Semantics Squad at BLP-2023 Task 1: Violence Inciting Bangla Text Detection with Fine-Tuned Transformer-Based Models

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

Abstract

This study investigates the application of Transformer-based models for violence threat identification. We participated in the BLP-2023 Shared Task 1 and in our initial submission, BanglaBERT large achieved 5th position on the leader-board with a macro F1 score of 0.7441, approaching the highest baseline of 0.7879 established for this task. In contrast, the top-performing system on the leaderboard achieved an F1 score of 0.7604. Subsequent experiments involving m-BERT, XLM-RoBERTa base, XLM-RoBERTa large, BanglishBERT, BanglaBERT, and BanglaBERT large models revealed that BanglaBERT achieved an F1 score of 0.7441, which closely approximated the baseline. Remarkably, m-BERT and XLM-RoBERTa base also approximated the baseline with macro F1 scores of 0.6584 and 0.6968, respectively. A notable finding from our study is the under-performance by larger models for the shared task dataset, which requires further investigation. Our findings underscore the potential of transformer-based models in identifying violence threats, offering valuable insights to enhance safety measures on online platforms.

1 Introduction

The global use of social media platforms has significantly increased due to the massive use of the internet and over the past few decades, internet use has rapidly expanded. People are now more able to share information and their opinions online because of easier access to the internet. Social media use has gained a curve that has been steadily increasing and shows no signs of stopping[1]. The study of (Tarannum et al., 2023) includes how some actors use social media platforms to spread false information that covers offensive language, cyber-bullying, cyber-aggression, rumors, and hate speech. As a result, spreading hate speech is now easy and common. This change also gives social media significant influence over our society. On these social media platforms, people are free to express different points of view on religion, politics, education, and other issues. The use of Bangla on social media platforms is growing because it is the seventh most spoken language[2], and internet usage is rising. While there has been much research on detecting hate speech in English, there is a knowledge gap in identifying hostile content in low-resource languages like Bangla. This type of information generated interest in identifying and flagging it due to its potentially misleading or dangerous nature, which might help stop its future spread. According to research by (Yin et al., 2009), machine learning algorithms are more accurate than keyword searches and textual content analysis for social media data. However, many of the proposed machine learning techniques are, in fact, topic-specific.

Even though there are some works on sentiment analysis for the Bangla language, there has not been much research done recently to identify abusive Bangla text on social networking sites. Due to the constantly changing nature of social media and the wide range of language used, identifying abusive text is difficult. Researchers tried to develop various methods to identify abusive or objectionable text (Nobata et al., 2016) to prevent abuse on online platforms.

In this study, we participated in BLP Shared Task 1: Violence Inciting Text Detection (VITD) (Saha et al., 2023b) and the dataset offered in the shared task has two columns (label, text), with the label in three categories (*Direct Violence*, *Passive Violence*, and *Non-Violence*) (Saha et al., 2023a). We conducted a series of Transformer-based experiments on the dataset provided by the organizers. In our blind submission with BanglaBERT large, we achieved a macro F1 score of 0.7441, securing the 5th position on the official leaderboard.

[1] https://www.pewresearch.org/internet/fact-sheet/social-media/

[2] https://www.ethnologue.com/insights/ethnologue200/

Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), pages 225–229
December 7, 2023 ©2023 Association for Computational Linguistics

The highest baseline for the task was a macro F1 score of 0.7879, while the best-performing system, developed by DeepBlueAI, achieved a macro F1 score of 0.7604. Subsequently, we re-ran the experiments employing m-BERT (Devlin et al., 2018), XLM-RoBERTa base (Conneau et al., 2019), XLM-RoBERTa large, BanglishBERT (Bhattacharjee et al., 2022), BanglaBERT (Bhattacharjee et al., 2022), and BanglaBERT large (Bhattacharjee et al., 2022) models. BanglaBERT, m-BERT, and XLM-RoBERTa base came quite close to the respective official baseline performance in these evaluations. Other models, including XLM-RoBERTa large, BanglishBERT, and BanglaBERT large, demonstrated noteworthy and commendable performance.

The structure of this paper is as follows: Section 2 summarizes the relevant works for this study. Section 3 reports the methodology. A detailed discussion of the results of our study is provided in Section 4. Finally, we state limitations and future work in Section 5.

2 Related Works

Social media has integrated itself into everyone's daily lives. It makes it possible to communicate quickly, share information easily, and receive opinions across geographical boundaries. However, this manifestation of freedom has also contributed to the development of harsh language and hate speech on social media platforms. Unfortunately, the detection of hate speech in Bangla social media has received very little attention. The main issue is the unavailability of sufficient data. In addition, the terminology used in hate speech is extremely diverse. Social media language frequently deviates significantly from that of traditional print media. Numerous linguistic characteristics are present in it. Therefore, it is difficult to recognize hate speech automatically.

