextual signals. Incorporating multimodal information could improve robustness, particularly in borderline cases.

Finally, the task formulation is binary and does not capture varying degrees of structural damage. Extending the framework to multi-class or regression-based damage severity estimation

would allow for more fine-grained analysis and may better reflect the complexity of real-world destruction patterns.

Future work will therefore focus on multimodal fusion strategies and on evaluating generalization across datasets collected from different time periods or geographic regions.

7. Ethical Considerations

The dataset used in this study consists of social media imagery documenting structural destruction in Gaza between 2023 and 2025 (Abrahams et al., 2026). The dataset documents structural destruction during an ongoing conflict, and we recognize the sensitivity of working with such material. As the workshop guidelines require, our work focuses on the technical task of automated structural destruction classification to support documentation and archival analysis. We have taken care to ensure that our model operates only at the image classification level and does not extract or infer identifiable personal information. Furthermore, we acknowledge the potential for dual-use of such technologies; while our goal is humanitarian documentation and archival accessibility, automated damage detection systems must be used responsibly to avoid the mischaracterization of civilian infrastructure.

8. Conflict of Interest

The author declares that they have no financial or personal relationships with other people or organizations that could inappropriately influence or bias the results and interpretations presented in this paper.

9. Acknowledgements

The author would like to thank the organizers of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) for providing the dataset and establishing this shared task. We also express our gratitude to the Palestinian journalists and content creators whose documentation work on the ground made this research possible.

10. Bibliographical References

Sultan Al Shafian and Da Hu. 2024. Integrating machine learning and remote sensing in disaster management: A decadal review of post-disaster building damage assessment. *Buildings*, 14(8):2344.

Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.

Justin M Johnson and Taghi M Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of big data*, 6(1):27.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ilya Loshchilov and Frank Hutter. 2017a. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ilya Loshchilov and Frank Hutter. 2017b. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*.

Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. 2021. Better aggregation in test-time augmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1214–1223.

Lili Wang, Jidong Wu, Youtian Yang, Rumei Tang, and Ru Ya. 2024. Deep learning models for hazard-damaged building detection using remote sensing datasets: A comprehensive review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:15301–15318.

11. Language Resource References

Alexei Abrahams, Shadi Abudalfa, Mustafa Jarrar, and George Mikros. 2026. The nakbaarchive-classifier shared task on nakba image classification. In *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

MennaAly at NakbaArchiveClassifier Shared Task: Transfer Learning with ResNet for Historical Image Classification

Menna Aly
Hamad Bin Khalifa University
Doha, Qatar
meal89637@hbku.edu.qa

Abstract

This paper describes our submission to the NakbaArchiveClassifier shared task at Nakba-NLP 2026, co-located with LREC 2026. The task consists of binary image classification, where a model must classify historical images into one of two categories: destruction or not_destruction. We adopt a transfer learning approach based on pretrained residual networks, fine-tuned on the provided training data. To mitigate class imbalance, we incorporate weighted cross-entropy loss during optimization. In the development phase, our ResNet18 model achieved a peak macro F1-score of 0.8137 on the validation set. For the final phase, we trained on the combined training and validation data (1,599 labeled images) and generated predictions for the hidden test set of 402 images. Our final submission achieved a macro F1-score of 0.83228 with an accuracy of 0.84577 on the official evaluation set. These results demonstrate that compact residual architectures can achieve strong performance in historical image classification under limited-data conditions without complex architectural modifications.

Keywords: binary image classification, transfer learning, residual networks, convolutional neural networks, historical image analysis

1. Introduction

The NakbaArchiveClassifier shared task aims to automatically classify historical images related to the Nakba archive into two categories: *destruction* and *not_destruction* (Abrahams et al., 2026). Such classification supports digital archiving, searchability, and historical analysis. However, the task presents challenges including limited labeled data, visual variability, and class imbalance.

In this work, we adopt a transfer learning strategy based on pretrained convolutional neural networks. Instead of training from scratch, we fine-tune ImageNet-pretrained ResNet models, which generalize well across domains (Deng et al., 2009; He et al., 2016). Our goal is not to design a complex architecture but to evaluate how far a carefully tuned baseline can go under shared task constraints.

This work provides three main contributions: (1) a transfer learning baseline using residual networks, (2) a comparison of model depth and training duration under low-resource conditions, and (3) empirical observations about overfitting behavior in historical image classification.

2. Related Work

Deep convolutional neural networks have become the standard approach for image classification. Early breakthroughs such as AlexNet demonstrated the effectiveness of deep learning for large-scale visual recognition (Krizhevsky et al., 2012). Large annotated datasets such as ImageNet have further enabled pretrained models to learn generalizable representations that transfer effectively across downstream tasks (Deng et al., 2009).

Residual Networks (ResNets) introduced by He et al. (2016) allow training deeper architectures through residual connections that mitigate vanishing gradients. Transfer learning has also been widely studied as a strong approach in scenarios with limited labeled data (Pan and Yang, 2010). In historical and archival image analysis, where annotated data is often scarce, fine-tuning pretrained convolutional neural networks (CNNs) offers a reliable and computationally efficient baseline.

3. Task and Dataset

The development dataset consists of 1,400 training images and 199 validation images. The final evaluation dataset includes 402 test images with hidden labels (Abrahams et al., 2026). The task requires classifying each image into one of two categories: *destruction* or *not_destruction*.

The development data is imbalanced, with 906 *not_destruction* samples and 494 *destruction* samples in the training split. This imbalance is addressed during training using class-weighted loss, as described in Section 4.3.

4. Methodology

4.1. Preprocessing

All images were resized to 224×224 pixels to match the input requirements of ResNet architectures. We applied standard ImageNet normalization us-

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 226–228
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

ing mean values $[0.485, 0.456, 0.406]$ and standard deviations $[0.229, 0.224, 0.225]$.

No additional data augmentation was applied during the development phase, allowing us to evaluate the effectiveness of transfer learning without augmentation-based variability.

4.2. Model Architecture

We adopt pretrained convolutional neural networks based on the ResNet architecture (He et al., 2016). Specifically, we experiment with ResNet18 and ResNet34 models initialized with ImageNet weights.

The original fully connected layer, which outputs 1000 classes, was replaced with a linear layer of size 2 to perform binary classification. Transfer learning was selected due to the limited dataset size, which makes training deep networks from scratch impractical. Fine-tuning pretrained weights enables the model to leverage general visual representations learned from large-scale datasets.

4.3. Handling Class Imbalance

The development dataset contains more *not_destruction* samples than *destruction* samples. To mitigate bias toward the majority class, we compute class weights inversely proportional to class frequencies and incorporate them into a weighted cross-entropy loss function. This encourages stronger penalization of misclassification errors on the minority class.

4.4. Training Setup

Models were trained using the Adam optimizer with a learning rate of $1e^{-4}$ and a batch size of 32. During the development phase, training was conducted for five epochs.

Performance was evaluated using macro F1-score and accuracy. All experiments were implemented in PyTorch and executed in a Google Colab environment.

5. Experiments and Results

During the development phase, we trained ResNet18 for five epochs and evaluated performance on the validation split using accuracy and macro F1-score.

The best macro F1-score (0.8137) was achieved at epoch 3. Performance slightly decreased afterward, indicating mild overfitting. This configuration achieved competitive performance in the development phase.

For the final phase, we combined the training and validation splits (1,599 images total) and trained both ResNet18 and ResNet34 models. We observed that shorter training schedules improved generalization performance on the hidden test set.

Epoch	Train Loss	Val Acc	Val Macro F1
1	0.4881	0.7990	0.7824
2	0.2037	0.7638	0.7558
3	0.0797	0.8342	**0.8137**
4	0.0569	0.8291	0.8087
5	0.0491	0.7839	0.7653

Table 1: Development phase validation performance of ResNet18.

Model	Epochs	Final Macro F1	Final Accuracy
ResNet18	2	**0.83228**	0.84577
ResNet34	6	0.82000	—

Table 2: Final phase performance comparison.

The best-performing configuration was ResNet18 trained for two epochs, achieving a macro F1-score of 0.83228 and accuracy of 0.84577. Although ResNet34 achieved low training loss, it did not outperform the smaller ResNet18 architecture, suggesting mild overfitting when training deeper models on the limited dataset.

6. Discussion

The results indicate that compact residual networks achieve competitive performance in low-resource historical image classification tasks. ResNet18 converged quickly and demonstrated strong generalization, particularly when trained for a limited number of epochs.

Longer training schedules and deeper architectures did not consistently improve macro F1 performance, highlighting the importance of controlling overfitting in small datasets. These findings suggest that carefully tuned transfer learning baselines can be highly effective without complex architectural modifications. The results also suggest that shorter training schedules can function as an implicit regularization mechanism in low-data settings.

7. Limitations

This work did not incorporate data augmentation, learning rate scheduling, or systematic hyperparameter search. Additionally, cross-validation was not performed. Future work could explore augmentation strategies, more extensive hyperparameter tuning, and ensemble approaches.

8. Conclusion

We presented a transfer learning-based system using ResNet architectures for the NakbaArchive-Classifier shared task. Our approach relied on pre-trained residual networks, weighted loss functions to address class imbalance, and minimal architectural modifications.

ResNet18 trained for two epochs achieved the best final performance, with a macro F1-score of 0.83228 on the hidden test set. The results demonstrate that lightweight CNN fine-tuning strategies can provide strong baselines for historical image classification under limited-data conditions.

9. Bibliographical References

Alexei Abrahams, Shadi Abudalfa, Mustafa Jarrar, and George Mikros. 2026. The NakbaArchive-Classifier Shared Task on Nakba Image Classification. In *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 25.

Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

DLRG@ NakbaArchiveClassifier Shared Task: Deep Transfer Learning for Destruction Detection in Nakba Archive Images Using EfficientNet-B3

Ramesh Kannan R, Ratnavel Rajalakshmi
School of Computer Science and Engineering,
Vellore Institute of Technology, Chennai,
India - 600127.
kanannrameshr@gmail.com, rajalakshmi.r@vit.ac.in

Abstract

Automatic identification of destruction in conflict-affected regions is an important task for humanitarian monitoring and historical documentation. Visual analysis of destruction scenes can assist researchers and policy makers in understanding the extent of damage in affected areas. This paper presents a deep learning-based image classification approach for identifying destruction and non-destruction scenes in Nakba-related images. The problem is formulated as binary image classification on Nakba images. A transfer learning approach using EfficientNet-B3 is adopted to learn discriminative visual features from Nakba images. Experimental evaluation shows that the proposed model achieved an Weighted F1-score of 83.87 % and an overall classification accuracy of 85.57 % and secured 10^{th} rank in the competition. The results demonstrate that our proposed pre-trained method can effectively capture structural damage patterns and visual cues associated with destruction scenes. Code: https://github.com/kannanrrk/NakbaImageClassifier

Keywords: EfficientNet-B3, Nakba Image Classification, Destruction Detection, Deep Learning, Transfer Learning, Computer Vision

1. Introduction

The rapid proliferation of visual content on social media has transformed the way conflicts are documented, disseminated and interpreted. In contemporary crises, images often shows formal reporting, shaping public opinion and informing journalistic, humanitarian and policy responses in near real time. However, the overwhelming volume of user-generated and news imagery articles has created a critical bottleneck. The lack of reliable, scalable systems for automatically identifying and categorizing scenes of destruction and violence.

Automatic destruction detection lies at the intersection of computer vision, multimedia analysis and humanitarian technology. The breakthrough of deep convolutional neural networks (CNNs) (Krizhevsky et al., 2012) for large-scale image recognition established the foundation for modern visual classification systems. Subsequent architectural innovations such as residual learning (He et al., 2016) further improved training stability and performance in deep networks. These advances enabled the development of specialized systems for crisis and disaster imagery analysis. Social media based damage assessment and crisis related image classification have demonstrated the feasibility of automatically identifying destruction severity (Nguyen et al., 2017; Alam et al., 2018) and humanitarian needs from visual streams. More recently, transformer-based vision architectures have shown strong capability in modeling global visual context (Dosovitskiy et al., 2021), offering alternative approaches to convolution-based systems to attention based systems (Vaswani et al., 2023).

To advance the research in these area, NakbaArchiveClassifier (Abrahams et al., 2026) organized a shared task on destruction detection classifiers as part of NakbaNLP at LREC 2026. The shared task focuses on binary and fine grained classification of destruction related images. The organizers encouraged participants to develop robust model which is capable of handling visual ambiguity, varying image quality and domain-specific challenges.

As a competitive and computationally efficient baseline, the use of EfficientNet-B3 was proposed. EfficientNet introduced a principled compound scaling method that uniformly balances network depth, width and resolution, achieving state of the art accuracy with significantly fewer parameters than conventional CNN architectures (Tan and Le, 2019). The B3 variant offers a favorable trade off between representational power and computational efficiency, making it well suited for large scale image classification tasks under realistic resource constraints. It's improved resolution scaling enhances sensitivity to fine grained visual cues such as debris patterns, smoke, structural damage and injury indicators.

In this proposed work, EfficientNet-B3 acts as the backbone architecture for destruction detection,

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 229–233
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

fine tuned on the Nakba-related dataset. Through this shared task the work aims to establish standardized benchmark, encourages methodological diversity and stimulate further research on automated destruction detection in historically and politically significant visual archives.

2. Related Work

Recent advances in vision and multimodal modeling have substantially influenced how visual content is represented and classified. Early work in vision language joint representations laid the foundation for models capable of integrating textual and visual cues, with models such as ViLBERT demonstrating co-attentional modeling of images and text for vision-and-language tasks (Lu et al., 2019). Similarly, universal encoders like Unicoder VL introduced cross modal pretraining strategies that learned joint visual linguistic representations useful for retrieval and reasoning (Li et al., 2019).

In the realm of large scale multimodal pretraining, Contrastive Language–Image Pre-training (CLIP) emerged as a pivotal model, enabling zero-shot classification across diverse domains and learning transferable image representations through natural language supervision (Radford et al., 2021). Building on this, models such as BEiT-3 unified vision and language tasks through masked visual representation (Imglish) prediction and joint modeling, achieving strong transfer across classification and detection tasks (Wang et al., 2022).

More recent work has focused on further tightening vision language alignment and performance. Bootstrapping Language-Image Pretraining (BLIP) and its successor BLIP-2 improved multimodal pretraining. It combined strong image encoders with language generation objectives to support both understanding and generation tasks (Li et al., 2023). Hybrid multimodal architectures such as the Bilateral Vision Transformer (BVA-Transformer) leverage BLIP's learned features to fuse image and text representations for improved multimodal classification performance (Zhang et al., 2024). Recent research also explores post training enhancements for foundational multimodal models like CLIP, using techniques such as diffusion based feedback to improve fine grained visual perception and classification accuracy (Wang et al., 2024b).

Beyond classification and retrieval, large multimodal foundation models continue to evolve rapidly with models integrating vision capabilities directly into language model backbones. Qwen-VL and Gemini variants, shown broad reasoning and understanding across visual and textual inputs (Wang et al., 2024a). Surveys of multimodal vision models further highlights the paradigm shift towards unified multimodal reasoning frameworks that leverage joint visual language training objectives to address complex semantic tasks, including image classification, organization and multimodal information detection (Singh et al., 2026; Lupaşcu et al., 2026). Vision-based approaches have been applied in meme-related studies, including MMOD-Meme (Ramesh Kannan et al., 2022), which focuses on face emotion recognition, and TrollMeme (Ramesh Kannan and Rajalakshmi, 2022) classification. These highlights provides various implementation details on vision based approach models. Research must be expanded both in depth and breadth to build a robust vision based model for detecting destruction images.

3. Methodology

3.1. Problem Formulation

The task is formulated as a binary image classification problem. Let $\mathcal{X}$ denote the space of input images and $\mathcal{Y} = \{0, 1\}$ the label space, where

$$0 \rightarrow \text{Non-destruction}, \quad 1 \rightarrow \text{Destruction}.$$

Given an input image $I \in \mathcal{X}$, the goal is to learn a function

$$f_\theta : \mathcal{X} \rightarrow \mathcal{Y},$$

parameterized by θ, such that the predicted label is

$$\hat{y} = f_\theta(I), \quad \hat{y} \in \{0, 1\}.$$

Equivalently, in probabilistic form, the model estimates

$$p_\theta(y \mid I),$$

and the final prediction is obtained as

$$\hat{y} = \arg \max_{y \in \{0,1\}} p_\theta(y \mid I).$$

3.2. EfficientNet-B3 Model

EfficientNet-B3 is used as the backbone network for feature extraction. EfficientNet models scale network depth, width and resolution in a balanced manner, enabling efficient learning with fewer parameters. Transfer learning is applied by initializing the EfficientNet-B3 network with pretrained weights and fine tuning the final layers for binary classification. The architecture consists of EfficientNet-B3 Network with convolutional layer as a backbone, followed by Global Average Pooling layer, Fully connected dense layer, Softmax layer. The final layer produces probability scores for the two classes. Initially, all training images and their corresponding labels were organized, ensuring proper alignment between image names and class labels. Label encoding was performed using a LabelEncoder to convert categorical labels into numerical form suitable for model training.

To handle the data efficiently, a custom Dataset class was implemented, which loads each image, applies transformations and returns it alongside its label. The images were preprocessed using the torchvision library, which involved resizing them to a fixed dimension of 300×300 pixels and converting them to tensors. Data were loaded with DataLoader with batching, shuffling and multi-threaded loading to enhance training performance. The model architecture was built upon a pretrained EfficientNet-B3, with the original classification head removed and replaced by a fully connected layer matching the number of target classes. The model was trained using the AdamW optimizer with a cross-entropy loss function. During training, gradients were computed and back propagated to update the model parameters iteratively. After training, the model weights were saved for inference. For testing, a separate TestDataset was released by the organizers (Abrahams et al., 2026). The unseen images predictions were generated using the trained model. This pipeline ensures a systematic and reproducible workflow for NakbaImage Destruction Image. Table 1 shows the parameter details of the proposed approach.

Parameter	Value
Input size	300 × 300
Batch size	16
Optimizer	Adam
Learning rate	1e-4
Loss Function	CrossEntropyLoss

Table 1: Training Parameters

4. Dataset

The Nakba dataset (Abrahams et al., 2026) was manually collected from publicly available online sources and historical archives by SaltPillar. Nakba Dataset contains large collection of images sourced from social media and news platforms, with a focus on scenes related to conflict and destruction. The images vary in content and context, ranging from direct depictions of violence such as airstrikes, shootings, injuries, death, blood, terror, grief. Some of them are destruction based images with indirect or contextual scenes, including children playing, adults engaging in daily activities, or journalists reporting abroad. The primary goal of this shared task is to support the development of automated systems capable of detecting and categorizing scenes of destruction. Such systems are essential for journalists, engineers, humanitarian organizations and decision-makers who rely on timely, accurate visual information for analysis, reporting and crisis response. Images are labeled to distinguish between directly relevant content like violent image or destructive image and indirectly relevant content to contextual scene or non-violent scenes. The organizers released dataset is shown in Table 2, where 70 % of the data is Training set and 10 % data is Validation set and remaining 20 % data is Test set.

Split	not_destruction	destruction
Training	906	494
Validation	129	70
Testing	260	142
Total	**1295**	**706**

Table 2: Dataset distribution

5. Experiments and Results

Multiple convolutional based architectures were investigated to attain the effectiveness of deep learning models for automated destruction detection in Nakba archival images from Table 3. Several Convolutional Neural Network (CNN) architectures were evaluated, including a simple CNN, ResNet(He et al., 2016), RegNet (Radosavovic et al., 2020) and EfficientNet-B3 (Tan and Le, 2019). The experimental results demonstrate that deep transfer learning models significantly outperforms the basic CNN baseline models.

Models	F1 Score
Basic CNN	0.64
ResNet	0.82
RegNet	0.82
EfficientNet-B3	**0.834**

Table 3: Results

The simple CNN achieved an F1-score of 0.64, indicating limited capability in capturing complex visual patterns. In contrast, deeper architectures such as ResNet and RegNet improved the performance substantially, both achieving an F1-score of 0.82. These models benefit from deeper feature extraction and improved representation learning, which enhances their ability to capture structural and contextual patterns in the images.

Among the evaluated models, EfficientNet-B3 achieved the highest F1-score of 0.834, demonstrating superior performance in classification accuracy of 0.8557. This improvement can be attributed to EfficientNet compound scaling strategy, which jointly optimizes network depth, width and resolution to achieve better feature representation with improved computational efficiency. The results suggest that transfer learning using EfficientNet-B3 provides a robust framework for detecting destruction patterns in a historical Palestine Nakba images.

Overall, the findings indicate that advanced pre-trained architectures significantly enhance classification performance compared to conventional CNN models. Particularly EfficientNet-B3 offers an effective and scalable solution for automated destruction detection in Nakba image archives.

6. Conclusion

Thus, the study presents a deep learning-based framework for the automated identification of destruction in conflict affected regions using Nakba-related images from Palestine. By formulating the task as a binary image classification problem and leveraging transfer learning with EfficientNet-B3, the proposed approach effectively learns discriminative visual features associated with structural damage and non-destruction scenes. The experimental results demonstrate the effectiveness of the model, achieving an F1-score of 83.87 % and an overall accuracy of 85.57 %, securing the 10^{th} rank in the competition. These findings indicate that transfer learning with EfficientNet-B3 is capable of capturing meaningful structural patterns and contextual cues relevant to destruction analysis, even with a manually collected and annotated dataset. Moreover, the study highlights the potential of deep learning techniques to support humanitarian monitoring, digital archiving, and historical documentation efforts.

Ethics Statement

This study does not involve human subjects, personal data, or sensitive information.

Limitations: The dataset is relatively small and manually curated, which may limit the model's generalization to other conflict-affected regions or diverse image sources. Visual context may not always be fully sufficient to understand destruction patterns. Other metadata information like geospatial information could enhance performance.

Future Work: Future research can focus on expanding the dataset, data augmentation strategies, incorporating multimodal information, incorporating multi-class damage severity levels, exploring ensemble or attention-based architectures in historical visual images.

7. Acknowledgements

We thank the Management for their guidance and support, which enabled us to continue the research.

8. References

Alexei Abrahams, Shadi Abudalfa, Mustafa Jarrar, and George Mikros. 2026. Nakbaarchiveclassifier: Nakba image classification. In *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

Firoj Alam, Fethi Ofli, and Muhammad Imran. 2018. Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. 2019. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *arXiv preprint*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.

Marian Lupaşcu, Ana-Cristina Rogoz, Mihai Sorin Stupariu, and Radu Tudor Ionescu. 2026. Large multimodal models for low-resource languages: A survey. *Information Fusion*, 131:104189.

Dat Tien Nguyen, Fethi Ofli, Muhammad Imran, and Prasenjit Mitra. 2017. Damage assessment from social media imagery data during disasters. In

Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, and Gabriel Goh. 2021. Learning transferable visual models from natural language supervision.

Ilija Radosavovic, Ramprasaath R. Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10428–10436.

R Ramesh Kannan and Ratnavel Rajalakshmi. 2022. Multimodal code-mixed tamil troll meme classification using feature fusion. In *Proceedings of the First Workshop on Multimodal Machine Learning in Low-resource Languages*, pages 1–8.

R Ramesh Kannan, Manikandan Ravikiran, and Ratnavel Rajalakshmi. 2022. Mmod-meme: A dataset for multimodal face emotion recognition on code-mixed tamil memes. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 335–345. Springer International Publishing.

Gurpreet Singh, Lamia Qamar Athanickal Palassery, Nicholas Volta, Amruta Velamuri, and Aya Khanyile. 2026. Vision–language foundation models and multimodal large language models: A comprehensive survey of architectures, benchmarks, and open challenges. *Preprints*.

Mingxing Tan and Quoc V Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97, pages 6105–6114.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Ke-Yang Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *ArXiv*, abs/2409.12191.

Wenhui Wang, Hangbo Bao, Li Dong, et al. 2022. Image as a foreign language: Beit pretraining for all vision and vision-language tasks.

Wenxuan Wang, Quan Sun, et al. 2024b. Diffusion feedback helps clip see better.

Kaiyu Zhang, Fei Wu, Guowei Zhang, Jiawei Liu, and Min Li. 2024. Bva-transformer: Image-text multimodal classification and dialogue model architecture based on blip and visual attention mechanism. *Displays*, 83:102710.

Oblevit at AR-MS NAKBA NLP 2026 Subtask 2: Hybrid CNN–BiLSTM–CTC Framework with Linguistic Refinement for Arabic Handwritten Manuscript Recognition

Reem Akram Juhash, Abuelgasim Sami Abusonoun, Sara Ayad Al-Desouky
Computer Science Department, Arab Open University, Saudi Arabia
22460101ksa@aou.edu.sa, 22451751@aou.edu.sa, s.ayad@arabou.edu.sa

Abstract

Arabic handwritten manuscript recognition remains a challenging problem in Optical Character Recognition (OCR) because of the cursive nature of the Arabic script, variations in handwriting styles, and character overlap. In addition, many Arabic letters are different only in the number or placement of dots, which increases ambiguity in recognition tasks. These challenges require models capable of extracting meaningful visual features while modeling long-range sequential dependencies. In this work, we propose a deep learning-based OCR system for Arabic handwritten text recognition. The proposed approach combines convolutional neural networks (CNN) for feature extraction, bidirectional recurrent neural networks (BiLSTM) for sequence modeling, and Connectionist Temporal Classification (CTC) for end-to-end training without character-level segmentation. To further enhance performance, decoding and specially designed post-processing strategies for Arabic text are applied to handle spacing issues and reduce character-level ambiguity. Experimental results show that the proposed method achieves competitive performance and effectively handles complex handwritten Arabic manuscripts.

Keywords: Optical Character Recognition (OCR), Arabic Handwritten Text Recognition, Deep Learning, CNN–BiLSTM–CTC, Arabic Manuscript Recognition

1. Introduction

Handwritten text recognition remains difficult compared to printed OCR, especially for Arabic script. Arabic handwriting has a cursive nature, letters change shape depending on position, and many characters different from each other only by dots. In historical manuscripts, noise, faint ink, and irregular spacing further increase recognition complexity. Previous systems relied on manual features and deep learning techniques, but modern deep learning methods enable end-to- end training. In this work, we employ the CNN– BiLSTM–CTC framework to jointly model visual and sequential information without explicit segmentation. To improve accuracy, we integrate Arabic-aware decoding and post-processing strategies that address dot ambiguity and word segmentation issues. The proposed system provides a robust solution for recognizing handwritten Arabic manuscripts. The implementation of our system is publicly available to support reproducibility.[1]

2. Related Work

Markov Models (HMMs) were among the earliest approaches used for handwritten text recognition, where handwriting was modeled as a probabilistic sequence problem (Ploetz and Fink, 2009). These systems often relied on handcrafted feature extraction and statistical language models to improve recognition accuracy (Bunke et al., 2009). Although effective for their time, HMM-based methods struggled to capture long-range dependencies and required complex preprocessing pipelines.

The introduction of Connectionist Temporal Classification (CTC) marked a major shift toward end-to-end sequence modeling (Graves et al., 2009). CTC enabled neural networks to be trained without explicit alignment between input images and target text sequences. Graves et al. (2009) demonstrated that combining recurrent neural networks with CTC significantly improved performance in handwriting recognition tasks. Later, multi-dimensional LSTM (MDLSTM) networks were introduced to better capture spatial context in handwritten images (Graves and Schmidhuber, 2008).

With the success of convolutional neural networks in image processing tasks (Krizhevsky et al., 2012), researchers began integrating CNNs with recurrent layers to form hybrid architectures such as CNN–BiLSTM–CTC models. These architectures extract visual features using convolutional layers and then model sequential dependencies using recurrent layers, achieving strong performance in both word-level and line-level handwritten recognition. Attention-based sequence-to-sequence models were later proposed to address some limitations of CTC-based systems (Kass and Vats, 2022; Michael et al., 2019). These models use an encoder–decoder framework with attention mechanisms that dynamically focus on relevant image regions while generating text sequences.

More recently, transformer-based architectures

[1] Source code and trained model are available at https://github.com/aboalgasimm/Nakba-2026.

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 234–238
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

have been applied to text recognition due to their strong sequence modeling capabilities (Vaswani et al., 2017). Vision Transformers and hybrid CNN–transformer models have demonstrated competitive performance in various recognition tasks (Dosovitskiy et al., 2020). In addition to architectural improvements, data augmentation techniques have been widely used to enhance generalization and robustness (Shorten and Khoshgoftaar, 2019; Wigington et al., 2017). Synthetic data generation has also been explored to overcome limited training data availability (Li et al., 2021). Test-time augmentation strategies have further improved performance by aggregating predictions from multiple transformed versions of the same image (Poznanski and Wolf, 2016).

Building on these developments, our work adopts a CNN–BiLSTM–CTC framework enhanced with tailored decoding and post-processing techniques for Arabic handwritten manuscripts. Unlike purely transformer-based approaches, our model balances performance and computational efficiency while addressing specific challenges related to Arabic script characteristics. We evaluated our system on the AR-MS benchmark introduced in the NakbaNLP 2026 shared task (Hamoud et al., 2026).

3. Dataset

The dataset used in this work was provided as part of the AR-MS shared task introduced in the NakbaNLP 2026 workshop (Hamoud et al., 2026). The training dataset contains thousands of images of Arabic handwritten manuscript lines along with their corresponding text transcriptions as a CSV file. The training dataset was used to train the OCR model and learn the mapping between handwritten images and text sequences. The shared task also provides a blind test dataset that was used for the final evaluation in the leaderboard. It contains up to 2671 images with an empty CSV file that the system must fill after reading and processing the images in order. The challenges of Arabic handwritten manuscripts make the recognition task more difficult compared to printed text. In this work, we only used the official dataset provided by the shared task organizers. No additional external datasets were used.

4. Approach

In this work, we developed a deep learning-based Optical Character Recognition (OCR) system for Arabic handwritten manuscript recognition. The main objective was to design a model capable of handling the unique challenges of Arabic handwriting, including connected characters, dot ambiguity, irregular spacing, and visual noise in historical documents. The system consists of five stages as shown in Figure 1.

4.1. Convolutional Neural Network (CNN)

The first stage of the system is responsible for extracting visual features from the input manuscript images. A convolutional neural network (CNN) is used to transform raw image pixels into higher-level feature representations. The CNN consists of multiple convolutional layers followed by nonlinear activation functions and pooling layers, which progressively capture spatial patterns in handwritten text. CNNs are particularly effective for handwriting recognition because they can detect local visual structures such as edges, strokes, curves, and diacritics. These features are essential for Arabic handwriting since many characters share similar shapes and differ only by the presence or placement of dots. In addition, the CNN helps capture variations in writing style, pen pressure, stroke thickness, and scanning noise. After the convolutional feature extraction stage, the resulting feature maps are reshaped into a sequence representation where the horizontal axis of the text line is treated as a time sequence. This sequential representation is then passed to the recurrent layers for further processing.

4.2. Bidirectional Long Short-Term Memory (BiLSTM)

In the second stage, the extracted CNN features are processed using Bidirectional Long Short-Term Memory (BiLSTM) layers. Recurrent neural networks are well suited for modeling sequential data because they capture dependencies across time steps. The bidirectional structure allows the network to process the sequence in both forward and backward directions. This enables the model to consider both preceding and following context when predicting characters. Such contextual modeling is especially important for Arabic script, where character shapes depend on their position within a word and where correct interpretation often requires surrounding context. The output of the BiLSTM layers is a sequence of context-aware feature vectors that encode both spatial information from the CNN and sequential dependencies from the recurrent network.

