Exploring the Potential of Large Language Models (LLMs) for Low-resource Languages: A Study on Named-Entity Recognition (NER) and Part-Of-Speech (POS) Tagging for Nepali Language

Bipesh Subedi[1], **Sunil Regmi**[1], **Bal Krishna Bal**[1], **Praveen Acharya**[2]
[1]Information and Language Processing Research Lab
[1]Department of Computer Science & Engineering, Kathmandu University, Nepal
[2]School of Computing, Dublin City University, Ireland
{bipeshrajsubedi@gmail.com, sunilregmi233@gmail.com, bal@ku.edu.com, acharyaprvn@gmail.com}

Abstract

Large Language Models (LLMs) have made significant advancements in Natural Language Processing (NLP) by excelling in various NLP tasks. This study specifically focuses on evaluating the performance of LLMs for Named Entity Recognition (NER) and Part-of-Speech (POS) tagging for a low-resource language, Nepali. The aim is to study the effectiveness of these models for languages with limited resources by conducting experiments involving various parameters and fine-tuning and evaluating two datasets namely, ILPRL and EBIQUITY. In this work, we have experimented with eight LLMs for Nepali NER and POS tagging. While some prior works utilized larger datasets than ours, our contribution lies in presenting a comprehensive analysis of multiple LLMs in a unified setting. The findings indicate that NepBERTa, trained solely in the Nepali language, demonstrated the highest performance with F1-scores of 0.76 and 0.90 in ILPRL dataset. Similarly, it achieved 0.79 and 0.97 in EBIQUITY dataset for NER and POS respectively. This study not only highlights the potential of LLMs in performing classification tasks for low-resource languages but also compares their performance with that of alternative approaches deployed for the tasks.

Keywords: Nepali language, Low-resource languages, Large Language Models (LLMs), NER, POS

1. Introduction

Large Language Models (LLMs), such as BERT (Bidirectional Encoder Representations from Transformers), have revolutionized NLP by developing the capability of understanding complex language patterns and excelling in tasks like language generation and sentiment analysis (Vaswani et al., 2017; Devlin et al., 2019). There is now a growing interest in leveraging these models to benefit low-resource and regional languages. By fine-tuning them for specific tasks like Named-Entity Recognition(NER) and Part-of-Speech(POS) tagging, these models can capture unique linguistic nuances and enhance performance, thereby supporting language preservation and facilitating communication within specific regions. Nepali, categorized as a low-resource language within the Indic language group, employs the Devanagari script, recognized as an abugida which consists of 33 consonants and 11 vowels, substantially increasing the complexity due to various character combinations and grammatical structures (Bal, 2004). This complexity poses a challenge for large language models to understand and adequately capture the representation of Nepali texts. Research has been conducted lately towards developing and fine-tuning the LLMs with an aim to enhance language understanding and analysis for low-resource languages like Nepali. However, this is far from adequate, with the major issue being the availability of data.

Despite some progress, working with low-resource languages and fine-tuning, LLMs still present challenges with a need to have the availability of quality linguistic resources and practical considerations like computing resources and infrastructures. This study evaluates the performance of eight LLMs: `multilingualBERT (mBERT)`, `Multilingual DistilBERT (mDistilBERT)`, `NepBERTa`, `NepaliBERT`, `RoBERTa`, `XLM`, `XLM-RoBERTa`, and `HindBERT-scratch` for NER and POS tagging in Nepali. Our objective is to conduct a thorough examination of these models and their capacity to capture linguistic nuances. To achieve this, we fine-tuned the models on the *ILPRL* and *EBIQUITY* datasets, assessing their proficiency in Nepali NER and POS tagging tasks. This research primarily seeks to explore the capabilities of large language models when applied to natural language processing tasks in regional or low-resource languages.

2. Related Works

Different methods like SVM (Bam and Shahi, 2014), HMM and Rule based (Dey et al., 2014) have been reported to have achieved an F1 score up to 92% for the NER task, while up to 96% F1 score has been achieved by using SVM (Maharjan et al., 2019). Similarly, (Singh et al., 2019) proposed a novel architecture called BiLSTM+CNN at the grapheme-level for the NER task. The results show that their novel neural-based model achieved a significant im-

LREC-COLING 2024, pages 6974–6979

provement, ranging from 33% to 50%, compared to a feature-based SVM model. Additionally, it outperformed existing neural-based models developed for languages other than Nepali, with up to a 10% improvement. After the pioneering research on Transformers by Vaswani et al. (2017), it was demonstrated that Transformers excel in language modelling by capturing contextual data, handling long-range dependencies, supporting transfer learning, enabling multilingual recognition, and capturing fine-grained tags and sequential dependencies.

