Mavericks at BLP-2023 Task 1: Ensemble-based Approach Using Language Models for Violence Inciting Text Detection

Saurabh Page[*] , **Sudeep Mangalvedhekar**[*], **Kshitij Deshpande**[*],
Tanmay Chavan[*] and **Sheetal Sonawane**[*]
Pune Institute of Computer Technology, Pune
{saurabhpage1,sudeepm117,kshitij.deshpande7,chavantanmay1402}@gmail.com,
sssonawane@pict.edu

Abstract

This paper presents our work for the Violence Inciting Text Detection shared task in the First Workshop on Bangla Language Processing. Social media has accelerated the propagation of hate and violence-inciting speech in society. It is essential to develop efficient mechanisms to detect and curb the propagation of such texts. The problem of detecting violence-inciting texts is further exacerbated in low-resource settings due to sparse research and less data. The data provided in the shared task consists of texts in the Bangla language, where each example is classified into one of the three categories defined based on the types of violence-inciting texts. We try and evaluate several BERT-based models, and then use an ensemble of the models as our final submission. Our submission is ranked 10th in the final leaderboard of the shared task with a macro F1 score of 0.737.

1 Introduction

In today's digital age, numerous social platforms play an important role in connecting individuals around the world. However, certain malicious elements resort to using these platforms to instigate riots, protests, and disturbances that lead to violence. The online posts and comments involve direct threats pertaining to resocialization, vandalism, and deportation while indirect threats involve derogatory language and abusive remarks. These texts which are thought to be a potential reason for instigating violence are called violence-inciting texts. Classifying them has become a major challenge and various techniques are used to implement it. These applications can be used to monitor social media websites and take precautions to avoid any mishaps. Thus the task boils down to text classification wherein we need to label such texts into predefined categories.

The shared tasks involve performing sentiment analysis and text classification. The BLP Workshop offers two shared tasks namely, Violence Inciting Text Detection (VITD) (Saha et al., 2023a) and Sentiment Analysis of Bangla Social Media Posts (Hasan et al., 2023). Our team, under the name *Mavericks* contested in the VITD task under the Codalab username *kshitij*. Our paper illustrates work on the VITD task where we have to classify text into predefined categories of violence. The dataset consists of text in the Bangla language with a length of up to 600 words.

Transformer-based models (Vaswani et al. (2023)) such as BERT (Devlin et al. (2019)) have brought revolution in NLP-related tasks and have proved their worth by attaining state-of-the-art (SOTA) results on several benchmarks (Lan et al., 2020). Large Language Models (LLMs) are increasingly used for text classification tasks (Liu et al., 2019). We use several transformer-based pre-trained models to achieve higher performance. Furthermore, we use ensembling techniques to produce better results. We present our results after experimenting with several models and ensembling techniques.

2 Related Work

Pang et al. (2002) considers classifying documents by overall sentiment and not just by topic. The three machine learning methods - Naive Bayes, Maximum Entropy Classification, and Support Vector Machines did not perform well on sentiment analysis. Warner and Hirschberg (2012) describes the definition of hate speech as the collection and annotation of hate speech corpus along with a mechanism for detecting some commonly used methods of evading common "dirty word" filters. Hammer (2014) automatically detects threats related to violence using machine learning methods. 24,840 sentences obtained from YouTube comments were manually annotated and were used to

[*] Equal contribution

Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), pages 190–195
December 7, 2023 ©2023 Association for Computational Linguistics

train and test the machine learning model. They suggest that the features that combine main words and the distance between those in the sentence attain the best results.

Hassan et al. (2016) provides a textual dataset in Bangla and Romanized Bangla language which can be directly used for sentiment analysis. The dataset was tested using Long Short Term Memory (LSTM) a type of Deep Recurrent Model. Two types of loss functions are used- binary cross-entropy and categorical cross-entropy. Emon et al. (2019) used Linear Support Vector Classifier (LinearSVC), Logistic Regression (Logit), Random Forest (RF), Artificial Neural Network (ANN), and Recurrent Neural Network with an LSTM cell. A Deep-learning-based algorithm using RNN beats all other algorithms by gaining the highest accuracy 82.20%.

In 2017, “Attention is all you need”(Vaswani et al. (2023)) introduces the concept of Transformers which transformed the Natural Language Processing (NLP) landscape. The paper introduced the concept of self attention. In 2019, a new language model called BERT (Bidirectional Encoder Representations from Transformers) was put forward by Devlin et al. (2019). BERT is designed to pre-train deep bidirectional representations from the unlabelled text by joint conditioning on both the left and right context in all layers. Pre-trained BERT can be used for numerous tasks including text classification by fine-tuning it.

