the_linguists at BLP-2023 Task 1: A Novel Informal Bangla FastText Embedding for Violence Inciting Text Detection

Md. Tariqzzaman, Md. Wasif Kader, Audwit Nafi Anam
Naimul Haque, Mohsinul Kabir, Hasan Mahmud, Md Kamrul Hasan
Systems and Software Lab (SSL)
Department of Computer Science and Engineering
Islamic University of Technology, Dhaka, Bangladesh
{tariquzzaman,wasifkader,audwitnafi,naimulhaque,mohsinulkabir,hasan,hasank}
@iut-dhaka.edu

Abstract

This paper introduces a novel informal Bangla word embedding for designing a cost-efficient solution for the task "Violence Inciting Text Detection" which focuses on developing classification systems to categorize violence that can potentially incite further violent actions. We propose a semi-supervised learning approach by training an informal Bangla FastText embedding, which is further fine-tuned on lightweight models on task specific dataset and yielded competitive results to our initial method using BanglaBERT, which secured the 7th position with an f1-score of 73.98%. We conduct extensive experiments to assess the efficiency of the proposed embedding and how well it generalizes in terms of violence classification, along with it's coverage on the task's dataset. Our proposed Bangla IFT embedding achieved a competitive macro average F1 score of 70.45%. Additionally, we provide a detailed analysis of our findings, delving into potential causes of misclassification in the detection of violence-inciting text.

1 Introduction

This study details our methods and results for the "Violence Inciting Text Detection (VITD)" task (Saha et al., 2023a), aiming to classify texts into three violence categories: Direct Violence, Passive Violence, and Non-Violence with a goal to identify texts that could lead to further violent actions. Unlike hate speech that targets groups based on attributes, violence-inciting texts advocate harm. The misuse of social media, especially in the Bengal Region, has escalated communal violence (Mathew et al., 2018), with hate speech being a primary cause. This task aims to understand and mitigate such violence.

Our study introduces a unique Bangla FastText(IFT) embedding trained on 3.8 million informal Bangla text samples collected from informal data sources such as Facebook and Youtube comments. We combine this with lightweight ML and DL models like Logistic Regression (LR), SVM, LSTM, BiLSTM, and GRU to detect violence-inciting texts and compare the performance with transformer models such as BanglaBERT, mBERT, XLM-RoBERTa. To the best of our knowledge, this is the first attempt to use FastText embeddings with lightweight models for detecting violence inciting texts in Bangla. Such methods have shown potential in various Bangla text classification methods in previous studies (Kowsher et al., 2022). Our contributions can be summarized as follows:

- An informal Bangla FastText(IFT) embedding trained on 3.8 million sample dataset with better vocabulary coverage on VITD dataset (Saha et al., 2023b) than the existing BanglaBERT's vocabulary.
- A cost-effective solution approach incorporating lightweight classification models and the proposed IFT embedding, that offers 17 times faster training and 1.54 times faster inference speed than BanglaBERT, while having only 4% lower macro-f1 score.
- Performance comparison of lightweight models like LR, SVM, LSTM, BiLSTM, GRU using the proposed IFT embedding with transformer models such as BanglaBERT, XLM-RoBERTa and mBERT.
- Analysis of the classification performance of all the models and how well IFT performs in detecting violence inciting text.

Our work is particularly noteworthy for its development of a versatile Bangla informal FastText embedding, which can have broader implications across various domains like Bangla text classification, token classification, sentiment analysis, etc. Both our informal FastText embedding and the

Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), pages 214–219
December 7, 2023 ©2023 Association for Computational Linguistics

training corpus will be made publicly available to advance Bangla research [1].

2 Related Work

We found several studies that addressed hate speech detection and analysis in under-resourced Bangla language. The concept of utilizing informal word embeddings is derived from the work of Romim et al. (2022) where they discovered that word embeddings generated from informal Bangla texts are quite effective in identifying hate speech in online comments, a finding further reinforced by the work of Karim et al. (2020) using an LSTM model. The potential of developing a Bangla word embedding model from a vast corpus of Bangla news articles and then using these embeddings to classify Bangla document was also discussed in the work of Ahmad and Amin (2016). Romim et al. (2020) presented a hate speech dataset comprising 30,000 user comments, underscoring the efficacy of SVM while observing issues of overfitting in deep learning models when utilizing BengFastText embeddings due to class imbalance. Romim et al. (2022) also introduced a dataset with 50,200 offensive comments, emphasizing linguistic diversity and the challenges of identifying hate speech targets. The study by Islam et al. (2021) focused on sentiment analysis of informally written Bangla texts, emphasizing the challenges posed by this ”noisy” text that includes various dialects, spelling errors, and grammatical inaccuracies. Additionally, it offered insights into the classification performance on informal texts using FastText embeddings.Karim et al. (2021) introduced DeepHateExplainer, where they utilized an ensemble transformer model for explainable hate speech detection, achieving an F1-score of 88%, while acknowledging potential overfitting due to limited dataset. Hate speech in romanized Bangla language on social media platforms was studied by Das et al. (2022). While there has been a considerable number of studies conducted for hate speech detection, notably less research has been dedicated to identify text that incites violence in the Bangla language.

