EmptyMind at BLP-2023 Task 1: A Transformer-based Hierarchical-BERT Model for Bangla Violence-Inciting Text Detection

Udoy Das, Karnis Fatema, Md Ayon Mia, Mahshar Yahan, Md Sajidul Mowla,
MD Fayez Ullah, Arpita Sarker, Hasan Murad
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
u1804{109, 052, 128, 007, 100, 094, 099}@student.cuet.ac.bd, hasanmurad@cuet.ac.bd

Abstract

The availability of the internet has made it easier for people to share information via social media. People with ill intent can use this widespread availability of the internet to share violent content easily. A significant portion of social media users prefer using their regional language which makes it quite difficult to detect violence-inciting text. The objective of our research work is to detect Bangla violence-inciting text from social media content. A shared task on Bangla violence-inciting text detection has been organized by the First Bangla Language Processing Workshop (BLP) co-located with EMNLP, where the organizer has provided a dataset named VITD with three categories: nonviolence, passive violence, and direct violence text. To accomplish this task, we have implemented three machine learning models (RF, SVM, XGBoost), two deep learning models (LSTM, BiLSTM), and two transformer-based models (BanglaBERT, Hierarchical-BERT). We have conducted a comparative study among different models by training and evaluating each model on the VITD dataset. We have found that Hierarchical-BERT has provided the best result with an F1 score of 0.73797 on the test set and ranked 9^{th} position among all participants in the shared task 1 of the BLP Workshop co-located with EMNLP 2023.

1 Introduction

The presence of violent language on social media has significantly increased in recent times which may lead to bigger crime in real life. Violent texts often result in cyberbullying in online communication. Government authorities and social media companies are very much concerned about such a critical issue. A significant amount of previous studies have been conducted on hate speech and toxic, and abusive text detection. The majority of related research works have been done in high-resource languages, like English (Lee et al., 2018). However, little has been done for low-resource languages such as Bangla. Since many social media users prefer using their regional languages, such as Bangla, it becomes a greater challenge to identify violent text content. Difficult lexemes and no specific pattern for tokens make Bangla violent word detection so hard.

Rule-based machine learning methods (Jia et al., 2019, Khalafat et al., 2021) for detecting violent text are considered insufficient nowadays. Therefore, applying rule-based lexical analyzers or parsing methods provides poor performance. Deep learning-based (Castorena et al., 2021) and transformer-based (Arellano et al., 2022,Ta et al., 2022) approaches provide better performance compared to traditional rule-based machine learning methods violence-inciting text detection. Transformer-based approaches have not been utilized for violent text detection in the Bangla language.

The primary objective of this paper is to detect violence-inciting text in Bangla on social media using a hierarchical transformer. The First Bangla Language Processing Workshop (BLP), co-located with EMNLP, has arranged a shared task, introducing a novel dataset called VITD, categorized into nonviolence, passive violence, and direct violence text, for the purpose of detecting Bangla violence-inciting text (Saha et al., 2023a,b).

To achieve this objective, we have employed a diverse range of models, including three machine learning models (RF, SVM, XGBoost), two deep learning models (LSTM, BiLSTM), and two transformer-based models (BanglaBERT, Hierarchical-BERT). We have conducted a comparative analysis by training and evaluating each model on the VITD dataset, ultimately determining that the Hierarchical-BERT model has outperformed the others with an impressive F1 score of 0.73797 on the test set. In the hierarchical-based transformer model, the first BERT model is em-

Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), pages 174–178
December 7, 2023 ©2023 Association for Computational Linguistics

ployed to differentiate between violence and non-violence text while the second BERT is used to classify direct and passive violence text.

The core contributions of our research work are as follows-

- We have developed a hierarchical transformer-based technique for detecting violent text.
- We have conducted a series of experiments on the dataset and provided a comprehensive analysis of their performance outcomes.

The implementation details have been provided in the following GitHub repository - https://github.com/ML-EmptyMind/blp-task1.

2 Related Work

The previous studies on Violence Inciting Text Detection (VITD) can be categorized under machine learning, deep learning, and transformer-based approaches.