Many versions of BERT (based on transformers) like mBERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2020), RoBERTa (Liu et al., 2019), and XLM(Cross Lingual Language Model) (Lample and Conneau, 2019) have significant influence on NLP with greater knowledge gathering capability from several languages (Wu and Dredze, 2019; Pires et al., 2019). Maskey et al. (2022) presented three different models, DistilBERT, DeBERTa, and XLM-RoBERTa for Nepali news text classification and achieved the highest accuracy of 88.93% where DeBERTa base performed best. In another study, Timilsina et al. (2022) developed NepBERTa, a BERT-based model trained on a monolingual Nepali corpus consisting of 0.8 billion words from 36 different Nepali news portal. NepBERTa outperformed other monolingual models such as NepaliBERT (Pudasaini et al., 2023) as well as multilingual models like mBERT (Devlin et al., 2019) and XLM-R base (lexis Conneau et al., 2020), proving the fact that complexity of low-resource and morphologically rich languages like Nepali can be significantly reduced by using these multi-lingual large language models. Similarly, Niraula and Chapagain (2022) explored different NER systems, including a rule-based baseline, BERT-based model (BERT-bbmu), and BLSTM-CRF models, for Nepali language NER with BERT-bbmu demonstrating the highest F1 score of 0.85, indicating its effectiveness.

From our study of the existing literature, we have found that a few LLMs have been employed for experimentation with Nepali languages on various tasks. Although the aforementioned studies show outstanding results, they don't necessarily incorporate a large number of LLMs in one place. Additionally, different models explored in these works utilize different datasets which makes it difficult to carry out a direct comparison between them.

3. Methodology

3.1. Data Collection

The *ILPRL* and the *EBIQUITY* are two different datasets used for Named Entity Recognition (NER) and Part-Of-Speech (POS) tagging. Both datasets provide labelled data for NER and POS tasks, but they differ in terms of the number of words/sentences and the tagging schemes they follow.

3.1.1. ILPRL Dataset

This dataset contains a sample size of over 11,000 words that have been manually annotated for NER and POS tagging. The annotations in the *ILPRL* dataset[1] follow the CoNLL-2003 IOB (Inside, Outside, Beginning) tagging format. The IOB format is commonly used for labelling consecutive entities in a text. The dataset includes a collection of 11 tags for NER and 56 tags for POS.

3.1.2. EBIQUITY Dataset

This dataset consists of 3,606 sentences with over 93,000 words. The annotations in the *EBIQUITY* dataset[2] follow the CoNLL formatted BIO tagging scheme, which is another commonly used format (Singh et al., 2019). The dataset includes a collection of 7 tags for NER and 68 tags for POS.

3.2. Dataset Preprocessing

Raw datasets require preprocessing prior to fitting them into the model due to compatibility issues. This involves developing a dataframe with columns for sentence_id, words, and labels from the original text. We employed the scikit-learn module, LabelEncoder, to encode sentence_id. This encoding assigns a unique identifier to each sentence. All the words within the same sentence share the same identifier. The labels column should contain NER or POS tags, depending on the task at hand.

3.3. Model Implementation and Fine-tuning

To implement our work, we trained the models using the simple transformers library[3] which offers a wide range of services for NLP tasks, including Text & Token Classification, Question Answering, Language Modeling & Generation, to name a few. With the support for various BERT models, it was well-suited for our specific use case. Our main focus was evaluating the performance of LLMs on NER and POS tagging tasks for the Nepali language by fine-tuning eight LLMs from HuggingFace: `multilingualBERT (mBERT)`, `Multilingual DistilBERT (mDistilBERT)`, `NepBERTa`, `NepaliBERT`, `RoBERTa`, `XLM`, `XLM-RoBERTa`, and `HindBERT-scratch`. All of these

[1] https://ilprl.ku.edu.np/

[2] https://github.com/oya163/nepali-ner/blob/master/data/ebiquity/stemmed/

[3] https://simpletransformers.ai/docs/installation/

models employ a self-supervised masked language modelling technique. The dataset was preprocessed, split into train and test sets, and used for training and evaluating the models individually. With the exception of RoBERTa, HindBERT-scratch, NepaliBERT, and NepBERTa, all other models underwent pre-training in multiple languages. RoBERTa focused primarily on English, while NepaliBERT and NepBERTa exclusively trained in Nepali. HindBERT-scratch was specifically trained in Hindi (Joshi, 2023). We fine-tuned these models for our tasks, without training from scratch, while optimizing parameters like batch size, learning rates, and epochs within our scope of study.

