SUST_Black Box at BLP-2023 Task 1: Detecting Communal Violence in Texts: An Exploration of MLM and Weighted Ensemble Techniques

Hrithik Majumdar Shibu†, **Shrestha Datta**†, **Zhalok Rahman**†
Shahrab Khan Sami, **Md. Sumon Miah**, **Raisa Fairooz**, **Md Adith Mollah**
Shahjalal University of Science and Technology, Sylhet, Bangladesh
{hrithik11804064, shresthadatta910, rahmanzhalok}@gmail.com
{shahrabkhan6620, iamsumon111, raisafairoozshafa, adibhasan35}@gmail.com

Abstract

In this study, we address the shared task of classifying violence-inciting texts from YouTube comments related to violent incidents in the Bengal region. We seamlessly integrated domain adaptation techniques by meticulously fine-tuning pre-existing Masked Language Models on a diverse array of informal texts. We employed a multifaceted approach, leveraging Transfer Learning, Stacking, and Ensemble techniques to enhance our model's performance. Our integrated system, amalgamating the refined BanglaBERT model through MLM and our Weighted Ensemble approach, showcased superior efficacy, achieving macro F1 scores of 71% and 72%, respectively, while the MLM approach secured the 18th position among participants. This underscores the robustness and precision of our proposed paradigm in the nuanced detection and categorization of violent narratives within digital realms.

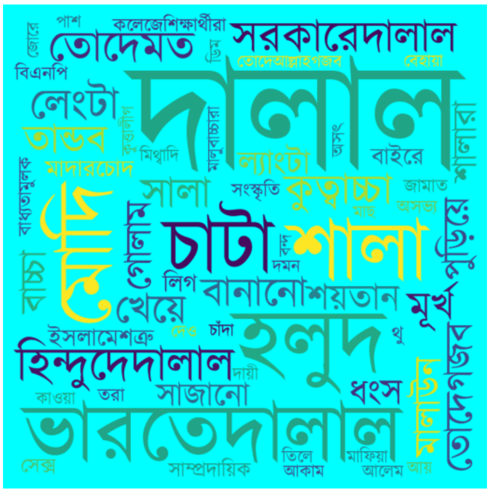

Figure 1: Words after exclusion of words in neutral class and discarding most of the positive and neutral words

1 Introduction

While fostering connections and facilitating information dissemination, social media has inadvertently become a platform for propagating hostility. Such hateful actions, encompassing communal violence, cyberbullying, and social platform attacks disrupt online communities and erode the foundational trust and safety intrinsic to such platforms Romim et al. (2021). By utilizing the latest advancements in artificial intelligence and natural language processing (NLP), we can effectively identify and prevent potential violent incidents, thus creating a safer environment. In this context, we will examine the BLP Shared Task 1: Violence Inciting Text Detection (VITD).

Recent advancements in the field have highlighted the potential of informal text embeddings in enhancing the accuracy of Hate Speech (HS) detection, evidenced by the work of (Romim et al., 2022). Furthermore, the advent of Masked Language Model (MLM) pre-training, exemplified by models such as BERT (Devlin et al., 2018), has revolutionized text classification tasks. The landscape of NLP has been significantly shaped by the adoption of transfer learning in recent years. Pioneering methodologies such as ULMFiT (Howard and Ruder, 2018; Khatun et al., 2020) have demonstrated the superiority of fine-tuning language models over traditional deep learning algorithms, especially when confronted with limited datasets and resources. This paradigm shift is further exemplified by models like BanglaBERT (Bhattacharjee et al., 2021), which builds upon the foundational BERT architecture, benefiting from extensive pretraining on diverse datasets. The burgeoning interest in Bangla text classification has catalyzed the development of several pivotal datasets and transformer-based approaches (Alam et al., 2020; Hasan et al., 2023; Islam et al., 2020), further enriching the ecosystem and setting the stage for our research.

Our approach to the VITD task (Saha et al.,

† These authors have equal contributions

2023a) is informed by these advancements, leveraging Transfer Learning (TL) and MLM training to incorporate informal texts into our models. During training, we have used a large volume of similar informal data collected from various domains (Islam et al., 2021; Kabir et al., 2023; Romim et al., 2022) using domain adaptation along with the VITD dataset. Utilizing our approaches, we have gotten better results than the benchmark models.

