Advancing Bengali Sentiment Analysis: A Benchmark Study with ELMo and XLNet Architecture

Prithwiraj Bhattacharjee
Department of CSE
Leading University, Sylhet

ABSTRACT

Even though Bengali is the sixth most spoken language in the world, many of its natural language processing domains remain underexplored, with sentiment analysis being one of the core areas. This study applies ELMo (Embeddings from Language Models) and XLNet (Transformer-XL) for the first time in Bengali sentiment classification. A dataset of over 40,000 user comments was used, which was collected from YouTube online platforms and used to train both models on binary classification with positive and negative labels and ternary classification with positive, neutral, and negative labels. Experimental results show that the ELMo two-class model achieved the highest accuracy of 71%, while the XLNet two-class model reached 56%. These findings highlight the potential of context-rich representations like ELMo and XLNet for Bengali sentiment analysis, while also revealing the challenges of more nuanced ternary classification. Overall, the research provides new insights into leveraging cutting-edge deep learning models for limited-resource languages such as Bengali.

General Terms

Computational Linguistics, Sentiment Analysis, Bangla Language.

Keywords

YouTube Comments, ELMo, XLNet, Deep Learning.

1. INTRODUCTION

Sentiment analysis, a significant part of NLP, aims to discern and categorize the underlying emotions in textual data. By analyzing the composition and interplay of words within sentences, sentiment analysis aims to understand the emotional undertones of texts, assigning those positions on a range of sentiments, from very negative to very positive. This range is critical for automating text interpretation and simplifying decision-making across various sectors. The evolution of internet communication has exponentially increased the volume of text data, necessitating the development of automated tools for efficient and accurate sentiment analysis. Despite its global relevance, the sentiment analysis application for Bengali, spoken by over 272 million people, still needs to be explored. This research addresses this gap by introducing a novel dataset derived from YouTube comments and employing two new models for Bengali sentiment analysis. These models are evaluated against a baseline to ascertain their effectiveness in identifying and categorizing sentiments within Bengali text, marking a significant stride toward advancing NLP capabilities for Bengali. The digital transformation of communication has led to a proliferation of online textual data, posing challenges in maintaining proper communication protocols and advancing linguistic research. This is particularly applicable for languages like Bengali, which, despite its vast number of speakers, lacks sufficient research in sentiment analysis. The primary challenge this research addresses is developing and applying sentiment analysis tools for Bengali, utilizing a dataset of over 40,000 online comments. This dataset includes diverse sentiments that reflect the complexity and richness of the Bengali language expression. The study aims to fill the research gap by categorizing these comments into three distinct sentiment categories: positive, neutral, and negative. The absence of sophisticated models tailored for Bengali sentiment analysis further motivates this study, which introduces and evaluates two new models designed to enhance our understanding and processing of Bengali textual sentiments. The purpose of this research is clear and aims to advance sentiment analysis for the Bengali language significantly by introducing and evaluating two novel models, marking the first application of these models to this domain. In the realm of computational linguistics, sentiment analysis represents a critical avenue for understanding subjective information within text data. This work embarks on an exploration within this domain, focusing on the Bengali language, a linguistic area that has seen limited exploration in sentiment analysis due to the scarcity of tailored resources and models. Recognizing this gap, our contributions to the field are as follows:

- For the first time, adaptation and implementation of XLNet and ELMo models for 2-class and 3-class Bengali sentiment analysis.
- Evaluation of the ELMo model outperforms Bengali state-of- the-art traditional ML approaches.