3.4. Performance Metrics

F1-score (a combined measure of precision and recall) was used to evaluate the model's ability to accurately capture named entities and assign POS tags. The emphasis was on the F1 score to assess the models' effectiveness in achieving precise NER and POS tagging outcomes.

In addition to F1-scores, confusion matrices were generated to delve deeper into the models' performance. Each confusion matrix provides a granular view of the models' strengths and weaknesses, facilitating a comprehensive analysis of their performance in NER and POS tagging tasks.

4. Experimentation

The experiments were conducted on Google Colab, with varying parameters like learning rate, batch size, and epochs to evaluate the models as shown in Table 1. The experiment considered a total

Parameter	Values
Batch size	4, 8, 16, 32
Learning rate	1e-3, 1e-4, 1e-5, 1e-6
Optimizer	AdamW
Weight decay	0.01
Epochs	5 to 20

Table 1: Fine-tuning parameters

of four different learning rates, four batch sizes, and a maximum of 20 epochs. For optimization, AdamW (Loshchilov and Hutter, 2019), a specialized variant of the Adam optimizer designed for training deep learning models with a weight decay of 0.01, was employed. This weight decay value, set at 0.01, is a common default across many deep-learning libraries. It mitigates overfitting by introducing a regularization term into the loss function, encouraging the model to maintain smaller weights. Initially, the datasets were divided into an 8:2 ratio for training and testing purposes. Within the training data, the simple transformers library automatically performed an additional internal split, following an 8:2 ratio, for training and validation. This internal split was implemented to monitor the model's performance during training and to identify potential overfitting issues. Subsequently, each model was trained using all of these specified parameters in order to determine the most optimal settings. Table 2 provides insight into the word counts within both the training and test sets. A learning rate of 1e-4

Dataset	Train & Valid	Test	Total
ILPRL	11067	2767	13834
EBIQUITY	75166	18792	93958

Table 2: Dataset distribution

yielded the best results in our experiments. A batch size of 8 showed the best average performance, while a batch size of 32 performed poorly. The models were trained for 20 epochs on the *ILPRL* dataset and 10 epochs on the *EBIQUITY* dataset as there were no improvements in performance beyond that mark. Training the model for 5 epochs on average yielded the best performance. The fine-tuning process took approximately 23 hours of GPU time. You can access the code implementation in our github repository[4].

5. Results and Discussion

Table 3 shows the F1-scores for the eight LLMs in Nepali NER and POS tagging on *ILPRL* and *EBIQUITY* datasets. The first four LLMs are trained in multiple languages including Nepali and all of them show promising results. However, XLM-RoBERTa and mBERT were slightly better compared to XLM and mDistilBERT. mDistilBERT which is a distilled or reduced version of mBERT focused mainly on efficiency and faster inference with reduced architecture, which can be the potential reason for its lower score. XLM-RoBERTa excelled with F1 scores of 0.69-0.75 for NER and 0.87-0.97 for POS tagging on *ILPRL* and *EBIQUITY* datasets respectively, leveraging from its self-supervised masked language modeling (MLM) technique that includes dynamic masking with additional cross-lingual objectives.

On the other hand, it is apparent from the results that NepBERTa, primarily trained in Nepali outperforms all other models with exceptional F1-scores of 0.76 for NER and 0.90 for POS on *ILPRL*, and 0.79 for NER and 0.97 for POS on *EBIQUITY*, demonstrating its efficacy in capturing intricate language nuances and contextual information. Surprisingly,

[4] https://github.com/bipeshrajsubedi/LLMs-for-NER-and-POS-Tagging-in-Nepali-context

Model	ILPRL		EBIQUITY	
	NER	POS	NER	POS
mBERT	0.67	0.86	0.73	0.96
mDistilBERT	0.61	0.85	0.72	0.96
XLM	0.63	0.86	0.72	0.96
XLM-RoBERTa	0.69	0.87	0.75	0.97
NepBERTa	**0.76**	**0.90**	**0.79**	**0.97**
NepaliBERT	0.52	0.77	0.70	0.96
HindBERT-scratch	0.65	0.82	0.74	0.97
RoBERTa	0.32	0.69	0.52	0.86

Table 3: Performance of the LLM models from our experiment

	Model/Methods	Task	F1
Singh et al. (2019)	BiLSTM+CNN(G)	NER	0.86
Niraula and Chapagain (2022)	BERT-bbmu	NER	0.85
Timilsina et al. (2022)	NepBERTa	NER	0.91
Timilsina et al. (2022)	NepBERTa	POS	0.95
Maharjan et al. (2019)	SVM	NER	0.96

