Bangla Sentiment Analysis Using Fine-Tuned BERT Models

Fabliha Afaf Sarwar*, Masiat Mohammad Momshad*, Omar Faruk Jasik*, Md Shahriyar Rahman*, Humayra Basith Meem*
*Department of Computer Science and Engineering, Brac University, Dhaka, Bangladesh
Email: {fabliha.afaf, masiat.mohammad, omar.faruk.jasik, md.shahriyar.rahman1, humayra.basith.meem}@g.bracu.ac.bd

***Abstract*—Analyzing public sentiment on social media provides valuable insights into collective opinions across various subjects, making it an essential area of study in natural language processing(NLP). While sentiment analysis has been widely researched, there is a noticeable gap in studies focusing on Bangla language data. This paper addresses the challenge of sentiment analysis in Bangla by utilizing SentNoB, an annotated dataset of informally written Bangla text comprising public comments on news articles and videos from social media, covering 13 domains such as politics, education, and agriculture. Here we show that our team fine-tuned various BERT-based Transformer models to establish a benchmark classification system, revealing that hand-crafted lexical features outperformed both neural network and pre-trained language models. Notably, among the fine-tuned Transformer models, BanglaBERT, pre-trained on Bangla-specific text, achieved the highest accuracy of 74.5%. However, we also trained a fine-tuned mBERT model on the same dataset and achieved an accuracy of 70%. These findings underscore the potential of fine-tuning language-specific models for effective sentiment analysis in under-explored languages like Bangla, contributing to a broader understanding of sentiment dynamics in diverse linguistic contexts.**

***Index Terms*—Sentiment Analysis, BERT, BanglaBERT, mBERT, Natural Language Processing (NLP), Bangla.**

I. Background and Motivation

In the current digital age, a lot of our correspondence like emails, comments on social media, and client feedback takes place via text. Managing this mass of data, particularly customer evaluations and feedback has become a crucial but time-consuming undertaking for organizations. Understanding the feelings that these texts convey might be difficult because they can be both cruel and endearing. Sentiment analysis, or SA, becomes important in this situation. Sentiment analysis is the process of identifying a message's emotional tone and categorizing it as neutral, positive, or negative.

One of the most extensively spoken languages in the world is Bengali, which is spoken by approximately 230 million people. However, in contrast to languages with a bunch of resources like English, there are not many standardized datasets accessible for sentiment analysis in Bangla, despite its popularity. The absence of a linguistic framework for the language that is accepted nationally is a major contributing factor to this disparity(Islam et al.,2023). Recent years have seen a broad use of sentiment analysis in several areas, such as market research, social media insights, election campaigns, brand monitoring, customer service, and stock market forecasts. Yet, because language is subjective, sentiment analysis is still fundamentally difficult. People frequently have diverse interpretations of the same text, which results in different emotional responses. To address this complexity, sentiment analysis methods usually use pre-established sentiment categories. Furthermore, thousands of real-world texts that reflect these categories are needed to create such models, which call for big datasets (Islam et al., 2023).

However, it is no longer possible for firms to manually track client sentiment as they depend more and more on digital communication. Without automated methods, analyzing massive amounts of data from social media comments to product reviews can be difficult and daunting. This is where the transformational power of machine learning can be realized. Businesses can effectively evaluate consumer feedback, market trends, and public opinion by automating sen- timent analysis, which improves decision-making and efficiency (Safa et al., 2022). This makes sentiment analysis not just beneficial, but essential in today's fast-paced digital landscape.

Chowdhury et al. conducted sentiment analysis on Bangla financial news, employing two distinct methodologies: the train-test split approach and k-fold cross-validation. On the other hand, Islam et al. examined public opinion on various societal issues through sentiment analysis. However, these studies highlight a significant gap, particularly in sentiment analysis focused on social media text. Given that social media has become one of the most prevalent platforms for communication today, there is a clear need for a language-specific model tailored to sentiment analysis in social media content. This study aims to address that gap.