In heterogeneous language-speaking countries like Ethiopia, data in Amharic was gathered and annotated to detect hate speech by (Mossie and Wang, 2020) and proposed a method for automatic detection of hate speech directed towards vulnerable minority groups on social media. The authors reported that RNN-GRU exhibits the best performance with an accuracy of 92.56% and an AUC of 97.85%. The accuracy of all algorithms improved using word embeddings like Word2Vec. Early research by (Kiilu et al., 2018) created a method for identifying and categorizing hate speech using content from self-identified hate communities on X (formally known as Twitter) and suggested that the Naive Bayes classifier greatly outperformed the existing approaches with 67.47% accuracy. Another study with GPT-3 to identify sexist and racist text passages, (Chiu et al., 2021) discovered that the model accuracy could reach as high as 85% with few-shot learning. (Romim et al., 2021a) created the HS-BAN dataset on hate speech with their benchmark system, and the best outcome was obtained by combining Bi-LSTM with FT(SG) or Bi-LSTM+FT(SG), which achieved an F1 score of 86.85%. (Ishmam and Sharmin, 2019) built a dataset with 5,126 Bangla comments from social media and got an accuracy of 70.1% using GRU-based models, which gave 18% higher accuracy than ML algorithms.

A recent study by (Alam et al., 2020) using several publicly accessible datasets for the experiments by fine-tuning multilingual transformer models for Bangla text classification tasks in several areas to improve accuracy upon the prior results between 5%-29% across different tasks. The dataset for this study was obtained from X (previously known as Twitter). According to (Das et al., 2022), the XLM-Roberta model has the highest accuracy on their developed annotated dataset which consists of 10K Bangla posts where 5K is actual and 5K is Romanized Bangla tweets. By preparing only a multi-modal hate speech dataset, after experiments (Karim et al., 2022) reported F1 scores of 78% and 82%, respectively, using Conv-LSTM and XLM-RoBERTa models, which scored best for texts. ResNet-152 and DenseNet-161 models produced F1 scores of 78% and 79% for memes, respectively. Concerning multi-modal fusion, XLM-RoBERTa + DenseNet-161 demonstrated the best performance, producing an F1 score of 83%. (Islam et al., 2021) took data from some controversial pages of social media and after evaluation, the maximum accuracy of 88% was achieved by SVM using the entire dataset. On the dataset of (Romim et al., 2021b), the authors ran baseline experiments, applied several deep learning models, and extensively trained Word2Vec, FastTest, and BengFastText models on Bangla words to facilitate future research opportunities, and the experiment showed that SVM had the best outcome with 87.5% accuracy.

3 Experimental Methodology

This section describes our experimental methodology. We start with a brief overview of the dataset, then talk about our pre-processing steps, and present in-depth explanations of the models used in this study.

3.1 Dataset

We utilized the dataset offered by the BLP-2023 Shared Task 1. The dataset consists of YouTube comments about the top nine violent incidents in Bangladesh and West Bengal over the last decade between 2013-2023. The dataset includes Bangla-language content with comments that can be up to 600 words long. The dataset contains three data classes: *Direct Violence*, *Passive Violence*, and *Non-Violence*.

Split	Samples	DV	PV	NV
Train	2700	15%	34%	51%
Dev	1330	15%	31%	54%
Test	2016	10%	36%	54%

Table 1: Overview of the Data and Splitting Procedure. NV: Non-Violence, PV: Passive Violence, DV: Direct Violence.

- **Direct Violence:** Comments directly promoting or inciting violence.
- **Passive Violence:** Comments indirectly endorsing or facilitating violent actions.
- **Non-Violence:** Comments that do not relate to violence.

The offered dataset in the shared task can help identify and classify threats associated with violence, potentially leading to further incitement of violent acts. Table 1 shows the dataset distribution.

3.2 Pre-Processing

Several pre-processing steps were carried out in preparing the BLP-2023 shared task 1 dataset for analysis and classification. The text data underwent an extensive cleaning phase, during which special characters, URLs, and punctuation were eliminated. Tokenization was then used to separate the text into individual words or tokens. Then we eliminated all of the stop words, which are generally low-content words that are used frequently in a language. When classifying documents, the elimination of stop words enables the classification algorithm to concentrate on the keywords. These pre-processing steps further enhanced the quality of the dataset for subsequent analysis and classification tasks.

