BLP-2023 Task 1: Violence Inciting Text Detection (VITD)

Sourav Saha † ♣, **Jahedul Alam Junaed** † ♣, **Maryam Saleki** ♠, **Mohamed Rahouti** ♠, **Nabeel Mohammad** ♢, **Ruhul Amin** ♠
♣ Shahjalal University of Science and Technology, Bangladesh,
♢ North South University, Bangladesh, ♠Fordham University, USA
{sourav95,jahedul25}@student.sust.edu, *
{msaleki, mrahouti, mamin17}@fordham.edu,
nabeel.mohammed@northsouth.edu

Abstract

We present the comprehensive technical description of the outcome of the BLP shared task on Violence Inciting Text Detection (VITD). In recent years, social media has become a tool for groups of various religions and backgrounds to spread hatred, leading to physical violence with devastating consequences. To address this challenge, the VITD shared task was initiated, aiming to classify the level of violence incitement in various texts. The competition garnered significant interest with a total of 27 teams consisting of 88 participants successfully submitting their systems to the CodaLab leaderboard. During the post-workshop phase, we received 16 system papers on VITD from those participants. In this paper, we intend to discuss the VITD baseline performance, error analysis of the submitted models, and provide a comprehensive summary of the computational techniques applied by the participating teams.

Warning: The paper examples and the corresponding dataset contain violent inciting, derogatory, abusive, and racist comments. .

1 Introduction

Social media's growth over the past decade has reshaped the distribution of information to the broader public (Ferguson et al., 2014). However, it has also surfaced as a potential breeding ground for provoking violence among different groups, from religious to ethnic to gender-based distinctions. In fact, many of the violent incidents of the recent past era can directly or indirectly be attributed to incitement from social media (Mengü and Mengü, 2015). Such platforms can act as catalysts for the incitement of violence and the radicalization of individuals or groups (Recuero, 2015). Extremist ideologies and hate speech can spread rapidly, leading to real-world acts of violence. Acts of violence, triggered or fueled by content shared on social media, can inflict physical harm to individuals and communities with dire consequences that include physical injuries, destruction of properties, and even loss of human lives.

In the recent past, numerous studies were conducted into areas like hate speech detection (Warner and Hirschberg, 2012; Waseem and Hovy, 2016; Davidson et al., 2017; Karim et al., 2020; Romim et al., 2021), abusive content identification (Nobata et al., 2016), and misinformation detection (Shu et al., 2017; Hossain et al., 2020), aiming to understand and prevent harmful social media activities. There have been several workshops that contributed datasets and organized shared tasks on online harmful content detection in different languages (Bosco et al., 2018; Fersini et al., 2018; Zampieri et al., 2019; Basile et al., 2019). However, to the best of our knowledge, there exists no research work on the violence incitement in the Bengal Region (Bangladesh and West Bengal in India), the residence of more than 272 million[1,2] people of many diverse background. Therefore, this shared task seeks to bridge this gap by contributing a novel dataset on VITD for the development of new systems and methodologies with the objective to advance our collective understanding and capabilities in this crucial domain. In this paper, we discuss the following:

1. **Dataset Overview:** VITD task presents an intriguing challenge centered around the catego-

† Authors have equal contributions

[1] https://en.wikipedia.org/wiki/Bangladesh
[2] https://en.wikipedia.org/wiki/West_Bengal

Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), pages 365–375

Category	Definition	Example
Direct Violence	It refers to killing, rape, vandalism, deportation, desocialization, and resocialization.	দোকানে আগুন জ্বালিয়ে দেওয়া উচিৎ (The shop should be set on fire)
Passive Violence	It refers to use of derogatory language, abusive remarks, slang or any form of justification for violence.	সরকারের দোষ, সরকারের দালালি বন্ধ কর (Blame the government, stop the government brokering)
Non-Violence	It refers to discussions about social rights or general conversational topics that do not involve any form of violence.	সত্য প্রকাশে যমুনা টিভিকে ধন্যবাদ (Thanks to Jamuna TV for revealing the truth)

Table 1: The Table depicts examples of 3 different categories: Direct Violence (Red), Passive Violence (Yellow), & Non-Violence (Green). We also show the English translation using Google Translator service.

rization of textual content into three distinct and vital categories: Direct Violence, Passive Violence, and Non-Violence. We discuss how this dataset was prepared for the task.