2 Dataset

The class distribution of training, validation, and test set of the VITD dataset by Saha et al. (2023b), facilitating the main classification task, is shown in Table 1. Each dataset contains three output classes namely Neutral(N), Passive Violence(PV), and Active Violence(AV).

Labels	Train	Validation	Test
Neutral	1,389	717	1,096
Passive Violence	922	417	719
Active Violence	389	196	201

Table 1: Class distribution of VITD datasets

In total, we have 2,700 instances in the final dataset for training and 1,330 instances for the development set. The mean text length of the instances is 17.51 ± 14.4 as shown in Figure 2 and detailed in Table 2.

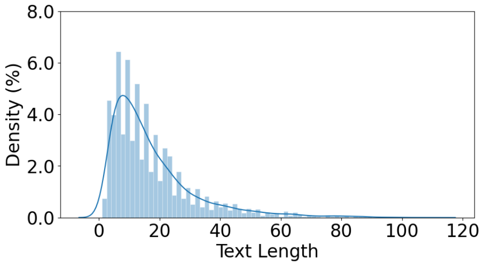

Figure 2: Histogram of text lengths of VITD dataset

Metrics	Values
Maximum Text Length	110
Minimum Text Length	1
Mean	17.51
Standard Deviation	14.4

Table 2: Some relevant metrics related to the length of the VITD dataset texts

In our endeavor to understand the linguistic nuances of the dataset, we constructed a word-cloud Filatova (2016) (as depicted in Figure 1). This was achieved by judiciously excluding words from the neutral class and systematically discarding a majority of the positive and neutral terms. This visualization offers insights into the specific linguistics that warrant detection. A salient observation from our analysis is the dataset's substantial inclusion of informal and colloquial expressions. Notably, such vernacular terms are often absent from the training corpora of widely recognized pre-trained models.

To get a deeper insight into the linguistic traits of the AV and PV classes, all words of the neutral class were excluded from the AV and PV classes. The resultant set of words of AV class and PV class is represented by the closed circular curve on the left and right respectively of the Venn diagram (Figure 3). This AV and PV set consists of 2702 and 8259 words respectively, while the intersection contains 245 words. From the word samples presented in the Venn diagram, the words unique to AV class(excluding the set of PV words from AV set) encompass most of the words that indicate violence of some form, and the words unique to PV class(excluding the set of AV words from PV set) hold most of the words related to dehumanization. While words common to both classes are predominantly linguistically dehumanizing, only a small portion of them consist of violence-inciting words. The ratio of dehumanizing-natured words within the intersection set significantly exceeds the ratio of such words in the exclusive PV class set. In these sets, neutral words also exist in a significant amount.

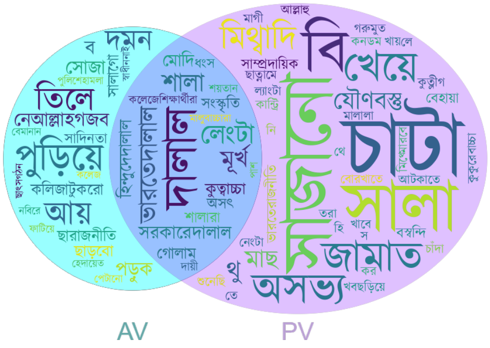

Figure 3: Venn diagram of AV and PV words set excluding neutral class words

3 System description

3.1 Dataset for MLM Training and TL

Task related datasets have been divided into three groups based on their usage, as shown in Table 3.

For MLM training, 9,674 text samples labeled as

'Negative' sentiment from the BanglaBook dataset by Kabir et al. (2023), 6,807 text samples labeled as 'Aggressive' from the BAD dataset by Sharif and Hoque (2022), 24,156 text samples labeled 'Hate Speech' from the BD-SHS dataset by Romim et al. (2022) and negative emotion-sentiment dataset from Alam et al. (2020). This group also includes the VITD (train and dev) dataset. Only the negative class samples have been taken for MLM training.

The BD-SHS containing 50,281 samples labeled 'Hate Speech' or 'Non Hate' has been used for TL.

