Score_IsAll_You_Need at BLP-2023 Task 1: A Hierarchical Classification Approach to Detect Violence Inciting Text using Transformers

Kawsar Ahmed, Md Osama, Md Sirajul Islam, Md Taosiful Islam, Avishek Das
and **Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u1804017, u1804039, u1804116, u1804041}@student.cuet.ac.bd
{avishek, moshiul_240}@cuet.ac.bd

Abstract

Violence-inciting text detection has become critical due to its significance in social media monitoring, online security, and the prevention of violent content. Developing an automatic text classification model for identifying violence in languages with limited resources, like Bangla, poses significant challenges due to the scarcity of resources and complex morphological structures. This work presents a transformer-based method that can classify Bangla texts into three violence classes: direct, passive, and non-violence. We leveraged transformer models, including BanglaBERT, XLM-R, and m-BERT, to develop a hierarchical classification model for the downstream task. In the first step, the BanglaBERT is employed to identify the presence of violence in the text. In the next step, the model classifies stem texts that incite violence as either direct or passive. The developed system scored 72.37 and ranked 14th among the participants.

1 Introduction

Social media and the internet have become crucial components of daily interactions. They can quickly spread information to millions of people. Thus, identifying and categorizing aggressive texts on social media is paramount in maintaining online safety, fostering positive digital interactions, and preventing dissemination of harmful or offensive content (Sharif and Hoque, 2022). Real-world problems like relational anger or even violence are significant problems since threats and insults made online can occasionally result in actual hurt. To keep us secure and calm, we must address this issue because it can make an impact in the short term and also in the long term on the victims (Ta et al., 2022). Different regions' governments try to prevent violations and ensure the safety of the nation's citizens due to social media (Kumar et al., 2021).

The BLP Shared Task 1, Violence Inciting Text Detection (VITD), was launched to address this problem (Saha et al., 2023a). This work presents us with the challenge of devising effective methods to identify diverse types of violent content within the text. The primary objective is to detect and avoid violence from internet remarks. The data used for this task was gathered from YouTube comments on violent incidents that have taken place in the Bengal region (Bangladesh and West Bengal) over the past ten years. We have tried to solve this problem of violence-inciting text detection with two significant contributions.

- Employed a hierarchical classification approach for detecting and classifying violent texts using transformer-based models.
- Explored the model's efficacy in detecting and categorizing violence-inciting texts through the developed model.

2 Related Work

Detecting violence-inciting text has become increasingly crucial in natural language processing. Numerous studies have already focused on identifying hate speech and aggression on social media comments (Badjatiya et al., 2017). Mustakim et al. (2022) employed classify emotions in Tamil text XLM-R model obtained the highest macro f1-score of 0.33. Riza and Charibaldi (2021) detected emotions in Twitter text using the LSTM and achieved an accuracy of 73.15% with both Word2Vec and FastText embeddings. This corpus was used in the DA-VINCIS (Ta et al., 2022) for detecting aggressive and violent incidents on Spanish social media. To train users' tweets on their text embeddings from previously learned transformer models, they employed a multi-task learning network and achieved the best f_1 of 74.80%. Plaza-Del-Arco et al. (2021) applied the transformer-based model to identify hate speech in Spanish tweets. Sharif et al. (2020) proposed a machine learning-based model that classifies Bangla texts into non-suspicious and

Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), pages 185–189

suspicious categories. This work attained the highest accuracy (84.57%) for the SGD classifier with TF-IDF features.

Sharif and Hoque (2020) developed a corpus containing 2000 texts and used several machine learning techniques (such as LR, NB, SVM, KNN, and DT) to classify the suspicious Bangla texts, where LR gained the best performance (accuracy=92%). Hossain et al. (2022) proposed a dataset (MUTE) containing 4158 memes with Bangla captions for identifying hateful memes, and they obtained the maximum f_1-score of 0.672 with the VGG16+Bangla-BERT model. Sharif et al. (2022) introduced a Bangla aggressive text dataset (M-BAD). Using a transformer-based technique (Bangla-BERT), they achieved top scores of 92% in identifying aggressive texts. A recent study (Sharif and Hoque, 2022) introduced a Bangla aggressive text dataset (BAD). Using a weighted ensemble of m-BERT, distil-BERT, Bangla-BERT, and XLM-R, they achieved top scores of 93.43% (coarse-grained) and 93.11% (fine-grained) in identifying and categorizing aggressive Bangla texts. As far as we are concerned, the research has yet to be conducted on identifying and classifying violence-inciting texts in Bangla. This work exploited a transformer-based model to detect violence-inciting texts and classify them into direct, passive, and non-violence targets.

