VacLM at BLP-2023 Task 1: Leveraging BERT models for Violence detection in Bangla

Shilpa Chatterjee*
IIT Kanpur
shilpa20@iitk.ac.in

P J Leo Evenss*
IIT Kanpur
leoevenss20@iitk.ac.in

Pramit Bhattacharyya*
IIT Kanpur
pramitb@cse.iitk.ac.in

Abstract

This study introduces the system submitted to the BLP Shared Task 1: Violence Inciting Text Detection (VITD) by the VacLM team. In this work, we analyzed the impact of various transformer-based models for detecting violence in texts. BanglaBERT outperforms all the other competing models. We also observed that the transformer-based models are not adept at classifying Passive Violence and Direct Violence class but can better detect violence in texts, which was the task's primary objective. On the shared task, we secured a rank of *12* with macro F1-score of *72.656%*.

1 Introduction

In the age of digital empowerment, microblogging sites and social media have ushered in a new era of unfettered expression, providing a global stage for individual voices to be heard like never before. However, this newfound freedom of speech has a darker side, one characterized by the rampant spread of hate speech, cyberbullying, and the toxic dissemination of prejudice across various online platforms. As the digital landscape evolves, so too does the challenge of striking a balance between enabling free expression and curbing the rising tide of online hostility. In this digital dichotomy, the need for innovative solutions to detect and combat hate speech in multiple languages has never been more pressing.

While significant progress has been made in identifying hate speech in languages with more resources, Bangla, despite being spoken by nearly 230 million people across the globe and characterized by its linguistic richness and diversity, faces a substantial shortage of computational resources, language models, annotated datasets and efficient methodologies needed for effective natural language processing(NLP) tasks. Transformer-based models that provide state-of-the-art results in various downstream tasks in European languages lag for Bangla (Bhattacharyya et al., 2023). In this paper, we tried to analyze the impact of transformer-based models on detecting violent inciting text (Saha et al., 2023b) (Saha et al., 2023a), specifically aiming to categorize communal violence on social media platforms in the Bangla language worldwide. BanglaBERT outperforms all the other competing models with a macro F1-score of 72.65% which helped us to secure a rank of *12* on the shared task. We observed that the transformer-based models misclassify Passive Violence as Direct Violence but there performance enhances in detecting violence in texts.

2 Related Works

Numerous methods have been proposed to effectively detect offensive and hateful statements across various platforms, primarily relying on traditional machine learning (ML) techniques, which heavily depend on manual feature engineering. However, ML-based approaches exhibit lower accuracy and also need to improve on scalability issues (Karim et al., 2020). In contrast, methods based on neural networks, particularly deep neural networks (DNNs), have the capability to learn more abstract features directly from raw text.

Prominent DNN architectures, including convolutional neural networks (CNN), long short-term memory (LSTM)(Staudemeyer and Morris, 2019), and gated recurrent unit (GRU) (Zhang et al., 2018), have their advantages. Some approaches have amalgamated CNN and LSTM into a unified network known as convolutional LSTM (ConvLSTM) (Karim et al., 2020). These hybrid models (Karim et al., 2020) have demonstrated superior classification accuracy compared to only neural networks. Additionally, pre-trained word embeddings, such as fastText (Grave et al., 2018) and Word2Vec (Mikolov et al., 2013), have been employed in conjunction with CNN, LSTM, or GRU

[*] These authors contributed equally to this work

Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), pages 196–200
December 7, 2023 ©2023 Association for Computational Linguistics

in recent years (Zhang et al., 2018). It's important to note that the majority of these methods have primarily been designed for well-resourced languages like English. Consequently, research in NLP for many underresourced languages, such as Bangla, is still in its early days.

In recent times, language models based on transformers, such as Bidirectional Encoder Representations from Transformers (BERT) built on attention mechanism and the Robustly Optimized Pretraining Approach (RoBERTa) (Liu et al., 2019), have achieved remarkable success in a multitude of natural language processing (NLP) tasks emerging as a natural and highly effective option for addressing the challenges in low-resource languages like Bangla. Other transformer-based language models such as GPT (Brown et al., 2020), DistilBERT (Sanh et al., 2019), ALBERT (Lan et al., 2019), ELECTRA (Clark et al., 2020) has also been proposed for Bangla. For Indian languages, including Bangla, multilingual BERT such as XLM-R (Conneau et al., 2020), multilingual BERT (mBERT) (Pires et al., 2019), IndicBERT (Kakwani et al., 2020) and MuRIL (Khanuja et al., 2021) are available. BanglaBERT (Bhattacharjee et al., 2022), BanglishBERT (Bhattacharjee et al., 2022), sahajBERT (Diskin et al., 2021) are BERT models made specifically for Bangla. BanglaBERT outperforms all other transformer-based models. BanglaBERT was trained on an extensive 40GB dataset derived from various internet sources, such as news articles, web discussions, blogs, government publications, TED Talks, subtitles, newspapers, and articles by crawling data from the web.