2. **Baseline Performance:** We present the Macro-F1 score of VITD using both multilingual and Bangla BERT models.

3. **Team Statistics:** We discuss the participant's demographics in terms of gender and background.

4. **Error Analysis:** We present a detailed error analysis of each model submitted by the 27 teams.

5. **Comprehensive System Summary:** We also discuss the computational techniques used by different teams for the shared task.

2 Dataset Overview

The Vio-Lens dataset addresses the challenges of Violence Incitement Text Detection (VITD). It comprises data from YouTube comments related to violent content from Bangladesh and West Bengal. The dataset categorizes violence incitement into three classes: *Direct Violence, Passive Violence, and Non-Violence*. The description of each category along with relevant examples is provided in Table 1. The dataset features 6046 samples: 786 samples for direct violence, 2058 for passive violence, and the remaining 3202 for non-violence. This distribution illuminates a discernible class imbalance within the dataset, underscoring the need for careful consideration when designing and implementing classification algorithms or methodologies. For a detailed description of the Vio-Lens dataset, we refer the reader to the dataset paper Saha et al. (2023)[3].

[3]The dataset is publically available in https://github.com/blp-workshop/blp_task1/tree/main/dataset

3 Task Description and Evaluation

3.1 Task Definition

The shared task provides a classification task on three categories of violence, *Direct Violence, Passive Violence, and Non-Violence*, as discussed below:

- **Direct Violence:** This category encompasses explicit threats directed towards individuals or communities, including actions such as killing, rape, vandalism, deportation, desocialization (threats urging individuals or communities to abandon their religion, culture, or traditions), and resocialization (threats of forceful conversion). The detection of direct violence is crucial due to its potential to have severe consequences in the future.

- **Passive Violence:** This category includes instances characterized by the employment of derogatory language, derogative terms, or abusive remarks aimed at individuals or communities. Moreover, any attempt to rationalize or justify violence is classified within this category. Acknowledging these nuanced forms of hostility is key to understanding the breadth of online aggression.

- **Non-Violence:** Content within this category addresses non-violent matters, ranging from discussions about social rights to general conversations that are free from any violent implications. It's crucial to distinguish these benign exchanges from those that carry a more harmful intent.

3.2 Task Organization

We ran our competition on the CodaLab [4]. platform. There were two primary phases: (i) the Trail

[4]https://codalab.lisn.upsaclay.fr/competitions/14620

phase started on 16 July 2023 and ended on 15 August 2023, and (ii) the Test Phase, which began on 16 August 2023 and ended on 18 August 2023. We provided a training phase with the text and label, while the test phase contained only text data.

Models	F1 Score (Macro)
Majority Voting	23.350
MBERT	63.282
DistillBERT	59.863
XLM-RoBERTa (base)	66.062
BanglaBERT (base)	**71.073**

Table 2: The table shows the outcomes (macro-F1) classification using majority voting, MBERT, DistillBERT, XLM-RoBERTa, and BanglaBERT for the test set. All the experiments used the same dataset and parameters for a fair evaluation. We observe that BanglaBERT achieved the best macro F1 score.

3.3 Evaluation Metrics and Baselines

We evaluated all participating systems with Macro-F1 score. We are providing five baseline models (see Table 2) to benchmark a range of simple to complex systems for VITD. The simplest baseline model is the Majority Baseline, where all the categories are predicted as the majority Non-violence class. We provided four other fine-tuned Large Language models: XLM-RoBERTa (Liu et al., 2019), MBERT (Devlin et al., 2019), DistillBERT (Sanh et al., 2019), and BanglaBERT (Bhattacharjee et al., 2021). The first two are Multilingual models, while the third were monolingual ones. We ran all the models using the following parameters: learning rate 1e-5, train batch size 8, evaluation batch size 8, epochs 50, evaluation steps 250, and early stopping patience 5. Among the four baselines, the monolingual BanglaBERT provided the best Baseline with the highest macro F1 score of 78.791 on the dev set and 71.073 on the test phase.

