Instant Messaging Platforms News Multi-Task Classification for Attitude, Sentiment, and Discrimination Detection

Denilson Barbosa[1], **Taras Ustyianovych**[2]
[1]Department of Computing Science, University of Alberta, Canada,
[2]Department of Artificial Intelligence Systems, Lviv Polytechnic National University, Ukraine
[1]denilson@ualberta.ca, [2]taras.o.ustyianovych@lpnu.ua

Abstract

In the digital age, geopolitical events frequently catalyze discussions among global web users. Platforms such as social networks and messaging applications serve as vital means for information spreading and acquisition. The Russian aggression against Ukraine has notably intensified online discourse on the matter, drawing a significant audience eager for real-time updates. This surge in online activity inevitably results in the proliferation of content, some of which may be unreliable or manipulative. Given this context, the identification of such content with information distortion is imperative to mitigate bias and promote fairness. However, this task presents considerable challenges, primarily due to the lack of sophisticated language models capable of understanding the nuances and context of texts in low-resource languages, and the scarcity of well-annotated datasets for training such models. To address these gaps, we introduce the TRWU dataset – a meticulously annotated collection of **T**elegram news about the **R**ussian **w**ar in **U**kraine gathered starting from January 1, 2022. This paper outlines our methodology for semantic analysis and classification of these messages, aiming to ascertain their bias. Such an approach enhances our ability to detect manipulative and destructive content. Through descriptive statistical analysis, we explore deviations in message sentiment, stance, and metadata across different types of channels and levels of content creation activity. Our findings indicate a predominance of negative sentiment within the dataset. Additionally, our research elucidates distinct differences in the linguistic choices and phraseology among channels, based on their stance towards the war. This study contributes to the broader effort of understanding the spread and mitigating the impact of biased and manipulative content in digital communications.

Keywords: News Messages, Dataset, Text Classification, Destructive content detection

1. Introduction

The proliferation of internet and web technologies has had an impact on public discourse, shaping opinions and perceptions. With a wealth of data available from diverse sources, ranging from factual information to personal opinions, navigating this informational landscape can be daunting (Adams et al., 2023; Mendoza et al., 2023). Therefore, the exploitation of information literacy and critical thinking can distort public understanding and opinion (Aslett et al., 2023). Traditional technological tools have proven difficult in addressing these complex challenges (Zakharchenko et al., 2021).

The complexity of discerning opinions from objective facts is compounded in politically charged scenarios, such as the Russian invasion of Ukraine in February 2022. The narratives surrounding such events do not merely shape public morale but also influence mental health, beliefs, and international perspectives on credibility and support (Haq et al., 2022). In this context, the automated classification of content based on its biases becomes a pivotal tool for fostering a more informed and trustworthy Web environment (Meel and Vishwakarma, 2020). Previous efforts have explored various computational approaches to address these challenges, including classification (Solopova et al., 2023), text summarization (Galeshchuk, 2023b), and topic modeling (Ustyianovych et al., 2023), particularly in Ukrainian and Russian contexts (Galeshchuk, 2023a). However, the development of robust, explainable, and efficient models capable of accurately identifying the biases of textual content remains a pressing and relevant challenge. Such models not only aid in filtering and understanding content but also play a vital role in educating users about the nuances of misleading information. Communication strategies and linguistics constantly evolve with new approaches developed to interact with and address the target audience. Therefore, technological means for processing and understanding natural language and communication contexts need to remain up to date to keep up with current issues.

Our research contributes to this field by presenting a novel annotated dataset related to the Russian aggression against Ukraine with a multi-task transformer-based model trained to identify geopolitical stance, sentiment, and the presence of hate or discrimination in the input message. By leveraging the capabilities of large language models (LLMs), we delve into the intricacies of textual data, seeking to unveil patterns that distinguish biased narratives. Our findings underscore the potential of these technologies to enhance our comprehension

The Third Ukrainian Natural Language Processing Workshop (UNLP) @LREC-COLING-2024, pages 30–40
25 May, 2024.

of biased content and, by extension, to promote a nuanced and critical engagement with information in the digital age.

2. Related Work

The study of information campaigns in digital environments has become increasingly pertinent with the advent of social networks and web technologies. These platforms are not solely conduits for the spreading of factual information; they also serve as sites for strategic communications aimed at influencing public opinion and garnering support within online communities. An illustrative example of how digital platforms can be utilized for such purposes is observed in the analysis of various information campaigns, including those conducted on social media platforms (Courchesne et al., 2022). This study investigates the dynamics of online activity associated with significant geopolitical events, highlighting the capacity of strategic communication efforts to engage with and influence digital communities. The analysis, which encompasses a broad dataset of social media accounts, reveals a marked increase in online activity coinciding with pivotal events and underscores the effectiveness of coordinated information dissemination strategies in capturing public attention and shaping narrative discourse.

The comprehensive examination of these social media activities, including a study of over 126 thousand accounts, illustrates the challenges faced by content moderation teams and the sophisticated nature of modern information campaigns. Such studies highlight the complexity of digital information verification and the need for advanced methodologies to understand and navigate the intricacies of information manipulation in the digital age. A recent study by Park et al. (2022) describes the VoynaSlov dataset that was collected from two social networks, Twitter and VKontakte, to analyze and detect media opinion manipulations related to the Russian war in Ukraine. It consists of more than 38 million posts based on Russian media statements and expressions. The authors focus on distinguishing sources into state-affiliated and independent. As expected, the usage of words and phases differs between these two categories along with the formed topics distribution. The study results highlight a spike in user engagement and the number of generated posts after the invasion began on February 24, 2022. This observation confirms how real-world events engage users in online activity and content creation.