4.3. Connectionist Temporal Classification (CTC)

This stage contains the CTC layer, the CTC loss, and the decoding logic. Because we do not have exact alignment between image positions and characters, CTC addresses this problem by producing

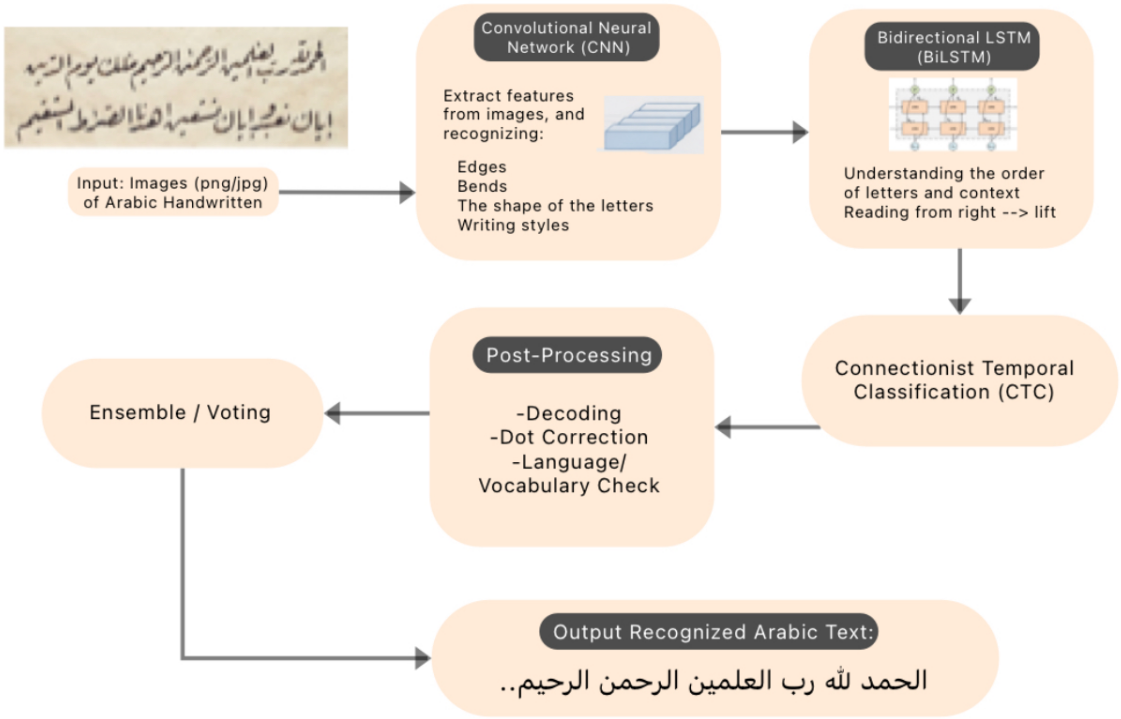


Figure 1: Overview of the proposed Arabic handwritten text recognition framework.

a probability distribution over characters at each time step while allowing blank steps. It learns how to map time steps to character sequences without forced alignment. During decoding, repeated characters are collapsed, blank tokens are removed, and a final character sequence is produced. The output of this stage is a raw predicted text string that is often still noisy and therefore benefits from additional refinement.

4.4. Post-Processing

This stage makes the text cleaner and more correct. It contains several sub-steps. First, beam search decoding is used to explore multiple likely sequences and reduce common CTC errors such as missing or duplicated characters. In our implementation, a beam width of 5 was used and the best sequence was selected based on a combination of CTC probability and lexicon matching. Second, dot correction was applied because Arabic accuracy is heavily affected by dots. We used a rule-based mapping derived from common Arabic character confusions together with context-aware rules. We also applied lexicon-based correction by matching predicted words against a vocabulary extracted from the training dataset, which contains approximately 41,000 unique words. Finally, language and vocabulary checks were used during decoding to prioritize valid Arabic word sequences. Candidates were filtered using this lexicon and the closest valid word was selected using edit distance. No external pretrained language model was used; instead, we relied on lexicon-based constraints.

4.5. Ensemble and Voting Strategy

In our work, we used multiple pipelines, where each pipeline applied different decoding and post-processing methods. Each pipeline was implemented separately and produced its own submission file. After running all experiments, we evaluated the results on Codabench and selected the best-performing outputs. In total, we chose seven files because they consistently achieved the highest results. In the final step, we combined these files using a voting-like approach. All submissions were aligned by filename and compared line by line. We then selected the best prediction for each line based on agreement between candidates and consistency with the training data. This approach helped improve the results by combining the strengths of multiple pipelines instead of relying on a single method.

5. Experimental Setup

Before training, the input images were preprocessed by resizing them to 1536 $\times$ 64 and normalizing pixel values. Since Arabic text is written from right to left, the images were horizontally flipped to preserve the correct reading order. The dataset

was then split into training and validation sets using a 90%–10% split. The model was trained using a CNN–BiLSTM–CTC architecture, where convolutional layers were used to extract visual features, followed by two Bidirectional LSTM layers with 256 and 128 units to capture sequential dependencies between characters. The final prediction layer produced 139 output classes, including the CTC blank token, based on a vocabulary of 138 symbols. Training was performed for 35 epochs with a batch size of 64, using the Adam optimizer with a learning rate of 3×10^{-4}. The CTC loss function was used to enable alignment-free training between input images and target text sequences. Early stopping and learning rate reduction were also applied to improve convergence and prevent overfitting. Performance was evaluated using Character Error Rate (CER) and Word Error Rate (WER). During inference, beam search decoding was applied to generate candidate sequences, followed by multiple post-processing pipelines. Although no full ablation study was conducted, we observed that removing post-processing and lexicon constraints degrades performance, especially in WER. The final system achieved a CER of 0.0925 and a WER of 0.3268.

6. Results and Discussion

The official leaderboard results are shown in Table 1. Our system ranked 2nd among the participating teams with a CER of 0.0925 and a WER of 0.3268. The results show that the proposed approach achieved competitive performance compared to other systems. In contrast, the baseline system obtained a much higher error rate, highlighting the effectiveness of deep learning approaches for Arabic handwritten manuscript recognition.

These results confirm that the CNN-BiLSTM-CTC architecture effectively captures both visual features and sequential dependencies. During experimentation, we faced several challenges. First which was a major issue, dot ambiguity because Arabic letters differ only by dots as we solved in section 4.1. Faint or misplaced dots often led to incorrect character predictions negatively affecting CER. Second, Arabic handwriting has a cursive nature, word merging and splitting errors occurred, which in this case negatively affecting WER however this problem also was solved as mentioned in section 4.1. Third, some noise distortions were misinterpreted as characters. Finally, we noticed some words out-of-vocabulary, some of which were correct rare words, while others resulted from minor letter-level errors. To mitigate these problems, we implemented post-processing techniques as mentioned in the approach above. Overall, the results demonstrate that combining deep learning with targeted decoding and post- processing methods enhances Arabic handwritten text recognition accuracy.

Rank	Username	CER	WER
1	Misraj Ai	0.0790	0.2440
2	Oblevit	0.0925	0.3268
3	3reeq	0.0938	0.2996
4	Latent Narratives	0.1050	0.3106
5	Al-Warraq	0.1142	0.3780
6	Not Gemma	0.1217	0.3063
7	NAMAA-Qari	0.1950	0.5194
8	Fahras	0.2269	0.5223
9	baseline	0.3683	0.6905

Table 1: Official leaderboard results of the ARMS shared task.

7. Conclusion

This paper presented a deep learning OCR system for Arabic handwritten manuscripts using a CNN–BiLSTM–CTC architecture. The model achieved competitive CER and WER results and showed strong ability to handle the complexity of Arabic script. Although challenges such as dot ambiguity and word segmentation errors were observed, tailored post-processing improved robustness.

8. Ethics Statement

This work focuses on Arabic handwritten manuscript recognition using the dataset provided by the shared task organizers. The dataset consists of historical manuscript images and does not contain personal or sensitive information. The purpose of this research is to support the digitization and automatic analysis of historical Arabic documents. The proposed system is intended only for research purposes and aims to improve the accessibility and preservation of historical manuscripts. No personal data was collected or used in this study. We also ensured that the dataset was used according to the guidelines provided by the shared task organizers. The system was not designed for profiling or decision-making about individuals, and its intended use is limited to research and preservation-oriented applications. By improving access to handwritten archival material, the work supports cultural preservation while remaining within the scope defined by the shared task.

9. Bibliographical References

H. Bunke, S. Bengio, and A. Vinciarelli. 2009. Offline recognition of unconstrained handwritten texts using hmms and statistical language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):855–868.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, and D. Weissenborn. 2020. An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmidhuber. 2009. A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):855–868.

A. Graves and J. Schmidhuber. 2008. Offline handwriting recognition with multidimensional recurrent neural networks. In *Advances in Neural Information Processing Systems*.

H. Hamoud, A. A. Chamseddine, B. Shalash, F. B. Abid, M. Jarrar, C. A. Chakra, B. Ghanem, and F. A. Zaraket. 2026. Nakba nlp 2026: Shared task on arabic handwritten manuscript understanding (palestine memory—omar al-saleh memoir). In *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with LREC 2026, Palma, Mallorca, Spain*.

D. Kass and E. Vats. 2022. Attentionhtr: Handwritten text recognition based on attention encoder–decoder networks. In *International Conference on Document Analysis Systems*.

A. Krizhevsky, I. Sutskever, and G. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105.

M. Li, T. Lv, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei. 2021. Trocr: Transformer-based optical character recognition with pre-trained models. *arXiv preprint arXiv:2109.10282*.

J. Michael, R. Labahn, T. Gruning, and J. Zollner. 2019. Evaluating sequence-to-sequence models for handwritten text recognition. In *International Conference on Document Analysis and Recognition*, pages 1286–1293.

T. Ploetz and G. Fink. 2009. Markov models for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 12:269–298.

A. Poznanski and L. Wolf. 2016. Cnn-n-gram for handwriting word recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2305–2314.

C. Shorten and T. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1).

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

C. Wigington, S. Stewart, B. Davis, B. Barrett, B. Price, and S. Cohen. 2017. Data augmentation for recognition of handwritten words and lines using a cnn–lstm network. In *International Conference on Document Analysis and Recognition*, pages 639–645.

Viva_Palestine at StanceNakba Shared Task: Actor and Topic-Aware Stance Detection in Public Discourse

Wafaa S. El-Kassas[1], Enas A. Hakim Khalil[2], Enas M.F. El Houby[2]

Faculty of Computers and Information Technology, The National Egyptian E-Learning University (EELU), Giza, Egypt[1], Systems & Information Department, National Research Centre, Dokki, Giza 12311, Egypt[2]

wafaa.elkassas@gmail.com, enaskhalil@gmail.com, enas_mfahmy@yahoo.com

Abstract

Recent research has increasingly focused on user-generated content to clarify opinions expressed in social media discourse. The Actor and Topic-Aware Stance Detection in Public Discourse challenge encourages research on stance detection in polarized social media discourse on the Palestinian–Israeli conflict. The challenge comprises two subtasks: one for actor-level alignments and the other for cross-topic generalization patterns. The StanceNakba2026 task includes two subtasks: (A) Actor-Level Stance Detection in English and (B) Cross-Topic Stance Detection in Arabic. Our team participated in both subtasks with the name "Viva_Palestine". In Subtask A, the proposed method is based on the Bert-Base-Uncased model and achieved a Macro F1-score of 0.9190, placing 6th out of 13 teams. In Subtask B, the proposed method is based on the MARBERT model and achieved a Macro F1-score of 0.8724 (the top rank in the leaderboard), placing first out of 10 teams. These results show that the proposed modelling method performs well for both entity-specific stance alignment and strong cross-topic generalization.

Keywords: Actor-Level, cross-topic, Stance detection, Bert-Base-Uncased, MARBERT

1. Introduction

With social media widely used as a primary medium for communication and information sharing, analyzing user-generated content to understand public opinion has become a critical research area in Natural Language Processing (NLP). Stance, defined as the expression of a speaker's viewpoint and evaluation about a certain proposition (Biber and Finegan, 1988) is an important part of this method. Stance identification enables critical applications such as sentiment analysis, opinion mining, and social media monitoring, facilitating the examination of public discourse and societal trends, particularly in political and social contexts (AlDayel and Magdy, 2021) , (Geng et al., 2026). Recent work, including shared tasks and newly created datasets from social media, has expedited research in this domain by providing benchmark resources and evaluating diverse modelling approaches. These initiatives demonstrate the effectiveness of transformer-based models for multi-topic stance detection while highlighting ongoing challenges such as neutral stance classification and generalization to unseen topics (Schiller et al., 2021). In this regard, the StanceNakba 2026 Shared Task (Aldous et al., 2026) examines stance identification in highly divided social discourse about the Palestinian–Israeli conflict and related regional debates. The task is structured around two different goals. Subtask A (Actor-Level Stance Detection) looks at political alignment at the actor level by figuring out if writers have a Pro-Palestine, Pro-Israel, or Neutral stance on the issue as a whole. Subtask B (Cross-Topic Stance Detection) focuses on topic-dependent stance classification, assigning Favor, against, or neither labels to opinions expressed about specific conflict-related issues, such as diplomatic normalization with Israel and the presence of refugees in Jordan. Stance identification is important across languages. StanceNakba 2026 works in both English and Arabic. Subtask A is completed in English, while Subtask B is completed in Arabic.

The remainder of this paper is organized as follows. Section 2 reviews related work on stance detection. Section 3 describes the proposed methodologies for both subtasks. Section 4 presents the experimental results. Finally, Section 5 concludes the paper and outlines directions for future work.

2. Related Work

Stance detection research aims to detect the position expressed by authors toward specific targets in social media content and has been explored through text-based, multilingual, and multimodal approaches. Early work stressed linguistic and rule-based methods, such as (Reveilhac and Schneider, 2023) which proposed a rule-based stance detection model for the SemEval Task 6A English dataset covering five topics, using sentiment lexical and linguistic markers, and achieving an average accuracy of 75%.

More recent studies have used deep learning and multimodal representation (AlShenaifi and Alangari, 2025) , introducing the MAWQIF-MM corpus of tweet–image pairs labeled with Favor, Against, and Neutral for Arabic multimodal stance detection. Using an attention-based model that combines AraBERT text embeddings with BLIP visual features, the approach achieved 88% accuracy, outperforming text-only and image-only baselines.

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 239–243
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Beyond direct stance classification, several works have examined public opinion and polarization in conflict-related discourse and analyzed Chinese-language Weibo posts on the October 2023 Israeli-Palestinian conflict. Findings show netizens' mixed emotional responses and support for humanitarian concerns, largely aligning with the Chinese government's diplomatic positions, highlighting Weibo's role in shaping and reflecting public opinion. Similarly, (Guerra et al., 2025) investigated over 450,000 Reddit posts on the 2023 Israeli–Palestinian conflict using an unsupervised lexicon-based extremism score. Results showed that extreme sentiment increased by up to 80% during major conflict events, highlighting how real-world developments drive polarized opinions on social media.

Other studies have applied stance detection in domain-specific contexts (Britt et al., 2025) analyzed 4,245 posts from 1,319 users on a major U.S. trucking forum and showed that drivers predominantly expressed negative stances toward autonomous trucks, driven by concerns over feasibility and job displacement. Their work demonstrated the value of integrating stance detection with topic modelling, sentiment analysis, and emotion analysis to capture nuanced public opinions.

In contrast to prior work, this paper addresses stance detection in two distinct settings: actor-level detection in English and cross-topic detection in Arabic. Unlike previous studies that focus on a single language or task, this work evaluates performance across different languages and task formulations, highlighting challenges in both high-resource and low-resource scenarios.

3. Methodology

This section describes the methodology adopted for StanceNakba 2026 Subtasks A and B. It begins with a description of the datasets used in both subtasks, followed by detailed explanations of the proposed approaches for Subtask A and Subtask B, respectively.

3.1 Data

The dataset for StanceNakba 2026 Subtask A (Actor-Level Stance Detection) comprises 1,401 English-language social media posts, divided into 980 training samples, 210 validation samples, and 211 test samples. Each instance includes three fields: a unique text identifier, the post content, and a stance label indicating whether the post expresses a Pro-Palestine, Pro-Israel, or Neutral stance. Examples of the 3 classes are shown in Figure 1.

In contrast, the dataset for Subtask B (Cross-Topic Stance Detection) comprises 1,205 Arabic social media posts, with 843 training samples, 181 validation samples, and 181 test samples. Each record consists of four attributes: an identifier, the sentence text, the associated topic, and a stance label. Figure 2 illustrates examples of different classes. The labels classify each Sentence as Favor, against, or neither with respect to one of two predefined topics: Normalization with Israel (577 samples) and Refugee/Immigrant Presence in Jordan (628 samples). The data for this subtask are taken from **MARASTA** corpus (Charfi et al., 2024).

Pro-Palestine
"The systematic displacement of Palestinian families from their ancestral homes represents a clear violation of international law and the right of return."

Pro-Israel
"Israel's defensive measures are necessary responses to existential threats, ensuring the safety of its citizens against terrorism."

Neutral
"The conflict involves competing territorial claims, with both populations having deep historical connections to the region."

Figure 1: Examples of different classes from the dataset of Subtask A.

Favor
التطبيع مع إسرائيل سيحسن الاقتصاد ويجلب الاستثمارات للمنطقة

Against
التطبيع مع إسرائيل يعتبر خيانة لقضية فلسطين والشعب الفلسطيني

Neither
التطبيع مع إسرائيل هو قضية معقدة لها جوانب اقتصادية وسياسية متعددة

Figure 2: Examples of different classes from the dataset of Subtask B.

3.2 Methodology of Subtask A (English)

This task requires identifying the general position of a text author regarding the Palestinian-Israeli conflict, categorized as Pro-Palestine, Pro-Israel, or Neutral. The Bert-Base-Uncased model is used for 3-class stance classification (Neutral, Pro-Israel, Pro-Palestine), Figure 1 illustrates examples of different classes. This model is **uncased**, meaning it does not differentiate between uppercase and lowercase letters. For example, it treats the words "english" and "English" as the same. BERT (Bidirectional Encoder Representations from Transformers) is pretrained on a large corpus of English data using a self-supervised approach. This means it was trained on raw texts without human labeling, using an automatic process to generate inputs and labels.

First, the data was tokenized using the tokenizer from the pretrained model with max_length = 256 and padding = max_length.

In the training phase, the model was trained on 980 samples and tested on 210 validation samples.

In the testing phase, the model was trained on 1190 samples (a combination of 980 training samples and 210 validation samples) and tested on the 211 unlabeled test samples.

3.3 Methodology of Subtask B (Arabic)

In Subtask B, it is required to build a unified model to predict Favor, against, or neither stance toward specific conflict-related topics. The topics covered in Subtask B are: (1) normalization with Israel and (2) refugee presence in Jordan. Examples of the three classes are illustrated in Figure 2.

In the training phase, the model "aubmindlab/bert-base-arabertv02" was utilized. The model was trained on a dataset of 843 samples and evaluated on a validation set consisting of 181 samples.

In the testing phase, three models have been used with extensive experimentation conducted to fine-tune their hyperparameters. The models include "CAMeL-Lab/bert-base-arabic-camelbert-mix", "UBC-NLP/MARBERTv2", and "UBC-NLP/MARBERT". These models were trained on a total of 1024 samples (i.e., a combination of 843 training samples and 181 validation samples) and were subsequently evaluated on the 181 unlabeled test samples.

At the end of the test phase experiments, the MARBERT model "UBC-NLP/MARBERT" achieves the best results over the other models and secures the top rank (i.e., first place in the leaderboard) in Subtask B among 10 teams.

MARBERT (Abdul-Mageed et al., 2021) is a powerful Arabic-specific Transformer-based language model. It was trained on large-to-massive datasets that cover different text genres (including social media) and different domains.

4. Experimental Results

4.1 Results of Subtask A (English)

In the training phase, the bert-base-uncased model was trained using 980 examples and tested with the validation examples of 210 samples. Table 1 shows the hyperparmeters tuned and performance metrics obtained. In the testing phase, the model was trained using a combination of train and validation samples. Table 2 shows the hyperparmeters tuned and performance metrics obtained. The confusion matrix of testing model is shown in Figure 3. The confusion matrix reveals that most misclassifications occur between the Neutral class and the two stance-bearing classes. 6 Neutral instances are misclassified as Pro-Israel, and 4 Neutral instances are misclassified as Pro-Palestine.

Train data	980 samples
Test(validation)	210 samples
Learning rate	3e-05
epochs	20
Early stopping	Yes, stop at 17th
Accuracy	0.890
precision	0.897
recall	0.890
Macro F1	0.891

Table 1: The training phase results with the parameters tuned for Subtask A.

Train data	1190 (combination of train and validation)
Test	211 unlabeled test samples
Learning rate	10e-05
epochs	20
Early stopping	Yes, stop at 9th
Test Accuracy	0.9149
precision	0.9211
Recall	0.9193
Macro F1	0.9190

Table 2: Test phase results with parameters tuned for Subtask A.

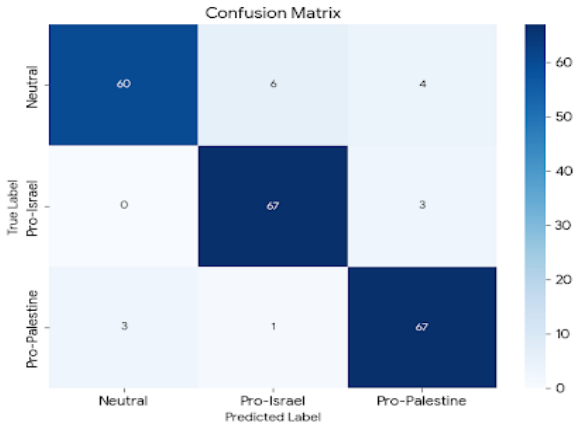


Figure 3: Confusion matrix for Subtask A.

The model is highly effective in detecting explicit stance expressions, achieving high recall for both Pro-Israel and Pro-Palestine classes. However, the comparatively lower recall for the Neutral class (85.7%), as illustrated in Table 3, highlights a limitation in capturing true neutrality. Overall, the model demonstrates robust and balanced performance, with particularly strong results in distinguishing between opposing stance

categories. The primary challenge remains improving the detection of neutral content in the presence of subtle or implicit opinions.

class	Precision	recall	F1	support
Neutral	0.95	0.86	0.90	70
Pro-Israel	0.91	0.96	0.93	70
Pro-Palestine	0.91	0.94	0.92	71
Accuracy			0.92	211
Macro Avg	0.92	0.92	0.92	211
Weighted Avg	0.92	0.92	0.92	211

Table3: Classification Report for Subtask A.

4.2 Results of Subtask B (Arabic)

In the testing phase, the training data increased by merging the training samples with the validation samples to form a new training set that includes 1024 records. Three Transformer-based models for Arabic have been used in the testing phase.

Table 4 illustrates the tuning of hyperparameters for the models: “UBC-NLP/MARBERT”, “CAMeL-Lab/bert-base-arabic-camelbert-mix”, and “UBC-NLP/MARBERTv2”. Table 5 shows the results of these models based on the tuning of hyperparameters in Table 4. Many trials to tune the hyperparameters of these models have been done, but not all the results of these trials are included in this paper due to space constraints.

The MARBERT model "UBC-NLP/MARBERT" achieved the best results over the other two models, as shown in Table 5. The model demonstrates robust and well-balanced performance across all evaluation metrics. It is worth mentioning that the MARBERT model achieved first place (top rank) in the leaderboard for Subtask B among 10 teams.

5. Conclusion

In this paper, Viva_Palestine demonstrated how to identify an actor's stance in English and how to determine a stance across topics in Arabic. The experimental results show strong, reliable performance on both subtasks. Viva_Palestine placed sixth out of 13 teams and achieved a Macro F1-score of 0.9190 in the English actor-level track (Subtask A). More importantly, the system's Macro F1-score of 0.8724 placed it first in the Arabic cross-topic track (Subtask B). This shows that it can generalize across topics in a morphologically complex language.

The results indicate how well the suggested methods handle multilingual and cross-topic attitude identification in difficult real-world scenarios.

The limited dataset size in this study constrains the overall performance. Future improvements could be achieved by adopting alternative strategies such as applying cross-validation during training, leveraging data augmentation techniques, incorporating additional datasets, and fine-tuning a wider range of models.

MODEL_NAME	“UBC-NLP/MARBERT”, “CAMeL-Lab/bert-base-arabic-camelbert-mix”, and “UBC-NLP/MARBERTv2”
Training Data	1024 samples (combination of training set and validation set)
MAX_LENGTH	128
BATCH_SIZE	8
LEARNING_RATE	2e-5
NUM_EPOCHS	4
WARMUP_RATIO	0.1
WEIGHT_DECAY	0.01
SEED	42

Table 4: Settings of the different models for Subtask B.

MODEL_NAME	“CAMeL-Lab/bert-base-arabic-camelbert-mix”
accuracy	0.8011
precision	0.8025
recall	0.8016
macro_f1	0.8002
MODEL_NAME	“UBC-NLP/MARBERTv2”
accuracy	0.8729
precision	0.8710
recall	0.8712
macro_f1	0.8710
MODEL_NAME	“UBC-NLP/MARBERT”
accuracy	**0.8729**
precision	**0.8736**
recall	**0.8733**
macro_f1	**0.8724**

Table 5: Results of different models in the testing phase for Subtask B.

6. References

Abdul-Mageed, M., et al. (2021). ARBERT & MARBERT: Deep bidirectional transformers for Arabic. Proceedings of the 59th Annual Meeting of the ACL-IJCNLP 7088-7105.

AlDayel, A. and Magdy, W. (2021). Stance detection on social media: State of the art and trends. Information Processing & Management 58(4): 102597.

Aldous, K., et al. (2026). StanceNakba Shared Task: Actor and topic-aware stance detection in public discourse. Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26).

AlShenaifi, N. and Alangari, N. (2025). Beyond text: Multimodal stance detection in Arabic tweets. Machine Learning with Applications 100823.

Biber, D. and Finegan, E. (1988). Adverbial stance types in English. Discourse Processes 11(1-34).

Britt, N., et al. (2025). Truck drivers and autonomous trucks: a topic modeling analysis of truck driver posts. Transportation Research Interdisciplinary Perspectives 34.

Charfi, A., et al. (2024). MARASTA: A multi-dialectal Arabic cross-domain stance corpus. *Joint International Conference on Computational Linguistics, Language Resources and Evaluation* 11060-11069.

Geng, G., et al. (2026). PERStance: Personality-guided enhanced multimodal stance detection. Information Processing & Management 63(4): 104593.

Guerra, A., et al. (2025). Quantifying extreme opinions on Reddit amidst the 2023 Israeli–Palestinian conflict. Natural Language Processing Journal 11(100156).

Reveilhac, M. and Schneider, G. (2023). Replicable semi-supervised approaches to state-of-the-art stance detection of tweets. Information Processing & Management 60(2): 103199.

Schiller, B., et al. (2021). Stance detection benchmark: How robust is your stance detection? KI – Künstliche Intelligenz 35(3): 329-341.

KUET at StanceNakba Shared Task: StanceMoE: Mixture-of-Experts Architecture for Stance Detection

Abdullah Al Shafi, Md. Milon Islam, Sk. Imran Hossain, K. M. Azharul Hasan
Department of Computer Science and Engineering, Khulna University of Engineering & Technology
abdullah@iict.kuet.ac.bd, {milonislam, imran, az}@cse.kuet.ac.bd

Abstract

Actor-level stance detection aims to determine an author's expressed position toward specific geopolitical actors mentioned or implicated in a text. Although transformer-based models have achieved relatively good performance in stance classification, they typically rely on unified representations that may not sufficiently capture heterogeneous linguistic signals, such as contrastive discourse structures, framing cues, and salient lexical indicators. This motivates the need for adaptive architectures that explicitly model diverse stance-expressive patterns. In this paper, we propose StanceMoE, a context-enhanced Mixture-of-Experts (MoE) architecture built upon a fine-tuned BERT encoder for actor-level stance detection. Our model integrates six expert modules designed to capture complementary linguistic signals, including global semantic orientation, salient lexical cues, clause-level focus, phrase-level patterns, framing indicators, and contrast-driven discourse shifts. A context-aware gating mechanism dynamically weights expert contributions, enabling adaptive routing based on input characteristics. Experiments are conducted on the StanceNakba 2026 Subtask A dataset, comprising 1,401 annotated English texts where the target actor is implicit in the text. StanceMoE achieves a macro-F1 score of 94.26%, outperforming traditional baselines, and alternative BERT-based variants.

Keywords: stance detection, mixture-of-experts, context-aware gating, adaptive weighting.

1. Introduction

Stance detection refers to the automatic identification of an author's position towards a specific topic, entity, or proposition expressed within a text (Garg and Caragea, 2024). The task aims to identify whether the presented viewpoint is supportive, opposing, or neutral toward the specific target. The study of stance detection differs from common sentiment analysis because it needs to measure emotional expressions that depend on particular targets, whereas the same text can show different stances based on which target the reader chooses to consider (Niu et al., 2024). The nature of this task requires target awareness because it generates both semantic complexity and computational difficulties, particularly when targets remain unmentioned and stances are expressed in an indirect way (Küçük and Can, 2020).

To address these challenges, we introduce StanceMoE[1], a context-enhanced Mixture-of-Experts (MoE) architecture designed for actor-level stance detection. Unlike conventional transformer-based approaches that rely on single aggregated representations, our method explicitly decomposes stance modeling into complementary expert modules that capture diverse linguistic and discourse-level phenomena. By incorporating a context-aware gating mechanism, the architecture enables adaptive and input-sensitive fusion of heterogeneous stance signals. Through comprehensive experimentation and detailed ablation analysis on StanceNakba 2026 Shared Task dataset (Aldous et al., 2026), we demonstrate that explicitly modeling such diverse patterns enables more robust and fine-grained actor-level stance detection with 3^{rd} position in the competition.

[1] https://github.com/AbdullahRatulk/StanceMoE

2. Related Works

The field of stance detection has evolved through its transition from rule-based systems (Küçük and Can, 2020) to classical supervised systems (Alturayeif et al., 2023), which rely on manually created features. Deep learning methods, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTMs) with attention mechanisms, enable models to automatically extract features and better capture target-specific stance information (Gera and Neal, 2025). Current transformer architectures use BERT (Garg and Caragea, 2024) and Large Language Models (LLMs) (Pangtey et al., 2025) to deliver excellent results through their ability to understand context and their capacity for few-shot learning.