Nuryani et al. (2023) proposes a BERT-based method for Aspect-based Sentiment Analysis that can identify and handle conflicting opinions. The method achieves better results on three-class and four-class classification tasks. Sarker et al. (2022) performs sentiment analysis of book reviews in Bangla. A dataset consisting of 5189 reviews was produced by crawling data. An investigation of several deep neural network models and three transformer models is performed. XLM-R outperforms all models, achieving a weighted F1-score of 88.95% on the test data. Anan et al. (2023) performs sarcasm detection using BERT and achieved 99.60% accuracy. A new dataset "BanglaSarc", consisting of comments from Facebook and YouTube was used. Prottasha et al. (2022) utilizes a deep integrated model "CNN-BiLSTM" for enhanced performance of decision-making in text classification.

Dataset	Number of Samples
Training	2700
Development	1300
Testing	2016

Table 1: Dataset statistics.

3 Data

We use the *Vio-Lens* dataset provided by Saha et al. (2023b) for the task. The dataset consists of YouTube comments related to nine violent incidents in the Bengal region (Bangladesh and West Bengal) within the past ten years. The comments are in the Bangla language with a length of up to 600 words. The dataset consists of two attributes: text, and label. The "text" column contains comments while the "label" column contains three values 0, 1, and 2 representing Non-Violence, Passive Violence, and Direct Violence respectively. The training dataset consists of 2700 samples out of which approximately 15% depict direct violence, 34% portray passive violence and the remaining 51% represent non-violent instances. The development dataset consists of 1330 samples out of which approximately 15% illustrate direct violence, 31% depict passive violence and the remaining 54% represent non-violent instances. The test dataset provided at the time of evaluation consists of 2016 samples as seen in Table 1.

4 System

This shared task discusses the problem of Violence Inciting Text Detection. This issue falls under the category of classification, for which transformer-based models have seen extensive application and have demonstrated outstanding performance. As a result, we use and experiment with a variety of such models and ensembling techniques in our research. In the section below, the approaches have been briefly discussed.

4.1 BERT-based Models

Khanuja et al. (2021) discusses how even the state-of-the-art models do not perform satisfactorily well in Indian languages and summarises the gaps found. To mitigate these gaps, they propose their model "MuRIL"[1] which is trained in 16 different Indian languages and English. As we deal with the Bangla

[1] Model link: https://huggingface.co/google/muril-base-cased

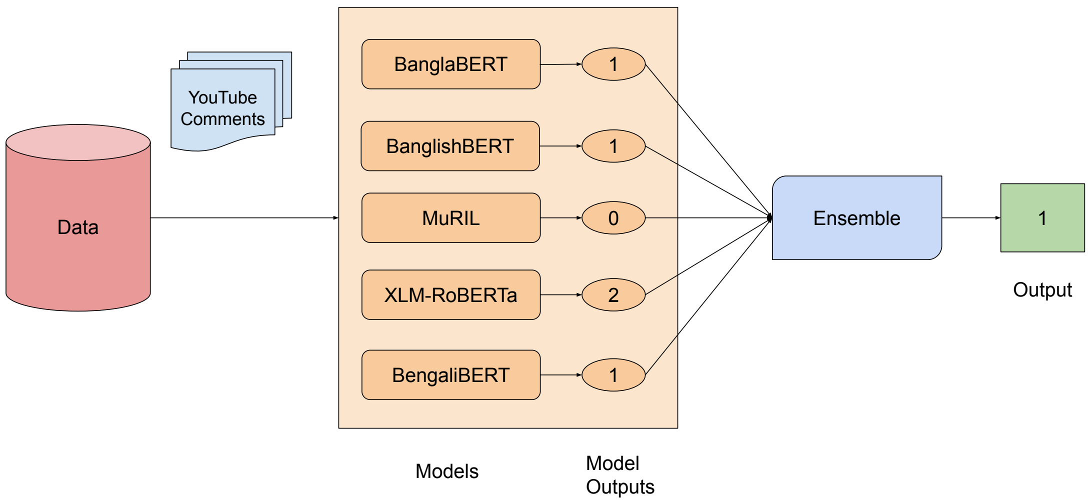

Figure 1: System Architecture

language in this task, MuRIL is specifically relevant. It is trained on two learning objectives, first - Masked Language Modeling, and second - Translation Language Modeling. The model has 236M parameters and a vocabulary of 197285.

Joshi (2022) states that even though multilingual BERT models are suitable for very low-resource languages, models trained on a single language outperform it when sufficient resources for a language are available. Based on this assertion, they propose several models for different languages. BengaliBERT[2] is of specific interest to us. Existing multilingual models are fine-tuned on the Bangla language corpus to create this model.