Fedushko et al. (2023) proposed innovative methods to support real-time decision-making about antagonistic user behavior on social networks. The proposed techniques showed significant results in decreasing the number of destructive content generated and shared, which contributed to more sustainable interactions in online communication. The developed models consider decisions, information environment, and decision-making criteria as the key processes for online community management. The methods were validated on a Facebook online community and showed an increase in user participation (and community size) in just one month after implementing the strategy for sustainable community development, indicating that it is possible to alert and guide users about the dangers of posting destructive comments online.

Threat detection in Web communication is another aspect that is worth attention and can be tackled with AI- and data-driven technologies. Semantic analysis combined with communication behavioral models is already successfully used to handle threats in social media discussions. Fedushko and Benova (2019) suggests a process for performing users' semantic analysis in an online environment, which improves the efficiency of threat detection by up to 40%.

Since a large number of discussions occur on social media platforms, it is crucial to understand the formed trends and patterns, especially in the context of specific subjects and objects. Visualization techniques might be efficiently used to investigate opinions and perform social communications mining. These methods were successfully used to analyze opinions appertained to such topics: 1) energy sources and 2) social network brands of academic institutions (Gutierrez et al., 2021). Our dataset and model contribute to the area of social media and instant messages analysis in order to have a full picture of the public stance towards specific topics, including sensitive ones.

Transformers and large language models have been effectively applied for the detection of unreliable information within news and online content. A case in point is the HQP dataset specifically collected to facilitate the identification of misinformation by incorporating 30 thousand tweets related to the war between Russia and Ukraine. This dataset is notable for its differentiated labeling approach, categorizing data into "high-quality" and "weak" labels. High-quality labels are distinguished by their validation through human review, ensuring the trustworthiness and accuracy of the data. In contrast, weak labels lack human validation, presenting a potential challenge to model accuracy (Maarouf et al., 2023). The methodology adopted for data labeling in the HQP dataset, and the subsequent application of pre-trained language models, showcases the critical role of high-quality labels in enhancing model performance. The achieved results highlight this, with models trained on high-quality labeled data achieving an Area Under the Curve (AUC) score of 92.25. This outcome indicates a significant improvement in the model's ability to detect untrustworthy content accurately, highlighting the importance of rigorously validated data during the design of effective detection systems.

Applications of few-shot learning and zero-shot classification are other promising areas discussed to improve the detection of harmful content and bias, and puzzle out related information trustworthiness tasks (Nayeon et al., 2021; Liew et al., 2023; Modupe et al., 2023; Yao et al., 2022).

3. Telegram War News Dataset

3.1. Dataset Collection

The Russian-Ukrainian war dataset has been collected from thoroughly selected pro-Russian and pro-Ukrainian Telegram channels. The selection of channels is based on the lists of reliable versus untrustworthy information sources provided by the Ukrainian Center for Countering Disinformation (for Countering Disinformation, 2022) and the Institute of Mass Information (of Mass Information, 2023). Telegram is an instant messaging application with 700 million monthly active users. It offers the option to create channels for broadcasting content to large audiences. Each message can contain media content, which makes it suitable for multimodal news analysis (Wang et al., 2022b). Users in a channel can leave comments and reach with emojis, which leads to another exciting area of research – online user engagement and behavior analysis (Fedushko et al., 2020). Telegram has an open API for extracting data from specific channels (based on their ID). We collected data from six news and blog-like channels regularly posting content about the Russian war against Ukraine. Statistics on the number of messages retrieved from each channel are given in Table 1.

The total number of messages collected is 252,677 from January 1st, 2022 until December 14th, 2023. At the time of writing, new data is being collected for further processing. Each message contains the channel name, timestamp, message ID for the selected channel, and the text of the message itself. The messages have been labeled using the `gpt-3.5-turbo-1106` large language model with a human-in-the-loop to ensure the reliability of the assigned labels. Additional data validation and normalization were accomplished to standardize the labels and meet the actual research purpose. Messages are labeled according to the channel's attitude mentioned in Table 1 and randomly split into training, validation, and testing sets with such percent ratios: 90%, 5%, and 5% correspondingly.

3.2. Dataset Statistics

The uniqueness of our dataset primarily derives from its comprehensive compilation process and focused applicability to the Russian-Ukrainian war. Unlike conventional datasets that predominantly source from widely used social media platforms like Twitter and Facebook, our dataset uniquely taps into the Telegram instant messaging platform. This choice was deliberate, given Telegram's distinct user base and communication style, which significantly differ from other platforms. Telegram channels offer a rich amount of data in varied tones—ranging from news and factual reports to blog posts and opinion pieces. This diversity not only improves the dataset but also makes it exceptionally versatile for Natural Language Processing (NLP) research, promoting a broad exploration of communication techniques and content types.

A pivotal aspect of our dataset's development was a thorough selection of sources, ensuring that each included channel introduced a clear stance (pro-Russian or pro-Ukrainian) regarding the war. This careful curation process guarantees the dataset's relevance and validity for studies focusing on sentiment analysis, manipulative content detection, and the examination of targeting tactics. Our research aims to analyze the sentiment of messages from a pro-Ukrainian perspective. It's important to consider that the same message can be interpreted differently by audiences based on their viewpoints and backgrounds.

Further distinguishing our dataset is the use of GPT-3.5 for initial labeling, tasked with extracting sentiment and filtering out irrelevant content. This step was augmented by human validation to ensure the accuracy and reliability of the labels assigned by the AI, addressing potential biases and inaccuracies inherent in automated processes.