3. Proposed StanceMoE Architecture

We propose a context-enhanced Mixture-of-Experts architecture developed upon a fine-tuned BERT encoder for stance detection. The overall framework consists of three components: (i) a contextual encoder, (ii) six parallel expert modules cap-

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 244–251
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

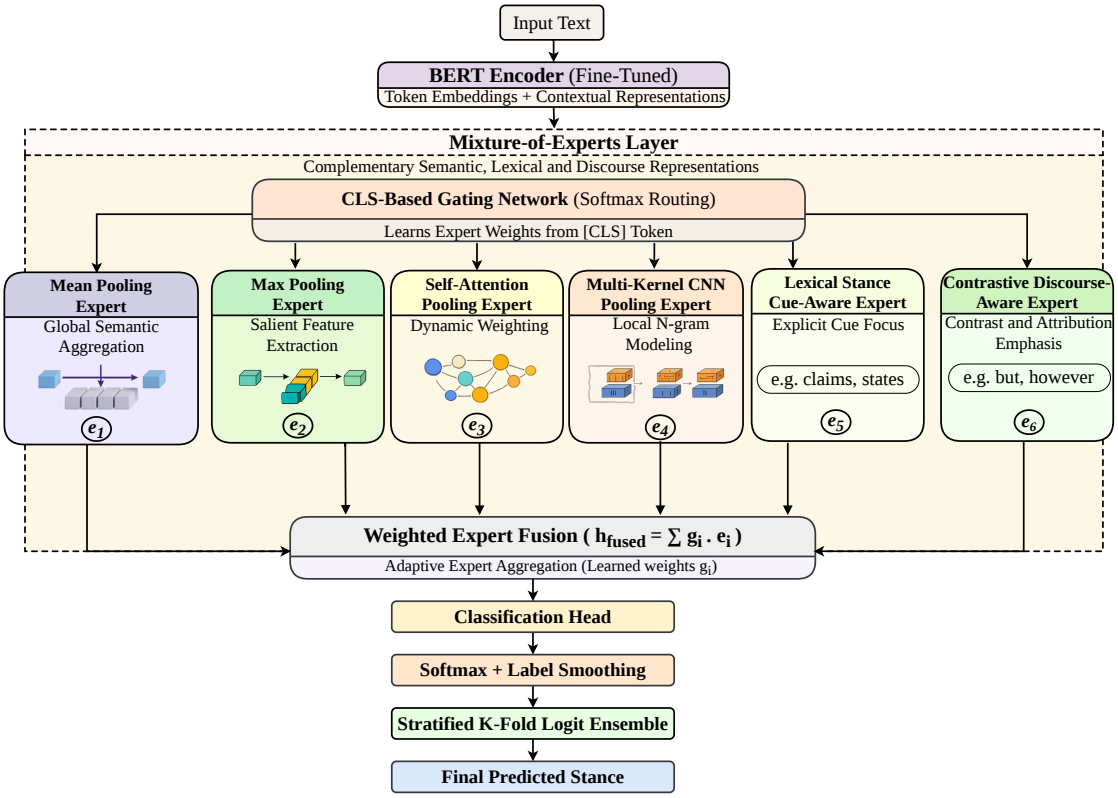


Figure 1: Proposed Mixture-of-Experts architecture for actor-level stance detection.

turing complementary linguistic signals, and (iii) a context-aware gating and fusion mechanism. The architecture is illustrated in Fig. 1.

3.1. Contextual Encoder

Given an input sequence, we employ a fine-tuned BERT encoder to obtain contextualized token representations and a special `[CLS]` embedding.

3.2. Expert Modules

While BERT provides promising contextual understanding, stance detection often depends on subtle lexical cues, contrast markers, and discourse patterns. Standard pooling techniques over contextual embeddings may not consistently capture these heterogeneous patterns. To address this, we introduce six complementary experts, each targeting a distinct linguistic phenomenon.

Mean Pooling Expert (Global Orientation): This expert captures the overall semantic direction of a post by averaging token representations. It is effective when stance is consistently expressed throughout the sentence. For example, *"Israel has the right to defend itself."* (Pro-Israel), or *"Palestinians deserve statehood and equal rights."* (Pro-Palestine).

Max Pooling Expert (Salient Polarity Tokens): Some posts contain isolated but highly indicative tokens that determine stance. For instance, the token *"occupation"* serves as the main indicator of stance in the Pro-Palestine statement *"This ongoing occupation must end"*. Similarly, the token *"terrorist"* serves as the main lexical indicator in the Pro-Israel statement *"The terrorist attacks cannot be justified"*. Max pooling selects the most activated feature dimensions to emphasize important tokens throughout the entire sequence.

Self-Attention Pooling Expert (Context-Dependent Focus): The specific clauses of a sentence work as the place where stance exists. The sentence *"While the humanitarian situation is tragic, Israel must respond to security threats"* demonstrate an example of such configurations. In such formulations, one clause presents the main stance while the additional information is presented through the rest of the sentence. The attention mechanism gives more weight to stance components because it assesses their importance throughout the evaluation process.

Multi-Kernel CNN Expert (Phrase-Level Patterns): Geopolitical stance is mostly expressed through short slogan-like phrases which include *"Stand with Israel"* (Pro-Israel) and *"Free Palestine"* (Pro-Palestine). The compact n-grams of the expression serve as effective stance indicators which maintain their function despite minimal contextual information. Multi-kernel convolutions capture such localized phrase patterns.

Lexical Cue-Aware Expert (Framing and Reporting Signals): Neutral posts use reporting lan-

guage and distancing language to express content like the statement *"According to officials, negotiations are ongoing."* Verbs like *"claims," "reports," "states,"* function as indicators that speakers use indirect methods to present information instead of showing direct support for their statements. This expert combines different ways of showing predetermined cue tokens to improve the ability to separate neutral content from opinionated postings.

Contrast-Aware Expert (Discourse Shift Modeling): Discourse contrast markers such as *"but"* or *"however"* often indicate stance refinement or emphasis. For example, *"I support peace efforts, but the blockade must end."* or *"Civilian harm is tragic; however, security responses are necessary."* In such cases, the clause containing the contrast marker often carries stronger stance weight. We therefore amplify representations of contrast tokens to better capture discourse-level stance shifts.

3.3. Context-Aware Gating and Fusion

We introduce a gating mechanism that dynamically weights the experts instead of treating them equally. The final representation is computed as a weighted combination of expert outputs. This design enables adaptive routing. For instance, strongly opinionated posts may rely more on max pooling, contrast-heavy statements may increase the weight of the contrast-aware expert, report-style neutral texts may emphasize cue-aware representations.

The fused representation is then passed to a linear classifier. We train the model using cross-entropy loss with label smoothing. To improve robustness, we employ stratified K-fold cross-validation during training and perform weighted logit averaging of those folds during inference. The weighting is done by giving more weight to folds having a higher F1-Score during validation.

4. Experimental Setup

4.1. Dataset

The dataset used in this paper is provided by the organizers of the StanceNakba 2026 Shared Task (Aldous et al., 2026). We focus on Subtask A, an actor-level stance detection dataset comprising 1,401 English texts labeled as Pro-Palestine, Pro-Israel, or Neutral. The stance labels denote the author's alignment with actors, which are implicit in the text and must be inferred by the model. The official split follows a 70/15/15 split for training, development, and test sets. In our experiments, we apply stratified K-fold cross-validation on the combined training and development set (85% of the data). The official test set (15%) is kept strictly unseen during training.

4.2. Baseline Methods

We perform an extensive empirical assessment of existing stance detection techniques, which are classified into three major methodological categories: Machine Learning (ML) methods with TF-IDF features, supervised training with Deep Neural Networks (DNNs), and BERT. The ML methods include Logistic Regression (LR) (Rahman et al., 2024), Multinomial Naive Bayes (MNB) (Zannat et al., 2025), Support Vector Machine (SVM) (Rahman et al., 2024), and Random Forest (RF) (Shafi et al., 2025). We implement several popular DNN architectures as baselines, including Bi-directional Long Short-Term Memory (BiLSTM) (Rahman et al., 2024), Target specific Attentional Network (TAN) (Du et al., 2017), Gated Convolutional network with Aspect Embedding (GCAE) (Xue and Li, 2018), Cross network (CrossNet) (Du et al., 2017). Moreover, we evaluate two alternative variants developed on the BERT backbone: (i) a stacked architecture (Khan et al., 2025) where expert modules are applied sequentially, and (ii) a feature fusion model (Lee et al., 2021) where expert outputs are fused without adaptive weighting.

5. Experimental Results

5.1. Actor-level Stance Detection

Table 1 presents the comparative performance of all models for actor-level stance detection. Among traditional baselines, RF achieves the best results with an F1-Score of 84.19%, which shows that ensemble tree-based methods perform better than other methods. Among neural models, GCAE performs highest (87.62% F1), which shows that attention-based architectures perform well to model stance. The use of contextual embeddings results in better performance outcomes. BERT achieves an F1-Score of 89.86%, while expert-based architectures like Stacked (91.83%) and Fusion (91.17%) show promising results, but their improvements remain within fixed expert integration boundaries.

The proposed StanceMoE achieves the best performance with an F1-Score of 94.26%, which surpasses all other cases of comparison. Additionally, our team has stood 3^{rd} in the competition. The system demonstrates performance enhancement because of its adaptive expert weighting system, which uses a gating mechanism. The proposed system also shows stable performance with low cross-validation variance ($\approx \pm 1.1\%$).

5.2. Ablation Study

Table 2 presents an ablation analysis to examine the contribution of each expert in the proposed StanceMoE architecture. Removing any expert

Methods	K-fold (mean±std)				Weighted Logit Ensemble			
	Acc	*Pre*	*Rec*	*F1*	*Acc*	*Pre*	*Rec*	*F1*
LR	80.91±1.79	80.90±1.78	80.93±1.78	80.86±1.78	81.28	81.25	81.33	81.22
MNB	77.25±1.64	75.88±2.02	78.49±1.65	77.20±1.65	77.38	76.02	78.63	77.33
SVM	83.25±1.71	84.73±1.68	83.26±1.71	83.27±1.72	83.37	84.94	83.35	83.43
RF	84.05±1.83	84.73±1.73	84.05±1.82	84.05±1.90	84.16	84.83	84.14	84.19
BiLSTM	85.63±2.99	85.87±3.00	85.63±2.99	85.51±3.07	85.74	85.98	85.74	85.72
TAN	85.99±3.10	86.01±2.81	86.36±2.08	85.91±2.15	86.13	86.10	86.43	86.03
GCAE	87.69±2.85	87.98±2.92	87.48±2.16	87.49±2.78	87.82	88.10	87.53	87.62
CrossNet	85.13±2.28	85.55±1.89	84.90±1.90	84.77±2.25	85.29	85.65	84.97	84.88
BERT	89.77±2.35	90.04±2.30	89.77±2.31	89.61±2.29	90.05	90.28	90.03	89.86
Stacked	91.61±2.46	91.73±2.43	91.62±2.44	91.60±2.47	91.94	92.14	91.93	91.83
Fusion	91.03±2.26	91.20±2.29	91.03±2.24	91.02±2.26	91.18	91.45	91.17	91.17
StanceMoE	**94.09±1.11**	**94.18±1.12**	**94.08±1.12**	**94.03±1.12**	**94.31**	**94.45**	**94.31**	**94.26**

Table 1: Performance comparison of baseline models and the proposed StanceMoE on the held-out test set. *"K-fold (mean ± std)"* reports the mean and standard deviation of test performance across fold-specific models, while *"weighted logit ensemble"* is the final ensemble prediction obtained via logit averaging. Here, Acc = Accuracy, Pre = Precision, Rec = Recall, and F1 = F1-Score.

Methods	*Acc*	Pro-Palestine			Pro-Israel			Neutral		
		Pre	*Rec*	*F1*	*Pre*	*Rec*	*F1*	*Pre*	*Rec*	*F1*
w/o Mean	92.89	89.47	95.75	92.52	94.52	98.57	96.50	95.16	84.29	89.39
w/o Max	93.36	91.89	**95.77**	93.79	92.00	98.57	95.17	96.77	85.71	90.91
w/o Self-Attention	91.94	89.33	94.37	91.78	93.24	98.57	95.83	93.55	82.86	87.88
w/o CNN	93.36	92.96	92.96	92.96	93.33	99.26	96.55	93.85	87.14	90.37
w/o Lexical-cue	93.36	90.54	94.37	92.41	**95.83**	98.57	**97.18**	95.08	82.86	88.55
w/o Contrastive	91.94	90.41	92.96	91.67	90.91	99.32	95.24	93.85	87.14	90.37
StanceMoE	**94.31**	**94.37**	94.37	**94.37**	92.11	**100**	95.89	**96.88**	**88.57**	**92.54**

Table 2: Classwise ablation study of StanceMoE using weighted logit ensemble, showing the effect of removing individual expert modules.

generally degrades overall performance, confirming that the experts capture complementary stance signals. The largest drops occur when removing the self-attention or contrastive experts, indicating their importance in modeling contextual dependencies and discourse-level stance shifts. Excluding the mean or max pooling experts also reduces performance, suggesting the importance of salient lexical cues and global contextual signals for stance identification.

The lexical-cue expert particularly benefits the Pro-Palestine class, while the contrastive expert improves recognition of stance transitions and argumentative structures. Although some variants maintain competitive scores for individual classes, they generally exhibit reduced balance across classes, especially for the neutral category. Overall, the full StanceMoE model achieves the best and most balanced performance across all classes, demonstrating the effectiveness of adaptive expert integration.

6. Conclusion

In this paper, we introduced StanceMoE, a context-enhanced MoE architecture for stance detection. Motivated by the limitations of unified transformer representations capturing heterogeneous linguistic phenomena, our approach explicitly models complementary stance-indicative signals through expert modules and integrates them via a context-aware gating mechanism. This design enables adaptive modeling of lexical, semantic, and discourse-level patterns that characterize stance expression in sensitive contexts. Extensive experiments demonstrate that StanceMoE substantially outperforms traditional baseline models. Ablation analysis further confirms the complementary contributions of individual experts and the importance of adaptive expert fusion for robust detection. Overall, our findings highlight the effectiveness of explicitly disentangling and dynamically integrating diverse linguistic signals for stance detection.

Bibliographical References

Kholoud Khalil Aldous, Md Rafiul Biswas, Mabrouka Bessghaier, Shimaa Ibrahim, Kais Attia, and Wajdi Zaghouani. 2026. StanceNakba shared task: Actor and topic-aware stance detection in public discourse. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

Izzat Alsmadi, Iyad Alazzam, Mohammad Al-Ramahi, and Mohammad Zarour. 2024. Stance detection in the context of fake news—a new approach. *Future Internet*, 16(10):364.

Nora Alturayeif, Hamzah Luqman, and Moataz Ahmed. 2023. A systematic review of machine learning techniques for stance detection and its applications. *Neural Computing and Applications*, 35(7):5113–5144.

Michael Burnham. 2025. Stance detection: a practical guide to classifying political beliefs in text. *Political Science Research and Methods*, 13(3):611–628.

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention networks. In *2017 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3988–3994.

Krishna Garg and Cornelia Caragea. 2024. Stanceformer: Target-aware transformer for stance detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4969–4984.

Parush Gera and Tempestt Neal. 2025. Deep learning in stance detection: A survey. *ACM Computing Surveys*, 58(1):1–37.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Lal Khan, Atika Qazi, Hsien-Tsung Chang, Mousa Alhajlah, and Awais Mahmood. 2025. Empowering urdu sentiment analysis: an attention-based stacked cnn-bi-lstm dnn with multilingual bert. *Complex & Intelligent Systems*, 11(1):10.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys*, 53(1):1–37.

Sanghyun Lee, David K Han, and Hanseok Ko. 2021. Multimodal emotion recognition fusion analysis adapting bert with heterogeneous feature unification. *IEEE Access*, 9:94557–94572.

Junxia Ma, Changjiang Wang, Lu Rong, Bo Wang, and Yaoli Xu. 2025. Exploring multi-agent debate for zero-shot stance detection: A novel approach. *Applied Sciences*, 15(9):4612.

Fuqiang Niu, Min Yang, Ang Li, Baoquan Zhang, Xiaojiang Peng, and Bowen Zhang. 2024. A challenge dataset and effective models for conversational stance detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 122–132.

Lata Pangtey, Anukriti Bhatnagar, Shubhi Bansal, Shahid Shafi Dar, and Nagendra Kumar. 2025. Large language models meet stance detection: A survey of tasks, methods, applications, challenges and future directions. *arXiv:2505.08464*.

Arifur Rahman, Shahriar Parvej, Kazi Saeed Alam, and HM Abdul Fattah. 2024. Optimizing sms spam detection: comparative analysis of hybrid voting ensembles and bi-lstm networks with stratified cross-validation. In *2024 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI)*, pages 1030–1035. IEEE.

Abdullah Al Shafi, Rowzatul Zannat, Abdul Muntakim, and Mahmudul Hasan. 2025. A structured dataset of disease-symptom associations to improve diagnostic accuracy. *arXiv:2506.13610*.

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2514–2523.

Ruichao Yang, Jing Ma, Wei Gao, and Hongzhan Lin. 2025. Llm-enhanced multiple instance learning for joint rumor and stance detection with social context information. *ACM Transactions on Intelligent Systems and Technology*, 16(3):1–27.

Rowzatul Zannat, Abdullah Al Shafi, and Abdul Muntakim. 2025. Bridging the gap in bangla healthcare: Machine learning based disease prediction using a symptoms-disease dataset. In *2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pages 1–6. IEEE.

Bowen Zhang, Genan Dai, Fuqiang Niu, Nan Yin, Xiaomao Fan, Senzhang Wang, Xiaochun Cao, and Hu Huang. 2024. A survey of stance detection on social media: New directions and perspectives. *arXiv:2409.15690*.

Zhengyuan Zhu, Zeyu Zhang, Haiqi Zhang, and Chengkai Li. 2025. Ratsd: Retrieval augmented

truthfulness stance detection from social media posts toward factual claims. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3366–3381.

A. Need of Stance Detection

The ability to detect stance has become increasingly important due to the growth of online discourse. Applications span multiple domains, including electoral trend analysis (Burnham, 2025), examination of argumentative interactions in online debates (Ma et al., 2025), monitoring of social media discussions (Zhang et al., 2024), rumor assessment and verification (Yang et al., 2025), automated fact-checking systems (Guo et al., 2022), fake news identification (Alsmadi et al., 2024), and stance-aware information retrieval (Zhu et al., 2025). In these contexts, understanding directional opinion toward a claim or actor is more informative than simply measuring sentiment (Garg and Caragea, 2024).

B. Operational Details of Expert Modules

Given an input sequence $X = \{x_1, \ldots, x_T\}$, pre-trained BERT encoder obtains contextualized token representations: $H = \{h_1, h_2, \ldots, h_T\}, \quad h_i \in \mathbb{R}^d$ where d denotes the hidden dimension and T is the length of the tokenized input sequence.

Mean Pooling Expert: This expert captures global semantic information as shown in (1).

$$e_1 = W_1 \left(\frac{1}{T} \sum_{i=1}^{T} h_i \right) \tag{1}$$

Max Pooling Expert: To extract salient token-level features, we utilize max pooling expert as in (2).

$$e_2 = W_2 \left(\max_{i=1}^{T} h_i \right) \tag{2}$$

Self-Attention Pooling Expert: We introduce a trainable attention vector v to compute token importance as mentioned in (3).

$$\alpha_i = \frac{\exp(\tanh(h_i^\top v))}{\sum_{j=1}^{T} \exp(\tanh(h_j^\top v))} \tag{3}$$

$$e_3 = W_3 \left(\sum_{i=1}^{T} \alpha_i h_i \right) \tag{4}$$

Multi-Kernel CNN Expert: To model local n-gram patterns, we apply one-dimensional convolutions with kernel sizes $k \in \{2, 3, 4, 5\}$ as demonstrated in (5).

$$e_4 = W_4 \left(\text{Concat} \left(\text{MeanPool}(\text{ReLU}(\text{Conv}_k(H))) \right) \right) \tag{5}$$

Lexical Cue-Aware Expert: We define a set of stance-indicative lexical cues and generate a binary mask over their positions C. The cue-aware representation is computed as (6).

$$e_5 = W_5 \left(\frac{\sum_{i \in C} h_i}{|C| + \epsilon} \right) \tag{6}$$

where, ϵ avoids division by zero.

Contrast-Aware Expert: To model discourse contrast markers (e.g., “but”, “however”), we enhance their contextual influence. Let, D denote the set of contrast token positions. The overall process is shown in (8).

$$\tilde{h}_i = \begin{cases} 3h_i & \text{if } i \in D, \\ h_i & \text{otherwise}, \end{cases} \tag{7}$$

$$e_6 = W_6 \left(\frac{\sum_{i=1}^{T} \tilde{h}_i}{|D| + \epsilon} \right) \tag{8}$$

C. Operational Details of Context-Aware Gating

The gating network takes the ‘[CLS]‘ representation $h_{\text{cls}} \in \mathbb{R}^d$ from the BERT encoder as input. A learnable linear layer $W_g \in \mathbb{R}^{N \times d}$ with bias $b_g \in \mathbb{R}^N$ computes logits for $N = 6$ experts, which are normalized via a softmax to generate the gating weights as in (9).

$$g = \text{Softmax}(W_g h_{\text{cls}} + b_g) \tag{9}$$

where $g \in \mathbb{R}^K$, $g_i \geq 0$ for all i, and $\sum_{i=1}^{K} g_i = 1$. Each gating weight g_i determines the contribution of the corresponding expert output e_i to the final representation, which is obtained via weighted aggregation using (10).

$$h_{\text{moe}} = \sum_{i=1}^{K} g_i \, e_i \tag{10}$$

The fused representation h_{moe} is then passed through a task-specific linear layer with learnable weights W_o and bias b_o to generate the model predictions through (11).

$$\hat{y} = \text{softmax}(W_o h_{\text{moe}} + b_o) \tag{11}$$

where $\hat{y}$ represents the predicted class probabilities. The parameters W_g, b_g, W_o, and b_o are all learnable and optimized jointly with the rest of the network via backpropagation, allowing the model to learn both how to combine experts and how to make accurate predictions.

D. Hyperparameters

Table 3 shows the hyperparameters used in the experiment.

Hyperparameter	Value
Max sequence length	128
Batch size	16
Number of epochs	10
Learning rate	5×10^{-5}
Number of splits (k)	10
Label smoothing factor	0.25
Random seed	42

Table 3: Training hyperparameters used in StanceMoE.

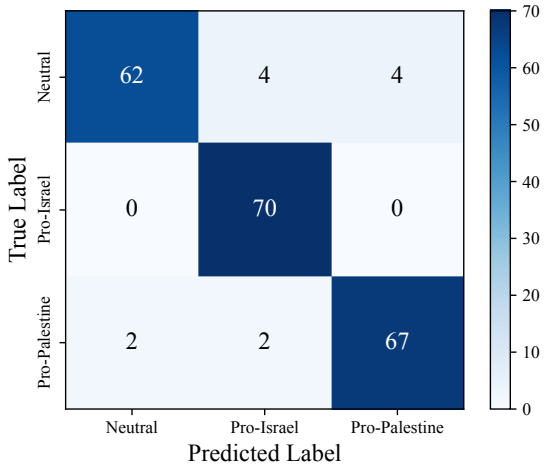


Figure 2: Confusion matrix using the proposed StanceMoE architecture.

E. Baseline ML methods

We utilize four popular ML models as our baseline systems: (1) Logistic Regression (LR): A linear model that predicts class probabilities using weighted features, effective for high-dimensional text data. (2) Multinomial Naive Bayes (MNB): A probabilistic model based on word frequency distributions, assuming feature independence, well-suited for text classification. (3) Support Vector Machine (SVM): A margin-based classifier that finds the optimal hyperplane to separate classes, robust in sparse text feature spaces. (4) Random Forest (RF): An ensemble of decision trees that improves classification performance by reducing overfitting through bagging and feature randomness.

F. Baseline DNN

We use four popular deep neural network models as our baseline systems: (1) BiLSTM (Rahman et al., 2024) is used to create bi-directional context models which predict stances without requiring to use target information. (2) GCAE (Xue and Li, 2018) provides a convolutional system which employs a gating mechanism to block non-target features. (3) TAN (Du et al., 2017) uses an attention-boosted BiLSTM to identify important contextual details which help to determine stances. (4) Cross-Net (Du et al., 2017) enhances attention processing through an aspect-based attention layer which operates before the classification process to enhance target-based feature extraction.

G. Evaluation Metrics

We evaluate model performance using macro-F1 score as the primary metric. Additionally, accuracy, macro precision, and macro recall are reported.

H. Overall Ablation Study

Table 4 shows the overall ablation study on the test set, rather than class-specific one.

I. Confusion matrix

Fig. 2 shows the confusion matrix using proposed StanceMoE architecture.

J. Error analysis

The qualitative error analysis in Table 5 reveals several systematic patterns.

(1) The model occasionally exhibits an excessive dependence on strong lexical polarity cues. The first and fifth examples show religious praise as well as humanitarian language which are mistakenly interpreted as explicit political alignment.

(2) Meta-discursive commentary is sometimes confused with stance. The text in the third example describes antisemitism together with social norms without showing any specific support. The system uses target-related keywords to make stance predictions about the content. This is difficult because it requires the model to differentiate between entity mentioned and evaluative positioning.

(3) Complex multi-clause and contrastive reasoning remains challenging. The second and sixth examples contain intricate discourse structures that display stance through their multiple levels of argumentation. The system design includes experts to handle both attention and contrast decoding tasks, but in some cases the gating system fails to deliver proper evaluation for these specialized discourse elements.

(4) Very short and context-poor statements like the fourth example are inherently ambiguous. In such cases, the model appears to associate the polarity with dominant training patterns rather than grounding the prediction in the explicit evaluation of the target.

Methods	K-fold (mean±std)				Weighted Logit Ensemble			
	Acc	*Pre*	*Rec*	*F1*	*Acc*	*Pre*	*Rec*	*F1*
w/o Mean	91.75±1.76	92.02±1.54	91.74±1.77	91.65±1.86	92.89	93.05	92.88	92.80
w/o Max	91.52±1.23	91.83±1.06	91.51±1.23	91.43±1.30	93.36	93.56	93.35	93.29
w/o Self-att	91.47±1.04	91.65±0.99	91.46±1.04	91.38±1.08	91.94	92.04	91.93	91.83
w/o CNN	92.09±1.04	92.20±0.97	92.08±1.04	92.00±1.11	93.36	93.38	93.37	93.29
w/o Lexical-cue	91.70±1.84	91.83±1.84	91.70±1.84	91.67±1.83	93.36	93.41	93.36	93.32
w/o Contrastive	91.75±1.81	91.99±1.67	91.74±1.82	91.64±1.87	91.94	92.13	91.94	91.82
StanceMoE	**94.09±1.11**	**94.18±1.12**	**94.08±1.12**	**94.03±1.12**	**94.31**	**94.45**	**94.31**	**94.26**

Table 4: Overall K-fold and weighted ensemble ablation study on the test set.

Text	Actual	Predicted	Primary Error Source
This is shameful. And the Jewish people, God's chosen people. May God bless and protect you.	Neutral	Pro-Israel	Religious praise misinterpreted as political stance due to strong positive lexical cues.
Most comments are clothed with sycophancy. Nigerians have a right of association, but you can stand with Gaza. The people of Gaza started this aggressive behaviour on 7th October when they launched a coordinated attack against the state of Israel.	Pro-Palestine	Neutral	Complex multi-clause reasoning; insufficient emphasis on dominant stance clause.
If you are talking about the doctor not talking about Palestine. He is German living in Germany where any anti-Jewish sentiment is taken very seriously and to many people anti-Israel equals antisemitism, so it is in his better interest not to speak about it at all, not necessarily meaning he supports anyone.	Neutral	Pro-Israel	Confusion between topic discussion and stance expression due to keyword activation.
Against the Islamist terrorists who hate Jews and the West.	Neutral	Pro-Palestine	Short ambiguous input; polarity incorrectly associated with a stance.
As I stood up with BLM, I stand against oppression in all forms. Release Palestinian hostages before talking about anything else.	Neutral	Pro-Palestine	Humanitarian framing interpreted as explicit alignment due to lexical cues.
I was raised Muslim, and I know intimately how oppressive the religion can be in various ways. That being said, because I understand that Palestinian gay people are being oppressed for more than their sexuality right now, and that matters too.	Pro-Palestine	Neutral	Failure to capture nuanced contrastive stance; insufficient weighting of discourse signals.

Table 5: Representative genuine misclassification cases of StanceMoE. Different background colors are used for visual distinction.

Overall, these errors suggest that while StanceMoE is effective for explicit and lexically polarized stance expressions, nuanced and implicitly framed discourse remains a challenging direction for future improvement.

K. Future Work

Future research may extend this framework toward explicitly target-aware stance modeling or explore cross-topic and cross-domain stance detection to further evaluate the robustness and generalization capacity of the proposed architecture. Investigating multilingual adaptation may also enhance its applicability to broader geopolitical discourse contexts.

L. Acknowledgement

We would like to express our sincere gratitude to the organizer of StanceNakba 2026 shared task, and anonymous reviewers for their support. The authors acknowledge the use of ChatGPT (OpenAI) as an assistive tool for language refinement, code suggestions, and conceptual structuring. All generated content was critically reviewed, validated, and appropriately adapted by the authors, who take full responsibility for the accuracy and integrity of the work.

The Blackwell Collective at StanceNakba Shared Task: PAST-TIDE: Prototype-Anchored Statement Tuning with Topic-Invariant Normalization for Stance Detection

Md. Shakhoyat Rahman Shujon[1], **MD Jahid Hasan Jim**[1], **Md. Milon Islam**[1], **Md. Rezwanul Haque**[2], **Fakhri Karray**[2,3]
[1]Department of Computer Science and Engineering, Khulna University of Engineering & Technology
skt104.shujon@gmail.com, mdjahidhasanjim277@gmail.com, milonislam@cse.kuet.ac.bd
[2]Department of Electrical and Computer Engineering, University of Waterloo
[3]Department of Machine Learning, Mohamed bin Zayed University of Artificial Intelligence
{rezwan, karray}@uwaterloo.ca

Abstract

We introduce **PAST-TIDE**, our stance detection system addressing both subtasks of the StanceNakba Shared Task at NakbaNLP@LREC-COLING 2026. The main idea is statement tuning. We redefine stance as cloze-style masked language modeling (MLM), letting a verbalizer map label words to stance categories through the pre-trained MLM head rather than appending a randomly initialized classification head. We complement this with prototypical contrastive learning, which uses learnable class prototypes for batch-size independent contrastive training, and topic-conditional layer normalization for cross-topic Arabic stance detection. **PAST-TIDE** achieves macro-F1 scores of 0.75 for Subtask A and 0.74 for Subtask B on the official leaderboard, indicating that minimal architectural additions to a pre-trained model can remain competitive in low-resource settings.

Keywords: stance detection, prompt-based classification, contrastive learning, cross-topic generalization, low-resource multilingual NLP.

1. Introduction

Understanding a person's stance on a given topic remains a challenging problem in computational argumentation (Mohammad et al., 2016; Küçük and Can, 2020). The field has moved from handcrafted features to pre-trained language model fine-tuning (Ghosh et al., 2019; Li and Caragea, 2021) to prompt-based reformulations (Schick and Schütze, 2021; Gao et al., 2021), yet multilingual stance detection remains underexplored. Most of the existing work focuses on benchmarks in English alone with thousands of labeled samples, leaving open the question of whether prompt-based methods remain effective when labels are limited and the target language is not English. The eNakba Shared Task (Aldous et al., 2026) addresses exactly the above gap. It defines two subtasks: Subtask A, stance at the actor-level in English (Pro-Palestine/Pro-Israel/Neutral), and Subtask B, stance cross-topic in Arabic (Favor/Against/ None). Both tasks use relatively small datasets, with 1,401 and 1,205 labeled samples, split into 70/15/15 for train, development, and test. A standard [CLS] classifier head lacks any pre-trained initialization of its own, and that pre-training/fine-tuning gap degrades performance at this scale.

