Kathmandu University Journal of Science, Engineering and Technology, Vol. 18, No. 2, December 2024

Kathmandu University
Journal of Science, Engineering and Technology

Evaluating Sentence Embedding Models for Nepali Sentiment Analysis: A Comparative Study

Abiral Adhikari[a], Samir Wagle[a], Reewaj Khanal[a], and Prashant Manandhar[a]

[a]Department of Computer Science and Engineering, Kathmandu University, Dhulikhel, Nepal

Abstract

Sentiment analysis for morphologically complex, low-resource languages like Nepali remains a developing field, where progress has been largely constrained by a reliance on traditional feature engineering and first-generation neural embeddings. This study confronts this methodological gap through a rigorous comparative analysis designed to decouple the influence of modern embedding representations from downstream architectural complexity. We benchmark four state-of-the-art multilingual sentence embeddings (BGE-M3, LaBSE, mE5-base, and DistilUSE) across three neural architectures of increasing complexity: a Multi-Layer Perceptron (MLP), a Residual MLP, and a Transformer network. BGE-M3, when paired with a simple MLP, achieved a notable accuracy of 82.49%, decisively outperforming the more complex Transformer-based classifiers. This result demonstrates conclusively that for this low-resource paradigm, the semantic richness of the input embedding is the dominant determinant of performance, eclipsing the architectural inductive biases of the downstream model. Our work not only establishes a powerful and resource-efficient benchmark for Nepali NLP but also provides a crucial insight for sentiment analysis in other low-resource languages.

Keywords: Sentiment Analysis, Nepali Language, Deep Learning, Sentence Embeddings, Transformers, Low-Resource NLP.

1. Introduction

The rapid rise of blogs, social media, and online digital news sites has created an explosion of user-generated data in the Nepali language. Extracting sentiment from such texts is valuable in many ways: businesses can assess what their customers truly think, public entities can observe changing public sentiment, and social scientists can highlight other evolving social trends. Fostering Natural Language Processing (NLP) techniques for Nepali sentiment analysis is particularly challenging due to its morphological richness and designation as a low-resource computational language. Previous studies in this domain, such as those by [1], [2], [3], [4], and [5], have explored limited embedding methods, such as TF-IDF, word2vec, mBERT, and skip-gram. While these efforts have laid an important foundation, they have not fully explored the comparative effectiveness of more recent and sophisticated sentence embedding models such as LaBSE, E5_base, DistilUSE, BGE_m3, etc., for Nepali NLP tasks such as sentiment analysis.

For the advancement of sentiment analysis capabilities in the Nepali language, a reliable evaluation of the embedding models is essential for building an NLP application. Despite the recent advancement towards multilingual embedding, there is a significant research gap in analyzing modern embedding methods and assessing their effectiveness and suitability for Nepali sentiment classification tasks. Addressing this gap is not merely a technical aspiration but also a necessary step towards enriching socio-cultural research through data-driven insights.

This paper addresses the aforementioned urgent need by presenting preliminary results. It aims to systematically evaluate the embedding quality of four state-of-the-art multilingual sentence embeddings on Nepali text. The objective is to identify the most effective embedding model based on the sentiment classification performance. By providing a systematic evaluation with a high level of evidence base, this study paves the way for supporting future development in Nepali NLP research and application.

2. Literature Review

The study and research of sentiment in Nepali has come a long way, especially considering it's a low-resource language. Some of the earliest works, like those by [1], [2], [3], [4], and [5], laid important groundwork for low-resource sentiment analysis. Early research in Nepali sentiment analysis relied on traditional machine learning techniques such as TF-IDF features and early neural methods such as word2vec. Some studies also explored multilingual models like mBERT as a part of these early explorations, but most of these approaches focused on word or token-level representations, which often miss the full meaning, especially in a complex language like Nepali.

Since the release of Transformer architectures [6], natural language processing has progressed at an impressive pace. Unlike earlier models such as LSTM and GRU, which processed sentences step by step, Transformers introduced a self-attention mechanism that allows models to consider the entire input at once. These improvements were beneficial for low-resource languages, where large labeled datasets were unavailable. Models trained on high-resource languages can now be generalized to other models having fewer data through multilingual pre-training. Models like mBERT and XLM-RoBERTa [7], which were trained on more than 100 languages, can produce language-agnostic embeddings. Strong performances were seen in these embeddings in multilingual sentiment analysis tasks [8], [9].