This naturally led us to choose BERT over other techniques, and we anticipated that incorporating additional training data could further enhance our approach.

3 DataSet

The dataset provided for BLP Shared Task 1 (Saha et al., 2023b) comprises YouTube comments primarily from social media discussions related to the nine most significant violent incidents in the Bengal region (encompassing Bangladesh and West Bengal) within the past decade. This dataset is characterized by its content being in the Bangla language, with individual comments extending up to 600 words. The dataset is categorized into three classes. They are:

1. **Direct Violence:** Explicit threats directed towards individuals or communities, including actions such as killing, rape, vandalism, deportation, desocialization (threats urging individuals or communities to abandon their religion, culture, or traditions), and resocialization (threats of forceful conversion) falls under this category. Earliest detection of direct violence is crucial because of its potential to yield severe consequences in future.

2. **Passive Violence:** Derogatory language, abusive remarks, slang targeting individuals or communities and justification for violence fall under this category.

3. **Non-Violence:** General conversational topic not involving any form of violence falls under this category.

The training dataset comprises 2,700 samples, with an allocation of around 15% for direct violence, 34% for passive violence, and the remaining 51% for non-violence instances. In the development dataset, which includes 1,330 samples, 15% pertain to direct violence, 31% to passive violence, and 54% to non-violence occurrences.

4 Dataset Preparation

We used the pre-trained models to train on the given dataset and then evaluate the test data provided. Since the training dataset only had around 2.7K sentences, we augmented the training dataset by integrating an additional dataset obtained from (Karim et al., 2020), consisting of 30,000 examples, with 10,000 categorised as violence. Our approach involved annotating these 10,000 hate speech examples into direct and passive violence categories. This annotation was done manually based on our observation from the original dataset, where sentences containing slang were classified as direct violence. We first cleaned different unicode characters to prepare the dataset and removed punctuations from the sentences (Bhattacharyya et al., 2023). Our next task involved identifying the top 200 words that contributed significantly to the direct violence class (directList) and the passive violence class (passiveList) from this original dataset.To form the top 200 word list, we first removed stop words from the original dataset and created a word dictionary consisting of each word and its count of occurrence,one word dictionary for each of the classes - direct and passive violence. We then selected the top 200 words that contributed to each of the classes. Subsequently, we compiled a corpus comprising slang words in the Bangla language. For each hate speech example within the new additional dataset, we assessed its likelihood of belonging to either the direct violence or passive violence class. If the sentence contained any word from the slang word corpus, it was immediately classified as direct violence. Otherwise, we evaluated each word in the sentence against the lists directList and passiveList. If a word was found in either of these lists, the corresponding score for direct violence or passive violence was incremented by 1. In cases where a word appeared in both directList and passiveList sets, the scores for both classes were incremented by 1.

The final classification was determined based on which class had the higher score. In instances of a tie, we labelled the example as passive violence. In this way, all of the 10,000 hate speech examples were categorised into direct and passive violence.

To maintain the same class proportions as the original dataset, with non-violence at 51%, passive violence at 34%, and direct violence at 15%, we selected an appropriate number of samples from each class in the newly annotated dataset.

5 Baseline Systems

We have used several pretrained models, for the BLP workshop Task 1 (Saha et al., 2023a).

5.1 MuRIL

Multilingual Representations for Indian Languages(MuRIL) supports 16 Indian languages and English and have shown significant gain over mBERT. So we selected MuRIL as our first baseline. We used pretrained MuRIL from Hugging Face.

5.2 IndicBert

IndicBert from Ai4bharat(Doddapaneni et al., 2022) was another choice for a baseline system. IndicBert supports 23 indic languages and english. It is a vanilla BERT which has been trained on IndicCorp with the MLM objective.

5.3 BanglishBert

BanglishBERT (Bhattacharjee et al., 2022) achieves state-of-the-art zero-shot cross-lingual transfer results in many of the NLP tasks in Bangla. It is an ELECTRA discriminator model which has been pretrained with the Replaced Token Detection (RTD) objective on large amounts of Bangla and English corpora.

5.4 BanglaBert

Our next system uses a pretrained BERT model which has been trained specifically on Bangla dataset, which fits perfect for the task in hand.BanglaBert(Bhattacharjee et al., 2022) can be used for a variety of tasks like sentiment classification, Named Entity Recognition,Natural Language Inference etc. and thus served perfect for our Violence Inciting Text Detection (VITD) task.