2. LITERATURE REVIEW

In computer linguistics, sentiment analysis becomes a very important aspect. It has seen significant advancements with the advent of deep learning techniques in recent years. While much of the research has focused on English, many areas of Bengali sentiment analysis are still underexplored. In 2020, Yadav et al. [1] conducted a comprehensive review of deep learning architectures for sentiment analysis. The authors highlighted a combination of two models that can capture sequential and spatial information in text data very nicely. The first one is recurrent neural networks (RNNs), and the second one is convolutional neural networks (CNNs). The study also emphasized the effectiveness of long short-term memory (LSTM) networks and gated recurrent units (GRUs) in modeling textual data for sentiment analysis. However, the authors noted that there is a lack of research specifically focusing on the application of these architectures to sentiment analysis in Bengali. Das et al. [2] present a comprehensive evaluation of machine learning and deep learning models for sentiment analysis in both English and Bangla languages. It focuses on reviews from the Bengali e-commerce platform, DARAZ. This research employs seven machine learning models. Those include advanced configurations like Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), and Convolutional 1D (Conv1D). Thus, the research systematically examines their effectiveness in text classification. The Support Vector Machine emerged as the most effective model. It achieves an accuracy of 82.56% for English data and 86.43% for Bangla texts. Among deep learning approaches, the Bi-LSTM model demonstrated superior performance. It achieved 78.10% accuracy for English and 83.72% for Bangla. Transfer learning and self-attention mechanisms have shown promise in enhancing the performance of sentiment analysis models for Bengali. A study by Khan et al. [3] implemented LSTM-based classifiers by leveraging transfer learning from Hindi to Bengali and introducing self-attention to better capture long-distance dependencies within texts. This approach underscores the utility of cross-lingual transfer learning and the effectiveness of self-attention mechanisms in improving model performance for under-resourced languages. The integration of convolutional neural networks (CNNs) with long short-term memory (LSTM) networks has been explored to address the challenges of Bengali sentiment analysis. Research presented by Karim et al. [4] employs a multichannel Convolutional-LSTM network. It aims to overcome the limitations of traditional machine learning methods. This architecture is designed to capture both local features through convolutional layers and long-term dependencies via LSTM while indicating a comprehensive approach to parsing Bengali texts. The study by Wahid et al. [5] explores the application of machine learning and deep learning models for sentiment analysis in cricket-related data. Khan et al. [6] focused on classifying sentiments from Bengali paragraphs as happy or sad using various machine learning algorithms. The Multinomial Naive Bayes provides the highest accuracy. This study underscores the importance of specific domain applications of sentiment analysis in Bengali. Salehin et al. [7] investigated the sentiment classification of Bengali Facebook posts using both ML and DL techniques. The study highlighted the superior performance of SVM for preprocessed corpora and the effectiveness of LSTM for unpreprocessed data. It offers insights into the application of different methodologies based on the nature of the data source. Islam et al. [8] investigated deep learning and BERT models for analyzing sentiments of Bengali social media posts. This research achieved an accuracy of 88.59% with a CNN-BiLSTM hybrid model using GloVe feature vectors. It underscores the effectiveness of combining deep learning architectures and advanced feature extraction techniques for sentiment analysis in Bengali. Deep learning-based models were proposed by Tripto et al. [9]. In this paper, Bengali sentences have been categorized into a five-class (strongly positive, positive, neutral, negative, and highly negative) and three-class (positive, negative, and neutral) sentiment label. They also developed a model that recognizes and categorizes the emotions in a Bengali sentence. Those emotions are happiness, surprise, anger, sadness, disgust, and fear. They collected the datasets from the most popular Bengali YouTube videos. They employed an LSTM with a DNN and CNN as its main layer for the opinion mining task. They also utilized SVM and Naive Bayes to identify sentiments and emotions. LSTM-based deep RNN architecture with a context encoder provides 85% accuracy in the election sentiment analysis from Bengali newspaper articles. This model clearly dominates other supervised classifiers like Naive Bayes, SVM, and Decision Tree [10]. Aspect-based sentiment analysis (ABSA) is another area of interest, where the sentiment is not only predicted for the entire sentence or document but also for specific aspects within the text. A manually annotated Bengali dataset, BAN-ABSA [11], has been introduced for this purpose. Then, deep learning models like CNN and Bi-LSTM have been evaluated on it. The latter one showed superior performance in terms of average F1-score. There is also a survey done by Wankhade et al. [12], which covers the evolution of sentiment analysis methodologies, from the inception of basic models to the integration of complex neural networks. Their comprehensive survey covers a wide array of sentiment analysis methods, applications, and challenges. It also provides a holistic view of the field's progression and the pivotal role of deep learning in advancing sentiment analysis technologies. Then, ELMo [13] and XLNet [14] were implemented for the English dataset from which the motivation has been taken for this study.