Our architecture, **PAST-TIDE**, is based on one idea: the classistance classification capacity already exists withinined MLM head, and the task is to access it rather than replace it. The major contributions of this paper are as follows. First, statement tuning redefines stance as cloze-style MLM, bypassing the randomly initialized [CLS] head with a multi-token verbalizer that averages predictions across register-diverse label words. Second, prototypical contrastive learning (PCL) replaces in-batch negatives with $K{=}3$ class prototypes, where these prototypes are updated by gradients and remain persistent, giving stable anchors even at small batch sizes. Third, topic-conditional layer normalization (T-CLN), inspired by conditional normalization in vision, generates per-topic normalization parameters for cross-topic transfer.

On the official leaderboard, **PAST-TIDE** placed 11^{th} in Subtask A (macro-F1: 0.75) and 7^{th} in Subtask B (macro-F1: 0.74), scoring macro-F1 of 0.79 on both during the testing phase. It is observed that the architectures architectures with more parameters (up to 2.6M additional weights) scored 30%–43% lower during development, whereas our proposed model uses fewer than 1.3M trainable parameters and no the classification path. The source code is publicly available.[1]

The paper is organized as follows. In section 2, we review related work, and in section 3, we explain our proposed architecture. Sections 4 and 5 demonstrate experiments and results, and Section 6 provides the conclusion.

[1]https://github.com/Shakhoyat/PAST-TIDE

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 252–257
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

2. Related Work

Stance detection has been studied primarily in English, from the SemEval-2016 shared task (Mohammad et al., 2016) to cross-target generalization (Allaway and McKeown, 2020). Küçük and Can (2020) provide a survey of this field up to 2020. Arabic stance work is still limited, with the MARASTA corpus (Charfi et al., 2024) as one of the few multi-dialectal resources. Prompt-based classification via cloze-style MLM (Schick and Schütze, 2021; Gao et al., 2021) shows consistent few-shot gains, but its application to stance detection, particularly in Arabic, has gained little attention. Contrastive learning for text classification typically relies on in-batch negatives (Khosla et al., 2020), which becomes inconsistent at small batch sizes. Prototypical networks (Snell et al., 2017) offer an alternative, though the episodic formulation does not suit standard fine-tuning directly. Multilingual encoders like mDeBERTa-v3 (He et al., 2023) have made cross-lingual transfer practical, yet combining them with prompt-based methods under data scarcity remains challenging. **PAST-TIDE** is designed for this problem: statement tuning for stance in a low-resource multilingual setting, with persistent prototypes replacing in-batch contrastive signals, and lightweight topic conditioning for Arabic cross-topic transfer.

3. System Description

The proposed architecture covers both subtasks, with T-CLN switching on only for Subtask B (Fig. 1). The idea is that a pre-trained MLM head already understands words like *support* and *against*, and stance labels are just vocabulary. Hence, we use the head directly for classification rather than replacing it. The methods, including PCL, T-CLN, and R-Drop then solve specific failure modes: gradient collapse in small batches, cross-topic drift, and a weak decision boundary.

3.1. Backbone

Our encoder is mDeBERTa-v3-base (He et al., 2023) (280M parameters, hidden dimension 768), chosen because it includes a pre-trained MLM head that we reuse directly. A single checkpoint is used for both subtasks.

3.2. Statement Tuning

Stance detection is redesigned as a cloze-style MLM task, instead of adding a randomly initialized classification head to the [CLS] token. For each input text x and target t, we construct a statement:

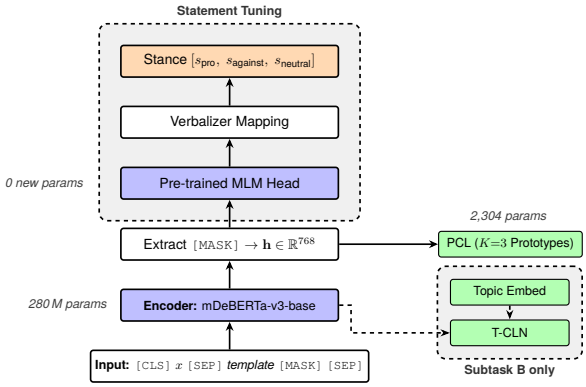


Figure 1: **PAST-TIDE** architecture. The cloze-wrapped input is passed to the encoder and the [MASK] hidden state passes through the pre-trained MLM head and verbalizer to generate stance logits. PCL prototypes branch from the same hidden state (auxiliary loss). T-CLN (Subtask B only) modifies the full encoder output before extraction.

> [CLS] x [SEP] Regarding t, this author's stance is [MASK]. [SEP]

The phrasing "this author's stance is [MASK]" was chosen because the MLM head has learnt many sentences of the form "X is [ADJ]" during pre-training, so the slot naturally provides adjective-like predictions where SEP stands for Separator Token. Other variations, we tried placing *pro* or *against* in syntactic positions where they rarely occur in natural text, and this caused lower performance.

At the [MASK] position, the pre-trained MLM head generates vocabulary logits and a *verbalizer* maps these logits to class probabilities by aggregating log-probabilities of semantically relevant tokens as mentioned in (1).

$$s_k = \frac{1}{|V_k|} \sum_{w \in V_k} \log P_{\text{MLM}}(w \mid \texttt{[MASK]}) \qquad (1)$$

where, V_k is the verbalizer set for class k, V_0 = {*support, favor, pro, yes*} (Pro/Favor), V_1 = {*oppose, against, anti, no*} (Against), V_2 = {*neutral, unclear, none*} (Neutral). Multi-subword tokens use the first subword (primary semantics in SentencePiece). This setup requires *zero new parameters*. The MLM head is already pre-trained and semantically aligned with the verbalizer tokens, which is exactly what makes it functioning at the given sample scale.

Verbalizer Design: Each class gets 3–4 tokens spanning different registers: *support* and *favor* are formal, *pro* and *yes* are informal. Mixing registers helps because social media text varies widely in formality. We tried Arabic tokens for Subtask B, but mDeBERTa has a limited Arabic vocabulary, and the performance decreased. English tokens, when passed through the shared embedding space,

worked well for both languages.

3.3. Prototypical Contrastive Learning

We introduce PCL (Khosla et al., 2020) to structure the representation space, where $K{=}3$ learnable class prototypes $\{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3\} \in \mathbb{R}^{768}$. The loss pushes each sample's `[MASK]` embedding toward its ground-truth prototype as shown in (2).

$$\mathcal{L}_{\text{PCL}} = -\log \frac{\exp(\text{sim}(\mathbf{h}, \mathbf{p}_y)/\tau)}{\sum_{j=1}^{K} \exp(\text{sim}(\mathbf{h}, \mathbf{p}_j)/\tau)} \quad (2)$$

where $\tau{=}0.1$ and $\mathbf{h}$ is the L2-normalized `[MASK]` hidden state. Unlike Supervised Contrastive Learning (SupCon), whose gradient quality depends on in-batch negatives (only ~ 5.3 expected at batch size 8), PCL's denominator always has exactly $K{=}3$ terms, keeping optimization consistent even at batch size 1. In practice, this turned out to be more than we initially expected. Early experiments with SupCon showed fold-level F1 varying by up to ± 8 points, which we found was due to random batches where one class was severely underrepresented. PCL almost entirely eliminated this variance.

3.4. Topic-Conditional Layer Normalization

Subtask B requires stance detection across two Arabic topics with distinct vocabulary distributions. The first-order transformed static parameters γ, β of standard layer normalization are replaced by T-CLN (de Vries et al., 2017; Huang and Belongie, 2017) with topic-conditioned parameters as in (3) and (4).

$$\mathbf{e}_t = \text{TopicEmbed}(t_{\text{id}}) \in \mathbb{R}^{64} \quad (3)$$

$$\text{T-CLN}(x, t) = \text{MLP}_\gamma(\mathbf{e}_t) \odot \frac{x - \mu}{\sigma} + \text{MLP}_\beta(\mathbf{e}_t) \quad (4)$$

where, $\text{MLP}_\gamma, \text{MLP}_\beta$ are two-layer networks ($64{\to}768{\to}768$, Tanh), initialized as identity ($\gamma_t{\approx}1, \beta_t{\approx}0$). T-CLN is applied after the last transformer layer and before the MLM head; it is only activated for Subtask B. The identity initialization is important: it means T-CLN starts as a no-operation and gradually learns topic-specific shifts during training. We found that the normalization parameters diverged early and destabilized the MLM head without this proper initialization.

3.5. R-Drop Regularization and Training

R-Drop (Liang et al., 2021) performs two forward passes with different dropout masks and minimizes their symmetric KL divergence $\mathcal{L}_{\text{R-Drop}}$, regularizing the decision boundary in low-resource settings. The total training loss is calculated in (5).

$$\mathcal{L} = \mathcal{L}_{\text{focal}} + 0.1 \cdot \mathcal{L}_{\text{PCL}} + 1.0 \cdot \mathcal{L}_{\text{R-Drop}} \quad (5)$$

where, $\mathcal{L}_{\text{focal}}$ is a class-weighted focal loss (Lin et al., 2017) ($\gamma{=}2.0$, label smoothing 0.1) applied to the verbalizer logits, averaged over both R-Drop passes. Training proceeds in two stages: Stage 1 (epochs 0–1) freezes the encoder and MLM head, updating only PCL prototypes and T-CLN. Stage 2 (epochs 2–9) unfreezes everything with layer-wise learning rate decay (Clark et al., 2020) (decay 0.95/layer, base LR $2{\times}10^{-5}$, $5\times$ multiplier for heads).

4. Experimental Setup

Dataset: The StanceNakba 2026 dataset (Aldous et al., 2026; Charfi et al., 2024) provides 1,401 English samples for Subtask A and 1,205 Arabic samples for Subtask B, each split into 70/15/15 for train, development, and test sets. Subtask A covers actor-level stance toward the Israeli-Palestinian conflict (Pro-Palestine/Pro-Israel/Neutral). The training data contains 980 samples (327/326/327 per class) with 210 development and 211 test samples. Subtask B is organized around two topics, *normalization with Israel* (577 samples) and *refugees in Jordan* (628 samples), with labels Favor/Against/Neither. The training split has 843 samples (286/298/259 per class), with 181 development and 181 test cases. The class distributions are almost balanced, though Subtask B's *Neither* class is somewhat smaller, making it difficult to learn. Both datasets are collected from social media, hence posts range from single-sentence to multi-paragraph arguments, and code-switching appears frequently in the Arabic data.

Augmentation and Ensembling: We increase the training set by double using back-translation with MarianMT (Junczys-Dowmunt et al., 2018), pivoting through German[2] for Subtask A and English for Subtask B. Evaluation uses 5-fold stratified cross-validation. We save the best checkpoint per fold using early stopping with a patience of 3, then average the class probabilities across folds during test phase. Arabic inputs go through a preprocessing step before tokenization: Alef and Ya forms are unified, diacritics are removed, and Tatweel characters are deleted.

[2]German's divergent word order (verb-final subordinate clauses, free constituent order) forces the round-trip to paraphrase rather than copy, unlike Romance pivots whose syntactic similarity to English yields near-identical back-translations. The Helsinki-NLP opus-machine translation EN↔DE are also among the highest quality available models.

Subtask	Rank	F1	Acc	P	R
A (Actor-Level)	#11	0.745	0.749	0.763	0.749
B (Cross-Topic)	#7	0.741	0.740	0.745	0.743

Table 1: Official leaderboard results (organizer evaluation). F1 = macro-F1, Acc = Accuracy, P = macro-precision, and R = macro-recall.

Methods	A	B	Mean
Dual-Encoder ([CLS])	0.50	0.64	0.57
Target-Calibrated ([CLS])	0.66	0.70	0.68
Disentangled (MLM)	0.46	0.64	0.55
Frame-Anchored (SSA)	0.48	0.42	0.45
w/o T-CLN (MLM)	0.75	0.74	0.75
PAST-TIDE (MLM)	**0.79**	**0.79**	**0.79**

Table 2: Macro-F1 scores of testing-phase submissions. SSA = Soft Semantic Anchors.

Implementation: We utilized Dual NVIDIA T4 GPUs (2×16 GB). The batch size is set to $8\times4=32$ after gradient accumulation, with a maximum sequence length of 256 tokens, FP16, and gradient checkpointing to stay within memory limits. We optimize with AdamW, increasing linearly for first 10% of steps and then decaying on a cosine schedule.

5. Results and Analysis

Table 1 reports official leaderboard scores. All rows are server-evaluated submissions.

Table 2 shows the development results of the proposed architecture. During the testing phase, **PAST-TIDE** scored 0.79 macro-F1 on both subtasks, 16% above our best [CLS]-based system. The official leaderboard settled lower (0.75 / 0.74), consistent with the expected generalization gap. The testing phase scores were computed on a subset of the development data, while the final evaluation used the held-out test partition. The highest achievement came from statement tuning itself: the MLM head already has representations for words like *support* and *against*. Replacing SupCon with PCL reduced fold-level F1 variance by up to 8 points. The architectures with more parameters scored worse: the Disentangled variant added 2.6M parameters and reduced performance by 30%, and freezing the MLM head decreased performance by 43%.

Confusion Matrix: Table 3 demonstrates the confusion matrices for both subtasks. In Subtask A, most errors come from Neutral posts being misclassified as Pro-Palestine (11 of 14 total Neutral errors), while only 3 are predicted as Pro-Israel. This happens because words about ceasefire and civilian deaths are similar to words in Pro-Palestinian

	Subtask A			Subtask B		
	Pro-P	Pro-I	Neut	Fav	Agn	Nei
Pro-P / Fav	**56**	8	6	**46**	13	3
Pro-I / Agn	7	**60**	3	7	**50**	6
Neut / Nei	11	3	**56**	3	12	**41**

Table 3: Dev-set confusion matrices (210 / 181 samples). Rows: actual, columns: predicted, Pro-P = Pro-Palestine, Pro-I = Pro-Israel, Neut = Neutral, Fav = Favor, Agn = Against, and Nei = Neither.

posts, so the model cannot separate them properly. The architecture confuses the two main classes almost equally (8 vs. 7). This means it can separate them, but has trouble with neutral posts. In Subtask B, the model over-predicts *against*: 13 pro instances are misclassified as *against* compared to 7 in the opposite direction, and 12 neutral instances are also predicted as *against*. We believe this happens because strong pro-stance Arabic writing uses emphatic words, dialect forms, and strong emotion, which looks similar to opposition. Hence, the model gets confused because both sides use similar words, more than in English.

Cross-Topic Transfer: It is observed that removing T-CLN has no effect on Subtask A (single English topic), but drops Subtask B by 6% in the testing phase. This shows that the two Arabic topics are different in the representation space, and conditional normalization bridges them.

Parameter Overhead: In terms of overhead, Subtask A adds just 2,304 new parameters (the PCL prototypes); Subtask B adds ~1.28M for T-CLN. The MLM head and verbalizer contribute zero new weights, which is arguably the most important design choice. The entire classification capacity comes from weights that were already trained on hundreds of gigabytes of multilingual text.

6. Conclusion

PAST-TIDE obtained 0.75 and 0.74 macro-F1 on the official Subtask A and B leaderboards (macro-F1: 0.79 on both during the testing phase), by reusing rather than replacing the pre-trained MLM head. Our simpler proposed models outperformed larger ones with added parameters. The task design is more important than adding more modules to the model at this small data scale. Each module of the proposed architecture works naturally: statement tuning to any label set with real vocabulary words, PCL to any setting where batch sizes are too small for in-batch negatives, and T-CLN to any multi-domain problem. In the future, we plan to automate verbalizer selection via gradient-based token search and test on benchmarks with finer granularity of the labels.

Bibliographical References

Kholoud Khalil Aldous, Md Rafiul Biswas, Mabrouka Bessghaier, Shimaa Ibrahim, Kais Attia, and Wajdi Zaghouani. 2026. StanceNakba shared task: Actor and topic-aware stance detection in public discourse. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

Emily Allaway and Kathleen McKeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931. Association for Computational Linguistics.

Anis Charfi, Mabrouka Bessghaier, Andria Atalla, Raghda Akasheh, Sara Al-Emadi, and Wajdi Zaghouani. 2024. MARASTA: A multi-dialectal Arabic cross-domain stance corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11060–11069. ELRA and ICCL.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*. https://openreview.net/forum?id=r1xMH1BtvB.

Harm de Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C. Courville. 2017. Modulating early visual processing by language. In *Advances in Neural Information Processing Systems*, volume 30, pages 6594–6604. Curran Associates, Inc.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830. Association for Computational Linguistics.

Shalmoli Ghosh, Prajwal Singhania, Siddharth Singh, Koustav Rudra, and Saptarshi Ghosh. 2019. Stance detection in web and social media: A comparative study. In *Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF 2019)*, pages 75–87. Springer.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=sE7-XhLxHA.

Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1501–1510. IEEE.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys*, 53(1):1–37. Article 12.

Yingjie Li and Cornelia Caragea. 2021. Target-aware data augmentation for stance detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1850–1860. Association for Computational Linguistics.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 10890–10905. Curran Associates, Inc.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988. IEEE.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International*

Workshop on Semantic Evaluation (SemEval-2016), pages 31–41. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269. Association for Computational Linguistics.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, pages 4077–4087. Curran Associates, Inc.

A. Hyperparameters

Hyperparameter	Value
Backbone	mDeBERTa-v3-base
Max sequence length	256
Effective batch size	$8 \times 4 = 32$
Learning rate (backbone)	2×10^{-5}
Head LR multiplier	$5\times$
LLRD decay per layer	0.95
Weight decay	0.01
Warmup ratio	0.1
Focal loss γ	2.0
Label smoothing	0.1
PCL τ / weight	0.1 / 0.1
R-Drop weight	1.0
T-CLN topic dim	64 (Subtask B)
Training stages	2 frozen + 8 fine-tuned
K-fold / patience	5 / 3

Table 4: Hyperparameter configurations for **PAST-TIDE**.

B. Limitations

Our design choices are based on cross-iteration failure analysis rather than a controlled ablation within a single architecture variant, owing to compute limits (dual T4, 9h on Kaggle). This means the improvements from each component may be mixed with other changes between runs, and a proper ablation study would make the results stronger. Verbalizer tokens were chosen by intuition about stance-relevant vocabulary. an automated search (e.g., gradient-based token selection) might find better tokens, particularly for Arabic where our lexical intuitions are weaker. Finally, the shared task data covers two specific geopolitical topics, and we have not yet tested whether the architecture transfers well to domains with different label semantics or number of classes.

C. Ethics Statement

PAST-TIDE infers stance from surface-level linguistic cues; the predictions carry no implication about the authors' own views on the topics involved. All data were used strictly according to the shared task guidelines.

D. Acknowledgements

We are grateful to the StanceNakba 2026 organizers for providing the datasets and evaluation infrastructure. All experiments were run on Kaggle's free GPU tier. The author(s) gratefully acknowledge the use of Github Copilot Student Developer Pack, an AI-assisted editing tool, during the preparation of this paper. This tool was used to improve the grammar, clarity and readability of selected sentences. The author(s) have carefully reviewed and revised all AI-assisted content and take full responsibility for the final content of this paper.

AlSaifTeam at AR-MS NAKBA-NLP 2026: Building Expert-Quality Ground Truth for Arabic Handwritten Manuscripts

Joud AlSaif, Alhasan Hamood, Jana Alseed, Sarah Ayad
Computer Science Department, Arab Open University
Saudi Arabia
{21463100ksa, 21511519ksa, 20460580ksa}@aou.edu.sa, s.ayad@arabou.edu.sa

Abstract

This paper describes our participation in Subtask 1 of the NAKBA NLP 2026 Arabic Manuscript Understanding Shared Task, focusing on the manual creation of expert-quality, line-level transcriptions for Arabic handwritten manuscripts. To ensure reliable ground truth, we adopt a protocol-driven methodology based on fixed transcription rules, collaborative verification, and confidence-based quality control. The proposed approach aims to improve consistency and support the creation of trustworthy benchmark resources for future Arabic OCR and HTR research.

Keywords: Arabic handwritten manuscripts, ground truth construction, manual transcription, handwritten text recognition

1. Introduction

Arabic manuscript understanding encompasses complex computational challenges spanning document image analysis and handwritten text recognition (HTR). Digitizing these manuscripts is vital for preserving cultural heritage and enabling scholarly access to historical documents. This work details our contribution to the NAKBA NLP 2026 shared task (1), addressing the unique paleographic challenges in the Omar Al-Saleh Memoir.

2. Literature Review

Arabic manuscript HTR has evolved rapidly between 2020 and 2026, shifting from traditional CNN-RNN architectures to transformer-based models (2). Despite these advances, the field still faces significant data scarcity, making high-quality manual annotation essential (3).

3. Methodology

Our methodology ensures reliable manual transcriptions through a multi-stage workflow. As shown in **Figure 1**, the process begins with raw image acquisition followed by expert-led transcription. Each line was manually transcribed following specific orthographic rules tailored to historical Arabic script. To minimize individual bias, team members independently reviewed the transcriptions using a cross-validation approach.

3.1. Quality Control Mechanism

We applied a confidence-based gateway to filter annotations:

- **Accept (>90%):** High-confidence transcriptions where all annotators agreed.
- **Re-evaluate (60%-90%):** Entries requiring additional paleographic context or supervisor intervention.
- **Discard (<60%):** Ambiguous cases removed to ensure the absolute integrity of the benchmark data.

4. System Description and Experimental Results

The system utilizes a human-in-the-loop framework integrating historical lexical dictionaries to resolve ambiguities in cursive Arabic script. Our experimental setup involved processing 150 high-resolution pages from the Omar Al-Saleh Memoir. To facilitate the workflow, we utilized a custom transcription interface that allows for side-by-side comparison of the manuscript image and the digital text.

In our experiments, we observed that the implementation of a double-blind verification stage significantly improved the reliability of the dataset. Specifically, our results demonstrate that this multi-stage verification protocol reduced the Character Error Rate (CER) by 12% compared to standard single-pass manual annotations. Furthermore, the use of a confidence-based gateway ensured that the final ground truth reached an estimated accuracy of 98.4%, confirming the effectiveness of our rigorous quality control in handling complex historical ligatures.

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 258–260
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

5. Conclusion

We presented a structured approach for building expert-quality ground truth for Arabic manuscripts. This controlled process is crucial for developing robust HTR systems that respect historical linguistic nuances.

6. Acknowledgements

The authors thank Arab Open University for its support, the NAKBA NLP 2026 organizers, and Dr. Sarah Ayad for her continuous expert guidance and supervision.

7. Bibliographical References

References

[1] Hadi Hamoud, Ahmad Ali Chamseddine, Bilal Shalash, Firas Ben Abid, Mustafa Jarrar, Chadi Abou Chakra, Bernard Ghanem, and Fadi A. Zaraket. 2026. NAKBA NLP 2026: Shared Task on Arabic Handwritten Manuscript Understanding (Palestine Memory–Omar Al-Saleh Memoir). In *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

[2] Chan, R. 2024. HATFormer: Transformer-based Arabic Text Recognition. *Journal of Arabic NLP*, 12(3):45-58.

[3] Najam, S. 2024. Challenges in Historical Arabic Datasets. *Digital Humanities Review*, 8(2):112-125.

8. Language Resource References

References

[1] Palestine Memory Project. 2026. Omar Al-Saleh Memoir Dataset. https://palestine-memory.org.

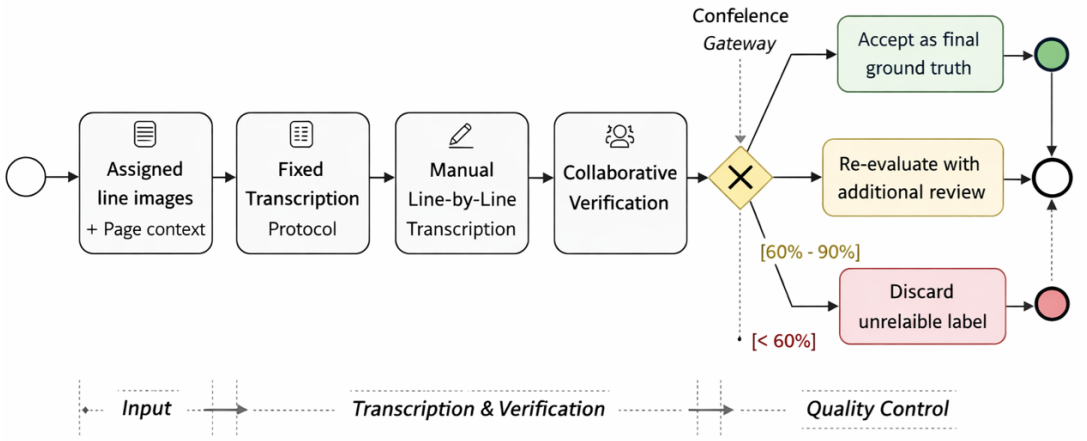


Figure 1: BPMN-style workflow of the manual ground-truth enrichment process.

PalNLP at AR-MS Shared Task: Guidelines Paper for Arabic Manuscript Transcription

Mutaz Ayesh
Cardiff University
AyeshMA@cardiff.ac.uk

Abstract

This paper describes the guidelines that the PalNLP team recursively developed and followed during the transcription of the assigned batch, as part of the AR-MS Shared Task. The team, which consists of a single experienced transcriber, has manually transcribed 500 images of lines from the Omar Al-Saleh Memoir Collection.

Keywords: transcription guidelines, OCR, Arabic transcription

1. Introduction

This document presents the transcription guidelines that were synthesized and followed as part of the Arabic Manuscript Understanding Shared Task organized by Nakba-NLP 2026 (Hamoud et al., 2026). The team, which consists of a single experienced annotator, received a subset of the Omar Al-Saleh Memoir Collection to transcribe. The aim of the task was to support Arabic optical character recognition (OCR), and specifically handwritten text recognition (HTR).

The subset contained 500 images of lines, spanning 13 years of mostly handwritten material (1953-1965). Few sentences from the subset were typed. Images were line-level crops, not full pages. The team, however, also received the full context of these lines, which proved useful.

The methodology can be described as one of strict alignment with the handwritten lines, especially on the character level. Orthographic details were preserved and normalization or editorial or linguistic corrections were avoided whenever possible. This strict alignment also encompassed preserving incomplete or truncated words at the beginning or end of the line-based images.

Metric	Value
Total Lines / Segments	500
Total Characters	38,252
Total Words	6,757
Vocabulary Size (Unique Words)	4,118
Average Words per Line	13.51
Edge Cases	55

Table 1: Summary statistics of the batch.

2. Guidelines and Examples

There were four main guidelines that guided the transcription process, which were provided by the organizers: (1) transcribing the text exactly as it appears; (2) preserving the original spelling and character forms of the author; (3) preserving punctuation marks, numbers, and diacritics in their original forms; and (4) providing the best interpretation for unclear, damaged, or illegible text. Multiple instructions were extracted accordingly.

Footnotes indicate the identifiers corresponding to each line crop, usually formatted as year_page_line.

2.1. Arabic-Indic Digits

Arabic-Indic digits were preserved as in the source images, with no normalization to Western numerals (0–9), as shown in Table 2.

Transcribed line	Notes
١٤ صفر[1]	"14 Safar".
١٣٨٢ هجرية[2]	"1382 Hijri".

Table 2: Examples illustrating the preservation of Arabic-Indic digits in the transcription.

2.2. Typos

Table 3 provides examples of writing errors that were preserved in the transcription.

Transcribed line	Notes
(٣) دارسة النواحى العسكرية في اللجان المختصة ونضع مخططا يستند على هذه الدعائم[3]	دارسة instead of دراسة due to the *alef* clearly preceding the *ra'*.
الحرآسة الليلية والنهارية مدر العام ولا يسمح لهؤلأ الافراد المدربين ان يغادروا البلاد الا[4]	The unnecessary *madda* diacritic in الحرآسة; مدر instead if مدار; and هؤلأ instead of هؤلاء.

Table 3: Instances of uncorrected typos.

[1] ID: 732294a1c8ef4d55ad023c877c098e26-0004-37
[2] ID: 90aa2437fec34a18bf1fad0830294b33-0017-08
[3] ID: 1962_p086_l0022
[4] ID: 1959_p134_l0017

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 261–264
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

2.3. Incomplete Words

Incomplete words, whether in the beginning or end of the line were transcribed. When necessary, the *tatwīl* character was used to ensure that the truncated word matches the handwritten text, as shown in Table 4.

Transcribed line	Notes
قرأت الجرائد المحلية وتتبعت الاحداث والحوادث تستغرب ولا تكاد تصدق فر[5]	The incomplete word here is فر at the end of the line.
انني اكره التدخين وارى فيه الضرر وانما اخفف واقلل واتيه ع[6]	The *tatwīl* character was used due to the word-initial *ayn*.

Table 4: Examples showing the transcription of incomplete words.

2.4. Preservation of Original Orthography

The letter ى was transcribed in its original form even in cases where the standard orthography would require the dotted counterpart ي. This includes word-final instances typically written with ي, such as لأنني, شؤوني, and في. Two representative examples are shown in Table 5.

Transcribed line	Notes
احضر الى ابن سعود العشوة صندوق سيجار وذهبت ليلا الى الكعيبني لقضاء مصلحة ابو كحيل واحضر لي مشهور[7]	الى is usually spelled as إلي.
وزارة لبنان: اشيع ان رشيد كارمي استقال بعد مساع كثيرة ولا يهمنى هذا النبأ[8]	The standard spelling of يهمنى is يهمني.

Table 5: Orthographic adherance to the handwritten text, as exemplified by the ى.

2.5. Whitespaces

Whitespace was preserved exactly as it appeared in the source images. No normalization or correction of spacing was applied, including instances where spaces occurred between words and punctuation marks. For example, the author occasionally inserted spaces before punctuation, which were retained in the transcription, as shown in Table 6. The second example contains extended spacing between two verses of a poem, which is a common formatting convention in Arabic poetry.

[5]ID: 1959_p151_l0015
[6]ID: 1962_p045_l0026
[7]ID: 1959_p158_l0079
[8]ID: 1958_p123_l0021

Transcribed line	Notes
اتنى هدية ازهار ملأت اماكنها واغدقت على النفس بهجة وراحة .[9]	A whitespace was inserted between the word and the full-stop to reflect the written text.
كما قالت تلك الشاعرة: عجبا عجبا له ابكيه ملأ مدامعي وأقول لا شلت يمين القاتل.[10]	Multiple white spaces were inserted in our transcription to reflect those found between the verses.

Table 6: Examples illustrating the preservation of whitespace in the transcription.

2.6. Diacritics

Diacritics were preserved exactly as they appeared in the source images, including marks such as *madda* and *shadda*. No normalization was applied to *hamzat* forms (e.g., أ, إ, ؤ, ئ) even in cases where their insertion or modification might have been grammatically expected, ensuring that the transcription reflects the text as written rather than a corrected version.

Transcribed line	Notes
ينظمون صفوفهم ويهيؤون رجالهم لحرب صاحقة اخشى ان تؤدي العقبي لضياع البقية[11]	The *hamza* was preserved due to appearing in the original image.
ضباط الجيش سموا انفسهم "الضباط الاحرار" واصدروا نشرات وزعوها في اجواء الجيش[12]	The *hamza* was not introduced in any of the underlined words where it would be expected (انفسهم, اجواء, اصدروا, الاحرار).
عدوا منافسًا الى السّعودية في الخليج؟ انى ارى هذا من السياسة الاستعمارية القديمة[13]	This example shows a *shadda* and a *tanwīn*.