Our motivation to create this dataset facilitates a nuanced analysis of communication patterns, enabling researchers to identify harmful and misleading content effectively. Its applicability extends to improving government accounting information systems, as demonstrated by related studies, showcasing its potential to influence a wide range of fields positively (Duan et al., 2023). By carefully curating, labeling, and validating our dataset, we have created a resource that stands out for its methodological stringency and direct relevance to current geopolitical events, offering invaluable insights into the dynamics of information dissemination and reception in the digital age.

According to Table 1, 152,502 (55.19%) of the content is retrieved from pro-Russian sources, whereas 123,812 (44.80%) entities belong to pro-Ukrainian channels. All the channels' sentiment most frequent value except *rian_ru* is negative, and for the latter it is neutral. A histogram with the col-

Channel	Stance	Count	Fraction	Mean token count	Mode sentiment
rian_ru	Pro-Russian	79,663	28.83%	28.26	neutral
ROSSIYA_SEGODNIA	Pro-Russian	69,238	25.05%	55.16	negative
uniannet	Pro-Ukrainian	67,727	24.51%	48.58	negative
radiosvoboda	Pro-Ukrainian	33,225	12.02%	108.63	negative
UkrPravdaMainNews	Pro-Ukrainian	22,860	8.27%	46.34	negative
ZE_kartel	Pro-Russian	3,601	1.30%	74.91	negative

Table 1: Number and percentage of messages per channel

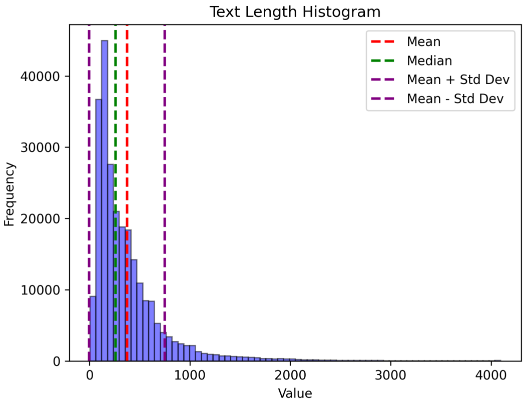

Figure 1: Text Length Histogram with mean, median, and standard deviation ranges.

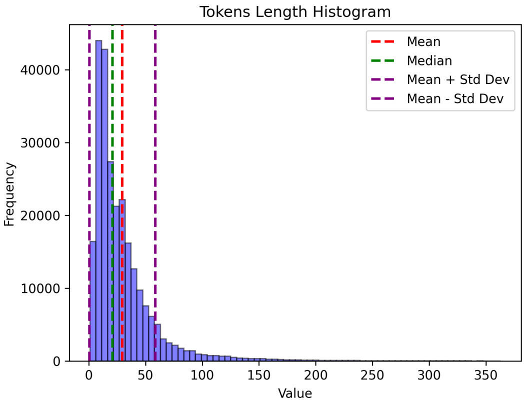

Figure 2: Histogram for the number of tokens with mean, median, and standard deviation ranges.

lected data text length is shown in Figure 1. The mean and median values are 373.47 and 259, respectively, and the standard deviation is 376.63. Also, 89.79% of the messages are below one standard deviation from the mean, meaning that their length is less than or equal to 750.10. After text preprocessing, the mean text length was reduced by 32%.

We provide a summary of the number of remaining tokens per message after applying preprocessing, which includes text cleaning, stopword removal, and lemmatization. The mean and median token count after text preprocessing are 29.52 and 21 respectively, and the standard deviation is 29. The obtained distribution pattern is similar to the one for text length and is shown in Figure 2.

3.3. Semantic Analysis

Analyzing data on the semantic level is crucial to extracting meaning from the text and understanding the critical features of the studied sources concerning word usage and style. To create a general comprehension of the text data after cleaning and lemmatization, we identified the most frequently used words: "Ukraine", "Russian", "warlike", "claim", "connection", "USA", "Putin", "Zelenskyi", "offensive", "sanction", "destroy", "weapon". There are 92,342 and 118,870 unique entities used in pro-Russian and pro-Ukrainian channels, respectively. This is an exciting finding since there are far more pro-Russian messages; nevertheless, the word usage within pro-Ukrainian sentences is significantly richer.

We observed a “separation” in vocabulary between the two sides: 57.06% of the unique words used by the pro-Ukrainian sources do not appear in pro-Russian channels; within Pro-Russian channels, this rate is 44.72%. We analyzed unique words within each side and found that they mainly include derogatory named entities against the opposite side, abbreviations, local areas and regions, and words with local and specific meanings. For example, unique words from pro-Russian channels contain the character "Z" which is known to be their symbol of the war. Also, the word "war" itself is replaced by "special military operation". Some pro-Ukrainian publications might contain Ukrainian words even though the text piece is written in Russian. This factor contributes to the number of unique words used between the sources and can help our model differentiate between these originating sources. The data presented in the figure 3 compares sentiment classification results from an automated method using OpenAI API `gpt-3.5-turbo-1106` model with human validation. The