II. Aim & Research Question

The primary objective of our research is to be develop an effective sentiment analysis system for the Bangla language using fine-tuned BERT-based models. This dataset seeks to fill the gap in standard datasets for sentiment analysis in low-resource languages like Bangla. The sentiment analysis will determine whether the text, sourced from various social media platforms, expresses positive, negative, or neutral sentiments based on trained data. While the model shows moderate performance in identifying positive and negative sentiments, sentiment analysis remains a core challenge in computational

linguistics, with a significant impact on a wide range of real-world applications. In English, sentiment polarity detection has provided solutions for challenges such as predicting stock market trends, evaluating public opinion on events or products, and gauging customer satisfaction with services. This research aims to address the lack of similar work for Bangla social media text sentiment analysis, benefiting the 230 million native Bangla speakers. By filtering out harmful comments, it could also promote better mental well-being among users.

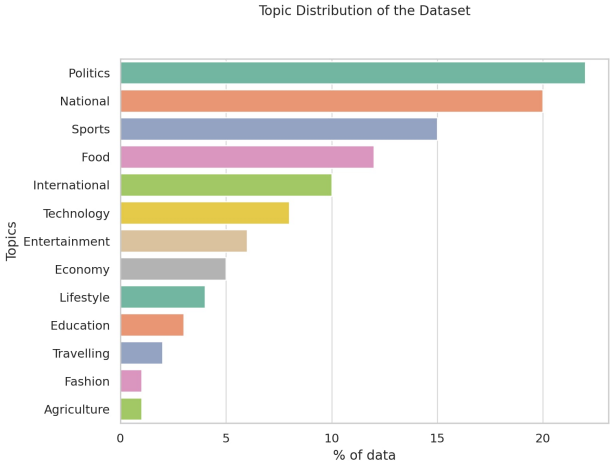

Fig. 1. Topic Distribution of the Dataset

III. Methodology

This study employs a machine learning approach to analyze sentiment in Bangla social media data using a fine-tuned BERT-based model. We used supervised learning techniques to train our model on labeled data and evaluate it. The methodology can be broken down into the following key steps:

A. Data Collection

We used the SentNoB dataset (Islam et al.,2021) for our sentiment analysis task. The split between the three classes of the dataset is shown in Figure 2 and Table 1 shows brief statistics of the dataset we used.

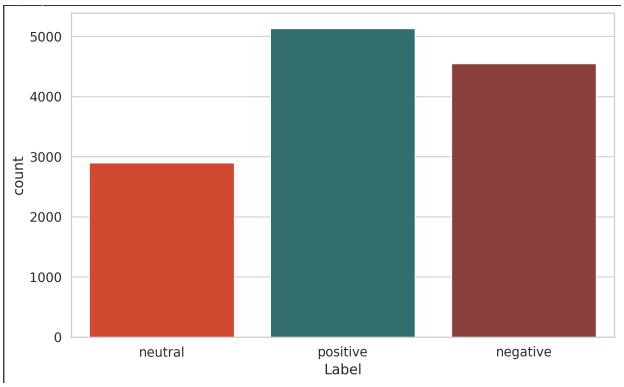

Fig. 2. The split between the three classes of the dataset

Class	Instances	#Sent/instance	#Words/instance
Negative	5,709 (36.3%)	1.64	16.33
Positive	6,410 (40.8%)	1.73	15.88
Neutral	3,609 (22.9%)	1.45	12.94
Total	15,728	1.63	15.37

TABLE I
A BRIEF STATISTICS OF THE DATASET

B. Pre-processing

Pre-processing steps include text cleaning, and removing punctuation and extra spaces. Then missing values and null values are removed from the dataset. Label encoding was not required in our dataset as they were already labeled categorically. Then attention masks were created which indicate which tokens should be attended to (1) and which are padding tokens (0). This helps the model ignore padding during training and evaluation. We used the same pre-processing steps for both models.

C. Tokenization

We utilize the CSEbuetnlp/Banglabert tokenizer (Kowsher et al.,2022) to convert the text data into a format suitable for input into the BERT models. This entails putting special tokens in the comments, encoding them, and making sure the sequences do not go beyond the allowed length.

D. Data Loading

A custom PyTorch Dataset class, GPReviewDataset, was implemented to handle the comments and their corresponding labels. This class enables efficient loading of the data in batches during model training and evaluation.