3 Task and Dataset Descriptions

The task organizer developed a benchmark corpus for the shared task 1 (Saha et al., 2023a). To perform the violence-inciting text classification, this task developed a dataset called *Violence Inciting Text Detection (VITD)* corpus[1] consisting of 6046 texts and 20199 unique words. This task focuses on classifying Bangla texts inciting violence into three categories: direct (DVio), passive (PVio), and non-violence (NVio). The definition of each class is illustrated in the following:

- **Direct violence (DVio):** This category encompasses texts that explicitly convey threats, thereby falling under the umbrella of direct violence.
- **Passive violence (PVio):** This violence pertains to texts that use abusive or derogatory language.
- **Non-violence (NVio):** This class is characterized by discussions conducted through texts that do not involve any form of violence in their content.

The VITD dataset (Saha et al., 2023b) was divided into training (2700 texts), validation (1330 texts), and test sets (2016 texts) for training and evaluation purposes. Table 1 shows the summary of the dataset statistics.

Table 1: Distribution of the dataset, where $\mathbf{W}_T$ denotes the total words.

Classes	Train	Valid	Test	$\mathbf{W}_T$
DVio	389	196	201	13071
PVio	922	417	719	38959
NVio	1389	717	1096	53838
Total	2700	1330	2016	105868

The dataset contains uneven distribution among the classes. The direct (contained 786 texts) and the passive (2058 texts) classes have fewer samples than the non-violence class (3202 samples). The maximum length of the data is 110 words, whereas the minimum and average data length are one and 18 words, respectively.

4 Methodology

This work exploited three pre-trained transformer-based models, XLM-R, BanglaBERT, and m-BERT, for classifying violence inciting text in Bangla. Specifically, we have used the 'xlm-roberta-base' (Conneau et al., 2019), 'csebuetnlp/banglabert' (Bhattacharjee et al., 2022) and 'bert-base-multilingual-cased' (Devlin et al., 2018) from Huggingface transformers[2] library and fine-tuned on the dataset. Figure 1 illustrates the schematic process of the proposed system.

4.1 Training

Instead of using the direct ternary classification method, we have used a hierarchical classification approach. In the first step, we split the dataset into two classes: 'violence' and 'non-violence.' The 'violence' class included text related to 'DVio' and 'PVio.' We finetuned 'Model 1' to differentiate between 'violence' and 'non-violence' classes. In the second step, we used the samples related to 'Direct violence' and 'Passive violence' to finetune 'Model 2'. All model's hyperparameters are tuned with the

[1] https://github.com/blp-workshop/blp_task1

[2] https://huggingface.co/docs/transformers/index

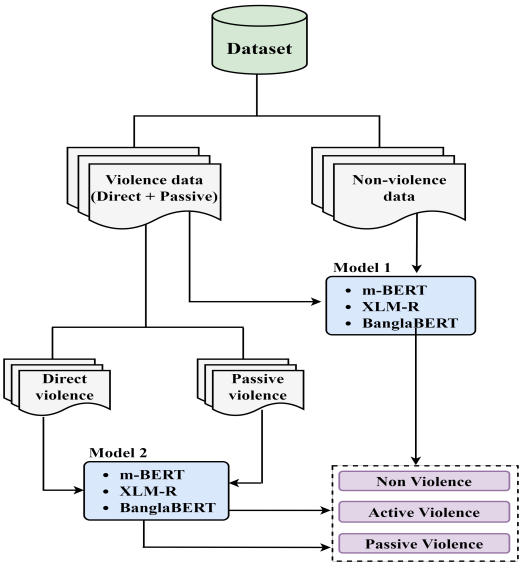

Figure 1: Schematic process of the proposed system.

training dataset. Table 2 shows the tuned hyperparameters of employed models. For BanglaBERT, we set a learning rate of 5e-05 for both Models 1 and 2. We executed training for Model 1 over eight epochs with a batch size of 32, while for Model 2 used five epochs with a batch size of 8.