Used for	Dataset(s)
MLM training	BanglaBook, BAD, BD-SHS, emotion-sentiment
Transfer learning	BD-SHS (with labels)
Fine-tuning	VITD (with labels)

Table 3: Used dataset groups

3.2 Masked Language Model Training

The Masked Language Model (MLM) is a pivotal neural network architecture in NLP that predicts omitted words within sentences. Leveraging hidden tokens minimizes the divergence between predicted and actual words while accounting for bidirectional context. To ensure that the linguistic representations align with specific domains, we have employed domain adaptation techniques to fine-tune the MLM. In our research, we have meticulously adjusted parameters such as learning rates, weight decay, and batch sizes and selectively frozen specific encoder layers for optimization. We have primarily used pre-trained BanglaBERT which is actually the **ELECTRA** model (Clark et al., 2020) for extensive contextual learning through **Masked Language Modeling** from our expansive dataset. In this model, tokens are replaced with feasible alternatives, enabling the model to distinguish between the original and substitute tokens. This discriminator model is quite effective and represents an intriguing development in NLP tasks.

3.3 Fine Tuning Pre-Trained MLM

Leveraging contextual linguistic knowledge, we have fine-tuned the pre-trained ELECTRA model from section 3.2 for improved text classification. Specifically, we froze the Encoder layers of the MLM-trained model to achieve desired classification results. Utilizing the best checkpoint from the pre-trained MLM and minimizing the difference between training and validation loss, we have obtained the highest macro F1 score.

3.4 Transfer Learning from BD-SHS dataset

We employed TL through downstream model training, leveraging the BD-SHS dataset, to train our model on the VITD dataset as our parallel approach for violence detection, which we refer to as TL approach. To address class imbalance, we upsampled the classes PV and AV by iteratively replicating samples until their sizes matched that of class N. To implement TL, initially we finetuned pre-trained BERT-based models with domain-related dataset as described in section 3.4.1. In the next step, as described in section 3.4.2, we further finetuned the model we had trained in the initial step (from section 3.4.1) keeping the embedding layer non-trainable. We utilized the models produced from section 3.4.2 for our validation and test on VITD dataset.

3.4.1 Training Transformers-based Models on BD-SHS Dataset

We have trained the BERT-based models, Monolingual BanglaBERT-base (sagorBERT) (Sarker, 2020), mBERT cased (Devlin et al., 2018), as well as XLM-RoBERTa (Conneau et al., 2019) on BD-SHS dataset keeping training epochs low. For each training epoch, we have randomized the order of our training data and implemented gradient clipping (Pascanu et al., 2013). We fine-tuned the pre-trained BERT variants using the Adam optimizer while limiting the input length to a maximum of 256 tokens. We took outputs experimenting with 1 and 2 layers of multi-head attention, followed by a linear layer as classification head.

3.4.2 Fine-Tuning Models Trained on BD-SHS dataset with VITD Dataset

We have discarded the classification head of the trained models on the BD-SHS dataset and added two tanh-activated nonlinear layers (for sagorBERT) and a linear layer as the new classification head for training on the VITD dataset. In the first training session, we had all the model layers frozen, including the embedding and encoder layers of the models except the classification head, and trained on the VITD dataset. In the second training session, we kept the classification head frozen and unfroze the encoder layers of the previously trained models. We have trained with gradient clipping on the upsampled dataset for both sessions by shuffling the data samples at each epoch.

3.5 Ensemble Approaches

In our study, we have primarily explored two ensemble techniques: Stacking (Wolpert, 1992) and Weighted Ensemble (WE). For stacking, we have incorporated four models: TL-based sagorBERT, mBERT, XLM-RoBERTa, and the MLM-trained BanglaBERT. Utilizing 60% of the VITD dataset's development set, we have trained a deep neural network comprising three non-linear ReLU layers, culminating in a softmax layer. This model was validated against the remaining 40% of the development set and subsequently evaluated on the test set. In our WE approach (Huber and Kim, 1996), we have selected seven models, all evaluated on the VITD dataset:

1. Four models have been trained solely on the training data including TL-based sagorBERT, mBERT, XLM-RoBERTa, and MLM-trained BanglaBERT.

2. Three models have been trained on both the training and validation data encompassing TL-based sagorBERT, mBERT, and XLM-RoBERTa. The optimal hyperparameters for these models were determined through rigorous validation.