Traditional machine learning (ML) techniques have been applied for violence text detection in online social media platforms (Khalafat et al., 2021). Machine learning algorithms such as Support Vector Machine (SVM), Naive Bayes (NB), and K-Nearest Neighbours (KNN) have been utilized where SVM provides the best result for violence-inciting text detection. In another previous work, a lexicon-based method has been used to retrieve violence-related microblogs and then a similarity-based method has been applied to extract sentiment words to detect violent text (Jia et al., 2019) which outperforms the previous SVM methods. However, hierarchically structured categories in text categorization have been used (Krendzelak and Jakab, 2015). They have emphasized the importance of considering the impact of hierarchy on machine learning approaches for improved text classification efficiency.

Compared to the traditional methods used for detecting violence-inciting text (VITD), deep learning (DL) based approaches are less dependent on explicitly defined features. Instead, these models learn patterns and features automatically. A deep learning neural network has been capitalized to detect gender-based violence (GBV) in Mexican tweets (Castorena et al., 2021). They have used techniques like CountVectorizer and a multilayer perceptron to design the model architectures.
A fine-tuned transformer named DistilBETO has been applied to detect aggressive and violent incidents from social media in Spanish (Arellano et al., 2022). Another approach utilizes GAN-BERT to detect violent text in the same dataset (Ta et al., 2022).

3 Dataset

We have utilized the violence detection dataset provided under shared task 1 (VITD) of the BLP Workshop @ EMNLP 2023 (Saha et al., 2023b). This dataset contains three categories Non-Violence, Passive Violence, and Direct Violence. The dataset is divided into three sets train, dev, and test with 2700, 1330, and 2016 samples. Each split contains 16-18 words on average. Each category contains 14-20 words on average. Table 1 shows that the provided dataset is imbalanced. The number of samples under the nonviolence category is considerably high (1389) whereas the number of samples under the direct violence category is significantly low (389).

Split	Nonviolence	Passive Violence	Direct Violence
Train	1389	922	389
Dev	717	417	196
Test	1096	719	201

Table 1: Category-wise distribution in the dataset

As this corpus has been built using YouTube comments, the input text contains several emojis and repeated punctuation. During the training and evaluation phase, several preprocessing steps have been performed on the dataset. We have removed all emojis, punctuation, extra spaces, URLs, ZWNJ, and ZWJ from the input text. However, as the numeric text is vital for semantic analysis, we do not remove the numeric text. At the final step of the preprocessing, we have normalized the text using a popular Bangla text normalizer library (Hasan et al., 2020).

4 Methodology

In this section, we provide an overview of the methods and techniques used on the dataset explained before. Initially, we have extracted features using different extraction techniques and applied various ML and DL algorithms. Moreover, different transformer models have been applied to develop the system shown in Figure 1.

Machine learning based approaches for detecting violence inciting text, we have applied traditional ML-based methods such as Random Forest and Support Vector Machine. We also have used XGBoost as an ensemble classifier to improve the performance. Here, we have used NLTKTokenizer to tokenize the dataset and applied Word2Vec to extract features from the dataset. For SVM, we have chosen the parameter C value of 1 for a soft margin in a hyperplane. For the ensemble method, we have specified the boosting rounds or number of decision trees to n_estimators = 100.

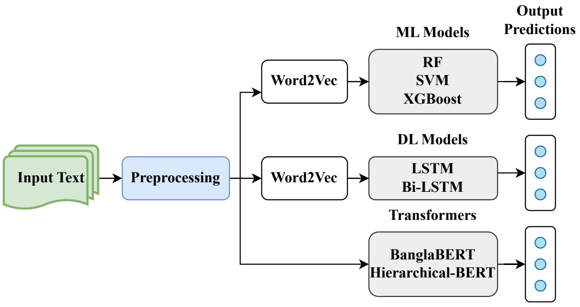

Figure 1: Abstract process of violence text detection

Deep learning-based approaches have been utilized for detecting violence-inciting text. We have implemented two LSTM-based models. In the first model, we have applied three bidirectional LSTM layers with different numbers of LSTM cells. The three directional layers consist of 32, 16, and 8 bidirectional LSTM cells respectively. In the second model, we have used four LSTM layers which consist of 32, 32, 16, and 8 LSTM cells respectively. Both models have been trained up to 10 epochs. We have used categorical cross-entropy as the loss function, and callback method to monitor validation loss during training to select the best model.
Transformer-based approaches are being used very widely in many aspects nowadays. We have employed BanglaBERT(Bhattacharjee et al., 2022) to address this task. As the dataset is imbalanced, we have used the hierarchical approach shown in Figure 2. In the hierarchical approach, we first classify the violence and non-violence text, then further classify the violence text into direct violence and passive violence. For both classification tasks, we have finetuned two BanglaBERT models.