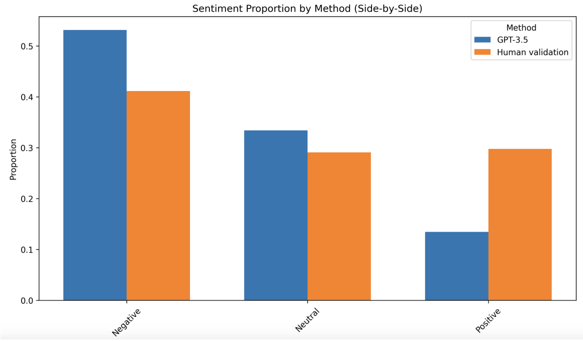

Figure 3: Sentiment proportion by method

figure illustrates the proportion of messages categorized as negative, neutral, and positive. The AI-based sentiment analysis results show that the majority of the dataset, 53.13%, exhibits a negative sentiment. Neutral sentiments, which may represent unbiased reporting, factual statements, or ambiguous content, constitute 33.30% of the dataset. Positive sentiments, indicative of optimism and supportive statements, account for 13.45%. A small fraction of the data (not represented on the figure), merely 0.12%, is categorized under mixed sentiments, highlighting texts that possibly contain conflicting emotions or viewpoints. In contrast, human validation results based on a randomly selected sample of labeled messages, are in a different sentiment distribution. While the proportion of negatively classified messages is similar to the GPT-3.5 results, human validation assigns a significantly lower proportion as neutral and a considerably higher proportion as positive. The discrepancy, specifically in the positive category, may be due to the detailed and contextual understanding that human validators bring to the task, which automated systems like GPT-3.5 may not fully capture, especially in the complex and sensitive context of war-related communications. The figure highlights the critical role of human oversight in sentiment analysis and the importance of multimodal validation for sensitive topics. The obtained sentiment values correspond with the stance and source channel of the message. The sentiment distribution underscores the complexity and variability of the sentiments expressed in the Web communication, offering valuable insights into the ruling attitudes and perceptions within the collected data. However, it should be noted that the AI-based labels offered sentiment classification from a prospective without favoritism to any side of the war. A refined version of the prompt with few-shot learning might improve the obtained results and make them suitable to identify the sentiment according to specific requirements.

We highlight the need to employ entity-level sentiment detection since distinct sentiments can be assigned to multiple entities represented in a piece of text (Rønningstad et al., 2022). This approach would contribute to the identification of the message's stance toward the war, and provide insights on the named entities represented within the text. Also, it offers a multi-faceted sentiment analysis compared to examining data from a single perspective.

Figure 4 shows 7-day window rolling average sentiment values by channel's attitude and applied methodology to detect sentiment over time. The usage of the rolling average sentiment score smooths out the noise, providing a clear view of the overall trends over time. Pro-Russian channels are represented with mostly negative sentiment scores according to the GPT-3.5 classification throughout the observed period. The sentiment scores for pro-Russian channels (shown in blue) demonstrate changes over time but with a generally less pronounced variance. In contrast, the sentiment scores for pro-Ukrainian channels (shown in orange) appear to follow a similar trend, maintaining lower average sentiment values compared to their pro-Russian counterparts. However, the sentiment values for pro-Ukrainian channels also fluctuate, suggesting that external factors and evolving news dynamics impact them. Therefore, when evaluating sentiment with technological means, it is important to consider the biases of these analytical tools being used. Our research results show that the GPT-3.5 model tends to interpret themes of war and conflict with a negative sentiment despite the evidence that some messages might be perceived differently by specific users. This is supported by the consistently negative sentiment scores given to massages of both channel viewpoints throughout the period studied. So, our finding highlights the importance of method selection in sentiment analysis studies and underscores the value of multiple analytical approaches comparison for a comprehensive view of trends in digital communication. Nevertheless, the employed GPT-based method proves the sentiment scores are mainly negative due to the nature of events.

3.4. Challenges and Limitations

Detecting biased, misleading, and manipulative content in such a dynamic environment as instant messaging platforms or social media is challenging because new data gets generated and shared in real-time, forming patterns unseen in historical data. So, usage of methods like incremental learning (Shan et al., 2020; Abdalla et al., 2022; Barve et al., 2022; Wang et al., 2022a) and well-established ML operations processes are becoming extremely helpful in these scenarios (Shukla and Cartlidge, 2022; Jarrahi et al., 2023; Mäkinen et al., 2021).

Furthermore, there is very little properly and publicly available labeled data to identify such con-

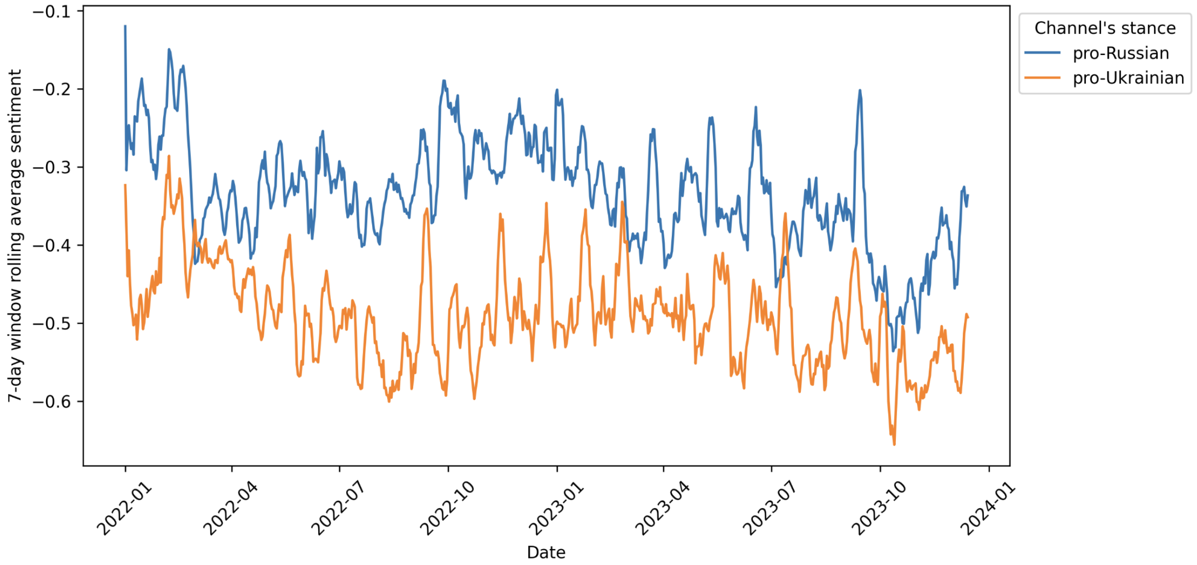

Figure 4: 7-day window rolling average GPT-3.5 sentiment score per channel stance