3. DATASET

For our experiments, we utilize a Bengali sentiment dataset introduced by Rahman et al. [15] (under review). The dataset originally contains five labels: Positive, Very Positive, Negative, Very Negative, and Neutral. The dataset comprises over 40,000 online comments collected from different public social media platforms. Basic preprocessing has been done for the removal of non-Bangla characters, normalization, and tokenization. To suit different classification tasks, the dataset has been reformatted as follows:

- Two-class classification: The original five labels were merged into two classes: Positive (including Positive and Very Positive) and Negative (including Negative and Very Negative). Neutral comments were excluded in this setting.
- Three-class classification: The labels were grouped into ”Positive,” ”Negative,” and ”Neutral.” ”Positive” label includes all the Positive and Very Positive entries, whereas the ”Negative” label includes Negative and Very Negative ones. Finally, Neutral corresponds to the original Neutral label.

The class distributions for both tasks are summarized in Table 1. This dataset has been used for evaluating the performance of deep learning models for Bangla sentiment classification. It is publicly available upon request and is concurrently submitted for publication (Rahman et al. [15], under review).

Table 1. Dataset Description for Both Classification

Class	#Sample (2-class)	#Sample (3-class)
Positive	20459	20459
Negative	13779	13779
Neutral	N/A	8868

4. METHODOLOGY

In this research, advanced natural language processing models have been applied to classify sentiment in Bengali text. The main focus was on two deep learning models, ELMo and XLNet, as these models capture the context of words better than traditional models. First, the dataset was preprocessed by cleaning the text and removing special characters. Then tokenize the sentences. After that, the original five-class sentiment labels were transformed into two- class and three-class schemes to simplify the classification tasks. The models were trained separately on the two-class and three-class datasets. We split the data into training, validation, and test sets to ensure the models generalize well to unseen data. During training, we used standard hyper parameters and optimization techniques for both ELMo and XLNet. For evaluation, metrics such as accuracy, precision, recall, and F1-score were calculated. These metrics provide a well-rounded understanding of model performance on the sentiment analysis tasks. This methodology allows us to analyze how effective ELMo and XLNet are when applied to Bengali sentiment classification with different levels of label granularity. Figure 1 illustrates the step-by-step procedure of the methodology we used.

4.1 ELMo ARCHITECTURE

A very famous deep contextualized word representation model is named ELMo. It introduces a significant advancement in natural language processing (NLP). It generates embeddings that consider the full context of words in a sentence. Traditional word embed-ding techniques generate a single word embedding for each word in the vocabulary. But ELMo analyzes words within the context of the sentences. It results in word representations that are dynamically informed by the surrounding text. This is achieved through a bidirectional LSTM. It was trained on a large text corpus using a language modeling objective. The idea is that the meaning of a word can significantly vary depending on its linguistic context. ELMo addresses this by producing embeddings that are a function of the entire input sentence. This has been done while capturing nuances such as syntax and semantics at different levels of abstraction. These embeddings are derived from the internal states of a two-layer bidirectional LSTM. This processes the text in both forward and backward directions and allows ELMo to capture both past and future context. For sentiment analysis, ELMo's context- aware embeddings offer a profound advantage. They enable models to discern the sentiment of homonyms with high precision based on their usage in a sentence. This is a notable challenge in sentiment analysis. Furthermore, by leveraging ELMo embeddings as input features, sentiment analysis models can achieve a deeper understanding of the text. Thus, it improves their ability to classify sentiments accurately in ambiguous scenarios.

4.1.1 Data Preprocessing

- Tokenization: The data column undergoes preprocessing for cleaning and normalization. This is crucial for the model to focus on meaningful content.
- Encoding: The label column is encoded using LabelEncoder. It converts categorical labels into a numeric format for model training and evaluation.