4.2 Testing and Prediction

If Model 1 classified a text as a 'violence' category, then Model 2 determines whether it is DVio or PVio. This hierarchical process helps us understand different aspects of violent text more effectively. Finally, the results from these two steps have been merged to get the final evaluation score. The predictions of Models 1 and 2 can be expressed by Eqs.1-2.

$$Y_{i_{logits}} = BERT(x) \tag{1}$$

$$M_{i_x} = \frac{e^{Y_{i_{logits_x}}}}{\sum_{p=1}^{p+1} e^{Y_{i_{logits_x}}}} \tag{2}$$

$$if\ M_{1_{(X=Vio)}}:$$
$$Prediction(X) := M_{2_{(X=DVio\ or\ PVio)}}$$

$$else:$$
$$Prediction(X) := NVio$$

The BERT model analyzes the input text x, yielding a result called Y_{logits}. We used a classification head that classifies the Y_{logits} using the softmax activation function into violence (Vio) and non-violence (NVio) classes. M1 (Model 1) represents the probability of violence (Vio) or non-violence (NVio). Subsequently, M2 (Model 2) assesses the likelihood of direct violence (DVio) and passive violence (PVio) within the subset categorized by M1 as violence (Vio). This two-step process helps refine the classification of violence in the text.

5 Results

The assessment of the models' performance relies on the macro F1-score (MF1) as a primary metric. In addition, we incorporated precision (P) and recall (R) metrics for analysis. Table 3 represents the performance of the employed models.

The evaluation encompassed BanglaBERT, XLM-R, and mBERT models in single-step multiclass classification. Among these models, the BanglaBERT achieved the highest macro f_1 (MF1) score, reaching 56.45. BanglaBERT emerged as the top-performing model in the hierarchical framework, surpassing all others with an impressive MF1 score of 72.37. Additionally, it is worth highlighting that the results in the hierarchical approach demonstrated a remarkable improvement of almost 28% over the single-step method.

5.1 Error Analysis

An extensive error analysis has been conducted, offering both quantitative and qualitative assessments. This in-depth examination furnishes valuable insights into the operational efficacy of the proposed model. We conducted a comprehensive quantitative error analysis on the proposed model, employing the confusion matrix depicted in Figure 2.

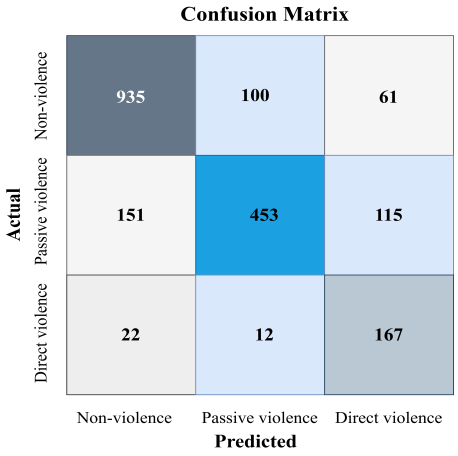

Figure 2: Confusion matrix of the top-performing model (BanglaBERT).

The proposed model misclassified 151 instances

Table 2: Summary of tuned hyper-parameters

Hyperparameters	XLM-R		BanglaBERT		m-BERT	
	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
Optimizer	Adam	Adam	Adam	Adam	Adam	Adam
LR scheduler	Linear	Linear	Linear	Linear	Linear	Linear
Learning rate	2e-05	3.20e-05	5e-05	5e-05	1e-05	1e-05
Epochs	10	4	8	5	5	5
Batch size	16	32	32	8	16	16

Table 3: Performance comparison of various models on the test set.

Approach	Classifier	P	R	MF1
Single-step	m-BERT	61.51	54.36	56.11
	XLM-R	45.03	48.92	46.65
	BanglaBERT	78.12	55.81	**56.45**
Hierarchical	m-BERT	61.52	65.73	61.90
	XLM-R	66.60	69.83	67.62
	BanglaBERT	71.08	77.13	**72.37**

of PVio as NVio and 115 instances of PVio as DVio. Additionally, the model erroneously labeled 100 NVio texts as PVio. These findings shed light on a notable difficulty faced by the model in distinguishing between PVio and NVio. We posit that the primary contributing factor to this challenge could be the class imbalance nature within the dataset.

Figure 3 illustrates a few predictions by the proposed model.

Text Sample	Predicted	Actual
Sample1: মাইজদী - চৌমুহনী - ফেনী মন্দিরে হামলা নিয়ে রিপোর্ট করুন। (Report on the attack on Maizdi - Chaumuhuni- Feni temple.)	NVio	NVio
Sample2: বিবিসি হলো সত্য কে বিনষ্টকারী আর মিথ্যা কে গ্রহণকারী।(The BBC is the destroyer of truth and the acceptor of falsehood.)	PVio	PVio
Sample3: বুধবার কি তোরা মারা গেছিলি বিবিসি বাংলা!(Did you die on Wednesday BBC Bangla!)	NVio	PVio
Sample4: আমরা হিন্দু রা কুরআন পূজা করি না। এটা সম্পূর্ণ চক্রান্ত, সঠিক বিচার চাই (We Hindus do not worship the Quran. This is a complete conspiracy, we want a fair trial.)	PVio	NVio
Sample5: শিক্ষা প্রতিষ্ঠান এ হিজাব নিশিদ্ধ হোক (Hijab should be banned in educational institutions)	DVio	DVio

Figure 3: Few instances of the predicted results generated by the proposed model.