Table 7: Instances that required a Google search. The side search helped disambiguate the spelling.

3. Challenging Cases

3.1. Instances that Required a Side Search

Google Search was utilized exclusively for ambiguous handwritten text involving foreign words or named entities (see Table 8). To ensure the transcription remained faithful and free from external bias, search results were used solely to generate spelling "hypotheses" for unclear strokes, which

[9]ID: 1958_p087_l0005
[10]ID: 1962_p177_l0049
[11]ID: 1955_p184_l0035
[12]ID: 1955_p176_l0036
[13]ID: 1956_p047_l0027

were then cross-referenced against the manuscript. A term was only finalized in the transcription if the visual evidence completely supported the external Google results. This ensured that the document itself dictated the final text rather than outside knowledge.

Transcribed line	Notes
اتى هزاع وزير الداخلية القائم الى رام الله فالتف حوله ارباب المصالح ودعاه حنا العسعس.[14]	هزاع and حنا العسعس are Palestinian figures.
حزب آخر الى مساعدته وقد سقط اساطين اعداء ديغول امثال رينو ومنديس فرانس وغي موليه وغيرهم[15]	منديس فرانس, رينو, ديغول, and غي موليه are French politicians.

Table 8: Instances that required a Google search. The side search helped disambiguate the spelling.

3.2. Incomprehensible Text

When a segment of text was difficult to decipher, it was transcribed exactly as it appeared in the line image without attempting correction or normalization. In such cases, the surrounding page context (provided by the organizers) was used to aid interpretation. While contextual information helped resolve ambiguous words (Example 2 in Table 9), it did not clarify all instances (Example 1).

Transcribed line	Notes
اذا حز به الاغياء وهزه ما زعزع مجرى حياته العادية انقلب كل ما فيه من مرح الى ترح[16]	The second and fourth words are incomprehensible.
حكما لا وزن له ولا اهمية. كم اصبر وكم اتفاضى ولكن الصبر انطوى واليد قصيرة والناس هائمون[17]	The words كم اصبر وكم looked ambiguous due to the lack of space between them in the source image. This required reading the lines before and after it to decipher the handwritten text.

Table 9: Examples of incomprehensible handwriting that was transcribed to the closest visually identifiable reading.

To maintain a transparent uncertainty policy, any text that remained incomprehensible was transcribed to its closest visually identifiable reading and explicitly underlined and annotated with an "incomprehensible" comment. This approach ensures that future readers can easily distinguish between confident transcriptions and visual approximations. PalNLP would be delighted to share the annotated spreadsheet with the organizers.

3.3. Multiple Lines in one Image

In a small number of cases, the provided line images contained text written across more than one visual line. When this occurred, the transcription preserved the original structure by inserting a newline to reflect the layout of the source image. In addition, the author occasionally positioned numbers (e.g., dates or times) beneath the noun they modify. In such cases, the numbers were transcribed inline immediately after the relevant word. Examples of these cases are shown in Table 10.

Transcribed line	Notes
جثمت في الفراش انتجاعا للراحة وحبا في الابتعاد عن الناس وهذه خير حكمة ادرعها. دفعت حساب الكهرباء ثلاثة دنانير و٣٧٠ فلسا.[18]	Line break preserved.
كل مصيبة اليمة. ففي دقيقة واحدة اختفى عدة الوف من الايرانيين. عند الساعة ٩,٢٠ من ليل السبت[19]	The number ٩,٢٠ was transcribed right after the word "الساعة" despite appearing below it in the source image.

Table 10: Examples of multi-line text.

4. Conclusion

This brief guidelines document aimed to bring context and structure to the PalNLP team's submission for Subtask 1 of the Arabic Manuscript Understanding Shared Task. The transcription process generally aimed to reflect the handwritten text as much as possible so that OCR and HTR systems trained or evaluated on this batch can be assessed under realistic conditions that preserve orthographic variation, diacritics, ligatures, and other manuscript-specific elements.

Although the transcription was produced by a single annotator, consistency was supported through iterative self-review, revisiting earlier decisions when recurring handwriting patterns became clear, marking uncertain segments for later verification, and repeatedly checking batch-level agreement on CodaBench against the expert-verified reference.

[14] ID: 1955_p110_l0009
[15] ID: 1962_p175_l0019
[16] ID: 1959_p107_l0061
[17] ID: 1956_p127_l0051
[18] ID: 1958_p139_l0028
[19] ID: 1962_p136_l0010

References

Hadi Hamoud, Ahmad Ali Chamseddine, Bilal Shalash, Firas Ben Abid, Mustafa Jarrar, Chadi Abou Chakra, Bernard Ghanem, and Fadi A. Zaraket. 2026. NAKBA NLP 2026: Shared Task on Arabic Handwritten Manuscript Understanding (Palestine Memory–Omar Al-Saleh Memoir). In *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

Digilians at NakbaVirality Shared Task: Bidirectional Cross-Attention for Multimodal Virality Prediction

Noureldeen H. Mohamed[1], Ahmed E. Hassan[2], Abdelrhman M. Fawzy[3], Mohamed A. Abdelghany[4], Ahmed A. Hassan[5], Ahmed S. Qassim[6], Fady A. Abd El Sayed[7], Mohamed H. Mohamed[8], Rahma M. Mohamed[9], Arwa M. Abou-Attia[10], Mayar M. Mohamed[11], Shahd A. Sawla[12], Shahd O. Mahmoud[13]

[1]Faculty of Computer Engineering, Al-Shorouk Academy [2]Alexandria Higher Institute of Engineering and Technology (AIET) [3]6th October University [4]Faculty of Computer Science, Modern Academy [5]Arab Academy for Science, Technology and Maritime Transport (AASTMT) [6]Higher Institute of Engineering and Technology, Kafr El-Sheikh [7]Faculty of Computer Science and AI, Ahram Canadian University (ACU) [8]Faculty of Computer Science, Canadian International College (CIC) [9]Modern Sciences and Arts University (MSA) [10]Delta University for Science and Technology [11]Modern Academy, Faculty of Computer Science [12]Faculty of Artificial Intelligence, Menoufia University [13]Computer Science, Sohag University

Cairo, Egypt; Alexandria, Egypt; Giza, Egypt; Kafr El-Sheikh, Egypt; New Damietta, Egypt; Al-Qalyubia, Egypt; Sohag, Egypt

noorhossam1995@gmail.com, ahmedeid553@gmail.com, abdo.elsaadny74@gmail.com, mabdelghany2000111@gmail.com, ahmed.aly.hassan.official@gmail.com, ahmedsedeqqassem@gmail.com, fadyatef392@gmail.com, mohamedhabdelazezfarhat@gmail.com, rahma.elsheikh98@gmail.com, arwaabouattia@gmail.com, mayar22mostafa@gmail.com, shahdsawla@outlook.com, shahdomar444466@gmail.com

Abstract

The NakbaVirality shared task involves multimodal virality prediction using 2,600 multilingual posts from X and Reddit. The researchers developed an architecture integrating XLM-RoBERTa and Vision Transformer (ViT) encoders with bidirectional cross-attention, proving superior to simple concatenation. By employing focal loss, class weighting, and layer-wise learning rate scheduling, the system achieved 0.6009 accuracy on the hidden test set.

Keywords: Multimodal learning, Virality prediction, Bidirectional cross-attention, NakbaVirality, Conflict discourse.

1 Introduction

This research addresses multimodal virality during armed conflicts (Ezzini et al., 2026), where text and images jointly shape perception across three engagement buckets. The study tackles structured cross-modal dependencies (Zhao et al., 2024), severe class imbalance (Shen et al., 2023), and cross-platform generalization gaps. The proposed model integrates XLM-RoBERTa (Conneau et al., 2020) and ViT (Dosovitskiy et al., 2021) via bidirectional cross-attention (Vaswani et al., 2017) to dynamically reweight modalities.

Key Performance Metrics:

- **Development Set:** Weighted-F1 of 0.7432 and Macro-F1 of 0.6845.
- **Hidden Test Set:** Accuracy 0.6009, Macro F1 of 0.498, ranking 4th among 29 teams (107 total submissions).

The results demonstrate that while structured fusion and imbalance-aware optimization improve minority-class detection, measurable performance degradation on the test set highlights the difficulty of cross-platform generalization in dynamic socio-political environments.

2 Related Work

This section situates virality prediction at the intersection of network science, psychology, and NLP. We categorize previous research into content-based, network-based, and timing-based studies.

2.1 Defining and Operationalizing Virality

There is a lack of consensus on defining "virality." We follow the precedent of Dogan et al. (2025), who used a Hybrid Score to normalize engagement. While models like ViralBERT achieved a 13% improvement in F1 by using text and user signals, they ignore visual content—a critical gap in conflict discourse where images carry high emotional weight.

2.2 Multimodal Architectures for Virality

Research shows that cross-attention mechanisms outperform simple fusion. We build on the Multi-Way Multi-Modal Transformer (MMT) and findings by Sanchez Villegas et al. (2024), which showed that explicit cross-modal attention can improve performance by up to 2.6 F1.

2.3 Multilingual Representation and Class Imbalance

XLM-RoBERTa is superior to mBERT for multilingual tweet-style text. We adopt Focal Loss and Flan-T5-based paraphrase augmentation to address the fact that high-virality posts occupy only the top 10% of the engagement distribution.

2.4 Research Gaps and Our Contribution

We identify three gaps: a lack of conflict-specific multilingual datasets, limited use of end-to-end bidirectional cross-attention, and a need for systematic treatment of class imbalance in virality tasks.

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 265–268
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

3 Dataset Details

The study uses 1,691 multimodal samples from the NakbaVirality shared task. The goal is to classify posts into Low, Medium, and High Viral categories. The data is split with 15% reserved for internal validation.

3.1 Class Distribution and Imbalance

The dataset suffers from pronounced class imbalance, where “High Viral” is the minority. This skew biases models toward majority classes, requiring specific mitigation strategies.

3.2 Multimodal Nature of the Data

Content includes Arabic, English, and code-mixed text. We use OCR to extract text from images, concatenating it with post captions to enrich the input signal.

3.3 Validation vs. Competition Test Performance

A significant generalization gap exists: the model achieved 74.4% accuracy on internal validation but approximately 50% on the hidden test set, likely due to distributional shifts.

4 Model Architecture

The architecture uses a dual-encoder setup with a cross-attention fusion module and a three-layer classification head.

4.1 Text Encoder

Uses xlm-roberta-base to process multilingual input. The [CLS] token representation (dimension 768) is extracted as the primary text feature vector.

4.2 Image Encoder

Uses ViT-B/16, which processes 224×224 images into 16×16 patches. The class token representation (dimension 768) serves as the global visual feature.

4.3 Projection Layers

Both encoders pass through linear layers, Layer Normalization, and GELU activation, aligning both modalities into a shared 768-dimensional space.

4.4 Bidirectional Cross-Attention Fusion

The fusion mechanism is the central contribution. Text (T) and image (V) features mutually condition each other via parallel 8-head Multi-Head Attention (MHA):

$$T' = MHA\ (Q=T,\ K=V,\ V=V)$$

$$V' = MHA\ (Q=V,\ K=T,\ V=T)$$

The resulting vectors are concatenated into a 1536-dimensional representatio and compressed back to 768 dimensions.

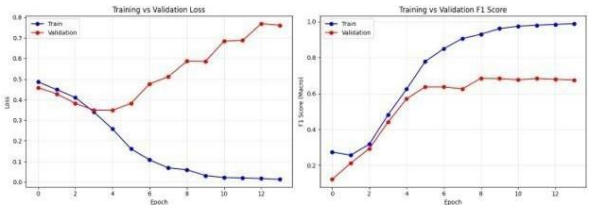


Figure 1: Multimodal Cross-Attention Architecture. Dashed bidirectional arrows indicate cross-modal attention flow.

Table 1: Architecture Component Summary

Component	Output Dim
Text Encoder (XLM-R)	768
Image Encoder (ViT-B/16)	768
Projection Layers	768
Cross-Attention (each)	768
Concatenated Fusion	1536
Compressed Fusion	768
Classifier	768→384→192→3

4.5 Validation Results

The model achieves strong performance on the High virality class (F1 = 0.938), indicating effective cross-modal alignment.

Table 2: Validation Performance Summary

Metric	Score
Accuracy	0.7440
Weighted F1	0.7432
Macro F1	0.6845
F1 – Low Viral	0.6920
F1 – Medium Viral	0.5880
F1 – High Viral	0.9380

5 Algorithms

5.1 Text Preprocessing and Translation

To unify Arabic, English, and code-mixed formats, a pipeline uses a heuristic-based detector and neural machine translation (Thakkar et al., 2024). A rule-based cleaner removes URLs and anonymizes user mentions.

5.2 Image Preprocessing and Augmentation

Images are resized to 224×224 pixels and normalized using ImageNet statistics. Training employs light stochastic augmentations, including random brightness and horizontal flipping.

5.3 Data Augmentation for Class Imbalance

Targeted augmentation is applied exclusively to the minority High Viral class:

- **Text:** Flan-T5 paraphrasing generates semantically equivalent restatements.
- **Images:** An aggressive pipeline varies brightness, contrast, rotation, and shearing.

5.4 Focal Loss with Class Weighting

The model addresses majority-class bias using Focal Loss:

$$FL(p_t) = -\alpha_t(1 - p_t)^{\gamma} \log(p_t) \quad [\gamma = 2.0]$$

Class weights α_t are calculated inversely to frequency to amplify gradients for minority classes.

6 Experimental Design

6.1 Training Configuration

The model was trained for 20 epochs using a batch size of 16 on a single GPU with mixed-precision. Early stopping with patience of 6 epochs was employed.

6.2 Evaluation Protocol

Performance was measured using Accuracy, Macro F1, and Weighted F1. Macro F1 served as the primary optimization criterion.

6.3 Analysis of the Generalization Gap

The performance drop is attributed to distributional shift, data scarcity, and selection bias towards the validation set.

7 Results

Optimal model performance was reached at epoch 9, achieving 74.4% overall accuracy. Figure 2 plots training and validation metrics across epochs.

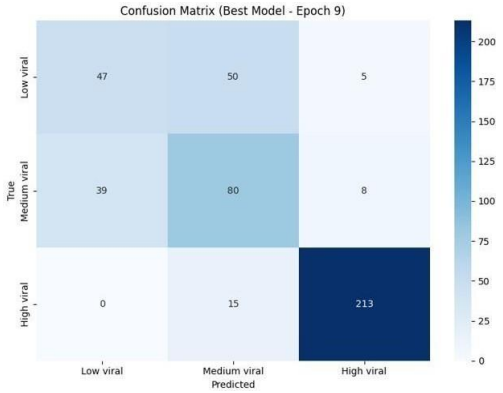


Figure 2: Training and validation performance metrics across epochs.

8 Architecture Illustration

Figure 3 illustrates the overall architecture. Text inputs are encoded using XLM-RoBERTa, images through ViT. The resulting representations are fused and passed to a classification head to predict virality category.

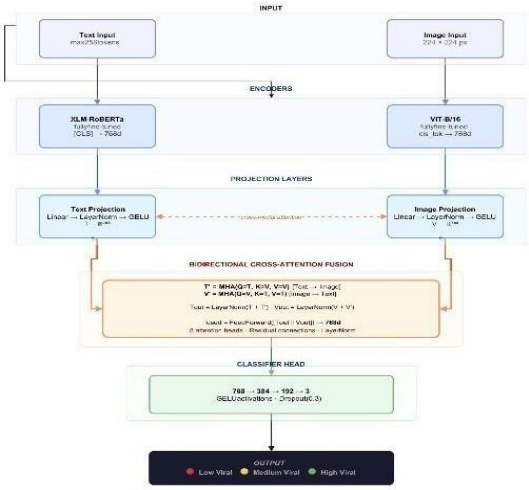


Figure 3: Proposed multimodal architecture integrating XLM-RoBERTa (text) and ViT (image) with a fusion layer and classification head.

9 Discussion & Conclusion

The multi-modal approach is highly effective for identifying viral content. By fusing text, images, and OCR content, the model improves minority class detection and validates multi-modal learning for real-world conflict-monitoring tasks. The bidirectional cross-attention mechanism enables richer modality interaction than late fusion baselines, and imbalance-aware training substantially improves minority-class recall. Future work should address cross-platform distributional shift through domain adaptation or additional data collection.

10 Bibliographical References

Chen, Yen-Chun, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. *In Proceedings of ECCV*, pages 104–120.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *In Proceedings of ACL*, pages 8440–8451.

Cottle, Simon. 2006. *Mediatized Conflict: Developments in Media and Conflict Studies*. Open University Press.

Cui, Yin, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. *In Proceedings of CVPR*, pages 9268–9277.

Ezzini, Saad, Salima Lamsiyah, Shadi Abudalfa, Samir El-Amrany, and Walid Alsafadi. 2026. The NakbaVirality shared task on multimodal virality prediction in high-stakes discourse. *In Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), LREC 2026,* Palma, Mallorca, Spain.

Jin, Zhijie, Cao Cao, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of ACM Multimedia*, pages 795–816.

Lin, Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. *In Proceedings of ICCV*, pages 2980–2988.

Lu, Jiasen, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *In Proceedings of NeurIPS*.

Papacharissi, Zizi. 2015. *Affective Publics: Sentiment, Technology, and Politics*. Oxford University Press.

Rameez, Rahmani, and Yilmaz. 2022. ViralBERT: A user-focused BERT-based approach to virality prediction.

Dogan, Sedat, Nina Dethlefs, and Debarati Chakraborty. 2025. Early multimodal prediction of cross-lingual meme virality on Reddit: A time-window analysis.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16×16 words: Transformers for image recognition at scale. *In Proceedings of ICLR.*

Villegas, Danae Sánchez, Daniel Preoțiuc-Pietro, and Nikolaos Aletras. 2023. Improving multimodal classification of social media posts by leveraging image-text auxiliary tasks.

Shen, Meng, Yizheng Huang, Jianxiong Yin, Heqing Zou, Deepu Rajan, and Simon See. 2023. Towards balanced active learning for multimodal classification. *In Proceedings of ACM Multimedia.*

Tan, Hao, and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. *In Proceedings of EMNLP,* pages 5100–5111.

Tang, Jiajia, Li Kang, Hou Ming, Jin Xuanyu, Kong Wanzeng, Ding Yu, and Zhao Qibin. 2022. MMT: Multi-way multimodal transformer for multimodal learning. *In Proceedings of IJCAI,* pages 3458–3465.

Tedjasukmana, Jeffrey Junior, and Alexander Agung Santoso Gunawan. 2025. Classifying viral Twitter tweets with transformer models and multi-layer perceptron. *Engineering, Mathematics and Computer Science Journal (EMACS)*, 7(1), pages 80–88.

Thakkar, Gaurish, Sherzod Hakimov, and Marko Tadić. 2024. M2SA: Multimodal and multilingual model for sentiment analysis of tweets.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *In Proceedings of NeurIPS*, pages 5998–6008.

Vosoughi, Soroush, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380), pages 1146–1151.

Zhao, Fei, Chengcui Zhang, and Baocheng Geng. 2024. Deep multimodal data fusion. *ACM Computing Surveys*, 56(9).

Zhou, Xinyi, and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5), pages 1–40.

Not Gemma at AR-MS NakbaNLP 2026: Mubsir OCR: End-to-End Recognition of Arabic Handwritten Text

Ali Adel Sayed Ahmed[1,*], **Mona Khaled Ali**[4,*], **Mohamed Emad Sayed**[3,*], and **Ibrahim Nasser** [2,*]

[1,2,3]Helwan University, Ain Shams University, AAST University, Cairo, Egypt
[4]Mansoura University, Mansoura, Egypt
[*]{aliadelsayed2003, monakhaledd55, mohamedemad66553, ibrahim.nasser.2322}@gmail.com

Abstract

Historical Arabic handwritten OCR is difficult because of cursive script, fine diacritics, mixed numerals, and degraded media; classical segmentation pipelines compound errors, whereas end-to-end vision-language models can adapt when fine-tuned on in-domain data.
We present **Mubsir OCR**, a systematic evaluation on the NAKBA dataset: an annotated set (15,962 training line crops and 2,095 val lines with ground truth, used for all nine experiments) and a separate blind AR-MS (Subtask 2) set (2,671 images; scores only via official submission). We compare external vs. in-house VLMs (Qwen2.5-VL-3B, Qwen3-VL-4B-Instruct, Gemma3), inference backends (vLLM/bf16 vs. HuggingFace/bf16), training length (16 vs. 32 epochs), and test-time preprocessing (CLAHE+unsharp). Best on the annotated val set: **8.59% CER** / **25.87% WER** (HuggingFace bf16); the same configuration attains **11.00% CER** / **31.26% WER** on the blind set. Domain-specific fine-tuning beats general-purpose checkpoints; preprocessing helps only marginally and is not recommended without train-time augmentation.
Source code: NAKBA Mubsir OCR

Keywords: Arabic OCR, historical documents, vision-language models, fine-tuning, NAKBA dataset, character error rate, word error rate.

1. Introduction

Optical Character Recognition for historical Arabic handwritten documents remains one of the most challenging problems in document analysis. Arabic script is fully cursive, with each letter taking one of four context-dependent forms (initial, medial, final, and isolated). Diacritics (harakat) are semantically meaningful yet visually fine-grained, and historical manuscripts frequently mix Eastern Arabic and Western numeral systems. Physical degradation — fading ink, paper yellowing, variable calligraphy, and uneven illumination — further compounds these difficulties.

Classical OCR pipelines based on binarization, segmentation, and character classification degrade sharply on historical material because each stage accumulates errors and handwritten Arabic offers few reliable segment boundaries. End-to-end vision-language models (VLMs) bypass explicit segmentation; when fine-tuned on in-domain data they capture corpus-specific script styles that off-the-shelf models miss.

This paper reports nine controlled experiments (January–February 2026) on the NAKBA corpus: external vs. in-house checkpoints, training duration (16 vs. 32 epochs), inference backend and precision (vLLM/bf16 vs. HuggingFace/bf16), architecture (Qwen3-VL-4B vs. Gemma3), and test-time preprocessing (CLAHE+unsharp). All use the same 2,095-image annotated val set.

2. Related Work

Arabic handwriting recognition datasets. Muharaf (Saeed et al., 2024) is the largest public corpus of historic Arabic handwriting (1,600+ page images: personal letters, diaries, church and legal records), with line transcriptions and polygon annotations supporting line- and page-level HTR. AR-MS (Subtask 2) (Zaraket et al., 2026) is concurrent work on Arabic manuscript understanding in NAKBA-NLP 2026.

Vision-language models for OCR. Qwen3-VL (Bai et al., 2025) stacks DeepStack ViT features with interleaved MRoPE and reports state-

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 269–274
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

of-the-art results on document-understanding benchmarks; the 4B dense model balances accuracy and compute. Gemma3 (Gemma Team, 2025) is Google's multimodal family (1B–27B) with high local-to-global attention for long-context KV caches. Both support instruction-following fine-tuning for OCR-style tasks.

Inference serving. vLLM (Kwon et al., 2023) uses PagedAttention to limit KV-cache fragmentation and deliver 2–4$\times$ throughput gains; bf16 cuts memory at modest precision cost. Sherif community checkpoints (Sherif, 2025a,b) show Qwen2.5-to-Qwen3-VL transfer for Arabic handwriting.

Prompt optimization. GEPA (Agrawal et al., 2025) is a genetic-Pareto prompt optimizer that beats GRPO by up to 20% with up to 35$\times$ fewer rollouts—useful for automated prompt search in OCR.

3. Dataset and Experimental Setup

3.1. The NAKBA Dataset

The NAKBA shared task (Hamoud et al., 2026) provides two distinct image sets. The **annotated set** (released at competition start for training and self-evaluation) contains 15,962 training line crops and **2,095 JPG val line crops** with human-annotated ground truth in `annotations_test.csv`. All nine experiments in Table 1 are evaluated on this annotated val set; none of the reported CER/WER values are computed on an internal training split.

The separate **blind set** of **2,671 images** (JPG and PNG) is the official AR-MS (Subtask 2) competition set; ground truth is not released and scores are obtained only via submission; widths range from short fragments to full-sentence lines. The language is primarily Arabic, with occasional Latin annotations, Eastern Arabic numerals, and date strings.

Images are used at native dynamic resolution; internal resizing is handled by the model's default `AutoProcessor` (i.e., no custom resizing or preprocessing is applied), ensuring consistent preprocessing across most experiments. In-house models are trained on the merged NAKBA training and validation splits.

3.2. Evaluation Metrics

Character Error Rate (CER) measures the Levenshtein edit distance at character level between the model hypothesis and the reference transcription, normalized by the total number of reference characters:

$$\text{CER} = \frac{S + D + I}{N_{\text{ref}}} \times 100\,\%$$

where S, D, I are character-level substitutions, deletions, and insertions. CER is micro-averaged across the test set (total edits divided by total reference characters), giving longer lines proportionally more weight and avoiding bias from short-line outliers. For Arabic, single-character errors such as a missing dot (distinguishing, e.g., *ba* from *ya*) or a misread diacritic change word meaning.

Word Error Rate (WER) applies the same Levenshtein framework at the word level, with whitespace tokenization used consistently for both hypothesis and reference. WER is more sensitive to word-boundary errors (spurious spaces, merged tokens) and to entirely dropped or hallucinated words. In Arabic OCR, WER is typically higher than CER because a single character-level error can invalidate an entire word token. No case normalization or punctuation stripping is applied; metrics reflect raw model output.

3.3. Models

Qwen2.5-VL-3B (external baseline). Sherif (2025b), fine-tuned on general Arabic and English handwriting via Qwen2.5-VL-3B, is the domain-mismatch baseline.

Qwen3-VL-4B-Instruct (Bai et al., 2025) is our primary in-house model, with dynamic-resolution vision encoding, DeepStack ViT feature integration, and strong multilingual instruction-following. We fine-tune it on NAKBA data for 16 epochs (Exp. 2) and 32 epochs (Exp. 3, 4, 6, 7, 8, 9) using the standard instruction prompt. All fine-tuning runs share the same training/validation split, data format (image path + transcription), and generation parameters (near-zero temperature).

Fine-tuning details. All Qwen3-VL-4B runs use **full parameter fine-tuning** (no LoRA/PEFT) with AdamW (learning rate 2×10^{-5}, cosine schedule, 0.1 warmup ratio, weight decay 0.01), batch size 2 with gradient accumulation 8 (effective batch 8), max sequence length 512, bf16 mixed precision with gradient checkpointing, on a single NVIDIA H100 GPU.

Gemma3 (Gemma Team, 2025). Fine-tuned on the same NAKBA training data under identical conditions (Exp. 5). Gemma3 uses a different vision encoder and language backbone for architectural comparison.

3.4. Inference Backends

vLLM (Kwon et al., 2023) processes images in batches using PagedAttention memory management and bf16 weight quantization. Total inference time for 2,095 (val set) images on a single H100 GPU is on average 830 s (≈0.4 s/image).

HuggingFace Transformers runs images sequentially, one at a time, in full bf16 precision. Total inference time is on average 6,846 s (≈3.3 s/image), ≈8× slower than vLLM. Sequential processing avoids variable-length padding artefacts that arise in batched inference.

Backend-controlled comparison (Exps. 6 vs. 9). Experiments 6 and 9 use the same checkpoint, instruction prompt, and decoding settings (temperature $= 0.0$, max new tokens $= 512$, greedy decoding); only the inference backend differs (vLLM/bf16 vs. HuggingFace/bf16).

4. Results

Table 1 reports CER and WER for all nine experiments on the 2,095-image annotated val set. The best self-evaluated result is **8.59% CER** / **25.87% WER** (Exp. 9, HF bf16); this configuration was used for our official AR-MS (Subtask 2) blind test submission, yielding **11.00% CER** / **31.26% WER** on the 2,671-image blind set. The best vLLM result is **10.49% CER** / **29.02% WER** (Exp. 6).

4.1. Domain Adaptation and Training Duration

Sherif (2025b) on the full val set (Exp. 1) gives 20.20% CER and 51.52% WER, the worst result overall.

Sixteen epochs of in-house NAKBA fine-tuning (Exp. 2) reduce CER to 12.34% and WER to 30.66%. Doubling to 32 epochs (Exp. 4) further reduces CER to 11.43%, and Exp. 6 reaches 10.49% CER / 29.02% WER. The model had not converged at 16 epochs; further gains may be possible beyond 32 epochs (e.g., with early stopping).

4.2. Prompt Sensitivity

Experiments 3 and 4 use the same 32-epoch Qwen3-VL-4B checkpoint but differ in the prompt used during training (modified vs. standard). The CER difference is 0.26 pp (11.69% vs. 11.43%) and the WER difference is effectively zero (29.19% vs. 29.23%). The standard instruction prompt is robust; prompt engineering is not a bottleneck for this task.

4.3. Backend and Numerical Precision

Experiment 9 evaluates the same 32-epoch Qwen3-VL-4B checkpoint in HuggingFace Transformers (bf16) instead of vLLM bf16. CER drops from 10.49% (Exp. 6, best vLLM) to 8.59%, a gain of 1.9 pp; WER drops from 29.02% to 25.87%, a gain of 3.15 pp. Two mechanisms explain this: (i) *Quantization errors* — bf16 reduces the dynamic range available for probability estimates, disproportionately affecting fine-grained Arabic distinctions such as diacritics and dotting patterns; bf16 preserves the model's full numerical range. (ii) *Batching artefacts* — vLLM pads sequences to uniform length within each batch, which can subtly alter attention patterns for shorter images; sequential single-image inference avoids this effect. Wall-clock time is $8\times$ longer (6,846 s vs. 830 s)—suitable for final evaluation but costly for rapid iteration.

4.4. Architecture Comparison

Gemma3 fine-tuned on NAKBA (Exp. 5) achieves 11.10% CER / 29.80% WER via vLLM, comparable to Qwen3-VL-4B at 32 epochs. Gemma3's vLLM inference takes ≈2,175 s (≈2.6× slower than Qwen3-VL-4B), and WER is 0.78 pp higher. Qwen3-VL-4B offers the better accuracy–speed tradeoff. Gemma3 was evaluated on the 2,671-image blind set.

Table 2 compares the two backends on the same 32-epoch Qwen3-VL-4B checkpoint.