3 Task Description

The primary objective of this task is to detect and categorize threats associated with violence, which have the potential to incite further acts of violence. The task features three distinct categories:

- Direct Violence: Explicit threats targeting individuals or communities, including murder, sexual assault, property damage, forced deportation, desocialization, and resocialization.
- Passive Violence: Violence expressed through derogatory language, abusive remarks, slang, or justifications for violence.
- Non-Violence: Content unrelated to violence, including discussions on social rights or general topics.

3.1 Dataset Description

The dataset (Saha et al., 2023b) employed for this task encompasses YouTube comments about the 9 most significant violent incidents occuring in the Bengal region, which includes both Bangladesh and West Bengal, in the last decade. The dataset contains text written in the Bangla language, with comment lengths of up to 600 words, and it is categorized as either Direct violence, Passive violence, or Non-violence. The dataset consists of the columns “text” and “label”, where the “text” column contains textual data extracted from social media, while the ”label” column assigns each sample a numerical value of 0, 1, or 2, representing non-violence, passive violence, and direct violence accordingly. Table 1 demonstrates a short instance of the dataset.

Label	Category	Example
DV	2	রক্ত যখন দিয়েছি রক্ত আরও দিবো তবুও নিউমার্কেটের আশেপাশে কোনো সাংবাদিকের মাথা না ফাটিয়ে ছাড়বো না ইনশা আল্লাহ!
PV	1	সরকারের সব লোক ভারতের দালাল মনে রাখিছ আল্লাহ ছাড় দেয় কিন্তু ছেড়ে দেয়না
NV	0	একজন বাবা কতোটা অসহায় হলে এই কথা বলতে পারে আল্লাহ তুমি বিচার করো

Table 1: Label Instances of Direct Violence (DV), Passive Violence (PV), and Non-Violence (NV)

[1] https://github.com/Tariquzzaman-faisal/VITD

4 System Description

The System proposed for the VITD shared task is based on IFT embedding that incorporates sub-word information, enabling effective handling of Out-Of-Vocabulary(OOV) words and capturing morphological patterns. We follow a semi-supervised methodology for training where the IFT embedding is created by our collected unlabelled data from social media comments. This embedding is then finetuned on the task specific VITD dataset (Saha et al., 2023b) and incorporated with lightweight models like Logistic Regression (LR), SVM, LSTM, BiLSTM, and GRU models. We carried out extensive experiments to validate the effectiveness of our method and utility of our proposed embedding. Our proposed system is illustrated in Figure 1. The configuration used for LSTM, BiLSTM, and GRU models are included in the Table 3.

4.1 Embedding Dataset Construction

We gather a large informal text dataset of 6.8 million samples from Facebook and YouTube, known sources of Bangla abusive content (Romim et al., 2020). To collect data efficiently, Facepager[2] was employed, using the Facebook Graph API. The preprocessing involves removal of redundant words, symbols, and non-Bangla content, which left us with a streamlined 3.8 million sample dataset. It's coverage on the VITD task's datasets is depicted in Table 2.

Dataset	IFT	BanglaBERT
Train	**58.32%**	35.00%
Dev	**62.45%**	40.49%
Test	**58.35%**	35.82%

Table 2: Vocabulary Coverage on task dataset

$$\text{Coverage} = \frac{|T \cap E|}{|T|} \tag{1}$$

In expression 1, $|T|$ denotes the total count of unique tokens in the task dataset, while $|E|$ represents the dataset of IFT embeddings and BanglaBERT's vocabulary in their respective columns as shown in Table 2. The term $|T \cap E|$ denotes the count of unique tokens common to both datasets. The term "Coverage" represents the proportion of unique tokens in the training dataset covered by the embedding dataset. It is evident that IFT provides better coverage compared to the existing vocabulary of BanglaBERT on this task.