To mitigate potential validation data leakage, models trained on both training and validation data were assigned minimal weights. Conversely, the model exhibiting the highest validation macro F1 score was accorded the maximum weight. After experimenting with diverse weight configurations on the validation set, we finalized the weights, opting for the label with the majority consensus.

4 Experimental Setup

We have presented our approach to strengthen a VITD model using a pre-trained MLM which is an ELECTRA model based on Transformers Network (Vaswani et al., 2017) which we have referred to as MLM approach. To facilitate the VITD model, first we have used pre-trained MLM on texts to comprehend contextual representations. The final classification has been done by freezing the 6 encoder layers of the ELECTRA model and fine-tuning the hyperparameters of the model. During both processes, we utilized a learning rate of 2e-5 and ran the model for 10 and 50 epochs respectively in which the epoch with the highest Macro F1-score is stored as the final result. For stacking, we have trained a deep neural network for 21 epochs with a learning rate of 0.03. For WE, we have assigned the fine-tuned BanglaBERT a weight of 3 and other TL-based models a weight of 1. We have used the mini-batch training paradigm for our experiments. Corresponding all the codes are publicly available at this repository.[1]

5 Results

We present our results using the macro F1 score for both the validation and test datasets, as detailed in Table 4. Notably, our SUST_Black Box's approach for BLP-2023 Task 1 achieved the highest macro F1 scores of 0.85 on the validation set and 0.72 on the test dataset. In Table 4, we delineate the methods, models, and their respective performances in terms of the macro F1 score. For the stacking approach, we incorporated TL models and MLM, as discussed in section 3.5. For the Weighted Ensemble (WE) method, TL models trained solely on the training dataset were termed TL models-1. Meanwhile, TL models trained on both the training and validation datasets were denoted as TL models-2. The MLM was assigned a weight of 3, as elaborated in 3.5.

Method	Model	Val	Test
Baseline	BanglaBERT	0.78	0.70
	sagorBERT	0.69	0.63
	mBERT (cased)	0.65	0.63
TL	sagorBERT	0.69	0.65
	mBERT	0.68	0.65
	XLM-RoBERTa	0.67	0.59
MLM	BanglaBERT	0.80	0.71
Stacking	TL models MLM	0.79	0.70
WE	TL models-1 TL models-2 MLM	0.85	0.72

Table 4: Validation and test macro-F1 score of each categorical models

In summary, as depicted in Table 4, our methods, particularly BanglaBERT with our MLM approach and WE, demonstrated superior performance on both validation and test sets. Notably, in both MLM and TL we have fine-tuned and used domain adaptation for linguistic representation. Afterward, we used these models in stacking and WE. In light of this, we discerned that MLM and WE methods spotted an impressive result. From Table 5 we see

[1] Github: https://github.com/Shibu4064/EMNLP

Method	Model	P	R	mF1
TL	sagorBERT	64	66	70
	mBERT	66	67	71
	XLM-RoBERTa	59	65	64
MLM	BanglaBERT	71	76	76
Stacking	TL models MLM	70	75	75
WE	TL models-1 TL models-2 MLM	72	74	76

Table 5: Macro Precision(**P**), Macro Recall(**R**) and Micro F1(**mF1**) score in percentage(%) for each categorical models on test data.

Model	Neutral	Passive	Direct
mBERT(t)	0.76	0.64	0.53
mBERT(t+v)	0.80	0.60	0.55
MLM(t)	0.84	0.67	0.61
RoBERTa(t)	0.73	0.58	0.48
RoBERTa(t+v)	0.77	0.40	0.52
sagorBERT(t)	0.77	0.64	0.53
sagorBERT(t+v)	0.79	0.64	0.53
WE(t+v)	0.84	0.67	0.64

Table 6: Individual class F1 score on test dataset

that MLM and WE methods achieved the highest micro F1 score of 76%, whereas, MLM achieved the best macro recall and WE best macro precision of 76% and 72% respectively. We have also presented individual class F1 scores from different TL and MLM approach models. N, PV and AV (2) these three class F1 scores are proffered in Table 6. Here (t) represents test sets and (t+v) represents both test and validation sets.