L	Acc	P	R	F1	F1-m
Multilingual BERT(m-BERT)					
NV		0.73	0.85	0.79	
PV	0.7138	0.77	0.52	0.62	0.6584
DV		0.51	0.63	0.57	
XLM-RoBERTa base					
NV		0.77	0.84	0.80	
PV	0.7376	0.76	0.59	0.66	0.6968
DV		0.55	0.72	0.63	
XLM-RoBERTa large					
NV		0.80	0.87	0.84	
PV	0.7679	0.80	0.61	0.69	0.7246
DV		0.55	0.78	0.65	
BanglishBERT					
NV		0.76	0.88	0.81	
PV	0.7321	0.81	0.50	0.62	0.7232
DV		0.52	0.78	0.62	
BanglaBERT					
NV		0.82	0.89	0.85	
PV	0.7867	0.82	0.64	0.71	**0.7441**
DV		0.58	0.79	0.67	
BanglaBERT large					
NV		0.81	0.88	0.84	
PV	0.7773	0.82	0.61	0.70	0.7344
DV		0.56	0.80	0.66	

Table 2: Comprehensive Breakdown of the Classification Results. Bold numbers indicate the best F1 score. NV: Non-Violence, PV: Passive Violence, DV: Direct Violence, L: Label, Acc: Accuracy, P: Precision, R: Recall, F1: F1 Score, F1-m: F1-macro.

.

3.3 Models

We employed several transformer-based models, including m-BERT, XLM-RoBERTa base, XLM-RoBERTa large, BanglishBERT, BanglaBERT, and BanglaBERT large. Each model was trained for five epochs, a duration sufficient for convergence on the test data. In order to enhance the model's performance, a batch size of 32 was utilized to accelerate the training procedure except for XLM-RoBERTa large, for which we employed a batch size of 16 due to resource limitations, wherein gradient accumulation was calculated following every 32 data samples. The selection of a learning rate of $2e^{-5}$ was based on the principle that this rate facilitates more efficient learning of parameter estimates by the algorithm.

4 Result Analysis & Discussion

In our study, we evaluated a wide range of models, such as m-BERT, XLM-RoBERTa base, XLM-RoBERTa large, BanglishBERT, BanglaBERT, and BanglaBERT large, to determine how well they perform identifying and categorizing threats related to violence in the BLP-2023 dataset. One of the most critical findings from our analysis was the unexpected performance of the smaller model, BanglaBERT, which outperformed larger models like m-BERT, XLM-RoBERTa large, BanglishBERT, and BanglaBERT large. This unexpected outcome highlights the importance of model architecture and the flexibility with which it can be adapted to the specifics of the dataset. Despite its smaller size, BanglaBERT outperformed other models, suggesting its ability to capture the subtleties of language and context related to violence within the Bangla dataset. This superior performance can be attributed to its training on a Bangla dataset, enabling it to excel in this specific linguistic and contextual domain. This finding emphasizes the significance of pre-training data and architecture, in addition to size, when choosing models for particular NLP tasks.

Table 2 illustrates that BanglaBERT achieved the highest accuracy, reaching 0.7876, surpassing all other models in our evaluation. BanglaBERT outperformed other models with precision and recall of 0.7996 and 0.7808, respectively. The official scoring metric for the BLP2023 shared task 1 was the macro F1 score. The baseline macro F1 scores for the shared task set by the organizers were 0.7879 for BanglaBERT, 0.7292 for XLM-RoBERTa base, and 0.6819 for BERT multilingual. BanglaBERT achieved a macro F1 score of 0.7441 in our study, quite close to the baseline. BERT multilingual base (cased) achieved a macro F1 score of 0.7068, surpassing the baseline, while the XLM-RoBERTa base achieved an F1 score of 0.7347, also surpassing the baseline. Additionally, other models, namely XLM-RoBERTa large, achieved an F1 score of 0.7246, BanglishBERT achieved an F1 score of 0.7232, and BanglaBERT large achieved an F1 score of 0.7344. The macro F1 score of BanglaBERT in our study closely matches the baseline, with a little difference, and it outperforms other models in our study for this specific dataset of shared task 1. Table 3 shows the performance on the official leaderboard of our works compared to the baselines and other works.

System	F1 Score	Rank
Our Work	**0.7441**	5^{th}
DeepBlueAI	0.7604	1^{st}
Baseline(BanglaBERT)	0.7879	–
Baseline(XLM-RoBERTa)	0.7292	–
Baseline(mBERT)	0.6819	–

Table 3: Official results on the test set and overall ranking of Task 1: Violence Inciting Text Detection (VITD). **Bold** indicates our systems.