4.2 Experimental Setup

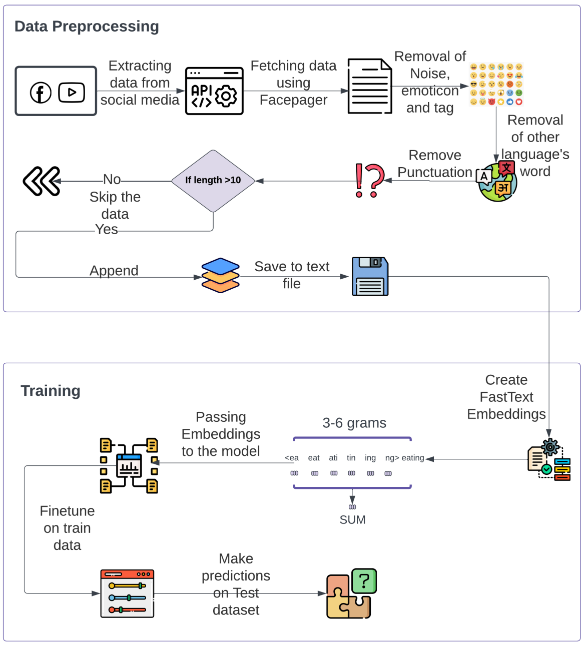
Data Preprocessing
Extracting data from social media
Fetching data using Facepager
Removal of Noise, emoticon and tag
Removal of other language's word
Remove Punctuation
If length >10
No
Skip the data
Yes
Append
Save to text file
Training
Create FastText Embeddings
3-6 grams
<ea eat ati tin ing ng> eating
SUM
Passing Embeddings to the model
Finetune on train data
Make predictions on Test dataset

Figure 1: Methodology of the Proposed System

The collected data is used to train a FastText model using a 300-vector length and character n-grams ranging from 3 to 6. The model employs the Continuous Bag of Words (CBoW) algorithm and specifies a minimum word count of 2 for the training procedure. CBoW is chosen over Skip-Gram because it efficiently learns word embeddings from the context in a more computationally efficient manner, making it faster for training on large datasets (İrsoy et al., 2021). Additionally, CBoW tends to perform better on downstream tasks like text classification when contextual information is not as critical. Its simplicity and ability to handle frequent words effectively make it a practical choice for our use case.

During the training process, the FastText model picked up the ability to represent words as continuous vector representations by taking into account the character n-grams that make up individual words as well as information about their context. The model was able to effectively capture the semantic and syntactic subtleties of the language after it leveraged the subword information and contextual signals that were included within the dataset. Significant consideration is given to the settings of the hyperparameters, which helped

[2] https://github.com/strohne/Facepager

to ensure that an optimal configuration is used, which in turn maximized the embedding's quality and performance. The process is shown in Figure 1. After creating the IFT embedding, it is integrated with LR, SVM, LSTM, BiLSTM, and GRU models. Then the models are trained on the labeled data containing non-violence, passive violence, and direct violence. To check the effectiveness of our proposed IFT embedding, we also train a separate version of each of the models with CBoW embedding. Apart from this, all the configurations are kept similar across these models.

Hyperparameter	Value
Max sequence length	256
Batch size	32
Units	150
Dropout	0.3
Learning rate	0.001
Optimizer	Adam
Loss	SCC
Embedding dim	300

Table 3: Hyperparameters of LSTM, BiLSTM and GRU

5 Results and Findings

Table 4 demonstrates the positive impact of the proposed IFT embedding on model accuracy. For comparison, the accuracy of the transformer models is also presented in the same table. To provide a comprehensive validation of this improvement, we assessed the precision, recall, and F1 scores, as detailed in Table 6. The macro F1 score, which gives equal consideration to each class, provides a holistic view of model performance, guaranteeing a fair assessment that accounts for potential dataset variations. Intriguingly, the BiLSTM model's performance not only aligns with the transformer models but even surpasses mBERT and XLM-RoBERTa in macro-f1 score. Among the transformer models, BanglaBERT emerges as a standout performer, showcasing superior accuracy and F1 scores compared to mBERT and XLM-RoBERTa. This underscores the potential of specialized models tailored for specific languages or regions. Our macro F1 score of BanglaBERT improved to 74.6% as shown in Table 6 due to better tuning of the parameters and the highest accuracy score of 78.67% on test dataset.