Building on this cross-lingual approach, a powerful alternative of using these models to generate sentence-level embeddings has emerged. Sentence-BERT [10] was one of the first to adapt Transformers to produce fixed-size vectors representing the full meaning of a sentence. More recent models, such as LaBSE [11] and BGE-M3 [12], extend this idea and provide strong, multilingual sentence embeddings. However, their performance on Nepali text remains underexplored. This study aims to fill that gap by evaluating and comparing these modern embedding models across several classification architectures to identify the most effective representation for Nepali sentiment analysis.

3. Methodology

Our experimental methodology is designed to be comprehensive and reproducible, systematically evaluating each architecture and embedding combination.

3.1. Data Collection

To construct a diverse dataset corpus for this study, the data were aggregated from three distinct public sources to include a representation of different domains. It includes general sentiment with more nuanced topics, a collection of tweets about national elections, and tweets related to COVID-19. The collection from these varied sources provides a comprehensive dataset that captures a wide range of vocabulary, slang, and contextual sentiment expressions presented in Nepali social media, which is ideal for sentiment analysis research. A total of around 150,000 tweets and social media comments were selected for further preprocessing.

Table 1: Raw vs. Preprocessed Comments

Raw Comment	Preprocessed Final Comment
#TheStruggleOfSaintRampalJi सत्यका लागि संघर्ष सन्त रामपालजी महाराजले विश्व कल्याणका लागि जागिरबाट राजीनामा दिएर आफ्नो सम्पूर्ण परिवारलाई भगवानको विश्वासमा छाडेर अनि अनेक अन्याय, अत्याचार सहँदै सद्ग्रन्थमा रहेको सच्चा ज्ञान जन जनसम्म पुऱ्याउनुभयो @SatlokAshramSojat https://t.co/uiELmZjNQB	सत्यका लागि संघर्ष सन्त रामपालजी महाराजले विश्व कल्याणका लागि जागिरबाट राजीनामा दिएर आफ्नो सम्पूर्ण परिवारलाई भगवानको विश्वासमा छाडेर अनि अनेक अन्याय अत्याचार सहँदै सद्ग्रन्थमा रहेको सच्चा ज्ञान जन जनसम्म पुर्याउनुभयो
@GokulPBaskota कमरेड ले जीत्छन हौ https://t.co/GMcjBF6nJk	कमरेड ले जीत्छन हौ
#पैसा_छ_भने पसलमा देवता किन्न पाईन्छ। अदालतमा न्याय किन्न पाईन्छ। नपढी विद्वान बन्न पाईन्छ। बुढेसकालमा जवानी किन्न पाईन्छ। चुनावमा टिकट किन्न पाईन्छ। मासुभातमा भोट किन्न पाईन्छ। नागरिकता किन्न पाईन्छ। #नोट_अरूपनि_धेरै_कुराहरू_पाईन्छ। #शुभबिहानी__शुभदिन	देवता किन्न पाईन्छ। अदालतमा न्याय किन्न पाईन्छ। नपढी विद्वान बन्न पाईन्छ। बुढेसकालमा जवानी किन्न पाईन्छ। चुनावमा टिकट किन्न पाईन्छ। मासुभातमा भोट किन्न पाईन्छ। नागरिकता किन्न पाईन्छ।

3.2. Data Preprocessing and Annotation

The final dataset, collected from different sources as mentioned in Section 3.1, was subjected to a preprocessing pipeline. It begins with the removal of platform-specific noise such as user mentions, hashtags, emojis, and URLs. After that, we removed all English characters and numbers, and a Unicode-based filter was applied to create a Devanagari-only corpus. We removed extra whitespaces to make the text more immaculate and consistent before tokenization. The full sentence structure was preserved by skipping stop-word removal, which is important for accurately capturing subtle sentiments. The rich morphological nature of the Nepali language motivated us to perform a final lemmatization step for reducing all words to their base dictionary form. The primary methodological challenge after preprocessing was the annotation of the large volume of unlabeled text aggregated from the topical datasets. To address this issue, we used a semi-supervised approach, "LLM-as-a-Judge", that automatically evaluates and labels the data as positive, negative, and neutral. This approach has demonstrated high-quality sentiment annotations capability comparable to human labels, in a study conducted by [13]. For our experiments, we selected the OpenAI o3 model for its strong multilingual performance. The sanity of the data annotation methodology implemented was validated by manual inspection of randomly selected data samples. The result indicated high accuracy in identifying negative instances; however, the model struggled to distinguish between neutral and positive labels.