Table 4: Performance of the existing works

despite being trained on a substantial Nepali corpus, NepaliBERT falls short of expectations, delivering lower F1-scores on both *ILPRL* and *EBIQUITY* datasets. It may be attributed to the fact that NepBERTa was trained on a relatively large dataset compared to NepaliBERT. Additionally, the bidirectional nature of BERT-based models, incorporating information from both sides of a sentence, likely contributes to their superior performance in tasks such as NER and POS, which heavily rely on bidirectional context. Conversely, masked language models (MLMs) like NepaliBERT might find suitability in applications where contextual knowledge from preceding words and generalizability play a crucial role.

Similarly, HindBERT-scratch performed relatively well on Nepali NER and POS tagging tasks although it was not able to outperform NepBERTa. The former performed at par with different multilingual LLMs which indicates its potential for executing tasks in the Nepali language. The possible reasons behind this could be that both Hindi and Nepali besides sharing the same Devanagari script also commonly share a large chunk of the technical vocabulary. Moreover, both Hindi and Nepali are free word order languages and follow a Subject Object Verb (S-O-V) structure in terms of grammatical structure at the sentence level. However, RoBERTa, primarily trained in English faces significant challenges when applied to Nepali. The language mismatch between training and evaluation data likely contributes to these suboptimal results.

Notably, the POS scores tend to be higher compared to the NER scores, mainly due to inconsistencies in the quantity of labelled tagged data for NER, whereas POS data exhibits more consistent labelling. Moreover, the superior scores achieved on the *EBIQUITY* dataset can be attributed to its significantly larger data size, surpassing the *ILPRL* dataset by over five times.

The analysis of NepBERTa, the best-performing model reveals distinct patterns in classification accuracy. In the NER task, the model performs better on the EBIQUITY test set compared to the ILPRL test set. The model has higher correct classification counts, notably for tags like B-PER, B-LOC, and B-ORG, although some tags exhibit significant misclassifications. Conversely, on the ILPRL test set the model struggles with classifications, particularly for tags like B-ORG and B-PER, and shows potential issues for I-LOC and I-ORG tags. However, the model is effective at classifying non-entity instances ('O' tag) for both datasets, with greater performance on the EBIQUITY test set among the two. In the POS task, the model struggles with classes like NN and JX experiencing significant misclassifications, while others like DDM, DJX, RD, RK, TT, YF, and YM exhibit strong classification performance. Similarly, the ILPRL test set faces misclassification challenges, especially with ADJ and PP classes, although certain classes like DUM, YB, YF, YM, and YQ consistently perform well. NN and JJ were also classified relatively well, however there were some instances where they were misclassified. These findings highlight the need for targeted improvements in classification accuracy, particularly for highly misclassified classes, across both datasets.

The misclassifications observed in both POS and NER tasks are likely influenced by the overall imbalance in the distribution of tags in the dataset. Some tags appear much more frequently than others, causing the model to prioritize learning these common tags over the less common ones. This imbalance leads to biased predictions, where the model struggles to accurately classify the less frequent tags due to limited exposure during training. Additionally, classes with very few instances in the test data may not provide enough information for the model to learn accurate representations, leading to inappropriate classification. These issues underscore the importance of addressing class imbalance, ensuring adequate representation for all tags, and developing strategies to handle complex patterns effectively in order to enhance classification accuracy. The analysis is based on

insights obtained from confusion matrices available in the GitHub repository associated with this project. These matrices offer a detailed understanding of classification performance in both NER and POS tasks for ILPRL and EBIQUITY datasets. Due to their large size, they are not included in the paper.

Table 4 presents the performance of some of the existing works and it is evident that the NER F1 scores in our research are lower than those reported in previous studies. A key reason for this difference lies in the language-specific challenges posed by the Nepali language, including the diversity of named entities and the scarcity of labelled data. While our study evaluates the performance of eight different LLMs, it suggests that LLMs don't consistently outperform shallow learning algorithms like SVM or hybrid models like BiLSTM+CNN in low-resource languages like Nepali. However, the potential of LLMs is underscored by the work of Timilsina et al. (2022), achieving high NER and POS scores, though not surpassing Maharjan et al. (2019) results. Our research confirms `NepBERTa`'s exceptional performance, specifically designed for Nepali, outperforming other LLMs.

