EmptyMind at BLP-2023 Task 2: Sentiment Analysis of Bangla Social Media Posts using Transformer-Based Models

Karnis Fatema, Udoy Das, Md. Ayon Mia, Md Sajidul Mowla, Mahshar Yahan, MD Fayez Ullah, Arpita Sarker, Hasan Murad
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
u1804{052, 109, 128, 100, 007, 094, 099}@student.cuet.ac.bd, hasanmurad@cuet.ac.bd

Abstract

With the popularity of social media platforms, people are sharing their individual thoughts by posting, commenting, and messaging with their friends, which generates a significant amount of digital text data every day. Conducting sentiment analysis of social media content is a vibrant research domain within the realm of Natural Language Processing (NLP), and it has practical, real-world uses. Numerous prior studies have focused on sentiment analysis for languages that have abundant linguistic resources, such as English. However, limited prior research works have been done for automatic sentiment analysis in low-resource languages like Bangla. In this research work, we are going to finetune different transformer-based models for Bangla sentiment analysis. To train and evaluate the model, we have utilized a dataset provided in a shared task organized by the BLP Workshop co-located with EMNLP-2023. Moreover, we have conducted a comparative study among different machine learning models, deep learning models, and transformer-based models for Bangla sentiment analysis. Our findings show that the BanglaBERT (Large) model has achieved the best result with a micro F1-Score of 0.7109 and secured 7^{th} position in the shared task 2 leaderboard of the BLP Workshop in EMNLP 2023.

1 Introduction

Nowadays, social media platforms produce a large amount of text data by posting, commenting, and messaging. Finding the sentiment of social media data is an active research area among practitioners due to its numerous practical applications. However, conducting sentiment analysis on social media data is a difficult task due to the natural variation of writing patterns among users.

A significant amount of effort has been devoted to analyzing sentiment in social media data for resourced enriched languages like English (Babu and Kanaga, 2022). However, we have found a limited number of relevant studies focused on sentiment analysis in the Bangla language due to the lack of a standardized annotated dataset of Bangla text sourced from social media platforms (Pran et al., 2020).

The main objective of this research work is to analyze sentiment on Bangla social media posts. Moreover, we conduct a comparative analysis among different ML, DL, and transformer-based models for Bangla sentiment analysis. To train and evaluate different models, we have utilized a dataset provided in a shared task named Sentiment Analysis of Bangla Social Media Posts organized by the First Workshop on Bangla Language Processing co-located with EMNLP-2023 (Hasan et al., 2023a,b; Islam et al., 2021).

Various ML models and DL models have been deployed for Bangla sentiment analysis. We have utilized three popular transformer-based model architectures named Bangla BERT Base, BanglaBERT, and BanglaBERT (Large) for the sentiment analysis model.

Among the machine learning models, SVM utilizing TF-IDF yields the best performance, achieving a micro F1-Score of 0.57. In the realm of deep learning, the BiLSTM + CNN model with Word2Vec attains the highest micro F1-Score at 0.61. The transformer-based BanglaBERT (Large) models (Bhattacharjee et al., 2022) outperform the rest, achieving an impressive micro F1-Score of 0.7109.

The main contributions of our research works are as follows -

- We have finetuned the transformer-based BanglaBERT and BanglaBERT (Large) models for Bangla sentiment analysis.
- We have conducted a comparative analysis among different ML, DL, and transformer-based models for sentiment analysis in the

Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), pages 300–304
December 7, 2023

Bangla language.

The implementation of our research work has been shared in the following GitHub repository - https://github.com/ML-EmptyMind/blp-task2.

2 Related Work

We divide all the previous works related to sentiment analysis into three different categories: ML approaches, DL approaches, and transformer-based approaches.

Machine learning techniques like SVM, Multinomial Naive Bayes, KNN, Logistic Regression, Decision Trees, and Random Forests are used for sentiment analysis. Among these, SVM and Multinomial Naive Bayes classifiers (Hassan et al., 2022) have demonstrated the best performance, with SVM achieving the highest accuracy scores. The dataset is subsequently transformed using a TF-IDF Vectorizer, and SVM is used as the classifier for data classification (Arafin Mahtab et al., 2018).

Numerous deep-learning techniques are employed for sentiment analysis as well. RNN with LSTM model is used (Wahid et al., 2019) for sentiment analysis to classify and categorize the sentiments of social media posts about cricket as positive, negative, or neutral. In order to analyze sentiment or opinion in Bangla, the attention mechanism is suggested (Sharmin and Chakma, 2020) in the study. It examines the difficulties with sentiment analysis and evaluation, particularly in the Bangla language.