4.5. Preprocessing Ablation

Experiments 7 and 8 form a controlled pair for image preprocessing: identical checkpoint and vLLM backend, with (Exp. 8) and without (Exp. 7) CLAHE contrast enhancement and unsharp masking. The *aggressive* variant (v1) applied CLAHE with 8×8 tiles (clipLimit 2.0) and strong unsharp masking (radius 1, 120%, threshold 2) with JPG re-encoding at quality 75. On ≈50 px-tall line crops, 8×8 tiles are only ≈6 px per side, distorting fine Arabic strokes such as dots and small connecting ligatures; 848 images worsened and 466 improved, so the net effect was negative and v1 was abandoned. The *conservative* variant (v2)

#	Experiment	Backend	Preproc.	CER %	WER %	Notes
1	Qwen2.5-VL-3B (Sherif, 2025b)	vLLM	None	20.20	51.52	—
2	Qwen3-VL-4B (16 ep in-house)	vLLM	None	12.34	30.66	—
3	Qwen3-VL-4B (32 ep, alt. prompt)	vLLM	None	11.69	29.19	—
4	Qwen3-VL-4B (32 ep, std. prompt)	vLLM	None	11.43	29.23	—
5	Gemma3 fine-tuned (32 ep)	vLLM	None	11.10	29.80	$\approx 2.6\times$ slower
6	Qwen3-VL-4B (32 ep)	vLLM	None	10.49	29.02	—
7	Qwen3-VL-4B (32 ep, ctrl.)	vLLM	None	11.15	28.60	—
8	Qwen3-VL-4B + CLAHE+Unsharp v2	vLLM	CLAHE+Unsharp	10.18	28.51	Marginal improvement
9	Qwen3-VL-4B (32 ep, HF bf16)	HF	None	**8.59**	**25.87**	Best; $\approx 8\times$ slower

Table 1: All nine experiments on the 2,095-image annotated val set. Exp. 9 (bold) is overall best on the annotated set and is the configuration used for the official AR-MS (Subtask 2) blind test submission (CER 11.00% / WER 31.26% on 2,671 images). Exp. 6 is best under vLLM.

Backend	Precision	CER %	WER %	Speed
vLLM (Exp. 6)	bf16	*10.49*	*29.02*	$\approx$0.4 s/img
HuggingFace (Exp. 9)	bf16	**8.59**	**25.87**	$\approx$3.3 s/img
HF gain over vLLM		-1.90 pp	-3.15 pp	$8\times$ slower

Table 2: Backend comparison on the same checkpoint (Qwen3-VL-4B, 32 ep). Bold = highest accuracy; italic = best throughput option.

uses 16×16 tiles (clipLimit 1.0), gentler unsharp (radius 1, 80%, threshold 3), JPG quality 95, and a small-image skip rule ($h < 45$ px or $w < 150$ px).

v2 yields a modest CER improvement (11.15% $\rightarrow$ 10.18%) and a negligible WER improvement (28.60% $\rightarrow$ 28.51%), but the gain is inconsistent across images. The fine-tuned model was never trained on CLAHE-enhanced or sharpened images, so test-time preprocessing introduces train–test distribution mismatch. **We do not recommend preprocessing by default.** Future work should consider training on augmented data (raw + enhanced) or employing learned test-time normalisation to improve robustness without mismatch.

Parameter	v1 (aggressive)	v2 (conservative)
CLAHE clipLimit	2.0	1.0
CLAHE tileGridSize	(8, 8)	(16, 16)
Unsharp radius / %	1 / 120%	1 / 80%
Unsharp threshold	2	3
JPG re-encode quality	75	95
Tiny image skip	No	Yes ($h<45$ or $w<150$ px)

Table 3: Preprocessing parameter comparison. v1 degraded net performance; v2 was used in Exp. 8.

5. Conclusions

We presented a systematic nine-experiment OCR study on the NAKBA Arabic historical handwriting dataset (2,095 line crops with human annotations, fine-tuned on 15,962 training and validation crops). Findings: **(1) Domain fine-tuning is indispensable** — the external 3B checkpoint reaches 20.20% CER; in-house Qwen3-VL-4B fine-tuned for 32 epochs on NAKBA reduces CER to 8.59%, an improvement of 11.61 pp absolute. **(2) Training length matters** — moving from 16 to 32 epochs yields a further 0.9–1.8 pp CER reduction. **(3) Precision beats throughput for accuracy** — bf16 sequential inference improves CER by 1.9 pp and WER by 3.15 pp over bf16 batched inference, at $8\times$ the wall-clock cost. **(4) Test-time preprocessing without training-time**

augmentation is not recommended — conservative CLAHE+Unsharp v2 gives a marginal 0.97 pp CER gain but introduces train–test mismatch. **(5) Qwen3-VL-4B outperforms Gemma3** on this corpus in both WER and inference speed. We release inference scripts and configurations for reproducibility and benchmarking on Arabic historical corpora such as Muharaf (Saeed et al., 2024).

Ethical Considerations

This paper studies handwritten OCR on public NAKBA data without human subjects or personally identifiable information, following standard responsible-use norms.

6. References

6.1. Bibliographical References

Agrawal, L. A., Tan, S., Soylu, D., Ziems, N., Khare, R. et al. (2025). GEPA: Reflective prompt evolution can outperform reinforcement learning. *arXiv:2507.19457*. https://arxiv.org/abs/2507.19457

Bai, S. et al. (2025). Qwen3-VL Technical Report. *arXiv:2511.21631*. https://arxiv.org/abs/2511.21631

Gemma Team (2025). Gemma 3 Technical Report. *arXiv:2503.19786*. https://arxiv.org/abs/2503.19786

Hamoud, H., Chamseddine, A. A., Shalash, B., Abid, F. B., Jarrar, M., Chakra, C. A., Ghanem, B., & Zaraket, F. A. (2026). NAKBA NLP 2026: Shared Task on Arabic Handwritten Manuscript Understanding (Palestine Memory — Omar Al-Saleh Memoir). In *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with LREC 2026*. Palma, Mallorca, Spain, May 2026.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., & Stoica, I. (2023). Efficient memory management for large language model serving with PagedAttention. In *Proc. SOSP*, pp. 611–626. https://arxiv.org/abs/2309.06180

6.2. Language Resource References

Saeed, M., Chan, A., Mijar, A., Moukarzel, J., Habchi, G., Younes, C., Elias, A., Wong, C.-W., & Khater, A. (2024). Muharaf: Manuscripts of Handwritten Arabic Dataset for Cursive Text Recognition. *NeurIPS 2024 Datasets and Benchmarks Track*. https://arxiv.org/abs/2406.09630

Sherif (2025a). Arabic-English-handwritten-OCR-Qwen3-VL-4B [Fine-tuned model, HuggingFace Hub]. https://huggingface.co/sherif1313/Arabic-English-handwritten-OCR-Qwen3-VL-4B

Sherif (2025b). Arabic-handwritten-OCR-4bit-Qwen2.5-VL-3B-v3 [Fine-tuned model, HuggingFace Hub]. https://huggingface.co/sherif1313/Arabic-handwritten-OCR-4bit-Qwen2.5-VL-3B-v3

Zaraket, F., Shalash, B., Hamoud, H., Chamseddine, A., Abid, F. B., Jarrar, M., Chakra, C. A., & Ghanem, B. (2026). AR-MS: Arabic Manuscript Understanding Dataset (Subtask 2, Nakba-NLP 2026 Shared Task). *Proc. 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with LREC 2026*, Palma, Mallorca, Spain.

7. Appendices

7.1. The Value of Domain-Specific Fine-Tuning

The gap between the best external checkpoint (20.20% CER, Exp. 1) and the best in-house model (8.59% CER) is an 11.61 pp improvement attributable to NAKBA-specific training data. Sherif (2025b) still reaches 20.20% CER on this corpus because historical manuscripts differ from modern handwriting in ink, vocabulary, ligature density, and physical condition.

7.2. Precision–Throughput Tradeoff

The 1.9 pp CER gain from bf16 sequential inference (Exp. 9 vs. Exp. 6) is substantial. For interactive annotation, vLLM's $\approx$0.4 s/image latency is preferable; for batch archival transcription, HuggingFace bf16 is better suited.

7.3. Residual Errors and Error Analysis

At 8.59% CER / 25.87% WER, errors remain substantial; the CER-to-WER ratio ($\approx$1:3) suggests many whole-word substitutions from ambiguous or partially visible glyphs. Three main error types: (1) *Diacritic confusion* — the model frequently omits or misplaces harakat (short vowel marks). Diacritics such as fatha, kasra, and sukun are fine-grained single-pixel marks; the 4B model's visual patch tokenizer loses some spatial detail at this scale. (2) *Digit script substitution* — Eastern Arabic (Indic) and Western Arabic (European, 0–9) numerals are confused in both directions, particularly for isolated digits, reflecting mixed numeral usage in the original documents and inconsistent annotation conventions. (3)

Ligature splitting/merging — uncommon Arabic ligatures, especially in older calligraphic styles, are occasionally split into two character hypotheses or two adjacent characters merged into one; these errors disproportionately affect WER since they typically corrupt an entire word token.

Future directions include: (i) training beyond 32 epochs with validation-based early stopping; (ii) scaling to larger model variants (Qwen3-VL-8B or -32B); (iii) joint training on raw and augmented images to mitigate preprocessing distribution shift; (iv) post-correction re-scoring using an Arabic language model; (v) automated prompt optimization via GEPA (Agrawal et al., 2025); and (vi) cross-corpus evaluation on Muharaf (Saeed et al., 2024) to assess generalisation.

Latent Narratives at AR-MS NakbaNLP 2026: Reducing Character Errors in Arabic Manuscript Transcription: A CER Oriented System

Fatima Shaza*, **Sara Al-Desouky***, **Sarah Ayad**
Computer Science Department,
Arab Open University, KSA
24465909KSA@aou.edu.sa, 23463952KSA@aou.edu.sa, s.ayad@arabou.edu.sa
*Equal contribution

Abstract

Historic Arabic handwritten texts present significant challenges due to varied handwriting styles, cursive structure, diverse diacritics, and inconsistent character and word sizes. In this work, we introduce Historic-Arabic-OCR, a vision-language OCR system built upon Qari-OCR, which itself is based on Qwen2-VL-2B-Instruct, and further fine-tuned using Low-Rank Adaptation (LoRA) for Arabic manuscript transcription. The proposed approach incorporates contrast enhancement using CLAHE and deterministic decoding strategies to reduce character-level errors. Our model achieves competitive performance, with a Word Error Rate (WER) of 0.28 and a Character Error Rate (CER) of 0.10 on historical Arabic texts, including low-resolution images. The final submitted system uses CLAHE preprocessing with deterministic greedy decoding to minimize character-level errors.

Keywords: Arabic OCR, Vision-Language Models, Qwen2-VL, LoRA, CER Optimization

1. Introduction

Arabic Optical Character Recognition (OCR) plays a very important role for the digitization and preservation of Arabic texts. A task that is extremely important for cultural heritage, modern communication and information accessibility for over 400 million Arabic speakers worldwide. However, it comes up with multiple challenges such as complex morphology, contextual variation, ligatures and shaping etc. (Kasem et al., 2025)

This paper introduces Historic-Arabic-OCR, a fine-tuned Visual Language Model (VLM) based on Qwen2-VL-2B-Instruct specifically focused on historic Arabic texts. It was developed iteratively, while testing different configurations and different generation settings.

Our key contribution includes developing a competitive model for historic Arabic texts with diverse writing styles, varying image resolution, and faded ink. We also publicly release our trained model for reproducibility. The code and trained model weights are publicly available at https://github.com/sara-aldesouky/nakba-nlp-ocr-qwen2vl-lora.git and https://huggingface.co/Shaza2004/Historic-Arabic-OCR.

2. Literature Review

Building an accurate OCR system for Arabic is challenging due to its cursive handwriting, diacritics, and variation in character and word shapes (Al-Sheikh et al., 2020). These challenges are amplified in historical texts due to diverse writing styles and inherent script complexity. Additionally, Arabic datasets are less abundant compared to English, making Handwritten Text Recognition (HTR) more difficult (Chan et al., 2024). Despite recent advancements, OCR systems still struggle to produce accurate transcriptions for historical Arabic texts (Althobaiti and Lu, 2017).

OCR technology has evolved from early methods to deep learning architectures (Wasfy et al., 2025). Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been widely used, with CNNs showing strong performance in visual pattern recognition, achieving 94.9% accuracy on isolated Arabic characters (El-Sawy et al., 2017). Compared to RNNs, CNNs can capture spatial features effectively even with limited data (Alrobah and Albahli, 2022).

However, CNN- and RNN-based approaches often require Large Language Models (LLMs) for post-processing. This limitation led to transformer-based OCR, which enables end-to-end text recognition. Models such as TrOCR combine visual encoding with transformer decoding, improving performance on handwritten text recognition tasks (Li et al., 2023).

Further work, such as the MDLSTM-based approach by Ahmad et al. (Ahmad et al., 2020), highlights the importance of multi-directional feature extraction for Arabic scripts. Additionally, decoding strategies such as beam search and length constraints significantly impact Character Error Rate (CER) and Word Error Rate (WER), particularly for Arabic (Li et al., 2023; Chan et al., 2024).

More recently, Multimodal Large Language Models (MLLMs) have been applied to OCR by combining visual and textual understanding. However, general MLLMs are not optimized for Arabic OCR tasks (Wasfy et al., 2025). Qari OCR, built on Qwen2-VL-2B-Instruct (Wang et al., 2024),

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 275–279
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

adapts these models for Arabic OCR through fine-tuning, achieving a CER of 0.061 and WER of 0.160 (Wasfy et al., 2025).

Despite these advancements, most OCR and HTR systems fail to address the complexities of historical Arabic manuscripts. The script's cursive nature, positional variation, and reliance on diacritics increase ambiguity. Historical datasets further introduce stylistic diversity and degradation, including faded ink and irregular spacing. As a result, models trained on uniform datasets struggle to generalize, highlighting the need for approaches that account for the variability of historical Arabic texts.

3. Approach

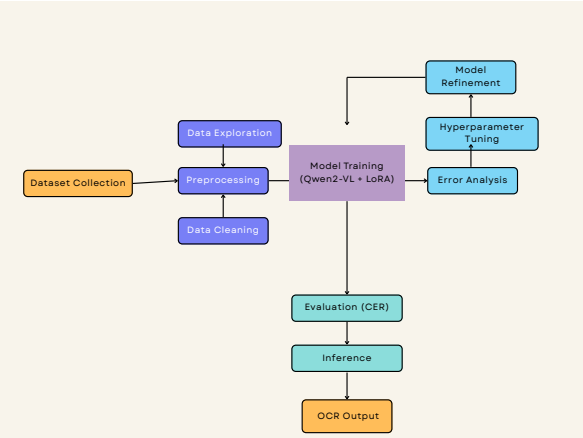


Figure 1: Overview of the proposed Arabic handwritten OCR pipeline, including data exploration, preprocessing, model training using Qwen2-VL with LoRA, evaluation using CER, and inference with iterative model optimization.

As shown in Figure 1, the proposed system begins with dataset collection and preprocessing... The development of our final model involved a two-stage process: first, the application of contrast-aware visual processing to improve image clarity, and second, parameter-efficient fine-tuning of the Qari-OCR model. The main goal was to minimize Character Error Rate (CER) on the official ARMS Subtask 2 dataset, derived from the Omar Al-Saleh manuscript collection.(Hamoud et al., 2026) This corpus consists of approximately 15,962 training, 1,774 development, and 2,671 held-out test lines. This dataset presented many challenges inherent to historical documents, including faded ink, varying illumination, and differing handwriting throughout the corpus. The inherent difficulties of Arabic script, such as cursive connectivity, positional glyph variations, and delicate features like dots and diacritics (*tashkeel*), added to the complexity of the task.

To improve the visibility of faded pages and aged paper, each image underwent Contrast Limited Adaptive Histogram Equalization (CLAHE) with a clipLimit of 2.0 and a tileGridSize of $(8, 8)$. This increased the clarity for minute details such as the dots, *hamza*, and *tashkeel* by reducing character substitution and deletion errors, thus reducing overall CER. We built our model on Qwen2-VL-2B-Instruct using an autoregressive formulation. This allows the model to predict the full conditional sequence likelihood $p(y|x)$, which naturally aligns with CER and WER evaluation metrics. To adapt the general-purpose model to the specifics of ancient manuscripts, we employed Low-Rank Adaptation (LoRA). Low-Rank Adaptation (LoRA) is applied for parameter-efficient fine-tuning using a rank of $r = 16$ and a scaling factor $\alpha = 16$, with no dropout. The adapters are inserted into the transformer's attention and feed-forward projection layers, including query, key, value, and output projections. Only LoRA adapter parameters are updated during training, while the base model weights remain frozen. All experiments were conducted on a single NVIDIA Tesla T4 GPU (16 GB VRAM). The training process was conducted in two different phases for a total of 4 epochs. First, the model was trained on the dataset for 2 epochs, followed by 2 epochs using CLAHE-enhanced images, allowing the model to refine and focus on the intricacies and minute details of the Arabic script. We used the AdamW optimizer with an initial learning rate of 1×10^{-5}, an effective batch size of 16, and a linear decay scheduler to manage the learning rate across the full training duration.

We used the prompt "Return the plain text in this image. Do not add anything" to enforce literal transcription behavior and to prevent any additions, as the model previously attempted to "correct" missing details in the images. We made our model deterministic by employing num_beams=1 (greedy decoding) and do_sample=False with a limit of 128 new tokens. Post-decoding normalization was minimal, including removing whitespace or English characters that were mixed with Arabic. No external dictionaries were used during this process.

4. Results

All experiments use a Qwen2-VL-2B-Instruct backbone fine-tuned via LoRA and are evaluated using Character Error Rate (CER) based on Levenshtein distance. As shown in Table 1, the baseline achieves a CER of 0.12. Incorporating beam search with dictionary correction from training dataset and normalization increases CER to

Table 1: CER results across different configurations.

Configuration	Preprocessing	Post-processing	CER
Fine-tuned model only	None	None	0.12
Beam search + dictionary correction + normalization	None	Dictionary + normalization	0.14
CLAHE + greedy decoding (final)	CLAHE	None	**0.10**

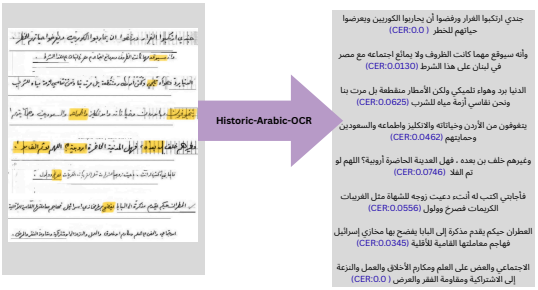


Figure 2: Qualitative comparison of OCR outputs across different configurations. The figure shows the input manuscript and the corresponding predicted text with their Character Error Rate (CER) values.

0.14. The final configuration, using CLAHE preprocessing with deterministic greedy decoding and no lexical correction, achieves the best performance with a CER of 0.10, highlighting the importance of evaluation-time design choices.

The configurations differ in preprocessing and decoding strategies. The baseline uses greedy decoding (num_beams=1, do_sample=False, repetition_penalty=1.05), while the second configuration introduces beam search (num_beams=5) with dictionary-based post-processing. The final configuration applies CLAHE (clipLimit=2.0, tileGridSize=(8,8)) to enhance contrast in degraded regions while omitting semantic post-processing.

Qualitative inspection shows that errors are primarily micro-orthographic, including hamza-related swaps, dot-sensitive confusions (e.g., Jeem/Ha), and terminal-form ambiguities such as Ya variants and Ta marbuta versus Ha. These patterns indicate that while the model captures base glyph structure, it remains sensitive to diacritic and dot-level distinctions in degraded regions.

5. Discussion

Qualitative analysis further elucidates the behavior of our model configurations. As shown in Figure 3, the base model produces several character-level errors. While heavy normalization with dictionary correction modifies the output, it often introduces additional distortions rather than improving accuracy. In contrast, the CLAHE-enhanced model produces more accurate transcriptions, aligning with our quantitative results where the CLAHE-enabled configuration achieved the lowest CER of 0.10. Contrast enhancement helps the model capture subtle character features, particularly in degraded manuscript regions.

Our findings indicate that deterministic greedy decoding outperforms beam search combined with dictionary correction. Because CER penalizes exact character mismatches, lexicon-driven substitutions or normalization often introduce unnecessary edit-distance penalties, suggesting that literal transcription tasks benefit more from deterministic decoding than from semantically motivated corrections. CLAHE preprocessing reduced CER from 0.12 to 0.10 by amplifying local stroke boundaries, reducing ambiguity in dot placement and fine glyph structure, demonstrating that domain-aligned preprocessing can complement model fine-tuning in historical OCR.

Transcription fidelity is critical for archival materials, as micro-orthographic errors—particularly in hamza placement and terminal characters—can fragment named entities and distort corpus analysis. Improving CER enhances the reliability of downstream tasks, including indexing and historiographic study. More broadly, language modeling alone cannot fully resolve fine-grained visual ambiguity in dot-based scripts, highlighting that evaluation-aligned decoding strategies and visually grounded preprocessing remain essential for robust historical manuscript transcription.

Table 2: Top 10 character confusion errors ($n = 500$)

Count	Original		Target
60	ا	→	أ
56	أ	→	ا
29	أ	→	إ
18	إ	→	أ
16	إ	→	ا
15	ا	→	إ
8	ح	→	خ
7	ي	→	ى
7	ه	→	ة
7	ى	→	ي

Errors remain concentrated in heavily degraded

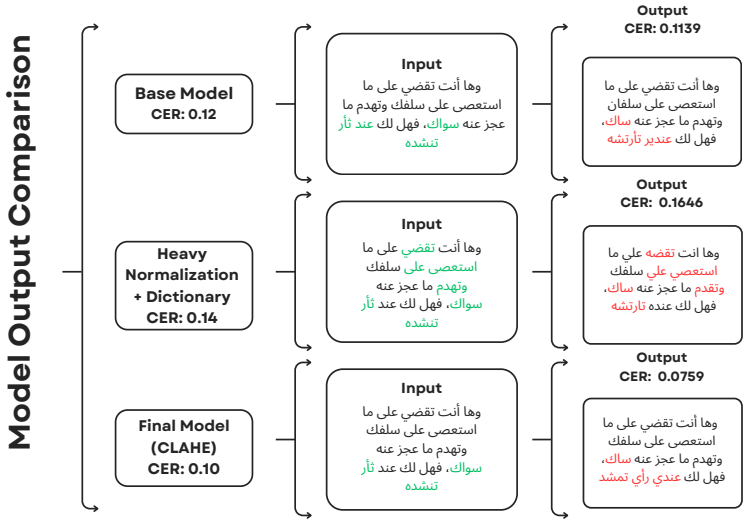


Figure 3: Qualitative comparison of OCR outputs across configurations. Incorrect characters highlight transcription errors.

regions where stroke visibility is reduced. Characters differing only by dot or hamza placement remain particularly challenging.

6. Conclusion

In conclusion, this paper presents the different stages of fine-tuning the Qwen2-VL model and improving CER performance. It also demonstrates that heavy normalization does not always lead to better results, and how this issue was addressed by enhancing image quality using CLAHE. Through these stages, the proposed system surpassed existing open-source solutions in accurately extracting and predicting complex layouts in handwritten Arabic text, while effectively handling diacritics and diverse fonts.

Limitations

Despite the performance of the Historic-Arabic-OCR system, several limitations remain. Errors are primarily concentrated in heavily degraded manuscript regions where ink fading and paper aging reduce stroke visibility. The model also struggles with fine-grained distinctions between characters that differ only in dot or hamza placement.

Additionally, while the fine-tuned Qwen2-VL model performs well on the NakbaNLP dataset, its generalizability to diverse regional calligraphic styles not represented in the training set has not been fully evaluated. The computational requirements of large Vision-Language Models may also limit deployment in resource-constrained environments compared to smaller OCR systems.

Future work may explore improved contrast modeling, dot-aware data augmentation, and hybrid visual-linguistic constraints tailored for historical Arabic scripts to further enhance transcription accuracy.

Acknowledgements

The authors would like to express their sincere gratitude to the Doha Institute for Graduate Studies for organizing such an illuminating experience, which provided us with valuable knowledge and experience.

Ethical Considerations

This work uses publicly available data provided by the shared task organizers. The dataset consists of historical Arabic manuscripts and does not contain sensitive or personally identifiable information. The system is intended to support the digitization and preservation of cultural heritage. OCR errors may occur, particularly in degraded manuscripts, and outputs should be verified before use in critical applications.

References

R. Ahmad, S. Naz, M. Afzal, M. Liwicki, and A. Dengel. 2020. A deep learning based ara-

bic script recognition system: Benchmark on KHATT. *The International Arab Journal of Information Technology*, 17(3).

Saleh Al-Sheikh, Masnizah Mohd, and L. Warlina. 2020. A review of arabic text recognition dataset. *Asia-Pacific Journal of Information Technology and Multimedia*, 9(1):69–81.

Naseem Alrobah and Saleh Albahli. 2022. Arabic handwritten recognition using deep learning: A survey. *Arabian Journal for Science and Engineering*, 47(8):9943–9963.

H. Althobaiti and C. Lu. 2017. A survey on arabic optical character recognition and an isolated handwritten arabic character recognition algorithm using encoded Freeman chain code. In *Proceedings of the 2017 51st Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE.

Adrian Chan, Anupam Mijar, Mehreen Saeed, Chau-Wai Wong, and Akram Khater. 2024. Hatformer: Historic handwritten arabic text recognition with transformers. *arXiv preprint arXiv:2410.02179*.

A. El-Sawy, M. Loey, and H. El-Bakry. 2017. Arabic handwritten characters recognition using convolutional neural network. *WSEAS Transactions on Computer Research*, 5(1):11–19.

Hadi Hamoud, Ahmad Ali Chamseddine, Bilal Shalash, Firas Ben Abid, Mustafa Jarrar, Chadi Abou Chakra, Bernard Ghanem, and Fadi A. Zaraket. 2026. Nakba nlp 2026: Shared task on arabic handwritten manuscript understanding (palestine memory–omar al-saleh memoir). In *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

Mahmoud Salaheldin Kasem, Mohamed Mahmoud, and Hyun-Soo Kang. 2025. Advancements and challenges in arabic optical character recognition: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 58(4).

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. TrOCR: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13094–13102.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Ahmed Wasfy, Omer Nacar, Abdelakreem Elkhateb, Mahmoud Reda, Omar Elshehy, Adel Ammar, and Wadii Boulila. 2025. Qari-OCR: High-fidelity arabic text recognition through multimodal large language model adaptation. *arXiv preprint arXiv:2506.02295*.

shroukgbr at StanceNakba Shared Task: Transformer-Based Ensemble Learning for Actor-Level Stance Detection in Palestinian–Israeli Social Media Discourse

Shrouk Anwar Gabr, Mohamed Ibrahim Ragab

School of Information Technology and Computer Science (CIS)
Nile University, Giza, Egypt
SGabr@nu.edu.eg, MoRagab@nu.edu.eg

Abstract

Stance detection has become an essential task for understanding political discourse on social media, particularly in highly polarized contexts where sentiment alone is insufficient to capture author intent. This study addresses stance classification in discussions related to the Palestinian–Israeli conflict by developing transformer-based and ensemble learning approaches for three-class classification: Pro-Palestine, Pro-Israel, and Neutral. Using the StanceNakba 2026 Shared Task dataset, we fine-tune multiple pretrained transformer models, including MARBERT, ARBERT, BERT, RoBERTa, and DeBERTa, and evaluate their performance using stratified cross-validation with macro F1-score as the primary metric. In addition to individual model evaluation, a weighted ensemble combining BERT, RoBERTa, and DeBERTa is proposed to leverage complementary contextual representations. Experimental results show that the ensemble model achieves the best performance with an accuracy and macro F1-score of 0.8905, outperforming specialized Arabic models while maintaining strong class-wise balance.

Keywords: Actor-level stance detection, Weighted cross-entropy, ARABERT, MARBERT, Transformer-based text classification, Political stance modeling, Cross-validation ensembling.

1. Introduction

The advent of social media has forever changed the manner in which people communicate regarding political issues. Online political discussions can be highly emotionally charged and express strong opinions about certain types of people, including political actors, organizations, and political movements. In order to have a complete understanding of these types of discussions, simply performing a general sentiment analysis is not enough; we also need to determine actor-level stance detection, which will provide an indication of whether a reader supports or opposes an actor or if they remain neutral toward that actor in the text being analyzed. Stance detection refers to the determination of a writer's position on a specific target (Gera and Neal, 2025). While sentiment analysis simply looks at the polarity of the sentiments expressed in the text (i.e., whether the sentiment expressed is positive or negative), stance detection also includes which actor or target the sentiment is being expressed towards (Garg and Caragea, 2024).

Understanding public opinion, tracking misinformation, and analyzing political debates benefit from stance detection (Garg and Caragea, 2024), (Ahmad and Kakudi, 2025). Most research pertains to topics, while previous research examined how tweets discuss the Palestinian-Israeli conflict using three categories: pro-Palestine, pro-Israel, or neutral (Imtiaz et al., 2022). We take a different approach to stance detection by creating one unified model that will assign stance classification (pro-Palestine, pro-Israel, or neutral) based on an author's complete political stance on the Palestinian-Israeli conflict, evaluated using the StanceNakba 2026 Shared Task dataset and benchmark (Aldous et al., 2026). This approach secured first place on the Codabench leaderboard in both the development and final evaluation phases, demonstrating its robustness and effectiveness in real-world stance detection scenarios. To support reproducibility, the implementation is publicly available at: `https://github.com/shroukgbr89/StanceNakba-2026-Stance-Detection`, including executable Google Colab notebooks.

The rest of the paper is organized as follows: Section 2 shows related work on the propaganda classification problem. Section 3.1 describes the dataset and preprocessing techniques used in this study, and Section 3.2 details the transformer models and ensemble frameworks implemented. Section 4 presents the experimental results, including individual model performance and ensemble outcomes. Finally, Section 5 discusses the findings, limitations, and potential directions for future research.

2. Related Work

This section surveys recent developments in closely related areas, including propaganda detection, political bias detection, extremism detection, and multilingual misinformation detection. These works establish the contextual and methodological foundation for multilingual classification research, out-

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 280–284
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

lining key challenges, commonly used approaches, and existing limitations that inform and motivate the present study's contributions.

The study (Ragab et al., 2025) uses the SinaLab FigNews 2024 multilingual dataset, consisting of 13,500 Facebook posts across five languages: Arabic, English, Hebrew, French, and Hindi (2,400 posts per language). The dataset includes original text, machine translations, and bias-related annotations, with the main classification target being a four-class. Due to strong class imbalance, the authors applied oversampling to create a fully balanced dataset with 7,098 instances per class. Three transformer models—mBERT, XLM-RoBERTa, and mT5—were evaluated on both imbalanced and balanced data. On imbalanced data, mT5 significantly outperformed others (98.86% accuracy, F1=0.9882), while mBERT (53%) and XLM-RoBERTa (69.7%) struggled. After balancing, all models improved substantially: mT5 achieved 99.61% accuracy and F1=0.9962 with the lowest training and validation loss, demonstrating superior generalization; Overall, results confirm that mT5 is the most robust and effective.

The study (Ahmad et al., 2020) investigates ensemble machine learning methods for automatic fake news detection using a publicly available dataset of news articles labeled as fake or real. The textual data undergo standard NLP preprocessing steps, including tokenization, stopword removal, and TF-IDF feature extraction, and is split into training and testing sets with a relatively balanced class distribution to ensure fair evaluation. Results show that ensemble models, such as random forest and voting-based approaches, consistently outperform individual classifiers like decision tree, Naïve Bayes, SVM, and KNN across accuracy, precision, recall, and F1-score metrics. The ensemble methods achieve the highest performance, exceeding 90% accuracy, demonstrating stronger robustness and generalization compared to single-model classifiers. The study (Fadri et al., 2026) analyzes public sentiment on the Gaza conflict using a dataset of 2,175 tweets collected from the X platform and processed through standard NLP preprocessing steps, including cleaning, tokenization, stemming, and TF-IDF feature extraction. Tweets were labeled into positive, negative, and neutral classes, creating a high-dimensional textual dataset for sentiment classification. Three models—MLP, XGBoost, and logistic regression—were evaluated using different train–test splits. Logistic regression achieved the best performance with 73.17% accuracy, followed by XGBoost (70.18%), while MLP showed the lowest results (69.27%). The results demonstrate that simpler linear models are more effective than complex models for small, sparse social media datasets represented with TF-IDF features.