tent in the context of the Russian-Ukrainian war. Ukrainian is considered a low-resource language, with few available tools and models (Gomez et al., 2023). Selecting relevant sources and designing comprehensive labeling methods is crucial for developing high-performing models in the future. With that in mind, care was taken to collect data that was clearly associated with either side of the war and represented the respective attitude in their Web publications. For instance, a Ukrainian-based Telegram channel *UkrPravdaMainNews* was chosen because it posts pro-Ukrainian news, whereas the Russian-based channel *rian_ru*, which is part of a well-known Russian news agency, was chosen because it contains pro-Russian publications. The application of GPT-3.5 to assign such labels as geopolitical stance, sentiment, and presence of discrimination in an input text allowed us to identify the most relevant to the subject matter messages. Additional human validation, which included exploratory data analysis and verification of the assigned labels, significantly improved the dataset.

Additionally, the usage of machine learning models, DNNs (deep neural networks) as well as statistical methods can confirm whether there is a statistical difference between these and other examined labels. Topic models can be applied to categorize the data in an unsupervised manner and provide insights about the subject matters they contain.

Our dataset contains both pro-Ukrainian and pro-Russian texts written in the Russian language. However, there is a shortage of pro-Russian publications in Ukrainian which complicates achieving the defined goal for this language. Data augmentation methods including transformer-based translation might become handy to overcome this challenge (Liu et al., 2023; Gong et al., 2022). This is also a promising way to develop a multilingual model in further research.

About 40% of the messages in the dataset are not related to the war between Russia and Ukraine, which has been identified with GPT-3.5 zero-shot classification and human verification. Controlling the percentage of these entities is crucial to keep an optimal balance between relevant and extraneous messages in order to accomplish the modeling part. So, employing means to denoise the data and extract the most informative samples is crucial to reach the target of this study.

4. Data Processing

The whole workflow is depicted in Figure 5. The diagram outlines a multi-stage process for analyzing and processing text data from Telegram messages, aimed at evaluating the dataset's predictive capabilities with conventional machine learning methods and fine-tuning language models. The former technique involved input text cleaning and preprocessing using spaCy `ru_core_news_lg` and `ua_core_news_lg` language pipelines, creating word embeddings with fastText, and vectors manipulation. The fastText model was trained with the following parameters: vector size of 300, window size of 5, minimum word frequency of 3, training algorithm was skip-gram, ten epochs, and four worker threads. Then, the formed word embeddings were passed to the XGBoost classifier for hyperparameter tuning and evaluation. The data processing required for performing the NLP transformer-based approach consisted of such steps: text data cleaning, prompts generation for zero-shot classification, extraction and standardization of the LLM's output, and data unification. The formed dataset contained the text messages with corresponding Telegram metadata (ID, datetime, channel) and assigned labels by `gpt-3.5-turbo-1106`. The following prompt instructions were provided to the large language model: "Analyze the following messages related to the war between Ukraine and Russia. For each message: 1. Determine the sentiment (positive, negative, neutral, etc.) expressed in the message. 2. Identify geopolitical attitude or hate/discrimination and in favor of what side it is expressed: indicate whether it's pro-Ukrainian, pro-Russian, or any other geopolitical stance. Take into account that messages might contain glorification, hate, and discrimination, which should be considered when classifying attitudes. 3. If the message lacks a geopolitical attitude or isn't related to the conflict, mark it as not applicable to geopolitical attitude. The output should be returned as a Python dictionary array with such keys: message ID, sentiment, detected favorable attitude, and whether a message contains hate or discrimination (yes or no)". Human validation was accomplished afterward to ensure data quality, standardized values for categorical variables, and accurate annotation. We employed exploratory data analysis of the GPT-based labels to find and correct abnormal or unexpected values, gather statistics, and correlate them to find mislabeled entities. A sample of the data was taken for manual validation, and accuracy scores between human and AI-based labels were calculated. The obtained human validation results show mediocre performance in determining the proper sentiment in the context of events like a war. On the other hand, the AI agent did more than 80% correct on the geopolitical attitude and identifying irrelevant content. The data was passed as input to language models for fine-tuning. The returned outputs by the AI-based agent were transformed and converted into a pandas data frame and joined with the original dataset to make it suitable for model training. This workflow is crucial for the methodological processing of raw Telegram messages into valuable information assets through advanced NLP techniques. Each step of the presented workflow is designed to enhance the overall predictive performance and capabilities of the models.

5. Text Classification

The modeling part was performed on the collected Telegram War News dataset, first to assess its predictive performance using the XGBoost classifier and, second, to build a robust multi-task language model capable of distinguishing between pro-Ukrainian and pro-Russian messages, their sentiment, and stance. Such a model will become extremely helpful in mitigating the consequences of bias and misleading content spreading through Internet resources with specific attitudes.