Alongside a deep learning model that utilizes multilingual BERT and transfer learning, the research incorporates datasets for two-class and three-class sentiment analysis that have been manually annotated in Bangla, as mentioned in (Islam et al., 2020). This model surpasses the current state-of-the-art algorithms in terms of accuracy, attaining a 71% accuracy rate for two-class sentiment classification and a 60% accuracy rate for three-class sentiment classification. The approach is also used to examine the tone of reader comments in an online daily newspaper, demonstrating that while comments on religious articles tend to be more positive than those on political or sports news, the former are more numerous for those topics.

The objectives of the study include finetuning the transformer-based models for Bangla sentiment analysis and providing a comparison analysis with the baseline models using ML and DL approaches.

3 Dataset

During our research work, we have capitalized the dataset provided using the shared task 2 (Sentiment Analysis of Bangla Social Media Posts) organized by the BLP Workshop @ EMNLP 2023 (Hasan et al., 2023a). The dataset used for this shared task consists of MUltiplatform BAngla SEntiment (MUBASE) (Hasan et al., 2023b) and SentNoB (Islam et al., 2021). The MUBASE dataset is a cross-platform collection of Facebook and Twitter posts that has been manually annotated with sentiment polarity. The SentNob dataset comprises user comments sourced from social media platforms in response to news articles and videos. The dataset covers several fields, such as politics, education, and agriculture. The provided dataset has three sentiment categories: Positive, Negative, and Neutral. This dataset has train, dev, and test split containing 35266, 3934, and 6707 texts respectively. In Table 1, statistics about the dataset are given with class-wise samples.

Classes	Train	Dev	Test
Positive	12,364	1,388	2,092
Negative	15,767	1,753	3,338
Neutral	7,135	793	1,277
Total	35,266	3,934	6,707

Table 1: Class-wise distribution of sentiment analysis dataset

The provided dataset contains URLs, emojis, and other symbols which are removed in the preprocessing step.

4 Methodology

In this section, we outline the methodology of our research. We establish baseline models by employing both ML and DL techniques. Subsequently, we enhance performance by incorporating a transformer-based model. Figure 1 shows an overview of our methodology.

Machine Learning Models

For machine learning algorithms, Word2Vec and TF-IDF word embeddings have been applied to extract the feature vector (Mikolov et al., 2013). Word2Vec embedding has been implemented with

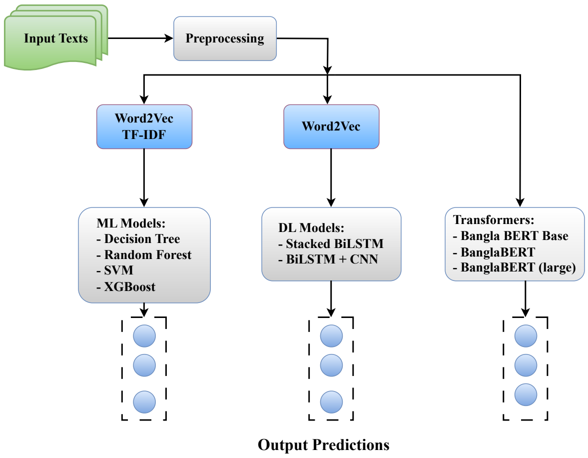

Figure 1: Conceptual process of sentiment analysis

a 100-dimension vector for each word in the vocabulary. We have explored different ML algorithms to set a baseline. We have trained the Decision Tree, Random Forest, SVM, and XGBoost models. Further, all the aforementioned algorithms have been investigated for both TF-IDF and Word2Vec word embedding.

Deep Learning Models

Three different DL models such as Stack of Bidirectional LSTM (BiLSTM), BiLSTM with CNN and dropout, and BiLSTM with CNN for Word2Vec embedding have been explored for Bangla sentiment analysis. We have padded the tokenized dataset by setting the max length value of text as 400 for three models. For the stacked BiLSTM model three consecutive layers of BiLSTM are stacked with 32, 16, and 8 neurons respectively.

Transformer Models

In recent ages, transformer models have gained popularity for their tremendous performance in NLP tasks. We use Bangla BERT Base (Sarker, 2020), BanglaBERT(Bhattacharjee et al., 2022), and BanglaBERT (large)(Bhattacharjee et al., 2022) pre-trained model to fine-tune on Bangla sentiment analysis dataset. The above three pre-trained models have been trained on the Bangla natural language dataset. Bangla BERT Base is a Bangla sentencepiece model containing vocab size 102025. BanglaBERT and BanglaBERT (large) are ELECTRA discriminator models that are pre-trained with the Replaced Token Detection (RTD) objective. Fine-tuned BanglaBERT (large) gives the best result for our case. We have used a pre-trained tokenizer and tokenized sample using 512 as the maximum length of the text. For training purposes, we have taken the help of trainer API.