Overall, BanglaBERT stands out as a dependable and competitive solution to the challenging problem of identifying violence threats in Bangla. Its capacity to closely match the baseline macro F1 score and its strong precision and recall metrics highlight its potential to strengthen safety measures in the online environment, where the detection of violent threats is of utmost importance. The effectiveness of transformer-based models, particularly BanglaBERT, in identifying violent threats is reaffirmed by this comprehensive viewpoint, which also provides invaluable insights for enhancing online security measures.

5 Conclusion and Future Work

In this study, we performed a comparative study on several transformer-based models to detect violent text. We used the dataset offered by shared task 1 of the BLP workshop for this study. We used such as m-BERT, XLM-RoBERTa base, XLM-RoBERTa large, BanglishBERT, BanglaBERT, and BanglaBERT large to compare their result. The result shows that BanglaBERT outperformed other models in terms of performance measures. Despite being larger models m-BERT, XLM-RoBERTa large, BanglaBERT large could not outperform BanglaBERT. One of the limitations of our work is that we were not able to reveal the specific reason why our large models are not performing as they are supposed to for this task.

To extend this study, we plan to employ transfer learning, efficient model designs, or model compression to improve the performance of large models such as m-BERT, XLM-RoBERTa, and BanglaBERT large. In the context of hate speech detection, we will investigate the ideal hyperparameters for transformer-based models, including learning rate schedules, model size, and optimization strategies. The development of more accurate, fair, and reliable hate speech detection algorithms may emerge from future research in these areas, thus, resolving the limitations of this study and enhancing the field of NLP.

References

Tanvirul Alam, Akib Khan, and Firoj Alam. 2020. Bangla text classification using transformers. *arXiv preprint arXiv:2011.04446*.

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*.

Ke-Li Chiu, Annie Collins, and Rohan Alexander. 2021. Detecting hate speech with gpt-3. *arXiv preprint arXiv:2103.12407*.

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.

Mithun Das, Somnath Banerjee, Punyajoy Saha, and Animesh Mukherjee. 2022. Hate speech and offensive language detection in bengali. *arXiv preprint arXiv:2210.03479*.

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.

Alvi Md Ishmam and Sadia Sharmin. 2019. Hateful speech detection in public facebook pages for the bengali language. In *2019 18th IEEE international conference on machine learning and applications (ICMLA)*, pages 555–560. IEEE.

Tanvirul Islam, Nadim Ahmed, and Subhenur Latif. 2021. An evolutionary approach to comparative analysis of detecting bangla abusive text. *Bulletin of Electrical Engineering and Informatics*, 10(4):2163–2169.

Md Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Md Shajalal, and Bharathi Raja Chakravarthi. 2022. Multimodal hate speech detection from bengali memes and texts. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 293–308. Springer.

Kelvin Kiema Kiilu, George Okeyo, Richard Rimiru, and Kennedy Ogada. 2018. Using naïve bayes algorithm in detection of hate tweets. *International Journal of Scientific and Research Publications*, 8(3):99–107.

Zewdie Mossie and Jenq-Haur Wang. 2020. Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, 57(3):102087.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Nauros Romim, Mosahed Ahmed, Md Saiful Islam, Arnab Sen Sharma, Hriteshwar Talukder, and Mohammad Ruhul Amin. 2021a. Hs-ban: A benchmark dataset of social media comments for hate speech detection in bangla. *arXiv preprint arXiv:2112.01902*.

Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam. 2021b. Hate speech detection in the bengali language: A dataset and its baseline evaluation. In *Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020*, pages 457–468. Springer.

Sourav Saha, Jahedul Alam Junaed, Arnab Sen Sharma Api, Nabeel Mohammad, and Mohammad Ruhul Amin. 2023a. Vio-lens: A novel dataset of annotated social network posts leading to different forms of communal violence and its evaluation. In *Proceedings of the 1st International Workshop on Bangla Language Processing (BLP-2023)*, Singapore. Association for Computational Linguistics.

Sourav Saha, Jahedul Alam Junaed, Nabeel Mohammed, Sudipta Kar, and Mohammad Ruhul Amin. 2023b. Blp-2023 task 1: Violence inciting text detection (vitd). In *Proceedings of the 1st International Workshop on Bangla Language Processing (BLP-2023)*, Singapore. Association for Computational Linguistics.

Prerona Tarannum, Md Arid Hasan, Firoj Alam, and Sheak Rashed Haider Noori. 2023. Z-index at checkthat! 2023: Unimodal and multimodal checkworthiness classification.

Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, Lynne Edwards, et al. 2009. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB*, 2(0):1–7.