If we focus on computational efficiency, table 5 shows the capabilities of our BiLSTM+IFT having an impressive 17 times faster training time than BanglaBERT and faster inference by a factor of 1.54. This remarkable speed, combined with competitive accuracy, positions BiLSTM+IFT as a cost-effective alternative for detecting texts that may incite violence. For clarity, our training spanned 6 epochs with 2,700 samples, while inference was executed on 2,016 samples. All tests were uniformly conducted on Google Colab using a T4 GPU.

Model	Without IFT	With IFT
LR	52.48%	70.29%
SVM	55.06%	72.02%
LSTM	**69.47**%	74.50%
BiLSTM	64.38%	**74.55**%
GRU	69.25%	74.45%
mBERT(base)	71.11%	-
XLM-RoBERTa(base)	72.22%	-
BanglaBERT(base)	78.67%	-

Table 4: Accuracy Comparison with and without the Proposed InformalFastText(IFT) Embedding

Model	Training	Inference
BanglaBERT	532.80	18.46
BiLSTM+IFT	**31.23**	**11.98**

Table 5: Speed comparison between BiLSTM+IFT and BanglaBERT in seconds

Key Observations:

- Incorporating IFT embeddings generally improves the performance across models. This is evident from the higher values in the rows with IFT as compared to their counterparts without IFT in table 4.
- BiLSTM with IFT has a macro F1 score of 70.5%, which is comparable to transformer models. Notably, it outperforms mBERT and XLM-RoBERTa, which have macro F1 scores of 65.8% and 67.4% respectively but falls short of BanglaBERT's 74.6%.
- BanglaBERT has the highest macro F1 score of 74.6% among all models, reinforcing its superior performance as observed in the accuracy Table.

Model	Non-violence			Passive Violence			Direct Violence			Macro Average		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
LR(CBoW)	55.4	85.5	61.2	43.7	15.9	23.4	10.0	2.9	4.6	36.4	34.8	31.7
LR(IFT)	69.8	89.9	78.6	77.0	46.2	57.7	57.6	49.3	53.1	68.1	61.8	63.1
SVM(CBoW)	54.5	97.0	69.8	49.3	3.2	5.9	22.2	1.0	1.9	39.0	33.7	25.9
SVM(IFT)	68.9	94.1	79.5	81.9	44.8	57.9	78.6	49.6	60.6	**76.5**	62.7	66.0
LSTM(CBoW)	69.2	92.0	79.0	79.9	44.8	57.4	62.4	48.8	54.7	70.5	61.8	63.7
LSTM(IFT)	73.5	89.9	80.9	79.8	55.9	65.7	66.7	56.7	61.3	73.3	67.5	69.3
BiLSTM(CBoW)	54.5	99.4	70.4	47.1	0.0	0.0	0.0	0.0	0.0	33.8	33.5	24.2
BiLSTM(IFT)	76.9	84.1	80.4	74.0	63.0	68.1	62.1	63.7	62.9	71.0	**70.3**	**70.5**
GRU(CBoW)	54.4	99.5	70.3	46.2	0.0	0.0	0.0	0.0	0.0	33.5	33.4	24.0
GRU(IFT)	73.9	90.3	81.3	82.1	52.9	64.4	60.3	64.2	61.2	72.1	69.2	69.3
mBERT	77.4	79.7	78.5	74.6	58.4	65.5	43.3	69.7	53.4	65.1	69.3	65.8
XLM-RoBERTa	80.2	80.8	80.5	74.2	57.2	64.6	44.7	79.6	57.3	66.4	72.5	67.4
BanglaBERT	88.0	82.5	85.1	63.3	82.7	71.7	83.1	56.2	67.1	**78.1**	**73.8**	**74.6**

Table 6: Model Performances with and without the Proposed InformalFastText(IFT) Embedding

- BanglaBERT offers the best accuracy and overall performance, while BiLSTM+IFT presents a compelling case as a cost-effective and efficient alternative, especially for applications where speed is crucial as it is 17 times faster in training and 1.54 times in inference for this particular task.

Our empirical findings indicate that the BiLSTM+IFT model exhibits a significant enhancement in performance upon the incorporation of IFT embeddings. Furthermore, this model not only demonstrates a marked cost-effectiveness compared to transformer architectures like BanglaBERT, mBERT, and XLM-RoBERTa, but it also achieves accuracy metrics that are competitive. This underscores the dual advantage of BiLSTM+IFT: its efficiency in computational resources and its competitive accuracy in the realm of NLP tasks.