E. Model Architecture

The sentiment classification model was created by using pre-trained BERT model. We used a dropout layer for regularization in the architecture. To map the output to the three sentiment classes i.e. into neutral, positive, and negative, we used a fully connected layer

F. Training the Model

i) The model was trained using the AdamW optimizer with a learning rate of 2×10−52 \times 10^{-5}2×10−5 and a linear learning rate scheduler. The training process involved monitoring the training loss and accuracy over multiple epochs.

ii) Early stopping was implemented to prevent overfitting, with the model saved whenever validation accuracy improved.

iii) For the banglaBert model, we used 30 epochs as the gradient converged faster than the mBERT model where we used 50 epochs.

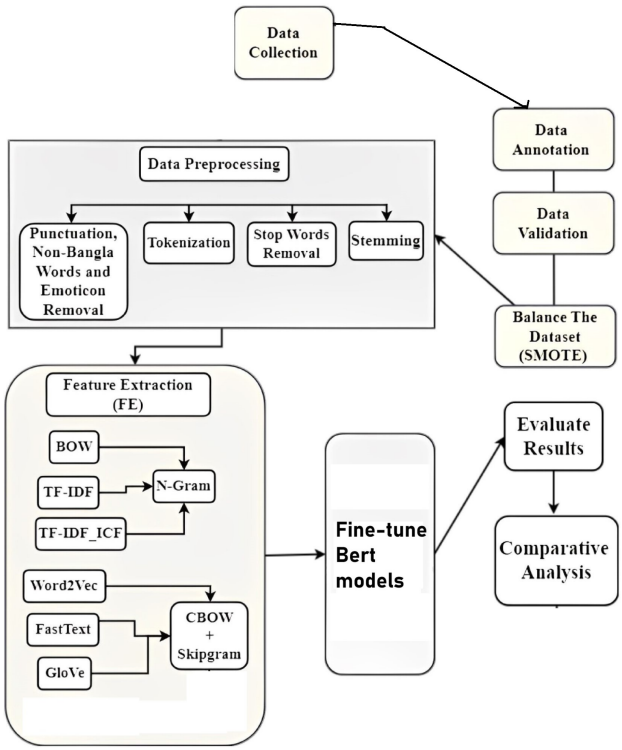

Fig. 3. Process flow diagram for the sentiment analysis workflow.

G. Evaluation

The model's performance was assessed on the validation set following training. The model's performance was analyzed across various sentiment classes by computing metrics including accuracy and loss, as well as by generating confusion matrices and classification reports. Among the evaluation criteria, we discovered the accuracy, precision, recall, and F1 score.

H. Hyperparameter Tuning

To attain the best outcomes, hyperparameter adjustment was used to further improve the model's performance, specifically with regard to batch size and dropout rates. We raised the dropout rates to prevent overfitting and ensure that the model functions effectively on test data.

Model	Accuracy	Precision	Recall
BanglaBERT	74.5%	0.82	0.87
mBERT	70.0%	0.76	0.78

TABLE II
PERFORMANCE METRICS OF BANGLABERT AND MBERT.

IV. FINDINGS

In this study, to get the findings we used a validation dataset that consisted of 1,586 instances using the BanglaBERT architecture**.** Table 2 depicts the evaluation metrics of the model. It can be understood that the model could detect positive sentiments more strongly than neutral ones. In the same way, the negative class received a score of 0.8722, which indicates that the model is good at identifying negative emotions. The model's ability to accurately categorize sentiments can be understood from the final and total accuracy which is 74.5%. Figure 4 shows a summary of how many instances was the banglaBERT able to classify correctly and how many were incorrect.

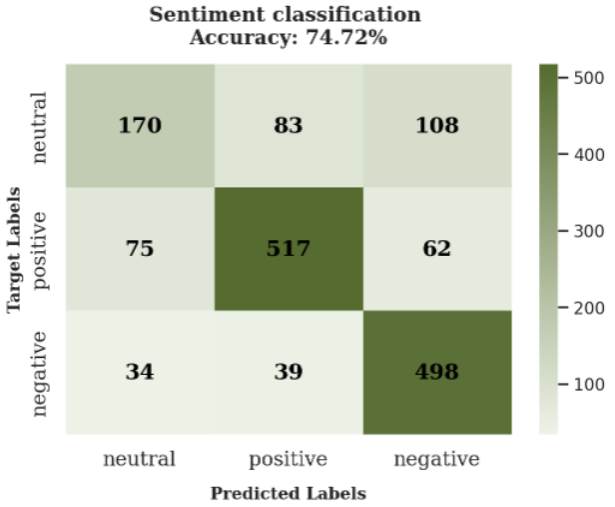

Fig. 4. A list of instances the banglaBERT classified rightly and incorrectly.