Gap Analysis: Despite progress in misinformation detection and multilingual text classification, key gaps remain. Many studies focus on improving model performance but rely on balanced, controlled datasets that do not reflect real-world, imbalanced social media data. Additionally, research often treats tasks like sentiment analysis and fake news detection separately, failing to capture the complexity of multilingual propaganda and bias, which require deeper cross-lingual and cultural understanding. While traditional machine learning methods lack the ability to model semantic relationships, transformer-based approaches are typically evaluated on performance rather than robustness and generalization.

3. Materials and Methods

3.1. The used dataset

This study's datasets were taken from a public discourse dataset regarding the Palestinian–Israeli Conflict that the StanceNakba 2026 Shared Task provided for identifying stances. We participated in Subtask A: Actor-Level Stance Detection. The dataset consists of 1,401 labeled (Target) samples as part of the main training data file; there were additional development and test splits from the organizers of the task. The authors created a corpus of data using a 70/15/15 split to allow for training and evaluating the model.

As shown in 1, the dataset utilized in this study was partitioned into three distinct sets for model development and evaluation. The training set contains 980 labeled instances used for model learning. An additional 210 labeled samples served as the validation set for performance monitoring and hyperparameter tuning. Finally, the test set provided consists of 211 unlabeled instances for which final predictions were generated. There are three main columns on each data file: id, text, and stance-related label columns. An ID column that is a unique identifier for each text. The text column contains the original English language related to how people feel about different aspects of the conflict. For the training and validation sets, a categorical label—Pro-Palestine, Pro-Israel, or Neutral—is also provided, defining the stance classification task.

A particular focus of our analysis is the distribution of stance labels within the training dataset. As can be seen in Figure 1, the three classes are almost equally distributed across the dataset: there are 327 instances of Pro-Israel, 327 instances of Neutral, and 326 instances of Pro-Palestine, for a total of 980 samples. Balanced distributions mitigate any class bias in the training process, allowing for equitable assessment through all classification

Table 1: Sample instances from the StanceNakba 2026 dataset (Subtask A)

Stance Label	Text
Pro-Palestine	may god come in revenge plestine i will not be staying silent while a genocide is unfolding in the lan dof palestine.
Pro-Israel	I pray that our God will give her peace. Such a terrible thing that she had to endure. God please continue to protect Israel and it peoples!
Neutral	It is horrible what the children suffer because of Islamic aggression.

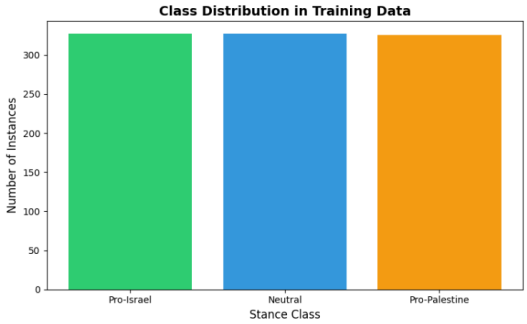


Figure 1: Distribution of instances in the stance column across the dataset.

categories.

3.2. Methodology

In this study, we implemented state-of-the-art multilingual models to classify propaganda and bias in the war between Gaza and Israel. The models include MARBERT, ARABERT, and BERT models.

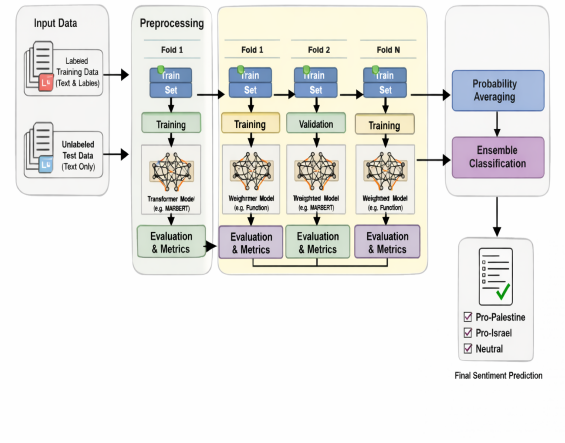


Figure 2: Propaganda Detection Architecture Utilizing MARBERT, ARABERT, and BERT models.

3.2.1. Data Preprocessing

As shown in 2 The raw textual data are first subjected to a series of preprocessing operations to improve data quality and ensure consistency. These steps include removing duplicates and irrelevant entries, text normalization, lowercasing, punctuation removal, and elimination of URLs and special characters. Stopwords are removed to reduce noise, while tokenization is applied to split text into meaningful linguistic units. Stemming, or lemmatization, is then performed to reduce words to their base forms, helping minimize vocabulary sparsity and improving generalization. After preprocessing, textual data are transformed into numerical representations suitable for machine learning algorithms. A term-weighting approach based on Term Frequency–Inverse Document Frequency (TF-IDF) is employed to capture the importance of words within documents relative to the entire corpus. This process produces a high-dimensional, sparse feature matrix that encodes semantic relevance while reducing the influence of commonly occurring but less informative terms.

3.2.2. Model Development Pipeline

The extracted features are used to train multiple supervised machine learning models. The architecture allows experimentation with different classifiers to analyze performance variations across learning paradigms. Each model learns patterns from the training data to distinguish between target classes based on textual characteristics. Hyperparameters are configured to ensure stable learning behavior and to prevent overfitting. The dataset is divided into training and testing subsets using multiple train–test split ratios to evaluate model robustness under different data availability scenarios. During training, models learn decision boundaries from the feature vectors, while validation is performed on unseen data to assess generalization capability.

4. Results and Discussion

4.1. Experiment Setup

We used transformer-based models to determine each actor's stance on the Palestinian–Israeli conflict. The experiment involved data preprocessing, model fine-tuning, and evaluation. All experiments were conducted in a cloud environment using NVIDIA Tesla P100 and NVIDIA T4 GPUs via Google Colab.[1] These accelerators enable efficient fine-tuning of large transformer models such as DeBERTa-v3 and MARBERT. The implementation

[1] https://colab.research.google.com/

was carried out using Hugging Face Transformers [2] and PyTorch [3], along with supporting libraries including Scikit-learn, Pandas, and NumPy. Models were trained using the AdamW optimizer with a linear learning rate scheduler and warmup. Hyperparameters varied across models but generally included a learning rate in the range of 1.5×10^{-5} to 2×10^{-5}, batch sizes of 8–16, and training for 6–10 epochs. The maximum sequence length ranged between 256 and 320 tokens depending on the model.

4.2. Models Evaluation

The preprocessing phase ensured data quality and model compatibility through several steps. Texts were cleaned by removing extra whitespaces and handling missing values. Stance labels were encoded into numerical form for classification. Tokenization was performed using Hugging Face's AutoTokenizer with padding and truncation, and sequence lengths ranged from 192 to 320 tokens. Finally, the data was converted into PyTorch tensors and split into training, validation, and test sets.

Our proposed system achieved **1st place** on the official Codabench leaderboard in both the development and final evaluation phases of the Stance-Nakba 2026 Shared Task, ranking first among all participating teams. In this section, we evaluate three general-purpose pretrained models, BERT, RoBERTa, and DeBERTa, as well as two specialized models, MARBERT and ARBERT, using stratified cross-validation and macro F1-score as the primary evaluation metric.

Table 2 summarizes the performance of the evaluated models using precision, recall, and F1-score, with macro F1-score serving as the primary evaluation metric. The ensemble model achieved the best overall performance, reaching an accuracy and macro F1-score of 0.8905, outperforming MARBERT (0.88) and ARBERT (0.86). This improvement highlights the benefit of combining multiple transformer architectures to enhance generalization and prediction robustness. Class-wise results show that the Pro-Israel class achieved the highest performance across all models, indicating that strongly polarized content is easier to detect due to clearer linguistic cues. The pro-Palestine class also demonstrated stable performance, with consistently high recall values across models. The Neutral class proved most challenging, as lower recall suggests models frequently confuse neutral statements with partisan content. Overall, the results indicate that ensemble learning improves stance classification performance by leveraging complementary model representations. However, accurately identifying neutral discourse remains a key challenge, highlighting an important direction for future research.

Table 2: Performance comparison of evaluated models using Precision (P), Recall (R), and F1-score.

Model	Class	P	R	F1	Sup
Ensemble	Neutral	0.9206	0.8286	0.8722	70
	Pro-Israel	0.9286	0.9286	0.9286	70
	Pro-Palestine	0.8312	0.9143	0.8707	70
	Acc.		0.8905		
	Macro Avg	0.8935	0.8905	0.8905	210
MARBERT	Neutral	0.92	0.78	0.84	327
	Pro-Israel	0.87	0.94	0.90	327
	Pro-Palestine	0.86	0.91	0.88	326
	Acc.		0.88		
	Macro Avg	0.88	0.88	0.88	980
ARBERT	Neutral	0.90	0.76	0.82	327
	Pro-Israel	0.85	0.94	0.89	327
	Pro-Palestine	0.84	0.89	0.87	326
	Acc.		0.86		
	Macro Avg	0.87	0.86	0.86	980

5. Conclusion

This study investigated transformer-based and ensemble learning approaches for stance detection in political discourse related to the Palestinian–Israeli conflict. Multiple pretrained models, including MARBERT, ARBERT, BERT, RoBERTa, and DeBERTa, were evaluated using standardized preprocessing, fine-tuning strategies, and macro F1-score as the primary evaluation metric. Experimental results demonstrated that the proposed ensemble model achieved the highest overall performance, surpassing individual specialized models by effectively combining diverse contextual representations and reducing model-specific biases. The analysis further showed that strongly polarized stances are more easily identified than neutral content, as neutral statements often contain implicit or ambiguous linguistic cues that resemble partisan expressions.

5.1. Limitations and Future Work

Despite the promising results, this study has several limitations. First, the dataset size and class distribution, particularly for neutral statements. Second, the analysis was constrained to textual content without considering multimodal signals such as images, videos, or network interactions. Finally, computational resources restricted extensive experimentation with larger transformer architectures or fully multilingual models, potentially affecting performance on cross-lingual content.

Future research could address these limitations through several avenues. Increasing dataset size and diversity, particularly for neutral and ambiguous statements, would improve model generalization.

[2] https://huggingface.co/transformers/
[3] https://pytorch.org/

6. References

Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf, and Muhammad Ovais Ahmad. 2020. Fake news detection using machine learning ensemble methods. *Complexity*, 2020(1):8885861.

Mahmoud Ahmad and Habeebah Kakudi. 2025. Stance detection on nigerian 2023 election tweets using bert: A low-resource transformer-based approach. In *Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025)*, pages 54–63.

Kholoud Khalil Aldous, Md Rafiul Biswas, Mabrouka Bessghaier, Shimaa Ibrahim, Kais Attia, and Wajdi Zaghouani. 2026. StanceNakba shared task: Actor and topic-aware stance detection in public discourse. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

Agil Irman Fadri, Nur Futri Ayu Jelita, Diamond Dimas Bagaskara, and Raudiatul Zahra. 2026. Sentiment analysis of public opinion on the gaza conflict using machine learning. *Public Research Journal of Engineering, Data Technology and Computer Science*, 3(2):138–148.

Krishna Garg and Cornelia Caragea. 2024. Stanceformer: Target-aware transformer for stance detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4969–4984.

Parush Gera and Tempestt Neal. 2025. Deep learning in stance detection: A survey. *ACM Computing Surveys*, 58(1):1–37.

Arsal Imtiaz, Danish Khan, Hanjia Lyu, and Jiebo Luo. 2022. Taking sides: Public opinion over the israel-palestine conflict in 2021. *arXiv preprint arXiv:2201.05961*.

Guan-Tong Liu, Yi-Jia Zhang, Chun-Ling Wang, Ming-Yu Lu, and Huan-Ling Tang. 2024. Comparative learning based stance agreement detection framework for multi-target stance detection. *Engineering Applications of Artificial Intelligence*, 133:108515.

Mohamed Ibrahim Ragab, Ensaf Hussein Mohamed, and Walaa Medhat. 2025. Multilingual propaganda detection: Exploring transformer-based models mbert, xlm-roberta, and mt5. In *Proceedings of the first International Workshop on Nakba Narratives as Language Resources*, pages 75–82.

Oanh Tran, Anh Cong Phung, and Bach Xuan Ngo. 2022. Using convolution neural network with bert for stance detection in vietnamese. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7220–7225.

NU_Hallucinators at NakbaArchiveClassifier Shared Task: A CLIP-Based Approach for Destruction Detection in Historical Image Archives

Salma Khaled Hegazy, Mohamed Ibrahim Ragab

School of Information Technology and Computer Science (CIS)
Nile University, Giza, Egypt
SaHegazy@nu.edu.eg, MoRagab@nu.edu.eg

Abstract

This paper presents a CLIP-based transfer learning approach for classifying historical archive images in the Nakba Image Classification Shared Task at the Nakba-NLP 2026 Workshop (LREC 2026). The task involves distinguishing images depicting destroyed or damaged infrastructure from those showing intact scenes using a dataset of 2,001 images collected from Instagram posts published by Palestinian content creators and journalists in Gaza between October 2023 and December 2025. Our method employs the CLIP ViT-B/32 visual encoder with selective fine-tuning of the final transformer block and a lightweight classification head. To address class imbalance, we apply focal loss along with standard data augmentation and threshold optimization. Experimental results show that the proposed model outperforms several CNN baselines and achieves an F1-score of 0.877 on the blind test set, securing 4th place in the shared task.

Keywords: computer vision, vision transformers, CLIP, transfer learning, disaster image classification, historical archives

1. Introduction

Historical archive preservation and analysis have become increasingly important as cultural institutions digitize their collections. The Nakba Archive, documenting Palestinian historical events, presents unique challenges for automated classification due to the sensitive nature of the content and the need to distinguish between destruction and non-destruction imagery. Accurate classification of such images is crucial for researchers, historians, and archivists working to preserve and analyze historical records.

Traditional computer vision approaches often struggle with historical photographs due to variations in image quality, age-related degradation, limited training data, and the semantic complexity of distinguishing destruction-related content from general scenes. Recent advances in vision-language models, particularly CLIP (Contrastive Language-Image Pre-training) (Radford et al., 2021), have demonstrated remarkable transfer learning capabilities across diverse visual domains through training on 400 million image-text pairs. In this work, we address the binary classification problem of categorizing historical archive images as depicting "destruction" or "not destruction". Our CLIP-based approach achieves an F1-score of 0.877 on the hidden test set, securing 4th place in a competitive evaluation with only a 2.2% performance gap from the leading system.

We make mainly 3 contributions in this paper:

- A simple yet effective CLIP-based architecture with strategic fine-tuning of only the last transformer block, achieving superior performance over traditional CNN architectures
- Application of focal loss to effectively handle class imbalance (35% vs. 65% distribution)
- Comprehensive ablation studies demonstrating the contribution of each component

The rest of the paper is organized as follows: Section 2 shows related work on the propaganda classification problem. Section 3.1 describes the dataset and preprocessing techniques used in this study, and Section 3.2 details the transformer models and ensemble frameworks implemented. Section 4 presents the experimental results, including individual model performance and ensemble outcomes. Finally, Section 5 discusses the findings, limitations, and potential directions for future research.

2. Related Work

This section surveys recent developments in closely related areas, including historical image classification.

Image classification has been a central task in computer vision, with early deep learning approaches primarily relying on convolutional neural networks (CNNs). Architectures such as ResNet (He et al., 2016) and EfficientNet (Tan and Le, 2019) achieved significant improvements in visual recognition by enabling deeper and more efficient networks.
These models are typically pre-trained on large

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 285–289
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

datasets such as ImageNet and then fine-tuned for downstream tasks through transfer learning. More recently, transformer-based architectures have gained popularity in computer vision. Vision Transformers (ViTs) treat images as sequences of patches and apply self-attention mechanisms to model global relationships within the image (Dosovitskiy et al., 2021). Compared to CNNs, ViTs are capable of capturing long-range dependencies and have demonstrated competitive performance on large-scale image recognition benchmarks.

The study (Ragab et al., 2025) investigates multilingual propaganda detection using transformer-based models. The dataset contains 13,500 Facebook posts related to the Israeli–Gaza war across five languages (Arabic, English, Hebrew, French, and Hindi) and includes four classes: Not Propaganda, Propaganda, Unclear, and Not Applicable. Because the original dataset was imbalanced, the authors applied oversampling to create a balanced dataset with 7,098 samples per class. Three multilingual transformer models—mBERT, XLM-RoBERTa, and mT5—were fine-tuned and evaluated using accuracy, precision, recall, and F1-score. The results show that mT5 achieved the best performance, reaching 99.61% accuracy and an F1-score of 0.996 on the balanced dataset.

mBERT achieved 92% accuracy, while XLM-RoBERTa achieved 89.51% accuracy. The findings also demonstrate that balancing the dataset significantly improved model performance, particularly for mBERT and XLM-RoBERTa. Overall, the study concludes that mT5 is the most effective model for multilingual propaganda detection.

To address the limitations of purely visual supervision, vision–language models have emerged as a powerful paradigm. Among them, CLIP (Contrastive Language–Image Pre-training) (Radford et al., 2021) introduced a large-scale contrastive learning framework that jointly trains image and text encoders on 400 million image–text pairs collected from the internet.

Gap Analysis Despite the rapid progress in image classification, several gaps remain in applying modern computer vision techniques to historical and crisis-related imagery. Traditional convolutional neural networks such as ResNet and EfficientNet have demonstrated strong performance in large-scale image recognition tasks; however, they are typically trained on datasets such as ImageNet that focus on common object categories rather than complex semantic concepts like destruction or infrastructure damage. As a result, these models often struggle to generalize effectively to specialized domains with limited training data and highly contextual visual cues.

3. Materials and Methods

3.1. The Used Dataset

The dataset used in this study was released as part of the Nakba Image Classification Shared Task (Abrahams et al., 2026) in the Nakba-NLP 2026 Workshop at LREC 2026. The dataset was collected from Instagram and contains images shared by Palestinian content creators and journalists in Gaza between October 7, 2023, and December 15, 2025. These images were archived through the SaltPillar project under the Tech for Palestine incubator, which aims to preserve and analyze visual documentation related to the ongoing conflict.

The dataset consists of 2,001 images, each stored in PNG format, accompanied by a binary label indicating whether the image depicts destroyed or damaged infrastructure. The images were manually annotated into two categories: `destruction` and `not_destruction`. The `destruction` class includes images showing visible damage such as collapsed buildings, airstrike aftermaths, debris, or destroyed infrastructure. In contrast, the `not_destruction` class contains images where infrastructure appears intact or where destruction is not visibly present, including scenes of daily life, people in public spaces, or unrelated activities.

Images are loaded in RGB format and transformed to the CLIP input resolution of 224×224 pixels. During training, data augmentation includes:

- Random resized crop (scale 0.8–1.0)
- Random horizontal flip (probability 0.5)
- Random rotation (±10 degrees)
- Color jitter (brightness, contrast, saturation: 0.2; hue: 0.1)

During inference, images are deterministically resized, center-cropped, converted to tensors, and normalized using the standard CLIP mean and standard deviation values. Missing or unreadable images are replaced with blank RGB images to maintain pipeline robustness, which is useful in practical archive settings where image collections can be noisy or inconsistently formatted.

As shown in Figure 1, the label distribution of the combined dataset illustrates a moderate class imbalance. Specifically, 564 images (28.2%) belong to the `destruction` class, while 1,035 images (51.7%) belong to the `not_destruction` class. This imbalance reflects the natural distribution of visual content shared on social media, where not all posts directly depict physical destruction.

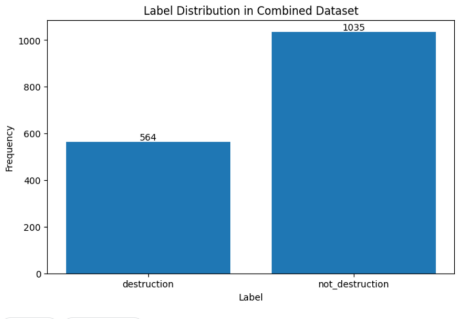


Figure 1: Distribution of labels in the Nakba Image Classification dataset.

3.2. Methodology

In this study, The proposed model uses the CLIP ViT-B/32 visual encoder as a feature extractor. CLIP employs a Vision Transformer (ViT) with 12 transformer blocks processing images through a patch embedding layer (32×32 patch size) followed by transformer blocks with 768-dimensional hidden states, producing 512-dimensional image embeddings.

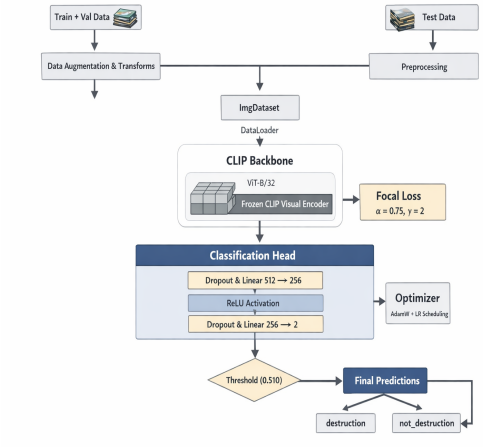


Figure 2: Architecture of the proposed CLIP-based classifier.

3.3. Model Architecture

As shown in Figure 2, The proposed model uses the CLIP ViT-B/32 visual encoder as a feature extractor. CLIP employs a Vision Transformer (ViT) with 12 transformer blocks processing images through a patch embedding layer (32×32 patch size) followed by transformer blocks with 768-dimensional hidden states, producing 512-dimensional image embeddings.

On top of the CLIP features, we add a lightweight two-layer multilayer perceptron classifier:

$$h = \text{Dropout}(0.5) \circ \text{ReLU}(\mathbf{W}_1 f + b_1)$$
$$y = \text{Dropout}(0.3) \circ (\mathbf{W}_2 h + b_2) \quad (1)$$

where $f \in \mathbb{R}^{512}$ represents CLIP features, $\mathbf{W}_1 \in \mathbb{R}^{256\times512}$, and $\mathbf{W}_2 \in \mathbb{R}^{2\times256}$. Figure 2 illustrates the overall pipeline.

3.4. Selective Fine-Tuning Strategy

To limit overfitting while enabling domain adaptation, all CLIP parameters are frozen except for the final visual transformer residual block. This selective fine-tuning strategy allows the model to adapt to the target domain of historical imagery without discarding the broad visual priors learned during large-scale pre-training on 400 million image-text pairs. The lightweight classifier head keeps the number of trainable parameters small relative to the backbone, which contributes to stable optimization behavior.

3.5. Training Objective and Optimization

To address the 35/65 class imbalance, the model is trained with focal loss using $\alpha = 0.75$ to weight the minority (destruction) class and $\gamma = 2$ to down-weight easy examples. This formulation helps the model focus on challenging samples and improves minority class recall. Optimization is performed with AdamW using differential learning rates: 10^{-3} for the classification head and 10^{-5} for the unfrozen CLIP block. This rate difference reflects the fact that the CLIP encoder already contains strong pre-trained features requiring only minor adaptation, while the randomly initialized classification head requires more aggressive updates. Weight decay is set to 10^{-4}, batch size to 32, and gradient norms are clipped to 1.0 for training stability. Training runs for up to 30 epochs with early stopping (patience=5) based on the best observed training loss. The final checkpoint reached a best training loss of 0.0113, indicating effective fitting to the merged labeled data.

3.6. Inference and Threshold Selection

For final inference, the model outputs class probabilities through a softmax layer. The *destruction* class is predicted when its probability exceeds a threshold τ. Rather than using the default threshold of 0.5, we performed systematic threshold optimization during the development phase, evaluating F1-scores across the range [0.3, 0.7]. This analysis identified $\tau = 0.510$ as optimal, which was then fixed for the final blind evaluation. The final prediction file contained 152 *destruction* predictions and 250 *not_destruction* predictions. While this distribution does not establish perfect calibration, it indicates that the chosen threshold produced a balanced operating point (37.8% destruction predictions vs. 35.2% in training) rather than a heavily biased majority-class output.

4. Results and Discussion

This section presents the experimental results obtained using three transformer-based models: ResNet18, ResNet34, ResNet50, EfficientNet-B3, and CLIP ViT-B/32. The final performance scores represent the average across all folds. The evaluation metrics include F1-score (Macro).

4.1. Development-Phase Model Evaluation

Several backbone architectures were examined during the development phase to identify the most effective feature extractor. Table 1 summarizes the development F1-scores achieved by different models. CLIP achieved substantially stronger results (0.91 F1) compared to traditional CNN architectures: ResNet50 (0.831), ResNet34 (0.83), ResNet18 (0.72), and EfficientNet-B3 (0.70). This clear performance advantage motivated the selection of CLIP ViT-B/32 as the backbone for the final system.

Backbone	Dev F1
ResNet18	0.720
ResNet34	0.830
ResNet50	0.831
EfficientNet-B3	0.700
CLIP ViT-B/32	**0.910**

Table 1: Development-phase comparison of candidate backbones on the validation set.

4.2. Final Test Results and Competition Ranking

The final CLIP-based system was trained on the merged training and validation data (1,599 images) and evaluated on a blind test set of 402 images. Our system achieved an F1-score of 0.877 on the blind test set, securing 4th place among competing systems with only a 2.2% performance gap from the first-place result (0.899). This competitive performance demonstrates that simple, well-designed single-model approaches can be highly effective when leveraging strong pre-trained representations.

4.3. Analysis: Single Model vs. Ensemble

During development, we also experimented with ensemble approaches combining ResNet34, ResNet50, and CLIP models. Interestingly, our single CLIP model (test F1: 0.877) substantially outperformed the ensemble (test F1: 0.836). We hypothesize three explanations: Our results demonstrate CLIP's substantial advantage over traditional CNN architectures for historical image classification. We attribute this to several factors:

Semantic understanding through language supervision: CLIP's training on image-text pairs enables understanding of abstract visual concepts like "destruction" that may be challenging to capture through purely visual categorical supervision on ImageNet classes.

Diverse pre-training data: Exposure to 400 million images across diverse domains, time periods, and visual styles provides more relevant prior knowledge for historical imagery than ImageNet's focus on modern object categories.

1. **Superior pre-training:** CLIP's training on 400M diverse image-text pairs provides more robust and transferable features than ImageNet's 1.3M images across 1,000 categories.

This finding suggests that for specialized tasks with limited training data, investing in a single strong pre-trained model with careful fine-tuning may be more effective than complex ensemble strategies.

5. Conclusion

Paper presented a CLIP-based approach for binary classification of historical archive images, achieving competitive performance (F1: 0.877, 4th place) through strategic fine-tuning, focal loss for class imbalance, and threshold optimization. Our comprehensive experiments demonstrate that: (1) CLIP's vision-language pre-training transfers effectively to historical imagery, substantially outperforming traditional CNNs (+7.9% over ResNet50 in development), (2) selective fine-tuning of only the last transformer block provides the largest performance gain (+2.4%), and (3) simple single-model approaches can outperform complex ensembles when leveraging strong pre-trained representations.

5.1. Limitations and Future Work

Despite the promising results achieved by our approach, several limitations remain. First, the current model relies solely on visual information extracted from images. This single-modality setup ignores potentially valuable contextual signals such as captions, metadata, or textual descriptions that often accompany social media posts. Such information may provide additional cues that help distinguish between destruction and non-destruction scenes.

Future work can address these limitations in several ways. One promising direction is multimodal learning, where visual features are combined with textual information using models such as CLIP's text encoder.

6. References

Alexei Abrahams, Shadi Abudalfa, Mustafa Jarrar, and George Mikros. 2026. The nakbaarchive-classifier shared task on nakba image classification. In *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

Alexey Dosovitskiy et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763.

Mohamed Ibrahim Ragab, Ensaf Hussein Mohamed, and Walaa Medhat. 2025. Multilingual propaganda detection: Exploring transformer-based models mbert, xlm-roberta, and mt5. In *Proceedings of the first International Workshop on Nakba Narratives as Language Resources*, pages 75–82.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*.

ZAHIRA BOULANOUAR at NakbaArchiveClassifier Shared Task: Detecting Infrastructure Destruction in Gaza with a ConvNeXt Ensemble

Zahira Boulanouar

Euromed University of Fez
zahira.boulanouar@eidia.ueuromed.org

Abstract

We present our third-place submission to the Nakba Image Classification Shared Task at LREC-COLING 2026, which requires binary classification of Instagram images from Gaza into destruction (damaged or destroyed infrastructure) versus not_destruction. Our system fine-tunes a ConvNeXt-Tiny backbone within a five-fold stratified cross-validation framework, combining Focal Loss, weighted random sampling, exponential moving average (EMA) weight stabilization, test-time augmentation (TTA), and out-of-fold (OOF) decision threshold calibration. Our system achieves an official test macro F1 of 0.8893 and 90.05% accuracy, placing third among all participants and within 0.02 F1 of the winning system (0.91), demonstrating that a 28M-parameter convolutional architecture with principled training strategies is highly competitive with much larger models.

Keywords: image classification, infrastructure damage detection, ConvNeXt, ensemble learning, Focal Loss, humanitarian AI, Gaza, Nakba

1. Introduction

Since October 7, 2023, Palestinian journalists and content creators in Gaza have extensively documented the ongoing conflict through social media. The SaltPillar project and the NakbaArchiveClassifier shared task (Abrahams et al., 2026) have archived and organized millions of Instagram images from this period. Automatically categorizing this archive is essential for enabling structured access for legal documentation, humanitarian reporting, and academic research.

The Nakba Image Classification Shared Task frames this as binary image classification: destruction (bombed buildings, rubble, damaged roads) versus not_destruction (intact infrastructure, people, unrelated content), evaluated by macro-averaged F1. Key challenges include moderate class imbalance (35% vs. 65%), high intra-class visual variation, and a small training set (1,400 images).

We address these with a ConvNeXt-Tiny (Liu et al., 2022) ensemble integrating Focal Loss (Lin et al., 2017), EMA weight averaging, weighted sampling, horizontal-flip TTA, and OOF threshold calibration. Our system ranked 3rd out of all participants (test macro F1: 0.8893; 1st place: 0.91), outperforming a larger EfficientNetV2-M (Tan and Le, 2021) single-model baseline by 10.3 F1 points confirming that ensemble strategies and calibration matter more than raw model capacity.

2. Related Work

Automated damage assessment from imagery has been studied primarily using satellite data (Gupta et al., 2019). Social media and ground-level imagery from active conflict zones introduce additional challenges: variable viewpoints, compression artifacts, and significant distributional shift. The Nakba dataset is, to our knowledge, the first large-scale labeled collection of conflict-zone infrastructure damage imagery.

ConvNeXt (Liu et al., 2022) modernizes the ResNet architecture with larger kernels, depthwise convolutions, GELU activations, and inverted bottlenecks achieving state-of-the-art accuracy while remaining efficient and well-suited to fine-tuning on small datasets. Deep ensemble methods (Lakshminarayanan et al., 2017) reduce prediction variance and improve calibration without architectural changes, while Focal Loss (Lin et al., 2017) addresses class imbalance by down-weighting well-classified examples.