5.1. Experimental Settings

We conducted hyperparameter optimization targeting the Area Under the Curve (AUC) score, complemented by a 3-fold cross-validation strategy for the XGBoost classifier on the training dataset. The evaluation of the optimized model was carried out on a separate testing set. Each input document was represented as a 300-dimensional vector. The search for optimal hyperparameters utilized the `hyperopt` package, with a defined search space that included the maximum depth of trees, learning rate, fraction of data used per iteration, minimum weight of child nodes, gamma as the regularization parameter, subsample ratio of features for constructing each tree, and the type of boosting model employed. We conducted a total of 35 trials, with the Tree of Parzen Estimators (TPE) algorithm chosen for the optimization process.

Fine-tuning of the multi-task language models was executed on computing instances equipped with NVIDIA Tesla V100 GPUs. We utilized the `google/mt5-base` and `xlm-roberta-base` for their multilingual capabilities in text tokenization and subsequent fine-tuning phases. The training phase involves fine-tuning the models on the multi-variable data frame with a custom PyTorch Dataset instance, which efficiently manages data fetching. The models, specifically MT5EncoderModel and XLMRobertaModel, were adapted with custom adjustments to their output layers and loss computation methods, assigning distinct weights to each predictive variable. The variables for prediction included: the channel's originating source attitude, sentiment, stance, presence of discrimination, combined channel's attitude and sentiment, and a merge of stance and sentiment. Tokenization restricted the text input to a length of 256 tokens. The training process spanned 10 epochs with batch sizes of 64 for both training and evaluation. Evaluation is conducted on a separate validation dataset to assess the model's accuracy and effectiveness in handling both tasks simultaneously, leading to its subsequent deployment for real-world applications.

5.2. Results

The optimal set of hyperparameters to build a robust XGBoost classifier for a message originating channel's attitude was: booster: 'gbtree'; colsample_bytree: 0.99837; gamma: 0.17946; learning_rate: 0.18935; max_depth: 17; min_child_weight: 14; and subsample: 0.89539. The final AUC scores on training and testing sets

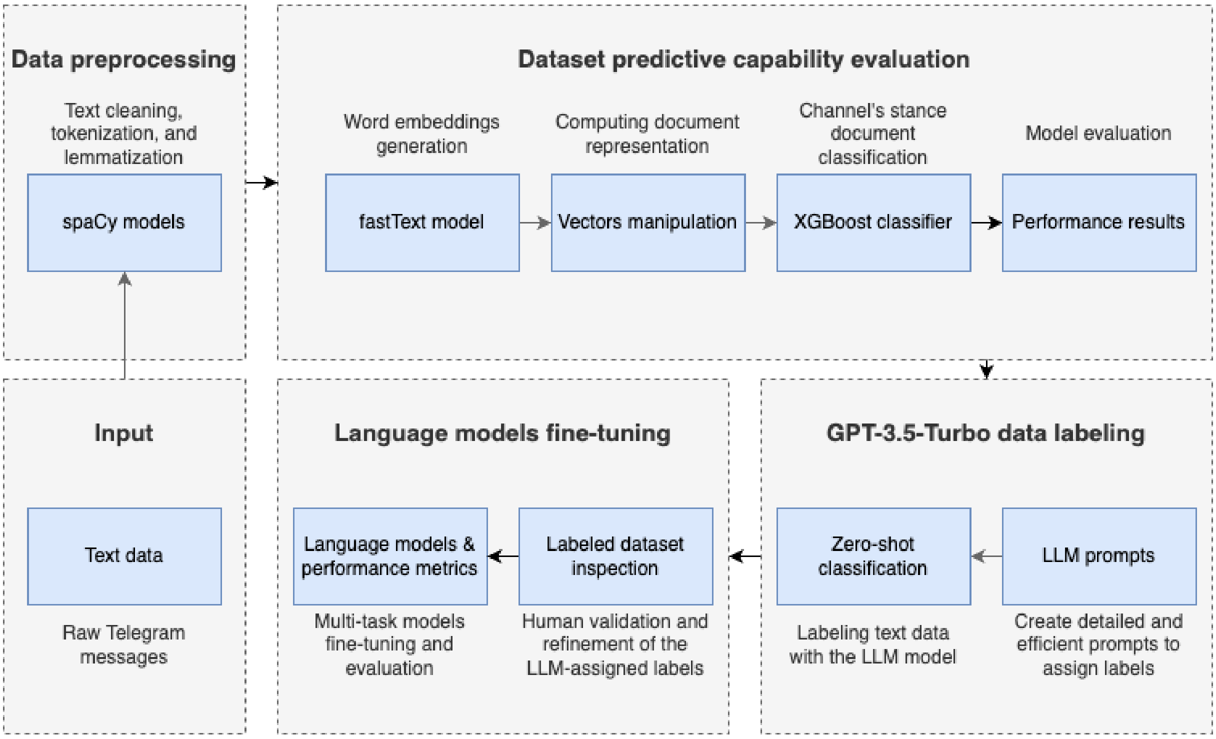
Data preprocessing
Text cleaning, tokenization, and lemmatization
spaCy models
Dataset predictive capability evaluation
Word embeddings generation
fastText model
Computing document representation
Vectors manipulation
Channel's stance document classification
XGBoost classifier
Model evaluation
Performance results
Input
Text data
Raw Telegram messages
Language models fine-tuning
Language models & performance metrics
Multi-task models fine-tuning and evaluation
Labeled dataset inspection
Human validation and refinement of the LLM-assigned labels
GPT-3.5-Turbo data labeling
Zero-shot classification
Labeling text data with the LLM model
LLM prompts
Create detailed and efficient prompts to assign labels

Figure 5: Telegram data inference pipeline

are 0.9715 and 0.9065, respectively. We computed such metrics as accuracy (0.9088), precision (0.9062), recall (0.8862), and F1 score (0.8961) as well.

