Semantics Squad at BLP-2023 Task 2: Sentiment Analysis of Bangla Text with Fine Tuned Transformer Based Models

Krishno Dey[1], **Md. Arid Hasan**[1], **Prerona Tarannum**[2], **Francis Palma**[1]
[1]SE+AI Research Lab, University of New Brunswick, Fredericton, Canada
[2]Daffodil International University, Dhaka, Bangladesh
krishno.dey@unb.ca, arid.hasan@unb.ca,
prerona15-14134@diu.edu.bd, francis.palma@unb.ca

Abstract

Sentiment analysis (SA) is a crucial task in natural language processing, especially in contexts with a variety of linguistic features, like Bangla. We participated in BLP-2023 Shared Task 2 on SA of Bangla text. We investigated the performance of six transformer-based models for SA in Bangla on the shared task dataset. We fine-tuned these models and conducted a comprehensive performance evaluation. We ranked 20th on the leaderboard of the shared task with a blind submission that used BanglaBERT Small. BanglaBERT outperformed other models with 71.33% accuracy, and the closest model was BanglaBERT Large, with an accuracy of 70.90%. BanglaBERT consistently outperformed others, demonstrating the benefits of models developed using sizable datasets in Bangla.

1 Introduction

Social networking sites' widespread use in the digital age has produced an unheard-of influx of user-generated content. These sites act as gathering places where people can publicly express their opinions and feelings. It has become popular to identify and measure the emotional tone in textual data through sentiment analysis (SA), a key component of Natural Language Processing (NLP).

While SA has been extensively studied for *resource-rich* languages like English, it is still largely unexplored for many *low-resource* languages like Bangla. Understanding public opinion is crucial for making well-informed decisions in democratic countries. Developing efficient SA tools for the Bangla language has not been possible due to the lack of SA resources, such as datasets and evaluation benchmarks.

This study is devoted to SA and focuses specifically on Bangla being the 7th most spoken language globally (Ethnologue, 2023), and its use on social media sites, particularly Facebook, X, and YouTube, has increased significantly. While much research has been conducted in SA, most attempts have been based on traditional machine learning (ML). Traditional ML techniques have drawbacks in feature engineering, representation learning, scalability, and handling sequential data. They perform best when working with structured data that has clearly defined features. In contrast, deep learning (DL) models like Transformers have excelled at a variety of tasks, especially when dealing with unstructured data like natural language text. Despite the enormous amount of data generated on social media platforms, not many Bangla benchmark datasets are available.

This study addresses this gap by concentrating on the SA of Bangla text in the context of social media. We employ multiple state-of-the-art pretrained transformer models: multilingual BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), XLM-RoBERTa (Conneau et al., 2019), BanglaBERT (Bhattacharjee et al., 2022), BanglishBERT (Bhattacharjee et al., 2022), fine-tuned for SA in Bangla.

We use the dataset provided in the shared task 2 of BLP-2023 (Hasan et al., 2023b) of Bangla text for SA in order to thoroughly assess the efficacy of these models. We measure and report the accuracy, precision, recall, and F1-score as important performance metrics used for SA evaluation. According to the performance matrices, monolingual models such as BanglaBERT BanglaBERT large outperform other transformer-based models.

We secured the 20th position with the submission of BanglaBERT Small of micro F1 score of 67.42%. The shared task followed a blind submission process, meaning the last submission was considered the final submission. Later, we reran the experiment with models including multilingual BERT, XLM-RoBERTa base, BanglishBERT, BanglaBERT Small, BanglaBERT and BanglaBERT large. The best-performing system

Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), pages 312–316

on the leaderboard achieved a micro F1 score of 73.1%, while BanglaBERT in our work achieved a micro F1 score of 70.62%, which is in close proximity to the best system's performance and significantly exceeds the baseline. BanglaBERT large also achieved an F1 score of 70.34%, and closely approaches the performance of the best system. Other models in our study did not achieve the same performance level as BanglaBERT and BanglaBERT large.

2 Related Work

There have been many attempts to address NLP tasks with traditional machine learning (ML). However, it has limitations related to feature engineering, representation learning, scalability, and handling unstructured data. In contrast, Transformer-based models can capture contextual information, rely on pre-trained representations, and can be applied to various languages and domains. Hence, we focus on NLP tasks that were addressed with deep learning (DL) and Language Models (LM).

Several attempts have been made by the researchers to develop resources for SA(Rahman et al., 2018; Tripto and Ali, 2018; Rezaul Karim et al., 2020; Patra et al., 2015), One of the most comprehensive and rigorous overviews of Bangla NLP tasks was conducted by (Alam et al., 2021) and (Hasan et al., 2020). They provided a comparative analysis of Bangla NLP tasks using both classical machine learning algorithms and transformer-based pre-trained models. Their study demonstrates that transformer-based pre-trained models outperform traditional machine learning algorithms.