5 Results and Analysis

In this section, we provide the outcomes obtained from our experimentation.

5.1 Parameter Settings

All parameters are kept identical for the TF-IDF and Word2Vec embedding. For random forest, we have selected *n_estimator* value 40. While training the SVM model, we have picked out C as 2 and kernel *rbf*. Lastly, *n_estimator* value 40 is chosen for XGBoost.

We have set *epochs* to value 30, *batch size* to value 32, *verbose* to value 1 along with callback having an accuracy threshold value of 0.99 for BiLSTM with CNN model which uses Word2Vec embedding. For all DL models, we have set *learning rate* as 0.001, *adam* as the optimizer, and *sparse_categorical_crossentropy* as a loss function. We have further investigated the DL model varying *epochs*, *batch size*, and *learning rate* to validate the consequence on the performance.

In our best performing transformer model, we set the 0.00005 as *learning rate*, 0.01 as *weight decay*, 0.1 as *warm-up ratio*, learning rate scheduler type to *linear*, 3 as training *epochs*, training *batch size* as 16, 2 as *radient accumulation steps* and *adafactor* as the optimizer. Moreover, we set the *dropout rate* to 50% to get the best result. We have evaluated the performance of the BanglaBERT (Large) model, training it without dropout and setting the learning rate to 0.01 for just 3 epochs, which has provided an F1 score of 0.7001. In another setup, we have introduced a 50% dropout rate and extended the training to 4 epochs, which has shown an F1 score of 0.7026.

5.2 Evaluation Metrics

We have applied micro F1-Score evaluation metrics according to guidelines set up by the organizer. Moreover, we also have evaluated precision and recall for all models.

5.3 Comparative Analysis

The performance of each model tested on the evaluation set is displayed in Table 2. We have determined the best-performing model based on the F1-score.

Approach	Classifier	Average		
		P	**R**	**F1**
ML	Decision Tree (TF-IDF)	0.48	0.48	0.48
	Random Forest (TF-IDF)	0.53	0.56	0.56
	SVM (TF-IDF)	0.54	0.57	0.57
	XGBoost (TF-IDF)	0.51	0.53	0.53
	Decision Tree (Word2Vec)	0.45	0.45	0.45
	Random Forest (Word2Vec)	0.50	0.52	0.52
	SVM (Word2Vec)	0.40	0.50	0.50
	XGBoost (Word2Vec)	0.50	0.52	0.52
DL	Stacked BiLSTM (Word2Vec)	0.57	0.57	0.57
	BiLSTM+ CNN (Word2Vec)	0.59	0.61	0.61
Transformer	Bangla BERT Base	0.63	0.63	0.63
	BanglaBERT	0.70	**0.71**	**0.71**
	BanglaBERT (large)	**0.71**	0.70	**0.7109**

Table 2: Performance of various systems on test set. Here P, R, and F1 denote weighted Precision, weighted Recall, and micro F1-Score respectively.

Among the ML models, SVM combined with TF-IDF word embedding has given the highest micro F1-score of 0.5688 while Decision Tree has provided a micro F1-score of 0.4839, Random Forest has shown an F1-score of 0.5555 and XGBoost has given an F1-score of 0.5264. In addition, using Word2Vec embedding, Decision Tree, Radom Forest, SVM and XGBoost model has given micro F1-score of 0.4471, 0.5167, 0.5008, and 0.5239 respectively.

The stacked BiLSTM model, which consists of an input layer with a text length of 400, a Word2Vec embedding layer, there BiLSTM layer, and finally one output layer, has provided a 0.5714 micro F1-score. The combination of BiLSTM along CNN has shown a micro F1-score of 0.6069 which surpasses all other DL and ML models in the evaluation.

Bangla BERT Base has provided a 0.63 micro F1-score which is better than the best performing DL model. In addition, BanglaBERT has shown a micro F1-score of 0.7100. Furthermore, BanglaBERT (large) pre-trained has archived the best score of 0.7109 for this task.

The findings suggest that the transformer-based models have delivered outstanding performance for the assigned task. By comparison, DL models have achieved better results than ML models. Moreover, in transformer-based models, BanglaBERT outperforms Bangla BERT Base. BanglaBERT (large) performs slightly better than BanglaBERT.