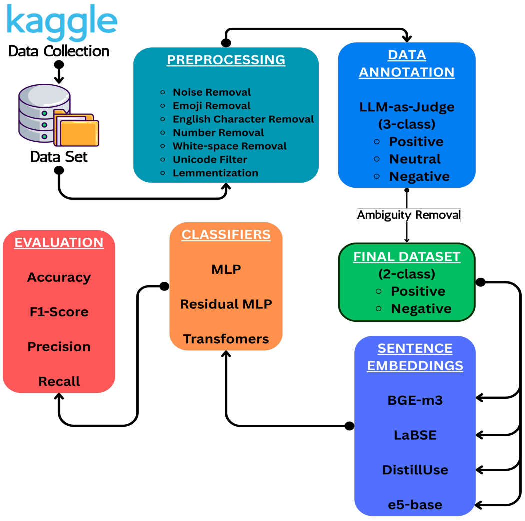

Figure 1: Overview of the experimental pipeline.

3.3. Data Analysis and Description

The final dataset after preprocessing and annotation comprised 101,996 social media comments, each labeled with 'Positive', 'Neutral', and 'Negative'. The initial finding of LLM-as-a-Judge's approach, struggling to segregate the ambiguous positive and neutral labels, motivated us to methodologically exclude the 'Neutral' class from the final dataset. This decision aimed to create a more distinct and robust binary classification task centered on the accurate assessment of the discriminative embedding performance of the sentence embedding models, with clear polarity. The bias due to class imbalance was mitigated by downsampling the dataset to 34,047 comments per class.

3.4. Text Embedding Models

A fixed-size vector representation of each Nepali text was obtained by employing pretrained sentence-level embedding models with the multilingual capacity to generate sentence embeddings, either by sentence transformers or by average pooling over token-level embeddings. The generated sentence semantic embeddings captured the semantic meaning of the entire sentence rather than

Table 2: Sentiment Classification Performance Across Different Embeddings and Model Architectures

Model architecture	Embedding	Accuracy (%)	Precision (Pos/Neg)	Recall (Pos/Neg)	F1-Score (Pos/Neg)
MLP	BGE-M3	82.49	0.81 / 0.84	0.85 / 0.80	0.83 / 0.82
	LaBSE	80.86	0.84 / 0.78	0.76 / 0.85	0.80 / 0.82
	mE5-base	80.26	0.83 / 0.78	0.76 / 0.85	0.79 / 0.81
	DistilUSE	70.83	0.73 / 0.69	0.67 / 0.75	0.70 / 0.72
Residual MLP	BGE-M3	82.13	0.82 / 0.82	0.82 / 0.82	0.82 / 0.82
	LaBSE	80.63	0.81 / 0.80	0.80 / 0.82	0.80 / 0.81
	mE5-base	80.31	0.82 / 0.79	0.78 / 0.83	0.80 / 0.81
	DistilUSE	70.75	0.72 / 0.70	0.69 / 0.73	0.70 / 0.71
Transformer	BGE-M3	81.25	0.84 / 0.79	0.77 / 0.85	0.80 / 0.82
	mE5-base	79.51	0.85 / 0.75	0.71 / 0.88	0.78 / 0.81
	LaBSE	79.73	0.80 / 0.79	0.79 / 0.81	0.80 / 0.80
	DistilUSE	71.72	0.73 / 0.71	0.69 / 0.75	0.71 / 0.73

individual words, which is crucial for languages with high morphological variance like Nepali. For our experiments, four distinct embedding models based on their demonstrated strong multilingual performance in previous studies were selected. The selection included:

1. **BGE-M3**, an extremely powerful multilingual model that supports over 100 languages and outperforms the state-of-the-art on cross-lingual benchmarks [12].

2. **mE5-base**, a great generalizer that has been adapted for multilingual purposes [14];

3. **LaBSE**, a model that creates language-agnostic sentence embeddings across 109 languages, and is well-suited to our task [11].

4. **DistilUSE**, a multilingual model designed to be efficient in computation by using knowledge distillation to maintain the performance of larger models, using the framework from [10].

The off-the-self models were implemented through Hugging Face[1] APIs without further fine-tuning to establish the baseline performance in Nepali language, serving as a benchmark for future works. The model embeddings serve as the input features for downstream classification tasks in our sentiment analysis pipeline.

3.5. Model Architectures

In this study, we evaluated the quality of fixed-size sentence embeddings generated from each embedding model using three distinct neural architectures. This choice was made to comprehensively evaluate the embedding quality for the classification task across different complexities. The selection was done to establish a strong baseline while ensuring appropriate architectural choice for the input data included:

1. A simple feed-forward Multi-Layer Perceptron (MLP) network was selected for its strong baseline performance in text classification tasks [15], which minimizes architectural biases.