(Bhowmik et al., 2022) used both DL and transformer-based models for SA of Bangla text. They used a domain-based categorical weighted lexicon data dictionary (LDD) (Bhowmik et al., 2021), which was developed for analyzing sentiments in Bangla from the original dataset (Rahman, 2018). They found that attention-based LSTM (HAN-LSTM), Dynamic routing-based capsule neural network with Bi-LSTM (D-CAPSNET-Bi-LSTM) and bidirectional encoder representations from Transformers (BERT) with LSTM (BERT-LSTM) outperformed other learning models. This study emphasized transformer models improve NLP tasks for languages with limited resources.

(Aurpa et al., 2022) addressed the growing issue of abusive comments in the Bangla language on social media platforms like Facebook. Using transformer-based models like BERT and ELECTRA (Clark et al., 2020), the study achieved a high accuracy of around 85% in identifying and classifying abusive comments from a novel dataset with more than 44k comments. (Rahman et al., 2020) conducted a study on Bangla text document classification using two transformer models, BERT and ELECTRA. The study highlighted the effectiveness of these models for accurately categorizing Bangla text documents, indicating their potential in NLP tasks. (Bhowmick and Jana, 2021) investigated the potential of multilingual BERT and fine-tuned XLM-RoBERTa for SA in Bangla as a low-resource language. The study demonstrated promising results, achieving a maximum accuracy of 95% across three different Bangla datasets, establishing itself as a valuable benchmark for this task. (Aurpa et al., 2022) addressed the growing issue of abusive comments in the Bangla language on social media platforms like Facebook. Using transformer-based models like BERT and ELECTRA, the study achieved a high accuracy of 85%.

In order to address the lack of high-quality Bangla SA datasets, (Hasan et al., 2023a) developed a dataset that focuses on attitudes toward the conflict between Russia and Ukraine. They fine-tuned various transformer-based models and achieved the best performance with 86% accuracy and 82% F-1 score using BanglaBERT. (Islam et al., 2020) introduced two manually tagged SA datasets and a DL model called BERTBSA.

3 Experimental Methodology

This section outlines our experimental methodology. We begin with an overview of the dataset, followed by a discussion of our pre-processing procedures, and conclude by presenting detailed descriptions of the models employed in our study.

Data: We used the dataset that was offered in the BLP shared task 2. The dataset employed for this shared task is a combination of Bangla text data from two distinct sources: MUBASE (Hasan et al., 2023c) and SentNob (Islam et al., 2021). SentNob is a compilation of public comments extracted from various social media platforms, spanning 13 domains such as politics, education, and agriculture, and manually annotated. The level of agreement among annotators for this dataset is moderate, with an agreement score of 0.53. On the other hand, the MUBASE dataset comprises a comprehensive col-

Split	# of Samples	Pos	Neg	Neu
Train	35,266	35%	45%	20%
Dev	3,934	35%	45%	20%
Test	6,707	31%	50%	19%

Table 1: Data Description and Split. Pos: Positive, Neg: Negative, Neu: Neutral

Model	Epochs	LR	Par
m-BERT	3	2e-5	180M
XLM-RoBERTa base	3	2e-5	270M
BanglishBERT	3	2e-5	110M
BanglaBERT Small	3	2e-5	13M
BanglaBERT	3	2e-5	110M
BanglaBERT large	3	2e-5	335M

Table 2: Training Parameters of Models. LR: Learning Rate, Par: Parameters

lection of multi-platform data, featuring manually labeled Tweets and Facebook posts, each categorized based on their sentiment polarity. This dataset presents a multi-class sentiment analysis (SA) challenge with three categories: *positive*, *negative*, and *neutral*. Table 1 show the overview of the data and splitting procedure.

Data Pre-Processing and Cleaning: Preprocessing for the Bangla text dataset offered in the shared task 2 of BLP-2023 entails several steps to ensure that the data is prepared for SA. First, standard text cleaning techniques like removing special characters, punctuation, extra white space, and URLs should be applied to the text data. Tokenization is then used to separate the text into tokens or single words. If stop words are present, they are typically eliminated to lower data noise. For modeling, it is crucial to convert the class labels into numerical values, such as 0 for negative, 1 for neutral, and 2 for positive.