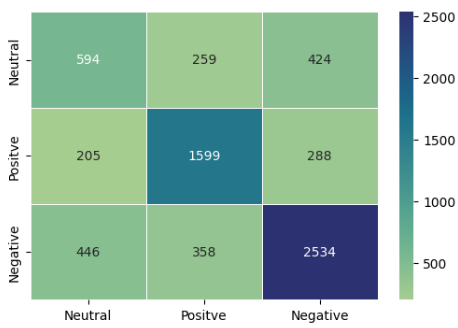

Figure 2: Confusion Matrix of best model

5.4 Error Analysis

Table 2 illustrates that BanglaBERT (large) has acquired the best performance for this task. An observational error analysis has been conducted for the best-performing model. From Figure 2, it has been observed that the model classifies 594 samples of Neutral class correctly and misclassifies 259 as Positive and 424 as negative. Furthermore, 446 samples of the negative class have been incorrectly classified as Neutral. The main reason behind this problem is due to the use of an imbalance dataset. Different size of text length has an impact on error. Sentences with just a few words are not classified correctly for all classes. In the case of neutral sentences, the model misclassifies as negative and for negative sentences model predicts as neutral on a large scale due to a rich set of inflections in the Bangla language, unable to capture all subword information.

6 Conclusion

In this research work, we have explored various transformer-based models for analyzing sentiment in the Bangla language. To train and evaluate different models, we have employed a dataset made available through the BLP Workshop in conjunction with EMNLP-2023. Additionally, we have conducted a comprehensive comparison among different ML, DL, and transformer-based approaches for Bangla sentiment analysis. We have found that the BanglaBERT (Large) model has outperformed the others, achieving the highest micro F1-Score of 0.7109.

In the future, we intend to investigate various architectures and employ ensemble methods to enhance model performance. Additionally, we will apply different techniques to address issues arising from the use of an imbalanced dataset.

Limitations

We have explored only 100-dimensional word embedding for ML and DL models. Other word embedding techniques and hyper-parameter tuning should be further analyzed. Hyper-parameter setting for the BERT model should be an option to investigate beyond. Removal of the impact of text length variation must be addressed.

Ethics Statement

In this paper, we have experimented with different models and techniques that have been ethically implemented. Our aim is to develop a system that finds the sentiment of Bangla text for the betterment of our society and culture. Moreover, we have shared the implementation details in a GitHub repository for reproducibility.

References

Shamsul Arafin Mahtab, Nazmul Islam, and Md Mahfuzur Rahaman. 2018. Sentiment analysis on bangladesh cricket with support vector machine. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–4.

Nirmal Varghese Babu and E Grace Mary Kanaga. 2022. Sentiment analysis in social media data for depression detection using artificial intelligence: a review. *SN Computer Science*, 3:1–20.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States. Association for Computational Linguistics.

Md. Arid Hasan, Firoj Alam, Anika Anjum, Shudipta Das, and Afiyat Anjum. 2023a. Blp-2023 task 2: Sentiment analysis. In *Proceedings of the 1st International Workshop on Bangla Language Processing (BLP-2023)*, Singapore. Association for Computational Linguistics.

Md. Arid Hasan, Shudipta Das, Afiyat Anjum, Firoj Alam, Anika Anjum, Avijit Sarker, and Sheak Rashed Haider Noori. 2023b. Zero- and few-shot prompting with llms: A comparative study with fine-tuned models for bangla sentiment analysis.

Mahmudul Hassan, Shahriar Shakil, Nazmun Nessa Moon, Mohammad Monirul Islam, Refath Ara Hossain, Asma Mariam, and Fernaz Narin Nur. 2022. Sentiment analysis on bangla conversation using machine learning approach. *International Journal of Electrical and Computer Engineering (IJECE)*, 12(5):5562–5572.

Khondoker Ittehadul Islam, Md. Saiful Islam, and Md Ruhul Amin. 2020. Sentiment analysis in bengali via transfer learning using multi-lingual bert.

Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. 2021. SentNoB: A dataset for analysing sentiment on noisy Bangla texts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3265–3271, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality.

Md. Sabbir Alam Pran, Md. Rafiuzzaman Bhuiyan, Syed Akhter Hossain, and Sheikh Abujar. 2020. Analysis of bangladeshi people's emotion during covid-19 in social media using deep learning. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–6.

Sagor Sarker. 2020. Banglabert: Bengali mask language model for bengali language understanding.

Sadia Sharmin and Danial Chakma. 2020. Attention-based convolutional neural network for bangla sentiment analysis. *AI & SOCIETY*, 36:381 – 396.

Md. Ferdous Wahid, Md. Jahid Hasan, and Md. Shahin Alom. 2019. 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.