5.5 Results

We first finetuned the pre-trained models – MuRIL, IndicBert, BanglishBert and BanglaBert on the training dataset provided and evaluated it on the test set. We cross-validated the hyperparameters and found that the best for a batch of 16 with Adam optimizer cross-entropy loss works the best for the task. The learning rate was set at $5 * 10^{-}5$. In addition, we combined the additional dataset to the train set, finetuned the same set of models, and evaluated them using the same metric. Results of these experiments are shown in Table 1. BanglaBERT outperforms all the other models for the task on the original dataset. It is also observed from Table 1 that adding new training points confused models more between Passive and Direct Violence classes, thereby degrading the F1-score.

Dataset Type	Model	F1-Score
Original	MuRIL	0.7026
Original	IndicBert MLM	0.7172
Original	BanglishBert	0.7239
Original	BanglaBert	**0.7265**
Augmented	MuRIL	0.6916
Augmented	IndicBert	0.6723
Augmented	BanglishBert	0.6939
Augmented	BanglaBert	0.7065

Table 1: Macro F1-Score of the models used on Test Data

On analysing the results, we observed multiple instances where the same words were used in various classes. For example, we observed that most sentences where the word “gajaba” was used denoted Passive or Direct Violence in the train set, however, it denoted Non-Violence in the development set. We also observed that there were occurrences of similar-meaning sentences labelled differently. "mithyā kathā āra kata balabē" and "ēi mēyēṭā mithyā kathā balachē" are similar meaning sentences but the former is labelled as "Non-Violence" whereas the later is labelled as "Passive Violence". These ambiguous words and sentences resulted in misclassification by the models, degrading the F1-score.

Our analysis also revealed that the models misclassified the Passive Violence class as the Direct Violence class. To confirm this claim, we further conducted an experiment that merged both direct and passive violence into a violence class, mapping it to a binary class classification problem. It was observed that the F1 scores of the models improved significantly. Table 2 reports the macro F1-score of different models on the binary classification task. It can thus be concluded that the models are good at detecting violence, which was the primary objective of the task.

Model	Macro F1-Score
IndicBERT MLM	74.26%
MuRIL	76.35%
BanglaBERT	81.86%

Table 2: Performance of models in detecting violence and non-violence texts.

6 Conclusion

We tried to leverage transformer-based models for violence detection in Bangla for the BLP shared task 1. Our analysis shows that the transformer-based models are not adept at segregating Direct Violence from Passive Violence but are good at detecting violence-inciting text. We would like to develop models that can accurately classify Passive violence in the future.

7 Limitations

Our approach suffers from the lack of a large number of data points essential for transformer-based models. Even after incorporating additional data, we acknowledge that this dataset is relatively small for such a vast language base. A substantial challenge arises from the need for suitable word embeddings for Bangla as used in social media, as the language used in social media significantly diverges from print media, featuring a multitude of misspellings, grammatical errors, and more. Furthermore, a significant portion of users frequently mix both Bangla and English in various contexts. The performance of transformer-based models on such data points lags for a low-resource language like Bangla.

References

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad Uddin, Kazi 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 North American Chapter of the Association for Computational Linguistics: NAACL 2022*.

Pramit Bhattacharyya, Joydeep Mondal, Subhadip Maji, and Arnab Bhattacharya. 2023. Vacaspati: A diverse corpus of bangla literature.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Michael Diskin, Alexey Bukhtiyarov, Max Ryabinin, Lucile Saulnier, quentin lhoest, Anton Sinitsin, Dmitry Popov, Dmitry V. Pyrkin, Maxim Kashirin, Alexander Borzunov, Albert Villanova del Moral, Denis Mazur, Ilia Kobelev, Yacine Jernite, Thomas Wolf, and Gennady Pekhimenko. 2021. Distributed Deep Learning In Open Collaborations. In *Advances in Neural Information Processing Systems*, volume 34, pages 7879–7897. Curran Associates, Inc.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreyansh Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2022. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages. *ArXiv*, abs/2212.05409.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Md. Rezaul Karim, Bharathi Raja Chakravarti, John P. McCrae, and Michael Cochez. 2020. Classification benchmarks for under-resourced bengali language based on multichannel convolutional-lstm network. In *7th IEEE International Conference on Data Science and Advanced Analytics (IEEE DSAA,2020)*. IEEE.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual Representations for Indian Languages.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A Lite BERT for Self-supervised Learning of Language Representations.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Sourav Saha, Jahedul Alam Junaed, Nabeel Mohammed, Sudipta Kar, and Mohammad Ruhul Amin. 2023a. 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.

Sourav Saha, Jahedul Alam Junaed, Maryam Saleki, Arnab Sen Sharma, Mohammad Rashidujjaman Rifat, Mohamed Rahout, Syed Ishtiaque Ahmed, Nabeel Mohammad, and Mohammad Ruhul Amin. 2023b. 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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter.

Ralf C. Staudemeyer and Eric Rothstein Morris. 2019. Understanding lstm – a tutorial into long short-term memory recurrent neural networks.

Ziqi Zhang, D. Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network.