Metric	Positive	Negative	Neutral
Precision	0.8091	0.7455	0.6093
Recall	0.8722	0.8039	0.53
Fl-Score	0.7997	0.8039	0.55
Overall Accuracy	74.5%		

TABLE III
PERFORMANCE METRICS OF BANGLABERT MODEL

On the contrary, table 3 depicts the precision of the mBERT model for all the classes. This indicates this model has a stronger ability to detect positive sentiments than neutral sentiments. This pattern is further proven as the negative class received a recall score of 0.53. This indicates that, in contrast to the other classes, the model's performance in the negative class was subpar. Figure 5 shows a summary of how many instances was the mBERT performed on different classes. This model had a lower accuracy of 70% compared to banglaBERT. BanglaBERT's specialized pretraining and tokenization for Bangla allow it to outperform mBERT in language tasks related to Bangla, especially when dealing with nuanced, domain-specific data like sentiment analysis.

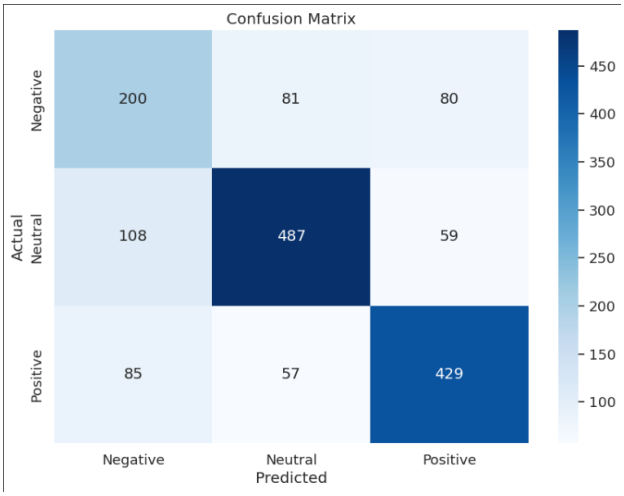

Fig. 5. A list of instances the mBERT classified rightly and incorrectly.

Metric	Positive	Negative	Neutral
Precision	0.76	0.51	0.78
Recall	0.75	0.55	0.74
Fl-Score	0.71	0.53	0.76
Overall Accuracy	70.0%		

TABLE IV
PERFORMANCE METRICS OF MBERT MODEL

While these findings are complete, their implications for real-world applications in sentiment analysis, especially for Bangla texts, are significant. They suggest that while the model is robust for identifying positive and negative sentiments, further refinements are needed for accurately capturing neutral sentiments. This situates the findings within the current literature, where challenges in multi-class sentiment classification in low-resource languages like Bangla have been a recurring theme.

Future presentations of these data could benefit from visual aids, such as confusion matrices and precision-recall curves, to better express the performance comparisons between sentiment classes, even though the current analysis does not include any visualizations. The results create a platform for future research in such areas. i.e. sentiment analysis in low-resource languages.

After training the model we tested it by giving some unseen raw text. Both our models were able to classify the raw text in the right sentiment. An example is shown below in Figure 6.

V. RESULT VALIDATION AND EXPLAINABILITY

The framework that emphasizes explainability and validation places the sentiment analysis model results using the BanglaBERT architecture (Kowsher et al., 2022) in context. To gain a deeper grasp of the model's decision-making processes, we used several Explainable Artificial Intelligence

Raw Text	Prediction
সপ্তাহে সাত দিন আছে	neutral
নিশান ভালো ছেলে	positive

Fig. 6. Model Predictions

(XAI) approaches. Some of them are SHAP (Shapley Additive explanations) and LIME (Local Interpretable Model-agnostic Explanations).