Transformer-Based Models: We employed a variety of transformer-based models to conduct SA on the dataset provided for Shared Task 2. Our approach involved taking our pre-processed dataset and fine-tuning it using multiple transformer models, including m-BERT, XLM-RoBERTa base, BanglishBERT, and BanglaBERT. To optimize model performance, batch size of 32 was employed to expedite the training process, meaning that gradient accumulation was computed after every 32 data samples. The choice of a learning rate of $2e^{-5}$ was predicated on the rationale that this rate allows the algorithm to more effectively learn parameter estimates. Three epochs were sufficient for the models to converge on the dataset and avoid model overfitting and under-fitting. These experiments were conducted to explore the effectiveness of different transformer models in capturing sentiment patterns within the dataset and achieve the most accurate SA results. Batch size 32 was used to speed up the training process, and we set gradient accumulation count set 1 which means the gradient accumulation was calculated after 32 data samples. The learning rate of a $2e^{-5}$ was due to the fact that at this pace algorithms learn the values of a parameter estimate in a better way. Table 2 provides an overview of the model parameters.

4 Results Analysis and Discussion

To determine which models were most effective and could be applied to real-life SA problems, we fine-tuned and applied the m-BERT, XLM-RoBERTa base, BanglishBERT, BanglaBERT Small, BanglaBERT and BanglaBERT large models. In particular, BanglaBERT consistently outperformed the other models in terms of various performance metrics.

Table 3 presents a comprehensive breakdown of the performances of all these models. From the table, we can see that BanglaBERT achieved the highest accuracy with 71.33% on the test set, and among other Bangla pre-trained models, BanglaBERT large was also quite close with an accuracy of 70.9%. The other two models, namely BanglaBERT Small and BanglishBERT, achieved 67.23% and 68.81%, respectively. On the other hand, the multilingual model XLM-RoBERTa achieved an accuracy of 68.81%, and m-BERT achieved an accuracy of 65.56%. From the perspective of accuracy, BanglaBERT outperforms the other models. However, in terms of precision, BanglaBERT and BanglaBERT large are very close, averaging 70.22% and 70.07%, respectively. Regarding the F1 score, BanglaBERT and BanglaBERT large also exhibit similar performance, with average F1 scores of 70.62% and 70.4%, respectively. Another pattern that emerges from the table is that the performance measures for all models in the neutral class are lower than those for both the positive and negative classes. In fact, the performance measures for the negative

CL	Acc	P	R	F1
Multi-lingual BERT(m-BERT)				
Negative		0.71	0.75	0.73
Neutral	0.6556	0.45	0.37	0.41
Positive		0.67	0.68	0.68
XLM-RoBERTa base				
Negative		0.73	0.78	0.75
Neutral	0.6826	0.49	0.35	0.41
Positive		0.69	0.73	0.71
BanglishBERT				
Negative		0.76	0.76	0.76
Neutral	0.6881	0.49	0.36	0.42
Positive		0.67	0.78	0.72
BanglaBERT Small				
Negative		0.72	0.79	0.75
Neutral	0.6723	0.47	0.28	0.35
Positive		0.67	0.73	0.70
BanglaBERT				
Negative		0.76	0.80	0.78
Neutral	0.7133	0.48	0.38	0.43
Positive		0.74	0.77	**0.76**
BanglaBERT large				
Negative		0.76	0.80	0.78
Neutral	0.7090	0.48	0.40	0.44
Positive		0.74	0.76	0.75

Table 3: Comprehensive Breakdown of the Classification Results. Bold numbers indicate the best F1 score with respect to positive class. CL: Class Label, Acc: Accuracy, P: Precision, R: Recall, F1: F1 Score

.

class are superior to those of the other two classes for all models. This likely stems from the significantly higher number of samples in the negative class. Nearly 50% of the samples in the training, development, and test sets belong to the negative class.

However, we were unable to extract insights into why BanglaBERT exhibited superior performance compared to m-BERT and XLM-RoBERTa models. It is possible that BanglaBERT's training on a substantial Bangla dataset provided a slight advantage over the other multi-lingual models. The superior performance of BanglaBERT indicates that models specifically trained on a sizable Bangla dataset have a natural advantage when identifying subtle sentiment nuances in Bangla text. This may be attributed to the fine-tuning process used by BanglaBERT, which allowed it to better comprehend the nuances of Bangla language and sentiment expression. However, despite being intended to be multi-lingual models, m-BERT and XLM-RoBERTa may not have fully adapted to the nuances of the Bangla language, which resulted in their comparatively poorer performance.

Although BanglaBERT outperformed the other models, our study could not pinpoint the precise causes of this performance disparity. For example, despite being larger and having three times more parameters than BanglaBERT, BanglaBERT large could not perform as expected. The observed behavior may be attributed to several potential factors within the context of the data provided for Shared Task 2 of BLP-2023. One likely contributor could be the inadequacy of the data structure for the models to perform optimally. Another possibility is that the pre-processing steps applied to the data may not have been sufficient to enable the models to achieve their expected levels of performance. Additionally, the choice of hyper-parameters for the models, including the fine-tuning process, might not have been optimal, potentially impacting their overall performance.