4.1.2 Model Architecture

- ELMo Embedding Layer: A customized ELMo embedding layer has been used with 1024 dimensions. It captures semantic meanings from the text.
- Dense Layers: Sequential application of dense layers with 512 and 256 units of neurons. Each with ReLU activation to learn complex patterns from embeddings.
- Batch Normalization: Incorporated after each dense layer and before dropout. It stabilizes the learning process by normalizing layer inputs.
- Dropout: Set at 0.5 to mitigate overfitting by randomly omitting a fraction of the neurons during training. It enhances generalization.
- Output Layer: Utilizes a softmax activation function tailored for multi-class classification. It uses exactly the number of units corresponding to the unique sentiment categories identified in the dataset.

4.1.3 Hyper parameters Configuration

- Loss Function: For multiclass sentiment, ELMo used categorical cross-entropy as a loss function.
- Optimizer: To facilitate robust convergence, as an optimizer, ADAM has been used.
- Batch Size: 32, defining the number of samples processed before the models internal parameters are updated.
- Epochs: Training is set to run for up to 100 epochs, with the flexibility to terminate early based on validation performance to prevent overfitting.
- Validation Split: 20% of the training data is reserved as a validation set to monitor loss and accuracy during training, ensuring the model's ability to generalize.

4.2 XLNet ARCHITECTURE

XLNet is a generalized autoregressive (AR) NLP model that leverages AR pretraining through permutation-based language modeling. AR models predict the next word in a sequence based on the previous one. They cannot handle deep bidirectional context, which is important for tasks like sentiment analysis. XLNet is ”generalized” because it integrates the idea of autoregressive models and bidirectional context modeling. This uses ”Permutation Language Modeling” (PLM). It refers to capturing bidirectional context by training an AR model on all possible permutations of words in a sentence. It predicts the masked tokens based on the tokens/words to its right as well as the left of the masked token. XLNet is based on the transformer architecture [16]. It learns the long-range token dependencies using the concept of attention. Another principal mechanism of XLNet is two-stream attention. It is a novel feature designed to enhance its understanding of text by processing content and positional information separately but in parallel. The first stream, known as the content stream, focuses on understanding the semantic content of each token in the input sequence. In contrast, the query stream is specifically designed to incorporate positional information. It ensures that the model can accurately predict a token's identity based on its position relative to others in the sequence.

4.2.1 Data Preprocessing

- Tokenization: Converting text data into tokens or words that the model can understand. Given the model's configuration, the XLNet tokenizer is used to convert the dataset's text into a format that is suitable for training.
- Encoding: This includes transforming text into numerical data that the model can process, including converting sentiment labels into a categorical format.

4.2.2 Model Architecture

- ELMo Embedding Layer: A customized ELMo embedding layer has been used with 1024 dimensions. It captures semantic meanings from the text.
- Output Hidden States: Set to False, indicating that only the final output is used for efficient classification of text.