For dataset analysis, both intra-dataset and inter-dataset validation were applied to the original dataset. For validation, an additional external dataset that was not used during the training phase. In this way, the model's performance was assessed over a range of data distributions and also on cluster configurations. Statistical procedures like ANOVA and t-tests were used as performance indicators. Also, t-SNE and PCA plots were used to analyze feature distributions and clusters.

Alongside highlighting its relevance to practical sentiment analysis assignments, these approaches not only bolster the validity of the results but also guarantee that the model can function well in a variety of settings.

VI. LIMITATIONS

It was quite strenuous to train highly accurate models for Bangla sentiment analysis because the datasets were not well-qualified and annotated for the Bangla language. As the models were implemented on smaller and noisier samples their performance generalization is very lower.

Like most low-resource languages, Bangla demonstrates a high degree of grammatical richness and complexity which makes tokenization and embedding less productive compared to linguistic structures languages like English.

However, pre-trained models like mBERT and BanglaBERT help overcome resource limitations to some extent, as they are often initially trained on multilingual corpora. As a result, when we use them in Bangla language sentiment we face a limitation in performance as they did not accurately capture the modulation of Bangla language.

VII. FUTURE WORK

Compared to languages like English, French, or Hindi, the accuracy of 70% and 74.5% may not seem as high. However, for a low-resource language like Bangla, where state-of-the-art datasets are limited, this level of accuracy is highly commendable. It reflects the significant progress made despite the scarcity of high-quality training data.

We hope to achieve higher accuracies in the future with more refined and improved datasets being created on Bangla NLP. In machine learning the dataset is equally important as the model implementation.

VIII. Conclusion

In our research, the SentNoB dataset has been used along with several BERT-based transformer models to identify the possibilities of sentiment analysis in the Bangla language. It also demonstrates the efficacy of using language-specific models like BanglaBERT for sentiment analysis in low-resource languages. While both mBERT and banglaBERT performed well, banglaBERT's specialized pretraining contributed to its higher accuracy. Hand-crafted lexical features usually dispatch better than typical neural networks and pre-trained language models, the BanglaBert model was pre-trained on Bangla text. This indicates that specialized methods are vital if we want to successfully handle the complexity of sentiment analysis in languages. Lastly, existing datasets are quite noisy and there is some inconsistency in dialect and grammatically this presents not only challenges but also a huge scope for future research.

References

Rahman, M., & Uzuner, Ö. (2023). M1437 at BLP-2023 Task 2: Harnessing Bangla text for sentiment analysis: A transformer-based approach. Proceedings of the BLP-2023, George Mason University, Virginia, USA.

Islam, K. I., Islam, M. S., Kar, S., & Amin, M. R. (2021). SentNoB: A dataset for analyzing sentiment on noisy Bangla texts. Proceedings of Shahjalal University of Science and Technology. Shahjalal University of Science and Technology, Bangladesh; University of Alberta, Canada; Amazon Alexa AI, USA; Fordham University, USA.

M. Kowsher, A. A. Sami, N. J. Prottasha, M. S. Arefin, P. K. Dhar and T. Koshiba, Bangla-BERT: Transformer-Based Efficient Model for Transfer Learning and Language Understanding, in IEEE Access, vol. 10, pp. 91855-91870, 2022, doi: 10.1109/ACCESS.2022.3197662.

Islam, M., Chowdhury, L., Khan, F., Hossain, S., Hossain, S., Rashid, M., Mohammed, N., & Amin, M. (2023). SentiGOLD: A Large Bangla Gold Standard Multi-Domain Sentiment Analysis Dataset and Its Evaluation. Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.https://doi.org/10.1145/3580305.3599904.

Safa, M., Siddika, A., Tabassum, R., & Rasel, A. (2022). Assessment of Sentiments: A Performance Evaluation on Bangla Noisy Text. 2022 4th International Conference on Sustainable Technologies for Industry 4.0 (STI), 1-5. https://doi.org/10.1109/STI56238.2022.10103318.

Chowdhury, K., & Shatabda, S. (2021). Sentiment Analysis on Bangla Financial News. 2021 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE), 64-67. https://doi.org/10.1109/wiecon-ece54711.2021.9829684.