6. Conclusion and Future Works

In this paper, we have fine-tuned and evaluated the large language models, particularly BERT variants for NER and POS tagging using the *ILPRL* and *EBIQUITY* datasets. The results suggested that NepBERTa performed best whereas RoBERTa performed the least among all the models tested. Our experiments also validate NepBERTa's effectiveness in achieving superior NER and POS results for Nepali, but other approaches remain equally capable. From this study, it can be suggested that large language multilingual models show promising results for NER and POS tagging in low-resource languages like Nepali especially if they are trained on Nepali corpus. The results indicate improved performance in language processing tasks for such models, highlighting their potential. Additionally, alternative approaches like BiLSTM+CNN, SVM, and rule-based methods prove to be highly effective and should not be underestimated in their capacity to perform such tasks efficiently. However, it is also important to note the limitations of this study. For future enhancements, we recommended incorporating larger and more diverse datasets, optimising hyperparameters, exploring different large language models, investigating transfer learning techniques, evaluating domain-specific texts, and extending the research to other low-resource languages.

7. Limitations

There are certain limitations that might affect the validity and applicability of the findings. One such limitation is the limited size of the datasets used, which could affect the models' performance, resulting in incomplete representations and difficulties in handling diverse inputs. Additionally, due to resource constraints, the fine-tuning process was constrained, potentially limiting the models' learning capacity. Although the investigation focused on popular BERT variants, the inclusion of other large language models could have provided further insights. Furthermore, we did not attempt to replicate the results from existing studies, which could have facilitated a more equitable and precise comparison. It is important to note that the models may not perform uniformly across all input types, particularly when encountering new or domain-specific texts. Hence, when interpreting the results for real-world applications, these limitations should be considered.

8. Acknowledgements

We would like to thank the Information and Language Processing Research Lab (ILPRL) and Singh et al. (2019) for providing datasets used in this work. We would also like to thank the anonymous reviewers for their feedback and comments. Mr. Praveen would like to acknowledge the financial support of the Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223.

9. Ethics Statement

In this study, we utilized datasets acquired from both internet sources and laboratory data. We want to emphasize that all the datasets used were either publicly available or obtained with proper permissions and licenses. Throughout the research process, we adhered to relevant guidelines and regulations, ensuring transparency and integrity in our approach. We recognize our responsibility to handle and use the collected data ethically, respecting the rights and interests of the data sources.

10. References

Bal Krishna Bal. 2004. Structure of Nepali Grammar. *PAN Localization, Madan Puraskar Pustakalaya*, pages 332–396.

Surya Bahadur Bam and Tej Bahadur Shahi. 2014. Named Entity Recognition for Nepali Text Using

Support Vector Machines. *Intelligent Information Management*, 06(02):21–29.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*.

Arindam Dey, Abhijit Paul, and Syam Purkayastha. 2014. Named Entity Recognition for Nepali language: A Semi Hybrid Approach. *International Journal of Engineering and Innovative Technology (IJEIT)*, 03(08):21–29.

Raviraj Joshi. 2023. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual Language Model Pretraining. *arXiv*.

lexis 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. *arXiv*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Gopal Maharjan, Bal Krishna Bal, and Santosh Regmi. 2019. Named Entity Recognition (NER) for Nepali. In *Creativity in Intelligent Technologies and Data Science: Third Conference, CIT&DS 2019, Volgograd, Russia, September 16–19, 2019, Proceedings, Part II 3*, pages 71–80. Springer.

Utsav Maskey, Manish Bhatta, Shiva Raj Bhatta, Sanket Dhungel, and Bal Krishna Bal. 2022. Nepali Encoder Transformers: An Analysis of Auto Encoding Transformer Language Models for Nepali Text Classification. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 106–111.

Nobal Niraula and Jeevan Chapagain. 2022. Named entity recognition for nepali: Data sets and algorithms. In *The International FLAIRS Conference Proceedings*, volume 35.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4969–5001. Association for Computational Linguistics.

Shushanta Pudasaini, Subarna Shakya, Aakash Tamang, Sajjan Adhikari, Sunil Thapa, and Sagar Lamichhane. 2023. Nepalibert: Pre-training of masked language model in nepali corpus. In *2023 7th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pages 325–330.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv*.

Oyesh Mann Singh, Ankur Padia, and Anupam Joshi. 2019. Named Entity Recognition for Nepali Language. In *2019 IEEE 5th International Conference on Collaboration and Internet Computing (CIC)*, pages 184–190. Springer International Publishing.

Sulav Timilsina, Milan Gautam, and Binod Bhattarai. 2022. NepBERTa: Nepali Language Model Trained in a Large Corpus. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 273–284.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *CoRR*.

Shijie Wu and Mark Dredze. 2019. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. 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 833–844. Association for Computational Linguistics.