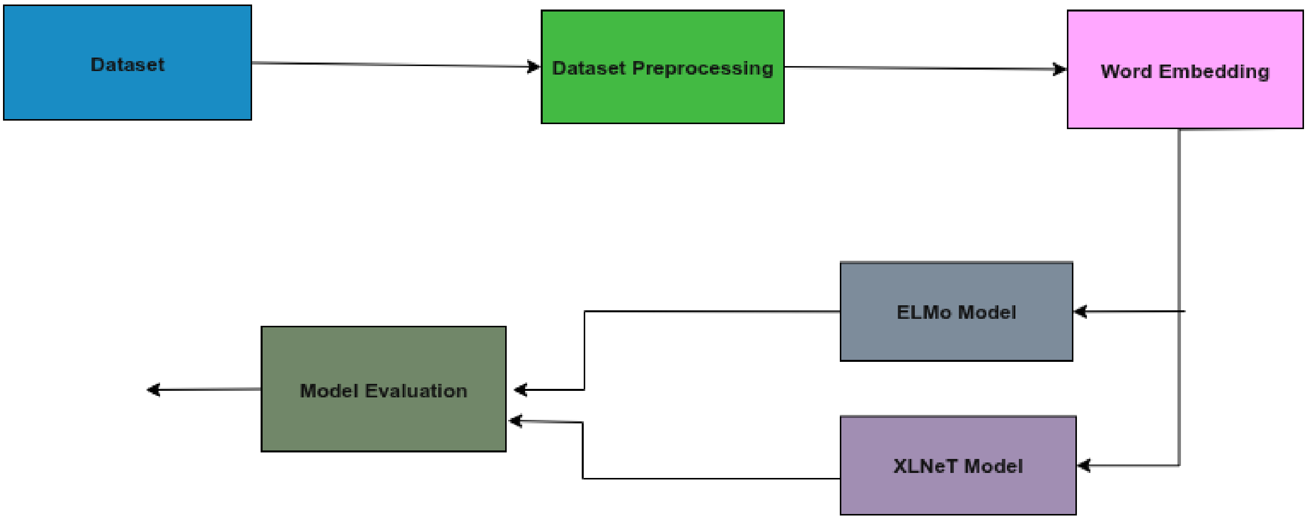

Fig 1: Illustration of the Methodology of this Research

- Dropout Rate: A dropout rate of 0.1 in the final layer reduces overfitting by randomly dropping units from the neural network during training.

4.2.3 Hyper Parameter Configuration

- Loss Function: There are multiple sentiment categories, and for this reason, categorical cross-entropy has been used as a loss function.
- Optimizer: An Adam optimizer has been used with a learning rate of 5e-05, an epsilon of 1e-08, and a decay of 0.01. The clip-norm set to 1.0 helps in managing gradients to prevent the exploding gradient problem.
- Batch Size: Set to 32, balancing the trade-off between training speed and memory usage.
- Epochs: Maximum of 100, with early stopping based on validation loss to prevent overfitting.

5. RESULT ANALYSIS

The evaluation has been done on all these models based on F1 score, precision, and recall. Accuracy has also been noted for each of the models. Table 2 illustrates the exact classification report of these models' average scores.

5.1 MODEL PERFORMANCE

- ELMo models, introduced in Bengali sentiment analysis research for the first time, show promising results. The 2-class model achieves 71% accuracy, while the 3-class model sees a reduction to 43%. The drop in performance in the 3-class scenario could reflect challenges in distinguishing between more nuanced sentiments or the need for further model.

- XLNet models, particularly the 2-class configuration, exhibit a unique distribution of precision and recall scores. Despite the 3 class model's lower accuracy (38%), it is noteworthy that this model and the ELMo models were pioneering efforts in the context of Bengali sentiment analysis.

5.2 CLASSIFICATION REPORT

High precision in the ELMo 2-class model (0.70 for positive sentiment and 0.73 for negative sentiment) suggests a strong ability to identify relevant instances of each sentiment accurately. On the other hand, the XLNet models show a disparity in precision and recall, particularly in the 2-class model, where precision for class 0 is high (0.79), but recall is notably lower (0.52). This indicates a conservative model that is accurate when it identifies a class but misses many instances of it. Table 2 depicts the classification report.

Table 2. Classification Report of the Four Models

Model	Accuracy	Precession	Recall	F1 Score
XLNet (2-class)	0.53	0.67	0.53	0.56
XLNet (3-class)	0.38	0.55	0.38	0.45
ELMo (2-class)	**0.71**	**0.71**	**0.71**	**0.71**
ELMo (3-class)	0.51	0.51	0.51	0.50

5.3 MODEL LIMITATION

The XLNet 3-class model's performance indicates significant challenges, particularly with class 0 showing zero precision and recall. This may point to difficulties in adapting transformer-based models to low-resource languages or specific sentiment analysis nuances without extensive fine-tuning or additional training data.