Observation: The challenge of distinguishing between passive and direct forms of violence is common across models as depicted in Table 6, likely due to the inherent textual similarity in violent content. Models struggle in these areas both with and without IFT embeddings. Yet, the incorporation of IFT embeddings shows a clear enhancement in classifying more challenging categories, supporting our claims of model performance improvement. The confusion matrix in Figure 2 highlights the predictive capabilities of our best model, BiLSTM, in per class classification and aiding a comprehensive analysis of its strengths and weaknesses in differentiating between violence forms.

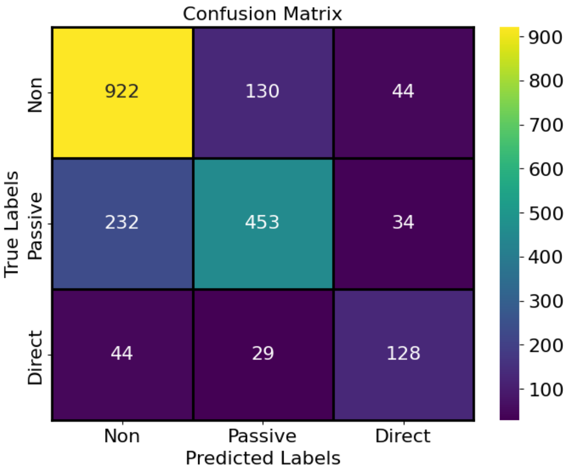

Figure 2: Confusion matrix for BiLSTM+FastText

6 Conclusion

This paper presents a cost-sensitive approach to the detection of violence-inciting text in Bangla using a semi-supervised method. Our results show that applying the proposed IFT embedding to lightweight models produces competitive performance compared to larger transformer models, all while maintaining cost-effectiveness. We believe that enhancing the dataset's size and coverage will lead to improved performance across various aspects when using IFT, thereby broadening the potential applications of our approach to other Bangla text classification tasks.

Limitations

Finding high-quality sources of diverse Bangla hate speech and violence inciting texts was a challenge for us. As a generalized informal embedding dataset, it shows the potential of enhancing the performance of detecting violence inciting texts. However, a larger dataset geared more towards violence inciting texts would yield better results. Furthermore, better bangla text preprocessing tools can also improve the overall scores of all the models. Also, the training data exhibited class imbalance where the neutral label had significantly more samples than the direct label. A more balanced dataset could potentially yield better results.

References

Adnan Ahmad and Mohammad Ruhul Amin. 2016. Bengali word embeddings and it's application in solving document classification problem. In *2016 19th International Conference on Computer and Information Technology (ICCIT)*, pages 425–430.

Mithun Das, Somnath Banerjee, Punyajoy Saha, and Animesh Mukherjee. 2022. Hate speech and offensive language detection in Bengali. 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 (Volume 1: Long Papers)*, pages 286–296, Online only. Association for Computational Linguistics.

Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. 2021. SentNoB: A dataset for analysing sentiment on noisy Bangla texts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3265–3271, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Md. Rezaul Karim, Bharathi Raja Chakravarthi, John P. McCrae, and Michael Cochez. 2020. Classification benchmarks for under-resourced bengali language based on multichannel convolutional-lstm network.

Md. Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Sagor Sarker, Mehadi Hasan Menon, Kabir Hossain, Bharathi Raja Chakravarthi, Md. Azam Hossain, and Stefan Decker. 2021. Deephateexplainer: Explainable hate speech detection in under-resourced bengali language.

Md Kowsher, Md. Shohanur Sobuj, Md Shahriar, Nusrat Prottasha, Mohammad Arefin, Pranab Dhar, and Takeshi Koshiba. 2022. An enhanced neural word embedding model for transfer learning. *Applied Sciences*, 12:2848.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2018. Spread of hate speech in online social media.

Nauros Romim, Mosahed Ahmed, Md Saiful Islam, Arnab Sen Sharma, Hriteshwar Talukder, and Mohammad Ruhul Amin. 2022. BD-SHS: A benchmark dataset for learning to detect online Bangla hate speech in different social contexts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5153–5162, Marseille, France. European Language Resources Association.

Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam. 2020. Hate speech detection in the bengali language: A dataset and its baseline evaluation.

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.

Ozan İrsoy, Adrian Benton, and Karl Stratos. 2021. Corrected cbow performs as well as skip-gram.