2. A modified feed-forward Residual MLP (ResMLP) network, including skip connections, was selected for its ability to capture complex patterns within the embedding space. It also helps to address the vanishing gradient problem, as demonstrated in this study [16].

3. A simplified Transformer-based architecture was selected for downstream classification to maintain consistency with the Transformer-based embedding model (e.g., LaBSE). Although the attention component does not provide its usual sequence-level benefits, the design aligns with the Transformer formulation [6] and supports more expressive representations in a feedforward-like manner.

The sequential model variants, such as LSTM, GRU, RNN, etc, were consciously and methodologically excluded. The usage of a single fixed-sized sentence vector embedding as input for each data point makes it non-sequential. The sequential models are unable to leverage such non-sequential representation, making them ineffective.

3.6. Experimental Setup and Evaluation

The training was conducted up to a maximum of 30 epochs with a batch size of 64 using Adam optimizer and binary cross-entropy loss at a train-test split of 8:2. Additional strategies of EarlyStopping(patience=5) and ReduceLROnPlateau(patience=3) were implemented to prevent overfitting. The performance measurement on the test set was conducted with: Accuracy, Precision, Recall, and F1 score.

4. Result and Experimental Analysis

The results of experiments presented in Table 2 compare the performance of implemented embeddings across selected architectures on the test set, based on evaluation metrics outlined in section 3.6.

The performance ranking of the embedding models was consistent as follows: BGE-M3 $>$ LaBSE $\approx$ mE5-base $>$ DistilUSE. BGE-M3, which consistently outperformed other embedding models across all architectures for all evaluation metrics, achieving accuracy up to 82.49% with MLP. While LaBSE and mE5-base performed slightly worse than BGE-M3, DistilUSE ranks lowest in every metric across every architecture, underperforming by 10% relative to BGE-M3, achieving a maximum accuracy of 71.72% with Transformers. With accuracy fluctuating by less than 1.3% for the same embedding model across different architectures, the suggested architecture choice had minimal impact on the performance. This highlights that a relatively simple architecture can achieve state-of-the-art performance with high-quality input features. These minor fluctuations in performance indicated that embedding quality is the determining factor for classification performance, rather than ar-

[1] https://huggingface.co/

chitectural choice, when using a single fixed-size sentence vector embedding for input. Even advanced architectures, such as Transformers, were unable to leverage the attention mechanism available for token-level embedding sequences to improve performance.

The reasons behind DistilUSE's underperformance, as with our results, were discovered by [10]. First, the model lacked sufficient training on low-resource datasets, and second, it had not undergone task-specific fine-tuning. On the other hand, the study by [12] demonstrated that BGE-M3 achieved better results on sentiment classification tasks by training on large-scale multilingual corpora with contrastive learning objectives, which enhanced its ability to capture distinct semantics. During the experiment, the increase in embedding dimensions of models was observed to have improved classification performance. The BGE-M3 model with an embedding dimension of 1024 performs significantly better than DistilUSE with an embedding dimension of 512, as it's larger dimensionality enabled it to better capture nuanced sematic relationships. While both LaBSE and mE5-base with the same embedding dimension of 768 offered similar performance.

5. Conclusion

In this paper, a comparative analysis of the performance of four distinct multilingual embedding models for the Nepali sentiment analysis task was conducted. The preliminary findings of this research highlight the BGE-M3 model as most effective among the selected models, including LaBSE, mE5-base, and DistilUSE, achieving a remarkable accuracy of 82.49% on the test set with an MLP architecture. The initial experiments with fixed-sized sentence embedding vector input highlighted that even a simple feed-forward MLP network can yield respectable performance against more complex Transformers architecture, when the latter is unable to capitalize its attention mechanism. This underscores the importance of robust embedding model selection over architectural complexity.

The embedding approach evaluated during this study offers valuable insights to future researchers on embedding model selection for Nepali sentiment classification and similar tasks. Motivated by our preliminary findings, we aim to fine-tune the BGE-M3 model on a large, domain-specific Nepali dataset to further enhance its contextual understanding. Additionally, we plan to expand the task to a multi-label classification by incorporating a 'neutral' sentiment class with the establishment of proper annotation guidelines. We also intend to conduct a comparative analysis of sentence versus word embedding performance on an extended dataset, including formal and diverse domain Nepali texts beyond social media. Additional embedding models will be incorporated in the experiments, to observe the performance difference in low-resource settings of the Nepali language.