6. CONCLUSION

To conclude, this work significantly contributes to the field of sentiment analysis for Bengali. It implements the first-ever application of the ELMo and XLNet models for Bengali sentiment analysis. This research also provides a comprehensive evaluation of these models' capabilities in sentiment classification. The ELMo model showed promising results in binary classification with a 71% accuracy, which affirms the potential of advanced NLP models in processing Bengali text. However, the XLNet model's performance in the 3-class model revealed challenges with an overall accuracy of 38%. This denotes the areas to explore for further investigation and model refinement. These variations in model performance in more complex classification tasks highlight the nuanced challenges inherent in Bengali sentiment analysis, such as model sensitivity to class imbalances and the need for a deeper understanding of linguistic features. This research fills a significant gap in Bengali linguistic analysis while setting the stage for future work in this area. It calls attention to the potential of leveraging advanced NLP techniques for low-resource languages. Finally, it underscores the importance of model adaptation and optimization as well

7. ACKNOWLEDGMENTS

My thanks to the Leading University Thesis Committee for allocating students and resources for this research because this research motivation has been found from a random experiment of a thesis work of final year attendees.

8. REFERENCES

[1] Ashima Yadav and Dinesh Kumar Vishwakarma. Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, 53(6):4335–4385, 2020.

[2] Rajesh Kumar Das, Mirajul Islam, Md Mahmudul Hasan, Sultana Razia, Mocksidul Hassan, and Sharun Akter Khushbu. Sentiment analysis in multilingual context: Comparative analysis of machine learning and hybrid deep learning models. *Heliyon*, 9(9), 2023.

[3] Fröhlich, Shahrukh Khan and Mahnoor Shahid. Hindi/bengali sentiment analysis using transfer learning and joint dual input learning with self-attention. *arXiv preprint arXiv:2202.05457*, 2020.

[4] Md Rezaul Karim, Bharathi Raja Chakravarthi, John P McCrae, and Michael Cochez. Classification benchmarks for under-resourced bengali language based on multichannel convolutional-lstm network. In *2020 IEEE 7th international conference on Data Science and Advanced Analytics (DSAA)*, pages 390–399. IEEE, 2020.

[5] Md Ferdous Wahid, Md Jahid Hasan, and Md Shahin Alom. Cricket sentiment analysis from bangla text using recurrent neural network with long short term memory model. In *2019 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–4. IEEE, 2019.

[6] Md Rafidul Hasan Khan, Umme Sunzida Afroz, Abu Kaisar Mohammad Masum, Sheikh Abujar, and Syed Akhter Hossain. Sentiment analysis from bengali depression dataset using machine learning. In *2020 11th international conference on computing, communication and networking technologies (ICCCNT)*, pages 1–5. IEEE, 2020.

[7] SM Samiul Salehin, Rasel Miah, and Md Saiful Islam. A comparative sentiment analysis on bengali facebook posts. In *Proceedings of the international conference on computing advancements*, pages 1–8, 2020.

[8] Samsul Islam, Md Jahidul Islam, Md Mahadi Hasan, SM Shahnewaz Mahmud Ayon, and Syeda Shabnam Hasan. Bengali social media post sentiment analysis using deep learning and bert model. In *2022 IEEE Symposium on Industrial Electronics & Applications (ISIEA)*, pages 1–6. IEEE, 2022.

[9] Nafis Irtiza Tripto and Mohammed Eunus Ali. Detecting multilabel sentiment and emotions from bangla youtube comments. In *2018 international conference on Bangla speech and language processing (ICBSLP)*, pages 1–6. IEEE, 2018.

[10] Baidya Nath Saha, Apurbalal Senapati, and Anmol Mahajan. Lstm based deep rnn architecture for election sentiment analysis from bengali newspaper. In *2020 International Conference on Computational Performance Evaluation (ComPE)*, pages 564–569. IEEE, 2020.

[11] Mahfuz Ahmed Masum, Sheikh Junayed Ahmed, Ayesha Tasnim, and Md Saiful Islam. Ban-absa: An aspect-based sentiment analysis dataset for bengali and its baseline evaluation. In *Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020*, pages 385–395. Springer, 2021.

[12] Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780, 2022.

[13] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. arxiv 2018. *arXiv preprint arXiv:1802.05365*, 12, 2018.

[14] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

[15] Md. Ashfaqur Rahman, Amer Mahbub, Bishwo Nikhil Paul, Prithwiraj Bhattacharjee, and Md. Akhter-Uz-Zaman Ashik. Sentifive: A multi-class bengali dataset for sentiment analysis. Manuscript under review, 2025.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.