The multi-task model displayed above-mediocre performance with an average accuracy of 0.74. It effectively identified the originating channel's attitude with a high accuracy of 0.95 and detected the presence of discrimination with an accuracy of 0.94. However, when the model was tasked with simultaneous detection of channel attitude and sentiment, the accuracy slightly reduced to 0.67, and further to 0.51 for combining geopolitical stance and sentiment. This indicates a more challenging scenario when the model is required to discern multiple nuanced aspects concurrently. These results show that while the model exhibits high accuracy with certain individual tasks, particularly in detecting the originating channel attitude and discrimination, there is a trade-off in performance when multitasking on sentiment and geopolitical stance. The obtained results highlight the complexity of multi-faceted analysis and point to opportunities for further improvement in multi-task modeling. It is worth paying detailed attention to data labeling and fine-tuning more complex language models. Also, the application of a single-task classification might improve the performance and design a specific targeted classification tool.

6. Conclusion and Future Work

Our research introduces the TRWU dataset, comprising texts from pro-Ukrainian and pro-Russian Telegram channels, featuring both factual and opinionated content. This dataset's uniqueness lies in its contemporaneous nature and thoroughly selected sources, delivering a comparative analysis of communication patterns. We used text mining to identify key lexical features and word usage across different channels. Our classification pipeline, which integrates spaCy, fastText, and XGBoost, was optimized to predict the stance of messages. We uncovered essential hyperparameters for optimal performance. We used zero-shot classification along with human validation for data labeling. The fine-tuned multi-task language model successfully classified the originating channel's attitude and presence of discrimination. Our findings indicate a need for enhanced sentiment detection tools for Ukrainian and Russian languages.

Future Work. Our proposed future work includes: 1) advancing stance and sentiment classification with rigorous labeling and model fine-tuning; 2) implementing vector databases for efficient document collocation; 3) context-based entity sentiment analysis, especially in conflict-related discourse; 4) pursuing excellence in model performance for both multi-task and single-task objectives; 5) further developing models for low-resources languages like Ukrainian (Laba et al., 2023).

7. Acknowledgements

This work was supported by a grant from the Canadian Institute of Ukrainian Studies (from the Mykola Klid Memorial Endowment Fund).

7.1. Ethical considerations

We are aware that the dataset we collected might contain harmful content because of the nature of the data. We have attempted to be unbiased in collecting the data from the selected channels and have not tried to censor any content. So, we will take respective precautions to warn users of this once the dataset is released. Therefore, ethical considerations are crucial when working this dataset for bias and manipulative patterns detection since content related to subjects like war can be sensitive, distorted, or unfair (Deepak, 2021). We strongly recommend evaluating the results with fairness metrics and using machine learning monitoring to improve observability and awareness of how such systems perform (Ashktorab et al., 2023). Utilizing tools for interpretability and explainability is essential to tackle this challenge and ensure transparency of the models.

8. Bibliographical References

H.B. Abdalla, A.M. Ahmed, S.R.M. Zeebaree, A. Alkhayyat, and B. Ihnaini. 2022. Rider weed deep residual network-based incremental model for text classification using multidimensional features and mapreduce. *PeerJ Computer Science*.

Zoë Adams, Magda Osman, Christos Bechlivanidis, and Björn Meder. 2023. (why) is misinformation a problem? *Perspectives on Psychological Science*, 18(6):1436–1463. PMID: 36795592.

Z. Ashktorab, B. Hoover, M. Agarwal, C. Dugan, w. Geyer, H. B. Yang, and M. Yurochkin. 2023. Fairness evaluation in text classification: Machine learning practitioner perspectives of individual and group fairness. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM.

K. Aslett, Z. Sanderson, W. Godel, N. Persily, J. Nagler, and J. A. Tucker. 2023. Online searches to evaluate misinformation can increase its perceived veracity. *Nature*, 625:548–556.

Y. Barve, J. R. Saini, K. Kotecha, and H. Gaikwad. 2022. Detecting and fact-checking misinformation using "veracity scanning model". *International Journal of Advanced Computer Science and Applications*, 13(2).

L. Courchesne, B. Rasikh, B. McQuinn, and C. Buntain. 2022. Powered by twitter? the taliban's takeover of afghanistan. ESOC Working Paper 30, Emperical Studies of Conflict.

P. Deepak. 2021. *Ethical Considerations in Data-Driven Fake News Detection*, pages 205–232. Springer International Publishing, Cham.

H. K. Duan, M. A. Vasarhelyi, M. Codesso, and Z. Alzamil. 2023. Enhancing the government accounting information systems using social media information: An application of text mining and machine learning. *International Journal of Accounting Information Systems*, 48:100600.

S. Fedushko and E. Benova. 2019. Semantic analysis for information and communication threats detection of online service users. *Procedia Computer Science*, 160:254–259. The 10th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN-2019) / The 9th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2019) / Affiliated Workshops.

S. Fedushko, K. Molodetska, and Yu. Syerov. 2023. Decision-making approaches in the antagonistic digital communication of the online communities users. *Social Network Analysis and Mining*.

S. Fedushko, T. Ustyianovych, Yu. Syerov, and T. Peracek. 2020. User-engagement score and slis/slos/slas measurements correlation of e-business projects through big data analysis. *Applied Sciences*, 10(24).

Center for Countering Disinformation. Ccd announces an updated list of infoterrorist channels operating in ukraine [online]. 2022.

S. Galeshchuk. 2023a. Abstractive summarization for the Ukrainian language: Multi-task learning with hromadske.ua news dataset. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 49–53, Dubrovnik, Croatia. Association for Computational Linguistics.