References

[1] Tamrakar S, Bal B K & Thapa R, Aspect based sentiment analysis of nepali text using support vector machine and naive bayes, *Technical Journal*, 2(1). https://doi.org/10.3126/tj.v2i1.32824. URL https://www.nepjol.info/index.php/TJ/article/view/32824.

[2] Pant P & Shakya S, Aspect based nepali text sentiment analysis using bilstm, *Journal of Ubiquitous Computing and Communication Technologies*, 4 (2022) 219--235. https://doi.org/10.36548/jucct.2022.4.001. URL https://irojournals.com/jucct/article/view/4/4/1.

[3] Tharu M, Pokhrel S & Lamichhane B, Sentiment analysis of nepali covid-19 tweets using bert-lstm, *Journal of Engineering and Sciences*, 2 (2023) 49--56. https://doi.org/10.3126/jes2.v2i1.60393. URL https://www.nepjol.info/index.php/jes2/article/view/60393.

[4] Shahi T, Sitaula C & Paudel N, A hybrid feature extraction method for nepali covid-19-related tweets classification, *Computational Intelligence and Neuroscience*, 2022. https://doi.org/10.1155/2022/5681574. URL https://www.hindawi.com/journals/cin/2022/5681574.

[5] Bade Shrestha B & Bal B K. Named-entity based sentiment analysis of Nepali news media texts. In: YANG E, XUN E, ZHANG B & RAO G, eds., *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications*. Association for Computational Linguistics, Suzhou, China (2020), pp. 114--120. https://doi.org/10.18653/v1/2020.nlptea-1.16. URL https://aclanthology.org/2020.nlptea-1.16/.

[6] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L & Polosukhin I. Attention is all you need (2023). URL https://arxiv.org/abs/1706.03762.

[7] Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L & Stoyanov V. Unsupervised cross-lingual representation learning at scale. In: Jurafsky D, Chai J, Schluter N & Tetreault J, eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online (2020), pp. 8440--8451. https://doi.org/10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747/.

[8] Pires T, Schlinger E & Garrette D. How multilingual is multilingual BERT? In: Korhonen A, Traum D & Màrquez L, eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy (2019), pp. 4996--5001. https://doi.org/10.18653/v1/P19-1493. URL https://aclanthology.org/P19-1493/.

[9] Nazir M K, Faisal C N, Habib M A & Ahmad H, Leveraging multilingual transformer for multiclass sentiment analysis in code-mixed data of low-resource languages, *IEEE Access*, 13 (2025) 7538--7554. ISSN 2169-3536. https://doi.org/10.1109/ACCESS.2025.3527710. URL https://ieeexplore.ieee.org/document/10835765.

[10] Reimers N & Gurevych I. Making monolingual sentence embeddings multilingual using knowledge distillation (2020). URL https://arxiv.org/abs/2004.09813.

[11] Feng F, Yang Y, Cer D, Arivazhagan N & Wang W. Language-agnostic BERT sentence embedding. In: Muresan S, Nakov P & Villavicencio A, eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland (2022), pp. 878--891. https://doi.org/10.18653/v1/2022.acl-long.62. URL https://aclanthology.org/2022.acl-long.62/.

[12] Chen J, Xiao S, Zhang P, Luo K, Lian D & Liu Z. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation (2024). URL https://arxiv.org/abs/2402.03216.

[13] Gu J, Jiang X, Shi Z, Tan H, Zhai X, Xu C, Li W, Shen Y, Ma S, Liu H, Wang S, Zhang K, Wang Y, Gao W, Ni L & Guo J. A survey on

llm-as-a-judge (2025). URL https://arxiv.org/abs/2411.15594.

[14] Wang L, Yang N, Huang X, Jiao B, Yang L, Jiang D, Majumder R & Wei F. Text embeddings by weakly-supervised contrastive pre-training (2024). URL https://arxiv.org/abs/2212.03533.

[15] Joulin A, Grave E, Bojanowski P & Mikolov T. Bag of tricks for efficient text classification (2016). URL https://arxiv.org/abs/1607.01759.

[16] Touvron H, Bojanowski P, Caron M, Cord M, El-Nouby A, Grave E, Izacard G, Joulin A, Synnaeve G, Verbeek J & Jégou H. Resmlp: Feedforward networks for image classification with data-efficient training (2021). URL https://arxiv.org/abs/2105.03404.