At first, we have divided the dataset into two groups, violence, and nonviolence. Under the violence category, we have assigned the remaining two categories- direct and passive violence. We have finetuned one BanglaBERT model named *nvBERT* to distinguish between violence and nonviolence texts. Then, we have finetuned another BanglaBERT model named *dpBERT* using violence-categorized text which further classifies the violent text into direct and passive violence. At the time of inference, when nvBERT recognizes a text as violence then we give that text as input to dpBERT to determine whether it is direct or passive violence.

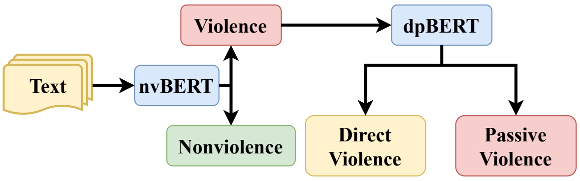

Figure 2: Hierarchical-BERT (HBERT)

Passive and direct violence texts are significantly different from non-violence texts. In passive or direct violence texts, we find the presence of violent words which is not the case for non-violence texts. That is the reason we have selected the combination where first nvBERT finds out the non-violence text and dpBERT finds out the passive and direct violence text.

5 Results and Analysis

In this section, we present performance comparisons among various machine learning, deep learning, and transformer-based approaches.

5.1 Parameter Setting

Table 2 shows parameter settings for different models.

Model	lr	optim	bs	wd	wr
nvBERT	$2e^{-5}$	adafactor	16	0.01	0.1
dpBERT	$2e^{-5}$	adafactor	16	0.01	0.1
BBERT	$6e^{-5}$	adafactor	16	0.01	-
LSTM	$1e^{-3}$	Adam	32	-	-
BiLSTM	$1e^{-3}$	Adam	32	-	-

Table 2: Parameter settings for different models

In Table 2, *lr*, *optim*, *bs*, *wd*, and *wr* represents *learning_rate*, *optimizer*, *batch_size*, *weight_decay*, and *warmup_ratio* respectively. Also, model name BBERT represents BanglaBERT.

Categories	SVM	RF	XGBoost	LSTM	BiLSTM	BanglaBERT	Hierarchical BERT
Nonviolence	0.72	0.72	0.74	0.72	0.78	0.84	**0.85**
Passive Violence	0.43	0.39	0.45	0.48	0.61	0.70	**0.71**
Direct Violence	0.13	0.11	0.30	0.30	0.52	0.65	**0.65**

Table 3: Category wise F1-Score based performance of various systems on test set

5.2 Evaluation Metrics

The performance of various models has been evaluated by calculating the precision (P), recall (R), and F1-Score on the test set.

5.3 Comparative Analysis

We have found that among the machine learning models, the XGBoost has achieved the highest F1 score (0.5). We have trained different deep learning-based models, where the stacked BiLSTM model has provided the best F1-score of 0.633. BanglaBERT has achieved the highest precision of 0.73 whereas Hierarchical-BERT has provided the highest F1-score of 0.73797 respectively. Table 3 highlights the classwise F1-score. Table 4 shows the performance of nvBERT and dpBERT on the test set. The nvBERT model performs slightly better than the dpBERT model in terms of classification. With a high margin compared to ML and DL-based approaches, Hierarchical-BERT performed better and slightly better than BanglaBERT securing 9^{th} rank in the leaderboard.

Classifier	Macro Average		
	P	R	F1
nvBERT	0.83	0.832	0.8310
dpBERT	0.81	0.893	0.8306

Table 4: Performance matrix for both BERTs of Hierarchical BERT on the test set. Here P, R, F1, TF denotes to Precision, Recall, F1-Score, Transfomrer.

	Classifier	Macro Average		
		P	R	F1
ML	RF	0.63	0.41	0.40
	SVM	0.63	0.63	0.43
	XGBoost	0.63	0.50	0.50
DL	BiLSTM	0.63	0.68	0.633
	LSTM	0.39	0.42	0.40
TF	BBERT	**0.75**	0.71	0.730
	HBERT	0.73	**0.79**	**0.738**