Svitlana Galeshchuk. 2023b. Abstractive summarization for the Ukrainian language: Multi-task learning with hromadske.ua news dataset. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 49–53, Dubrovnik, Croatia. Association for Computational Linguistics.

Frank Gomez, Alla Rozovskaya, and Dan Roth. 2023. A low-resource approach to the grammatical error correction of ukrainian. In *Proceedings* *of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 114–120.

H. Gong, X. Li, and D. Genzel. 2022. Adaptive sparse transformer for multilingual translation.

C. A. Gutierrez, Whittaker, A. Whittaker, K. M. Patenio, J. Gehman, L. L. M. Lefsrud, D. Barbosa, and E. Stroulia. 2021. Analyzing and visualizing twitter conversations. In *Proceedings of the 31st Annual International Conference on Computer Science and Software Engineering*, CASCON '21, page 4–13, USA. IBM Corp.

E.-U. Haq, G. Tyson, T. Braud, and P. Hui. 2022. Weaponising social media for information divide and warfare. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, HT '22, page 259–262, New York, NY, USA. Association for Computing Machinery.

M. H. Jarrahi, A. Memariani, and S. Guha. 2023. The principles of data-centric ai. *Commun. ACM*, 66(8):84–92.

M. Korobov. 2015. Morphological analyzer and generator for russian and ukrainian languages. In *Analysis of Images, Social Networks and Texts*, pages 320–332, Cham. Springer International Publishing.

Yurii Laba, Volodymyr Mudryi, Dmytro Chaplynskyi, Mariana Romanyshyn, and Oles Dobosevych. 2023. Contextual embeddings for Ukrainian: A large language model approach to word sense disambiguation. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 11–19, Dubrovnik, Croatia. Association for Computational Linguistics.

X. Y. Liew, N. Hameed, and J. Closand J. E. Fischer. 2023. Predicting stance to detect misinformation in few-shot learning. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*, TAS '23, New York, NY, USA. Association for Computing Machinery.

X. Liu, J. He, M. Liu, Z. Yin, L. Yin, and W. Zheng. 2023. A scenario-generic neural machine translation data augmentation method. *Electronics*, 12(10).

A. Maarouf, D. Bär, D. Geissler, and S. Feuerriegel. 2023. Hqp: A human-annotated dataset for detecting online propaganda.

S. Mäkinen, H. Skogström, E. Laaksonen, and T. Mikkonen. 2021. Who needs mlops: What data scientists seek to accomplish and how can mlops help? *CoRR*, abs/2103.08942.

P. Meel and D.K. Vishwakarma. 2020. Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Systems with Applications*, 153:112986.

Marcelo Mendoza, Sebastián Valenzuela, Enrique Núñez-Mussa, Fabián Padilla, Eliana Providel, Sebastián Campos, Renato Bassi, Andrea Riquelme, Valeria Aldana, and Claudia López. 2023. A study on information disorders on social networks during the chilean social outbreak and covid-19 pandemic. *Applied Sciences*, 13(9).

A. Modupe, T. Sindane, and V. Marivate. 2023. Zero-shot transfer learning using affix and correlated cross-lingual embeddings. *Authorea*.

L. Nayeon, B. Z. Li, S. Wang, P. Fung, H. Ma, W. Yih, and M. Khabsa. 2021. On unifying misinformation detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5479–5485, Online. Association for Computational Linguistics.

Institute of Mass Information. Online media that have become the highest quality: white list of the second half of 2023 [online]. 2023.

C.Y. Park, J. Mendelsohn, A. Field, and Yu. Tsvetkov. 2022. Challenges and opportunities in information manipulation detection: An examination of wartime Russian media. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5209–5235, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Egil Rønningstad, Erik Velldal, and Lilja Øvrelid. 2022. Entity-level sentiment analysis (ELSA): An exploratory task survey. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6773–6783, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

G. Shan, S. Xu, L. Yang, S. Jia, and Y. Xiang. 2020. Learn#: A novel incremental learning method for text classification. *Expert Systems with Applications*, 147:113198.

R.M. Shukla and J. Cartlidge. 2022. Challenges faced by industries and their potential solutions in deploying machine learning applications. In *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0119–0124.

Veronika Solopova, Christoph Benzmüller, and Tim Landgraf. 2023. The evolution of pro-kremlin propaganda from a machine learning and linguistics perspective. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 40–48, Dubrovnik, Croatia. Association for Computational Linguistics.

T. Ustyianovych, N. Kasianchuk, H. Falfushynska, S. Fedushko, and E. Siemens. 2023. Dynamic topic modelling of online discussions on the russian war in ukraine. In *Proceedings of International Conference on Applied Innovation in IT*, pages 81–89.

R. Wang, T. Yu, H. Zhao, S. Kim, S. Mitra, R. Zhang, and R. Henao. 2022a. Few-shot class-incremental learning for named entity recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 571–582, Dublin, Ireland. Association for Computational Linguistics.

Zh. Wang, X. Shan, X. Zhang, and J .Yang. 2022b. N24News: A new dataset for multimodal news classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6768–6775, Marseille, France. European Language Resources Association.

P. Yao, T. Renwick, and D. Barbosa. 2022. WordTies: Measuring word associations in language models via constrained sampling. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5959–5970, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A. Zakharchenko, T. Peráček, S. Fedushko, Yu. Syerov, and O. Trach. 2021. When fact-checking and 'bbc standards' are helpless: 'fake newsworthy event' manipulation and the reaction of the 'high-quality media' on it. *Sustainability*, 13(2).