Notably, the proposed model accurately forecasts text samples 1, 2, and 5, aligning with their labels. In contrast, text samples 3 and 4 are challenging as they are not accurately classified. Text sample 3 is erroneously categorized as NVio when its actual class is PVio, while text sample 4 is misclassified as PVio instead of its actual class, NVio. These prediction disparities may be attributed to class imbalance concerns, mainly stemming from the limited number of DVio instances, totaling just 201 samples within the dataset.

6 Conclusion

This paper developed a transformer-based model to address the task of identifying and classifying violence-inciting texts in Bangla. The experimental investigation demonstrated that the BanglaBERT model outperformed the other transformer models (XLM-R and mBERT) by obtaining the highest macro f_1-score (0.72). We plan to investigate the task with the advanced transformer-based model (such as GPT). Additionally, we aim to explore various ensemble techniques of transformers to enhance the model's performance.

Limitations

Model 1 should better identify violence-inciting texts in the proposed two-step hierarchical approach. The success of the entire system hinges directly on the performance of Model 1. If Model 1 fails to deliver accurate results, it will inevitably lead to subparity of the overall system performance. This dependency on Model 1 underscores the critical nature of achieving optimal performance at the initial classification stage, as any shortcomings will adversely affect the outcomes of the developed approach. This limitation emphasizes the need for continuous refinement and enhancement of Model 1 to ensure the effectiveness of the suggested hierarchical system. A fundamental weakness of the proposed solution stems from the imbalanced dataset, with relatively small instances of direct violence (DVio). This imbalance may have influenced the prediction disparities. Additionally, variations

in vocabulary and context within DVio texts, compared to the majority class (NVio), could have contributed to these prediction anomalies. It is worth noting that the dataset's limited size is another constraint.

References

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

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.

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. *CoRR*, abs/1810.04805.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. Mute: A multimodal dataset for detecting hateful memes. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 32–39.

Ritesh Kumar, Bornini Lahiri, and Atul Kr Ojha. 2021. Aggressive and offensive language identification in hindi, bangla, and english: A comparative study. *SN Computer Science*, 2(1):26.

Nasehatul Mustakim, Rabeya Rabu, Golam Md. Mursalin, Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. CUET-NLP@TamilNLP-ACL2022: Multi-class textual emotion detection from social media using transformer. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 199–206, Dublin, Ireland. Association for Computational Linguistics.

Flor Miriam Plaza-Del-Arco, M Dolores Molina-González, L Alfonso Ureña-López, and María Teresa Martín-Valdivia. 2021. A multi-task learning approach to hate speech detection leveraging sentiment analysis. *IEEE Access*, 9:112478–112489.

M Alfa Riza and Novrido Charibaldi. 2021. Emotion detection in twitter social media using long short-term memory (lstm) and fast text. *International Journal of Artificial Intelligence & Robotics (IJAIR)*, 3(1):15–26.

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.

Omar Sharif and Mohammed Moshiul Hoque. 2020. Automatic detection of suspicious bangla text using logistic regression. In *Intelligent Computing and Optimization*, pages 581–590, Cham. Springer International Publishing.

Omar Sharif and Mohammed Moshiul Hoque. 2022. Tackling cyber-aggression: Identification and fine-grained categorization of aggressive texts on social media using weighted ensemble of transformers. *Neurocomputing*, 490:462–481.

Omar Sharif, Mohammed Moshiul Hoque, A. S. M. Kayes, Raza Nowrozy, and Iqbal H. Sarker. 2020. Detecting suspicious texts using machine learning techniques. *Applied Sciences*, 10(18).

Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2022. M-BAD: A multilabel dataset for detecting aggressive texts and their targets. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 75–85, Dublin, Ireland. Association for Computational Linguistics.

Hoang Thang Ta, Abu Bakar Siddiqur Rahman, Lotfollah Najjar, and AF Gelbukh. 2022. Multi-task learning for detection of aggressive and violent incidents from social media. In *Proceedings of the 2022 Iberian Languages Evaluation Forum, IberLEF*.