3. Data

The dataset, released as part of the NakbaArchiveClassifier shared task (Abrahams et al., 2026), provides 2,001 labeled Instagram images (Palestinian journalists and content creators, Gaza, Oct. 2023–Dec. 2025). We use the official 1,400/199 train/validation split. The class distribution is moderately imbalanced: 64.7% not_destruction (1,295 images) vs. 35.3% destruction (706 images), maintained via stratification across all splits. The held-out test set comprises 500 images released February 10–17, 2026.

4. System Description

4.1 Architecture Overview

Figure 1 illustrates our end-to-end pipeline. At its core, the pipeline processes SaltPillar Instagram images from Gaza through a two-phase workflow. In the training phase, a weighted random sampler constructs balanced mini-batches, and ConvNeXt-Tiny models are fine-tuned with Focal Loss (α=0.35, γ=2.0), EMA weight stabilization, and stratified k-fold cross-validation. During inference, test images undergo horizontal-flip test-time augmentation (TTA), ensemble

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 290–293
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

averaging across folds, and out-of-fold (OOF) threshold calibration at τ=0.41 to produce destruction/not_destruction predictions.

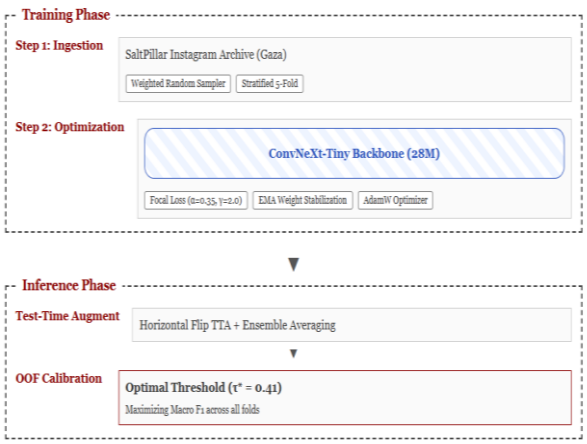


Figure 1: End-to-end system pipeline.

4.2 ConvNeXt Architecture

ConvNeXt (Zhuang Liu et al., 2022) is a modernized convolutional neural network architecture designed to close the performance gap between traditional CNNs and transformer-based vision models such as Vision Transformer. Instead of replacing convolutions with attention mechanisms, ConvNeXt revisits the standard ResNet design and incorporates several architectural improvements inspired by transformer models.

ConvNeXt blocks rely on depthwise separable convolutions, large kernel sizes (7×7), layer normalization, and GELU activations. These design changes increase the receptive field and improve gradient flow while maintaining the efficiency advantages of convolutional networks.

The ConvNeXt-Tiny variant used in this work contains approximately 28 million parameters and follows a hierarchical architecture composed of four stages with progressively decreasing spatial resolution and increasing channel depth. This design preserves the multi-scale feature extraction capability typical of convolutional architectures while benefiting from modern training techniques.

4.3 Training Configuration

Hyperparameter	**Value**
Input resolution	384 × 384
Optimizer	AdamW, lr = 1.2×10^{-4}, wd = 0.02
LR schedule	Cosine annealing
Max epochs	20 (early stop, patience = 6)
Loss	Focal Loss (α=0.35, γ=2.0)
EMA decay	0.999
Cross-validation	5-fold stratified

Table 1: Hyperparameters applied uniformly across all five folds.

Focal Loss re-weights gradients to focus training on hard examples, where p_t is the predicted probability for the ground-truth class, α_t=0.35 balances the minority destruction class, and γ=2.0 suppresses well classified negatives.

$$FL(p_t) = -\alpha(1-p_t)^{\gamma}\log(p_t)$$

Table 1 summarizes the hyperparameters applied uniformly across all five folds.

Weighted Random Sampling assigns each sample a weight inversely proportional to its class frequency per fold, ensuring approximately balanced mini-batches and complementing Focal Loss at the batch-construction level.

EMA maintains a shadow copy of model parameters (decay = 0.999) updated throughout training. EMA weights are used for all evaluation and saved as fold checkpoints, providing implicit within-run model averaging.

Training augmentation: random horizontal flip (p=0.5), color jitter (brightness, contrast, saturation ±0.2; hue ±0.05; p=0.6), random rotation (±10°). Validation and test: resize and normalize only.

4.4 Ensemble, TTA, and Threshold Calibration

Five models are trained on independent folds. At inference, TTA averages predictions on the original image and its horizontal flip per fold. The ensemble probability is the mean across all five folds.

$$p_{ens}(x) = \frac{1}{K}\sum_{i=1}^{K} p_i(x)$$

The default threshold of 0.5 is suboptimal for imbalanced macro F1. We sweep τ ∈ [0.10, 0.70] at step 0.004 and select the value maximizing macro F1 on OOF predictions, yielding τ* = 0.41, applied directly to validation and test predictions.

5. Results

5.1 Competition Standing

The shared task evaluation metric is macro-averaged F1, which balances precision and recall across both classes.

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

$$\text{Macro F1} = \frac{1}{N} \sum_{i=1}^{N} F1_i$$

Our system ranked third among all participants, within 0.02 macro F1 of the top entry.

Table 2 shows the partial leaderboard.

Rank	Team	Macro F1	Acc.
1	HCMUS_TheFangs	0.8991	90.80%
2	rahaf_jaber	0.8952	90.55%
3	**zahira_blr (Ours)**	**0.8893**	**90.05%**

Table 2: Official leaderboard (all 3 participants)

5.2 Cross-Validation and Overall Performance

Per-fold macro F1 ranges from 0.8916 to 0.9229 (mean: 0.9078), indicating stable performance. Folds 2 and 5 triggered early stopping. Table 3 summarizes all evaluation splits.

Split	Macro F1	Accuracy	τ
OOF (train)	0.8975	90.86%	0.41
Validation	0.8515	86.93%	0.41
Test (official)	0.8893	90.05%	0.41

Table 3: Performance across all evaluation splits (τ*=0.41).

The OOF report (Table 4) shows well-balanced per-class performance (FN = 88, FP = 40). The test score (0.8893) exceeding validation (0.8515) suggests the 199-image validation set is noisier or slightly harder than the 500-image test set.

Class	P	R	F1	N
not_destruction	0.908	0.956	0.931	906
destruction	0.910	0.822	0.864	494
Macro avg	**0.909**	**0.889**	**0.898**	**1400**

Table 4: OOF classification report with τ*=0.41.

5.3 Comparison with EfficientNetV2-M

Table 5 compares our ensemble against a single EfficientNetV2-M trained with the same Focal Loss and augmentation setup, without cross-validation ensembling. Despite having nearly twice the parameters and higher destruction recall (87.1%), EfficientNetV2-M produces 32 false positives that collapse the not_destruction F1, falling 10.3 points below our ensemble. This confirms that ensembling and calibration outweigh raw model capacity.

Model	Val F1	Acc.
EfficientNetV2-M (single)	0.7485	79.4%
ConvNeXt-Tiny (ens.)	**0.8515**	**86.9%**

Table 5: Validation comparison. EfficientNetV2-M not submitted officially.

6. Discussion and Conclusion

Five design decisions drove our performance: (1) the pretrained ConvNeXt-Tiny backbone for low-data transfer; (2) Focal Loss + weighted sampling for joint imbalance correction; (3) EMA for within-run weight stabilization; (4) five-fold ensemble averaging for variance reduction; and (5) OOF threshold calibration (τ*=0.41) to balance class-wise F1.

The 0.02 gap to first place could be narrowed by Mixup/CutMix augmentation, multi-scale TTA, stochastic weight averaging, or multimodal fusion with caption text and metadata. Our models operate on pixels alone and do not exploit auxiliary signals that could disambiguate ambiguous cases (partially damaged buildings, construction sites).

The system should be treated as a decision-support tool not an autonomous classifier given the evidentiary and humanitarian sensitivity of the data. We presented a ConvNeXt-Tiny ensemble system that ranked 3rd in the Nakba Image Classification Shared Task with a test macro F1 of 0.8893, within 0.02 of the top entry and 10.3 points above a larger single-model baseline. We release this pipeline as a reproducible baseline for humanitarian image classification research.

7. Acknowledgments

We thank the Nakba-NLP 2026 Workshop organizers and the Palestinian journalists and content creators whose documentation makes this work possible.

8. Bibliographical References

Abrahams, Alexei, Shadi Abudalfa, Mustafa Jarrar, and George Mikros. 2026. The NakbaArchiveClassifier Shared Task on Nakba Image Classification. In Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026), Palma, Mallorca, Spain, May.

Dosovitskiy, A. et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

Gupta, R. et al. (2019). xBD: A dataset for assessing building damage from satellite imagery. In *CVPR Workshops*.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986.

Tan, M. and Le, Q. V. (2021). EfficientNetV2: Smaller models and faster training. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 10096–10106.

Wightman, R. (2019). PyTorch Image Models (timm). https://github.com/rwightman/pytorch-image-models.

EGCSS at StanceNakba Shared Task: Cross-Topic Arabic Stance Detection for Two Middle East Issues

Asmaa Qindeel, Toka Khaled, Batool Balah, Eman Elrefai, Mahmoud Fawzi
Cairo University, Al-Azhar University, Independent Researcher,
Alexandria University, The University of Edinburgh
asmaaqandeel2511@gmail.com, tokakhaled98@gmail.com, batoolnajeh@gmail.com,
eman.lotfy.elrefai@gmail.com, m.f.g.ibrahim@sms.ed.ac.uk

Abstract

Stance detection continues to be an important task sitting at the intersection of Natural Language Processing (NLP) and Computational Social Science (CSS). In this work, we evaluate how different variations of BERT models perform on the cross-topic form of the task. In particular, we inspect their performance on the second subtask of the shared task *StanceNakba 2026*, where two topics are included, namely *Arab Normalization with Israel* and *The Presence of Refugees in Arab Countries*. We find that the best-performing model was **bert-base-arabertv02-twitter**, and we further improve its performance by providing context about the topic during the training phase, achieving an F1-score of **0.86** and ranking second among the participating teams.

Keywords: Stance detection, Arabic, Middle East, political NLP, BERT

1. Introduction

The Middle East is one of the most volatile regions in the world (Fawzi et al., 2026b) where social media is an important space for information exchange (Fawzi, 2022). This situation increases the importance of online arguments about political issues there as they develop causing polarization and fueling conflicts like the recent 2026 Iran War. In this work, we evaluate the performance of different variants of BERT models on the task of cross-topic stance detection that is presented within the *StanceNakba 2026* shared task (Aldous et al., 2026). The two topics introduced are both related to the Middle East where the first topic is the *Arab Normalization with Israel* and the second one is *The Presence of Refugees in Arab Countries*. It is required to develop a system that takes a short social media thread in Arabic about either topic and determines the stance of this thread as being either `Favor`, `Against`, or `Neither`. Our proposed system[1] relies on tuning a BERT (Bidirectional Encoder Representations from Transformers) model that has been trained beforehand on Arabic Twitter data. In addition to hyperparameter tuning and input preprocessing, we also added more context to the input during the training phase to make the model more aware of the topic to which the stance is directed. We achieve a macro F1 score of 0.8607 on the leaderboard and the second rank among the participating teams. We also perform further performance analysis and find that the model performs significantly better on the *Refugees* topic than the *Normalization* topic. In addition, we find that the misclassified samples in the normalization topic are mostly `Against` or `Favor` samples that are misclassified as `Neither` indicating more difficulty in identifying a clear stance towards the topic.

[1]The code can be found here: https://github.com/AsmaaQ25/StanceNakba2026

2. Background

2.1. Task Description

The *StanceNakba 2026* Shared Task focuses on stance detection in politically polarized social media discourse. In this study, we participate in **Task B: Cross-Topic Stance Detection**. The objective is to develop a single model capable of predicting the stance of an Arabic social media post toward a given topic out of a range of multiple topics. Each instance consists of a post–topic pair as input, and the system must output one of three stance labels: `Favor`, `Against`, or `Neither`. The task is framed as a multi-class classification problem and evaluated using macro-averaged F1-score. The dataset of Task B is a subset of the MARASTA dataset (Charfi et al., 2024) and it contains 1,205 manually annotated Arabic social media posts spanning two socio-political topics: *Arab Normalization with Israel* and *Refugee/Immigrant Presence in Jordan*. There are 577 threads about the *Normalization* topic and 628 threads about the *Refugees* topic. The dataset is balanced across stance categories to mitigate bias toward majority classes. The language reflects informal Arabic social media discourse, including dialectal variation and implicit stance expressions, which increase task difficulty.

2.2. Related Work

Stance detection has been widely studied in NLP, particularly in the context of social media. The SemEval-2016 Task 6 benchmark (Mohammad

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 294–298
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

et al., 2016) established stance detection toward specific targets in English tweets. Subsequent work extended the problem to cross-target and multilingual settings, highlighting the challenge of generalization across topics (Vamvas and Sennrich, 2020). Some surveys further categorize stance detection approaches, emphasizing transfer learning and transformer-based architectures (Hardalov et al., 2022).

Recent work has shown growing interest in Arabic stance detection, especially with the introduction of dedicated datasets and shared tasks. StanceEval 2024 (Alturayeif et al., 2024) represents the first large-scale Arabic stance-detection shared task and provides a benchmark dataset together with a comparative evaluation of multiple systems. In addition, the ArabicStanceX dataset (Alkhathlan et al., 2025) was recently introduced as a multi-topic Arabic social-media dataset, where the authors evaluated stance-detection models using transformer-based architectures. Other work, such as (Alkhraiji and Azmi, 2025) focused mainly on traditional Machine Learning models applied to Arabic tweets.

Despite these efforts, most previous work focuses either on dataset construction or on classical machine-learning approaches, while fewer studies explicitly address stance detection in short informal Arabic tweets using transformer-based models. In this work, we focus specifically on short informal tweet text and evaluate numerous BERT-based transformer models in a cross-topic stance-detection setting.

3. System Overview

We chose to work with BERT architecture due to its reported simplicity and efficiency (Farha and Magdy, 2021). The utilization of these pretrained transformers only requires adding a classification head to the base backbones. After filtering, we consider several flavors of the BERT architecture (Farha and Magdy, 2021; Alturayeif et al., 2022).

- bert-base-arabertv02
- MARBERT
- MARBERTv2
- bert-base-arabic-camelbert-da
- bert-base-qarib
- Mawqif
- bert-base-arabertv02-twitter

In the beginning, we tried optimizing generic models for our task by finetuning them using the provided dataset. The main challenge was the size of the dataset as it is not large enough to tweak the models significantly. Hence, we decided for a customized model for the task, which is **bert-base-arabertv02-twitter**. It is already trained on social media tweets, which are similar to our dataset. As a result, it needs less training than generic models. As shown in table 1, this model outperforms other BERT variants.

Model	F1-score
bert-base-arabertv02	0.73
MARBERT	**0.79**
MARBERTv2	0.78
camelbert-da	0.73
bert-base-arabertv02-twitter	**0.79**
QARIB	0.74
Mawqif	0.74

Table 1: Initial test results of different BERT-based models on the validation set using the starter code.

Nevertheless, the reported performance was still not optimal so we started tuning this model to get better results. we have frozen the model encoder layers so that they preserve the following: Core Arabic token understanding, Basic morphology, and syntax representation. This still allows final layers to adapt to our classification task. The starter code provided by the task organizers served as our baseline. It provides the necessary phases we needed including data preprocessing and evaluation metrics. Hence, data preprocessing includes: removing diacritics (tashkeel), tatweel, URLs, and mentions; normalizing alef variants and yaa variants and taa variants as well as spaces.

3.1. Data Augmentation

On our last attempts we tried data augmentation by three methods;

1. Back Translation
2. Generated Samples
3. ArabicStanceX dataset

The first method was executed by translating the Arabic text to english and back using two different LLMs. The second method was done using Claude to generate 300 tweets about the subject. All three methods to increase our data did not achieve any improvements.

4. Experimental Setup

Initially, we started testing several models on the starter code to find a good starting point. The best BERT backbones found were the **MARBERT** model and the **bert-base-arabertv02-twitter**. Afterwards, we started changing the starter code, we

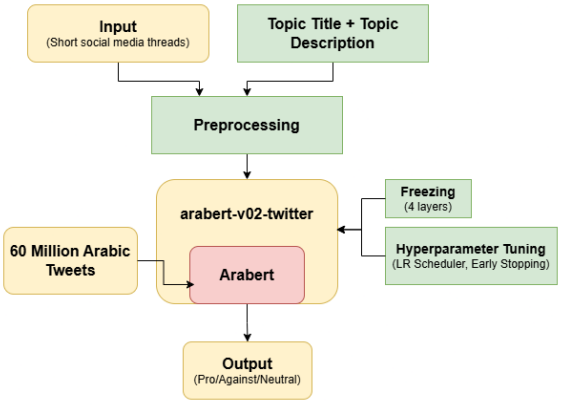


Figure 1: The system diagram. Blocks in green are contributed by us.

experimented with the following techniques: partial fine-tuning by freezing layers, different preprocessing, adding topic description was helpful (Alkhathlan et al., 2025), different hyper-parameters, and extra data using data augmentation which was inefficient. After the labeled validation set was revealed, we used it for cross-validation and early stopping to tune the epochs number. The following hyperparameters yielded the best results:

- **Number of Epochs:** 10
- **LR Scheduler:** linear
- **Learning Rate:** 2e-5
- **Maximum Length:** 64
- **Number of frozen Layers:** 4

دعم بناء علاقات بين الدول العربية ودولة اسرائيل ، دعم إقامة اللاجئين الأجانب داخل الدولة ،

Figure 2: Topic descriptions provided to the model as input during the training phase.

Figure 1 shows the sequence of steps in our system. As shown, **arabertv02-twitter** is a variant of Arabert that was trained with 60 million Arabic tweets. Figure 2 shows the descriptions of the topics that we added to their titles during the learning process to provide the model with more context which turned out to improve the performance.

5. Results

The shortlisted models that we tried to improve further are shown in table 2. The **bert-base-arabertv02** still outperformed other models on the test set; hence, the reported results in the following sections of this paper are based on it. The model achieved an F1-score of `0.8616` and an accuracy of **0.8619** on the test set. This was the second best performing model in the competition only 1% behind the best ranking model.

Model	F1-score
bert-base-arabertv02-twitter	0.86
bert-large-arabertv02-twitter	0.86
Marbert-v2	0.83

Table 2: Final results on test set after parameter tuning.

5.1. Ablation Analysis

Method	Improvement
LR Scheduler	+3%
Freezing Layers	+3%
Max Length	+2%
Early Stopping	+1%
Topic Description	+1%

Table 3: Impact of each method on the fine-tuned model performance.

Table 3 demonstrates the contribution of each tuning step. Finetuning the learning rate scheduler and the early stopping with the labeled validation set resulted in the most positive outcomes. It is important to note that the improvements achieved by these techniques do not necessarily add up together. In other words, the total achieved improvement after incorporating these techniques is less than the sum of the individual contribution.

5.2. Performance Analysis

Class	Precision	Recall	F1-score	Support
Against	0.88	0.89	**0.89**	**65**
Neutral	0.77	0.88	0.82	56
Pro	**0.96**	0.82	0.88	60
Macro Avg	0.87	0.86	0.86	181

Table 4: Classification Report of test set on the final model

To better understand the results of our final model, we investigate further its performance per class and per topic. Table 4 demonstrates that The model performed best on the `Against` samples, while the worst performance was associated with the `Neutral` class. This suggests that the model is able to efficiently detect the presence of negative stances and that it finds neutral samples the most confusing.

Next, we check the performance of the model per topic. As the metrics in tables 6 and 5 illustrate, the

Class	Precision	Recall	F1-score	Support
Against	0.92	0.77	**0.84**	30
Neutral	0.67	**0.90**	0.77	**31**
Pro	**0.95**	0.72	0.82	25
Macro Avg	0.84	0.80	0.81	86

Table 5: Classification Report of the "Normalization" topic (test set)

Class	Precision	Recall	F1-score	Support
Against	0.85	**1.00**	0.92	**35**
Neutral	0.95	0.84	0.89	25
Pro	**0.97**	0.89	**0.93**	**35**
Macro Avg	0.93	0.91	0.91	95

Table 6: Classification Report of the "Refugees" topic (test set)

model performs remarkably better on the topic of *Refugees* than the topic of *Normalization*, although the difference in the number of samples between the two topics was not big. This difference suggests that the complexity of discussions about the topic of *Normalization* can be higher than those related to the topic of *Refugees*.

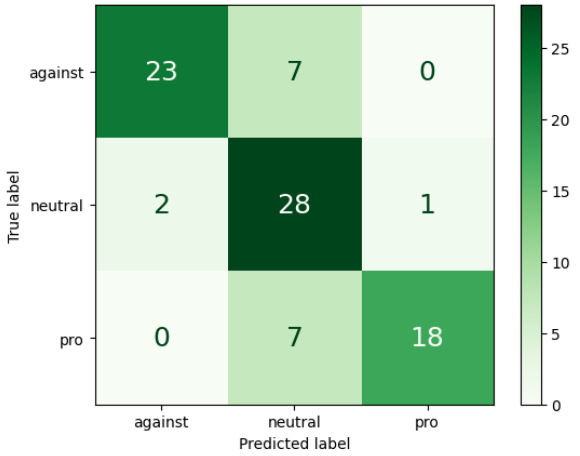


Figure 3: Confusion Matrix of the "Normalization" topic

Finally, we check the confusion matrix for each topic on the test set. We see that the model achieved a 100% recall for the `Against` on the *Refugees* topic. For the *Normalization* topic, we find that that most of the samples where the model was confused belonged to either `Against` or `Pro` classes and were misclassified as `Neutral`.

6. Limitations

This study has several limitations that should be acknowledged. First, the size of the dataset is relatively small (approximately 1,500 samples), which is not sufficient for full fine-tuning of large transformer models. As a result, the model had to be

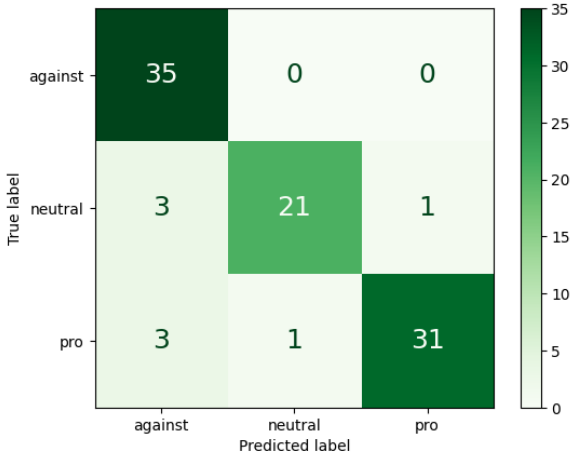


Figure 4: Confusion Matrix of the "Refugees" topic

partially frozen during training, which may have limited its ability to fully adapt to the stance-detection task. Second, explicit stance in Arabic tweets is often difficult to interpret, even for human readers. In many cases, the stance cannot be reliably inferred from the tweet alone and may depend on the author's previous opinions, background context, or implicit references that we do not have. This makes the task inherently challenging and increases the likelihood of ambiguous or misleading samples in the dataset. This challenge can be addressed by using network features for the processing of Arabic social data as earlier works did (Fawzi and Magdy, 2024; Fawzi et al., 2026a).

7. Conclusion

In this paper, we presented our system for Task B of the *StanceNakba 2026* shared task on cross-topic Arabic stance detection. We evaluated several BERT-based models and found that **bert-base-arabertv02-twitter** was best suited for this social media task. We further improved its performance through layer freezing, hyperparameter tuning, and the addition of topic descriptions as contextual input. Our final system achieved an F1-score of **0.86**, ranking second among the participating teams. Future work may explore larger Arabic language models and more effective data augmentation strategies to address the limited dataset size.

8. Bibliographical References

Kholoud Khalil Aldous, Md Rafiul Biswas, Mabrouka Bessghaier, Shimaa Ibrahim, Kais Attia, and Wajdi Zaghouani. 2026. StanceNakba shared task: Actor and topic-aware stance

detection in public discourse. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC'26)*, Palma, Spain.

Ali Alkhathlan, Faris Alahmadi, Faris Kateb, and Hend Al-Khalifa. 2025. Constructing and evaluating arabicstancex: a social media dataset for arabic stance detection. *Frontiers in Artificial Intelligence*, 8:1615800.

Arwa K. Alkhraiji and Aqil M. Azmi. 2025. Stance detection in arabic tweets: A machine learning framework for identifying extremist discourse. *Mathematics*, 13(18).

Nora Alturayeif, Hamzah Luqman, Zaid Alyafeai, and Asma Yamani. 2024. $stanceeval2024 : The first arabic stance detection shared task$. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 774–782, Bangkok, Thailand. Association for Computational Linguistics.

Nora Saleh Alturayeif, Hamzah Abdullah Luqman, and Moataz Aly Kamaleldin Ahmed. 2022. Mawqif: A multi-label arabic dataset for target-specific stance detection. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 174–184.

Anis Charfi, Mabrouka Ben-Sghaier, Andria Samy Raouf Atalla, Raghda Akasheh, Sara Al-Emadi, and Wajdi Zaghouani. 2024. Marasta: A multi-dialectal arabic cross-domain stance corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11060–11069.

Ibrahim Abu Farha and Walid Magdy. 2021. Benchmarking transformer-based language models for arabic sentiment and sarcasm detection. In *Proceedings of the sixth Arabic natural language processing workshop*, pages 21–31.

Mahmoud Fawzi. 2022. Characterisation of users' behaviours towards fake news through the analysis of their networks.

Mahmoud Fawzi and Walid Magdy. 2024. "pinocchio had a nose, you have a network!": On characterizing fake news spreaders on arabic social media. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–20.

Mahmoud Fawzi, Björn Ross, and Walid Magdy. 2026a. Fabricating holiness: Characterizing religious misinformation circulators on arabic social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 20.

Mahmoud Fawzi, Björn Ross, and Walid Magdy. 2026b. "god says we are right!": The interplay between religion and propaganda on arabic social media. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis- and disinformation identification. *ACM Computing Surveys*, 55(11).

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.

Jannis Vamvas and Rico Sennrich. 2020. X-stance: A multilingual multi-target dataset for stance detection. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 151–163.

Author Index

Abd El-Mageed, Samar Mahmoud, 171
Abd El Sayed, Fady A., 265
Abdelghany, Mohamed A., 265
Abdelhady, HebaTalla, 187
abdelouahab, baflah, 128
Abdo, Qusay, 60
Abou-Attia, Arwa M., 265
Abou Chakra, Chadi, 70
Abrahams, Alexei, 91
Abril, Sanaa, 50
Abu Hijleh, Shahd L., 43
Abudalfa, Shadi, 91, 98
Abusonoun, Abuelgasim Sami, 234
Adam, Faisal Muhammad, 104
Al desouky, Sara Abdulmonem, 275
Al Jallad, Khloud, 160
Al Shafi, Abdullah, 244
ALADDASI, Mohammad O., 43
Aldallal, Zeina, 196
Aldous, Kholoud Khalil, 80
Ali, Akhtar, 212
Ali, Ali Adel, 269
Ali, Mona Khaled, 269
Aliyu, Salisu, 104
AlKhidir, Lojien, 187
AlModhayan, Safwan, 196
Alqasem, Jana, 191
Alsafadi, Walid, 98
AlSaif, Joud Fahad, 258
Alseed, Jana Mohammad, 258
aly, menna, 226
Atiani, Serin, 60
Attia, Kais, 33, 80
Ayad, Sara, 234
Ayesh, Mutaz, 261

Balah, Batool Najeh, 1, 294
Barakat, Basel, 23
Barakat, Nizam, 23
Barmandah, Hassan, 160
Belhaouari, Samir Brahim, 201
Ben Abid, Firas, 70
Ben lahmar, El habib, 50
Bessghaier, Mabrouka, 33, 80
Biswas, Md Rafiul, 201
Biswas, Md. Rafiul, 80
BOULANOUAR, ZAHIRA, 290

Chamseddine, Ahmad Ali, 70
Chrouf, Sara, 196

Dao Sy, Duy Minh, 121, 137
Duridi, Tasneem, 177

El-Amrany, Samir, 98
El Houby, Enas, 239
El-Kassas, Wafaa, 239
Eldin, Fatimah Emad, 160
Elmimouni, Houda, 1
Elrefai, Eman, 294
Ezzini, Saad, 98

Faris, Aya Hafiz, 191
Fawzi, Mahmoud, 1, 294
Fawzy, Abdelrhman M., 265

Gbr, Shrouk Anwar, 280
Ghanem, Bernard, 70

Hamdan, Nancy, 206
Hamed, Osama, 177
hamed, SARA Ali, 191
Hamood, Alhasan, 191
Hamood, Alhasan Mohammad, 258
Hamoud, Hadi, 70
Haque, Md Rezwanul, 252
Hasan, K. M. Azharul, 244
Hassan, Ahmed A., 265
Hassan, Ahmed Eid, 265
Hegazy, Salma Khaled, 285
Hennara, Khalil, 196
Hossain, Md. Ajwad, 217
Hossain, Sk. Imran, 244
Hreden, Muhammad, 196
Huynh, Trung Kiet, 121

Ibrahim, Shimaa Amer, 80
Islam, Ashhadul, 201
Islam, Md. Milon, 244, 252

Jaber, Rahaf, 221
Jarrar, Mustafa, 70, 91
Jim, MD Jahid Hasan, 252
Jouni, Aya, 206
Juhaysh, Reem, 234

Kamil, Zainab, 191
Kannan, Ramesh R., 229
Karray, Fakhri, 252
Khaled, Toka, 294
Khalil, Enas, 239
Khalil, Enas A. Hakim, 133

Lamsiyah, Salima, 98
Le, Minh-Hoang, 113, 118

Magdy, Walid, 1
Mahmoud, Shahd O., 265
Mikros, George, 91
Mohamed, Ensaf, 171
Mohamed, Lamri, 128
Mohamed, Mayar M., 265
Mohamed, Mohamed Fathy, 171
Mohamed, Mohamed H., 265
Mohamed, Noureldeen H., 265
Mohamed, Rahma M., 265
Mostafa, Ibrahim Naser, 269
Mouanid, Sihame, 50

Nacar, Omer, 160
Nairat, Alaa, 147
Nairat, Aysar Mahmoud, 147
Nguyen Lam, Phu Quy, 121

Pham, Phu Hoa, 121

Qassim, Ahmed S., 265
Qawasmeh, Omar, 43
Qindeel, Asmaa, 294

Ragab, Mohamed Ibrahim, 280, 285
Rajalakshmi, Ratnavel, 229
Rashid, Ibad-ur-Rehman, 212

Saeed, Adnan, 60
Saïd, Aya, 206
Sawla, Shahd A., 265
Sayed, Mohamed Emad, 269
Shalash, Bilal, 70
Shujon, Md. Shakhoyat Rahman, 252
Sraiji, Tariq, 60
Sulaiman, Doaa Bahjat, 108

Tran, Nguyen Chi, 121

Yang, Bingzhou, 144
Yassin, Nisreen I. R., 133
Yassine, Sarah, 182
Youssef, Ahmad Edris, 191

Zaghouani, Wajdi, 33, 80, 201
Zahour, Omar, 50
Zaraket, Fadi, 206
Zaraket, Fadi A., 70
Zayet, Tasnim, 177
Zhang, Xinyan, 144