5 Conclusion and Future Work

This study conducted a comprehensive evaluation of fine-tuned transformer-based models for sentiment analysis (SA) in Bangla text. The importance of models specifically trained on large Bangla datasets for SA tasks is highlighted by BanglaBERT's consistent and superior performance across a variety of performance metrics. The advantage that BanglaBERT showed over the multi-lingual models, m-BERT and XLM-RoBERTa, suggests that a deeper comprehension of the Bangla language and sentiment expression is crucial for obtaining accurate SA results. The precise linguistic and contextual factors contributing to BanglaBERT's superior SA abilities need to be further investigated. In our future research endeavors, we aim to delve deeper into why transfer-based multi-lingual models struggled to compete with BanglaBERT, further enhancing our understanding of their performance disparities.

References

Firoj Alam, Md Arid Hasan, Tanvir Alam, Akib Khan, Janntatul Tajrin, Naira Khan, and Shammur Absar Chowdhury. 2021. A review of bangla natural language processing tasks and the utility of transformer models. *arXiv preprint arXiv:2107.03844*.

Tanjim Taharat Aurpa, Rifat Sadik, and Md Shoaib Ahmed. 2022. Abusive bangla comments detection on facebook using transformer-based deep learning models. *Social Network Analysis and Mining*, 12(1):24.

Abhik Bhattacharjee, Tahmid Hasan, Kazi Mubasshir, Md. Saiful Islam, Wasi Ahmad Uddin, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. Banglabert: Lagnuage model pretraining and benchmarks for low-resource language understanding evaluation in bangla. In *Findings of the North American Chapter of the Association for Computational Linguistics: NAACL 2022*.

Anirban Bhowmick and Abhik Jana. 2021. Sentiment analysis for bengali using transformer based models. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 481–486.

Nitish Ranjan Bhowmik, Mohammad Arifuzzaman, and M Rubaiyat Hossain Mondal. 2022. Sentiment analysis on bangla text using extended lexicon dictionary and deep learning algorithms. *Array*, 13:100123.

Nitish Ranjan Bhowmik, Mohammad Arifuzzaman, M Rubaiyat Hossain Mondal, and MS Islam. 2021. Bangla text sentiment analysis using supervised machine learning with extended lexicon dictionary. *Natural Language Processing Research*, 1(3-4):34–45.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. Cite arxiv:1810.04805Comment: 13 pages.

Ethnologue. 2023. The most spoken languages worldwide in 2023. https://www.ethnologue.com/insights/ethnologue200/. [Online; accessed 09-September-2023].

Mahmud Hasan, Labiba Islam, Ismat Jahan, Sabrina Mannan Meem, and Rashedur M Rahman. 2023a. Natural language processing and sentiment analysis on bangla social media comments on russia–ukraine war using transformers. *Vietnam Journal of Computer Science*, pages 1–28.

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

Md. Arid Hasan, Jannatul Tajrin, Shammur Absar Chowdhury, and Firoj Alam. 2020. Sentiment classification in bangla textual content: A comparative study. In *23rd International Conference on Computer and Information Technology (ICCIT)*.

Khondoker Ittehadul Islam, Md Saiful Islam, and Md Ruhul Amin. 2020. Sentiment analysis in bengali via transfer learning using multi-lingual bert. In *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pages 1–5. IEEE.

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.

Braja Gopal Patra, Dipankar Das, Amitava Das, and Rajendra Prasath. 2015. Shared task on sentiment analysis in indian languages (sail) tweets-an overview. In *Proc. of MIKE*, pages 650–655. Springer.

Atik Rahman. 2018. Bangla absa datasets for sentiment analysis. `https://github.com/atik-05/Bangla_ABSA_Datasets`.

Md Rahman, Emon Kumar Dey, et al. 2018. Datasets for aspect-based sentiment analysis in bangla and its baseline evaluation. *Data*, 3(2):15.

Md Mahbubur Rahman, Md Aktaruzzaman Pramanik, Rifat Sadik, Monikrishna Roy, and Partha Chakraborty. 2020. Bangla documents classification using transformer based deep learning models. In *2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI)*, pages 1–5. IEEE.

Md Rezaul Karim, Bharathi Raja Chakravarthi, Mihael Arcan, John P McCrae, and Michael Cochez. 2020. Classification benchmarks for under-resourced Bengali language based on multichannel convolutional-lstm network. *arXiv*, pages arXiv–2004.

Nafis Irtiza Tripto and Mohammed Eunus Ali. 2018. Detecting multilabel sentiment and emotions from bangla youtube comments. In *Proc. of ICBSLP*, pages 1–6. IEEE.