6 Discussion

In this section, we present the results of our experiments with MLM and TL methodologies, which have outperformed the base models. The primary reason for this improvement is the inclusion of informal words that were previously absent in the pretraining datasets of the pre-trained models. To further optimize our results, we used ensemble techniques. We prioritized MLM within the Weighted Ensemble (WE) framework by assigning it the highest weight, recognizing its superior accuracy. Interestingly, we found that integer weights of WE predominantly excelled in AV class detection, despite our initial expectations of learned weights from stacking yielding superior outcomes. This also highlights the importance of the inclusion of models in WE trained on both validation and training sets. To improve our outcomes further, we integrated upsampling, which, in certain instances, led to improved outcomes. During training with upsampled data, our approach of freezing the embedding layer throughout the training process and selectively freezing and unfreezing different layers at various stages of training lessens the chance of overfitting. Lastly, being dominated by the majority neutral class, the micro F1 score is considerably higher compared to the macro F1 score. This indicated that, as backed up by individual class F1, the finetuned models were able to classify between neutral and non-neutral classes more rigorously.

7 Conclusion and Future work

Our experiments aimed to explore various approaches to integrating informal words, and we found that the MLM and WE methods performed the best. Our MLM and TL approaches are still unexplored for all BERT baseline models, including exploring based on the same models. Discovering the effects of our approaches and their comparison will lead to promising future research directions and help improve our methods' robustness and scalability. The effect of freezing embedding layers and, selectively freezing and unfreezing other layers on overfitting due to upsampled data still needs in-depth study. As we worked to familiarize pretrained models with the nuances of informal words for the VITD task in Bangla, we hope to contribute to safer online spaces for everyone and unlock new frontiers in NLP.

8 Limitations

Several approaches were applied for the improvement of VITD. However, we encountered challenges such as a highly imbalanced dataset, limited computational resources, and a relatively small dataset size. During MLM training for the MLM approach and downstream model training on the BD-SHS dataset for the TL approach, although increasing the training time helps the models to adapt to the datasets, it also increases the knowledge decay of the models as we are not training with a huge dataset. The initial phases of our approaches also demand a huge amount of data from similar domains. Although freezing parameters reduce the chance of overfitting due to upsampling but the chance still remains.

References

Tanvirul Alam, Akib Khan, and Firoj Alam. 2020. Bangla text classification using transformers. *arXiv preprint arXiv:2011.04446*.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Kazi Samin, Md Saiful Islam, Anindya Iqbal, M Sohel Rahman, and Rifat Shahriyar. 2021. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. *arXiv preprint arXiv:2101.00204*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators.

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.

Olga Filatova. 2016. More than a word cloud. *Tesol Journal*, 7(2):438–448.

Md Arid Hasan, Shudipta Das, Afiyat Anjum, Firoj Alam, Anika Anjum, Avijit Sarker, and Sheak Rashed Haider Noori. 2023. Zero-and few-shot prompting with llms: A comparative study with fine-tuned models for bangla sentiment analysis. *arXiv preprint arXiv:2308.10783*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Gary A Huber and Sangtae Kim. 1996. Weighted-ensemble brownian dynamics simulations for protein association reactions. *Biophysical journal*, 70(1):97–110.

Khondoker Ittehadul Islam, Md Saiful Islam, and Md Ruhul Amin. 2020. Sentiment analysis in bengali via transfer learning using multi-lingual bert. In *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pages 1–5. IEEE.

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.

Mohsinul Kabir, Obayed Bin Mahfuz, Syed Rifat Raiyan, Hasan Mahmud, and Md Kamrul Hasan. 2023. Banglabook: A large-scale bangla dataset for sentiment analysis from book reviews. *arXiv preprint arXiv:2305.06595*.

Aisha Khatun, Anisur Rahman, Md Saiful Islam, Hemayet Ahmed Chowdhury, and Ayesha Tasnim. 2020. Authorship attribution in bangla literature (aabl) via transfer learning using ulmfit. *Transactions on Asian and Low-Resource Language Information Processing*.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318. Pmlr.

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. *arXiv preprint arXiv:2206.00372*.

Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam. 2021. Hate speech detection in the bengali language: A dataset and its baseline evaluation. In *Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020*, pages 457–468. Springer.

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.

Sagor Sarker. 2020. Banglabert: Bengali mask language model for bengali language understanding.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

David H Wolpert. 1992. Stacked generalization. *Neural networks*, 5(2):241–259.