Conneau et al. (2019) demonstrates how cross-lingual understanding can be improved by pre-training multilingual models on a large scale. XLM-RoBERTa[3] is pre-trained on 2.5TB of filtered CommonCrawl data, which included 100 different languages. It is trained with the multilingual Masked Language Modeling objective. We use the base-sized model in our experiments, XLM-RoBERTa-base which has 270M parameters.

4.2 ELECTRA-based Models

In Bhattacharjee et al. (2022), authors propose training Large Language Models on a dataset specifically tailored for pre-training transformer models useful for Natural Language Processing tasks in the Bangla language. The authors observe that instead of the Masked Language Modeling(MLM) pre-training approach used to train BERT-based models, using ELECTRA and its Replaced Token Detection (RTD) objective provides significant performance improvements while at the same time using significantly less compute power for pre-training. Two Large Language Models are pre-trained namely BanglaBERT[4] and BanglishBERT[5].

The dataset used for pre-training these models was collected by crawling 110 Bangla websites. The total size of the dataset is 27.5GB consisting of 5.25 million documents.

BanglaBERT, introduced in Bhattacharjee et al. (2022), is trained using the ELECTRA pre-training approach consisting of a 12 layer Transformer encoder with 768 embedding size and 12 attention heads. The batch size used is 256 and it is trained for a total of 2.5M steps.

BanglishBERT introduced in Bhattacharjee et al. (2022), is a bilingual model trained on Bangla and English data. It acts as the generator model in the pre-training phase of the ELECTRA approach. BERT pretraining corpus is used along with Bangla data which is upsampled to have equal participation of both languages.

BanglaBERT outperforms other multilingual models such as mBERT and XLM-R (base) on

[2] Model link: https://huggingface.co/l3cube-pune/bengali-bert

[3] Model link: https://huggingface.co/xlm-roberta-base

[4] Model link: https://huggingface.co/csebuetnlp/banglabert

[5] Model link: https://huggingface.co/csebuetnlp/banglishbert

Model	Pre-Training Approach	Macro F1-Score
BanglaBERT	**ELECTRA (RTD)**	**0.791**
BanglishBERT	ELECTRA (RTD)	0.742
MuRIL	MLM	0.753
XLM-RoBERTa	MLM	0.743
BengaliBERT	MLM	0.739
Ensemble - Hard Voting	**-**	**0.782**

Table 2: Results on the development dataset.

a Bangla-specific benchmark introduced by the authors - Bangla Language Understanding Benchmark (BLUB). BanglaBERT achieves impressive results while having better convergence and thus being more compute-efficient than other previously pre-trained multilingual models.

The batch size used for training all the models is 16. The learning rate used is $1e$-5. We use the AdamW optimizer and the Cross-Entropy Loss. We train the models for 10 epochs. All of the models we use in the experiments are freely available on HuggingFace. We have tagged the models with their respective HuggingFace model links in the footnotes. We use the tokenizers recommended by the model developers provided along with the HuggingFace models.

5 Ensembling

Ensembling is a technique that combines the results of various models to generate the system's eventual result. Statistical as well as non-statistical methods are used for this purpose. Ensembling is useful as it helps generate results that are better than the results given by the individual models.
Amongst several methods leveraged for ensembling, we use the "hard voting" ensemble technique. In hard voting, the majority vote or the "mode" of all the predictions is selected as the final prediction. It helps improve the robustness of the system and minimizes the variance in the results. The ensembling mechanism is illustrated in figure 1.

In the post-evaluation phase, we experiment with the weighted ensemble keeping in mind the varied performances of the underlying models. We give higher weights to the models which perform better. We experiment with different weights for models and choose the weights which provide the best results. We also explore different subsets of the 5 mentioned models and form an ensemble of the models to generate predictions. However, the ensembles of the subsets did not provide improvements to our system's predictions.

Model	Pre-Training Approach	Macro F1-Score
BanglaBERT	**ELECTRA (RTD)**	**0.733**
BanglishBERT	ELECTRA (RTD)	0.662
MuRIL	MLM	0.720
XLM-RoBERTa	MLM	0.705
BengaliBERT	MLM	0.690
Ensemble - Hard Voting	**-**	**0.737**
Weighted Ensemble	**-**	**0.745**

Table 3: Results on the test dataset.

6 Results

This section discusses the findings of our experiments. Table 3 contains our results for the models and ensembles. The macro F1 score is the shared task's official score statistic for the Violence Inciting Text Detection task.