Table 5: Performance of various systems on test set

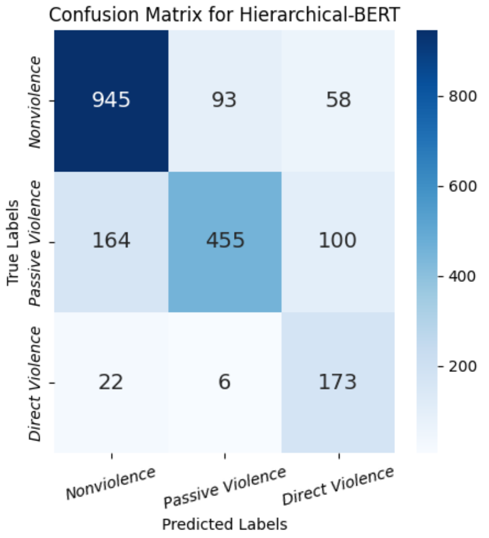

Figure 3: Confusion matrix of Hierarchical-BERT model

5.4 Error Analysis

Table 5 shows that the Hierarchical-BERT model has outperformed all models in terms of classifying violence-inciting text. To get further insights about the system, a confusion matrix (Figure 3) is used. We notice that the model achieves the highest True Positive Rate (TPR) of 86.22% for the nonviolence category and 86.07% for the direct violence category. However, the model provides the lowest TPR of 63.28% for the passive violence category.

Our model has misclassified text of passive violence category as nonviolence or direct violence. Non-violence text does not contain any direct violence words. Passive violence is treated as nonviolence words because often passive violence does not contain any violent words. Thus nvBERT treats passive violence as non-violence text due to the lack of direct violent words. Therefore, it leads to misclassification between non-violence and passive. Table 1 indicates imbalances between three classes and therefore leads to misclassification.

6 Conclusion

In this research work, we have conducted a comparative study among different machine learning, deep learning, and transformer-based models for Bangla violence-inciting text detection in social media content. During the training and evaluation of the different models, we utilized the VITD dataset provided in a shared task. We have found that the Hierarchical-BERT model has outperformed all other models with an F1-score score of 0.73797. The error analysis shows that our trained models become biased toward the majority class. In the future, we will address the issue by incorporating different strategies to address the class imbalance in the VITD dataset.

Limitations

Several limitations can be noted in our work. First, the provided dataset is quite small and highly imbalanced. The impact of the dataset on model development is visible in the result and analysis section. Secondly, our employed model shows limitations in efficiently detecting the category of passive violence text. Future work should explore advanced techniques and the robustness of passive violence text classification.

Ethics Statement

In this study, the tools and technologies used to perform data analysis and development of the model have been ethically and responsively employed. The aim of our work is to develop a system that detects violence-inciting text for the greater good of our society and culture. As per our belief, knowledge should be shared and we are committed to sharing our findings and contributing to the development of violence-inciting text detection in the Bangla language.

References

Luis Joaquín Arellano, Hugo Jair Escalante, Luis Villaseñor Pineda, Manuel Montes y Gómez, and Fernando Sanchez-Vega. 2022. Overview of da-vincis at iberlef 2022: Detection of aggressive and violent incidents from social media in spanish.

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.

Carlos M Castorena, Itzel M Abundez, Roberto Alejo, Everardo E Granda-Gutiérrez, Eréndira Rendón, and Octavio Villegas. 2021. Deep neural network for gender-based violence detection on twitter messages. *Mathematics*, 9(8):807.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2612–2623, Online. Association for Computational Linguistics.

Yun-Fei Jia, Shan Li, and Renbiao Wu. 2019. Incorporating background checks with sentiment analysis to identify violence risky chinese microblogs. *Future Internet*, 11(9):200.

Monther Khalafat, S Alqatawna Ja'far, Rizik Al-Sayyed, Mohammad Eshtay, and Thaeer Kobbaey. 2021. Violence detection over online social networks: An arabic sentiment analysis approach. *iJIM*, 15(14):91.

M. Krendzelak and F. Jakab. 2015. Text categorization with machine learning and hierarchical structures. In *2015 13th International Conference on Emerging eLearning Technologies and Applications (ICETA)*, pages 1–5.

Ho-Suk Lee, Hong-Rae Lee, Jun-U Park, and Yo-Sub Han. 2018. An abusive text detection system based on enhanced abusive and non-abusive word lists. *Decision Support Systems*, 113:22–31.

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.

Hoang Thang Ta, Abu Bakar Siddiqur Rahman, Lotfollah Najjar, and Alexander Gelbukh. 2022. Gan-bert: Adversarial learning for detection of aggressive and violent incidents from social media. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), CEUR Workshop Proceedings*.