BanglaBERT achieves the best result with a macro F1 score of 0.733 among the individual models as seen in table 3. This performance can be attributed to the fact that BanglaBERT is trained on a carefully curated dataset of the Bangla language, unlike other multi-lingual models such as MuRIL and XLM-RoBERTa whose training corpus consists of numerous other languages. It also uses the ELECTRA approach for pre-training which involves using the Replaced Token Detection (RTD) objective instead of the Masked Language Modeling (MLM) objective used in other multilingual BERT models; this allows BanglaBERT to achieve a better performance whilst also converging faster. The performance of MuRIL and XLM-RoBERTa is limited by the quantity and quality of Bangla text they used in pre-training, although it is worth noting that the models will perform much better in a multilingual setting.

BanglaBERT performs marginally better on the development dataset as seen in Table 2, than the ensemble of the five mentioned models but underperforms on the test dataset. We can attribute this slight difference to variations in the performance of individual models on different data samples and the ensemble's stable and high performance across different data samples. We chose ensembling as the final approach for our final submission owing to its better generalizability and low variance in its predictions. Our final submission to the task using the hard voting ensembling mechanism achieves a macro F1 score of 0.737.

Our post-evaluation phase experiments yielded better results with the weighted ensembling technique. The weighted ensemble achieves a macro F1 score of 0.745 on the test dataset, thus outperforming the hard voting-based ensembling approach.

7 Conclusion

We present our approach for the shared task in the First Workshop on Bangla Language Processing through this paper. We experiment with several BERT and ELECTRA-based models as a part of our efforts. We observed that the ELECTRA-based BanglaBERT model has the best performance, followed by MuRIL. We can see that the ELECTRA-bSased models have similar performances compared to their BERT-based counterparts, despite being smaller in size. Our final submission consists of predictions generated by ensembling the evaluated models and has a macro F1 score of 0.737, placing us tenth on the shared task leaderboard. Our experiments have shed light on several further avenues for improvement. Larger pre-training datasets are required for better low-resource models. More sophisticated ensembling techniques can better utilize the performance of individual models and need to be researched further.

Acknowledgment

We thank the Pune Institute of Computer Technology's Computational Linguistics Research Lab for providing us guidance and support necessary for the work. We are grateful for their help.

Limitations

The models that have been utilized are compute-intensive and thus can pose a challenge in real-world applications. Also, it must be considered that the pre-training and evaluation datasets, although of high quality, might possess certain implicit biases and thus might not fully model real-world situations.

References

Ramisa Anan, Tasnim Sakib Apon, Zeba Tahsin Hossain, Elizabeth Antora Modhu, Sudipta Mondal, and MD. Golam Rabiul Alam. 2023. Interpretable bangla sarcasm detection using bert and explainable ai.

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*.

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. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Estiak Ahmed Emon, Shihab Rahman, Joti Banarjee, Amit Kumar Das, and Tanni Mittra. 2019. A deep learning approach to detect abusive bengali text. In *2019 7th International Conference on Smart Computing Communications (ICSCC)*, pages 1–5.

Hugo Lewi Hammer. 2014. Detecting threats of violence in online discussions using bigrams of important words. In *2014 IEEE Joint Intelligence and Security Informatics Conference*, pages 319–319.

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

A. Hassan, M. R. Amin, N. Mohammed, and A. K. A. Azad. 2016. Sentiment analysis on bangla and romanized bangla text (brbt) using deep recurrent models.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

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. 2020. 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.

Nuryani Nuryani, Ayu Purwarianti, and Dwi Hendratmo Widyantoro. 2023. Identification of conflict opinion in aspect-based sentiment analysis using bert-based method. In *Proceedings of the 2022 International Conference on Computer, Control, Informatics and Its Applications*, IC3INA '22, page 276–280, New York, NY, USA. Association for Computing Machinery.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques.

Nusrat Jahan Prottasha, Abdullah As Sami, Md Kowsher, Saydul Akbar Murad, Anupam Kumar Bairagi, Mehedi Masud, and Mohammed Baz. 2022. Transfer learning for sentiment analysis using bert based supervised fine-tuning. *Sensors*, 22(11).

Sourav Saha, Jahedul Alam Junaed, Maryam Saleki, Mohamed Rahouti, Nabeel Mohammed, 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.

Gobinda Chandra Sarker, Kazi Md Sadat, and Aditya Das. 2022. Book review sentiment classification in bangla using deep learning and transformer model. In *2022 4th International Conference on Sustainable Technologies for Industry 4.0 (STI)*, pages 1–6.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada. Association for Computational Linguistics.