3.4 Team Statistics

Our contest attracted 27 teams containing members from around the world. Among the contestants, 69 were male and 19 were female (Figure 1). The contest attracted participants including undergraduate students, graduate students, and professionals containing 13 undergraduates majority, 7 graduates majority, and 7 professionals majority teams.

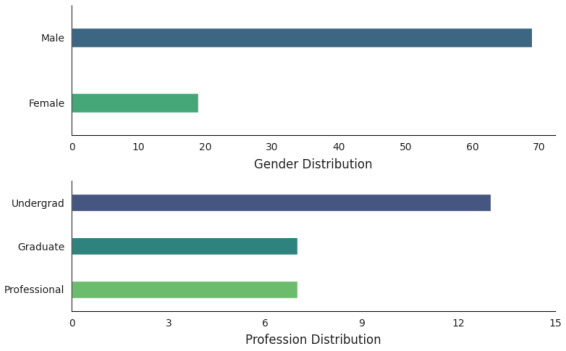

Figure 1: The figure shows gender distribution among the contestants and professions of each category of participants.

4 Participants Results

The baseline model with the best performance, BanglaBERT (Bhattacharjee et al., 2021), was outperformed by 16 teams. We display the ranking and best-performing models performance for each team in Table 3. We also report precision, recall, and F1 score for each category. Team DeepBlueAI achieved the highest overall performance, obtaining the Macro-F1 score of 76.044.

We observe that the highest precision, recall, and F1 score were reported for the **Non-Violence** category and worst on the *Direct Violence* category - indicating potential challenges in identifying explicit content. This may be due to the data imbalances in the dataset. Specifically, *Non-Violence* occupies 51.44%, 53.90%, and 54.37% of data on the train, validation, and test sets, respectively. On the other hand, *Direct Violence* is represented in only 14.41%, 14.74%, and 9.97% of the corresponding sets. In terms of team performance, a total of 20 teams surpassed the benchmark F1 score for the *Direct Violence*, and 17 teams achieved that for *Non-Violence*, while only 11 teams were found to cross the benchmark for *Passive Violence*. In particular, three teams: DeepBlueAI, Aambela, and NLP_CUET, exhibited high F1 scores across all three categories.

4.1 Error Analysis

A total of 27 teams participated in the VITD task. Among the 2,016 test samples, 506 unique samples were accurately predicted by all participating teams. There are a total of 72 samples that were incorrectly predicted by all the 27 teams. Additionally, there are a total of 214 unique samples that were incorrectly predicted by exactly one of the 27

Rank	Team	F1 score (macro)	Direct			Passive			Non-Violence		
			P	R	F1	P	R	F1	P	R	F1
1	DeepBlueAI	76.044	56.811	85.075	68.127	85.634	63.839	73.147	83.800	90.146	86.857
2	Aambela	76.041	59.286	82.587	69.023	84.404	63.978	72.785	82.872	90.055	86.314
3	NLP_CUET	74.587	61.004	78.607	68.696	73.745	71.488	72.599	83.868	81.113	82.468
4	Team Embeddings	74.418	52.761	85.572	65.275	81.122	66.342	72.992	84.755	85.219	84.986
5	Semantics Squad	74.413	57.664	78.607	66.526	81.607	63.561	71.462	82.149	88.595	85.250
6	NLP_BD_PATRIOTS	74.313	54.276	82.090	65.347	78.537	67.177	72.414	85.141	85.219	85.180
7	the_linguists	73.978	54.485	81.592	65.339	80.000	65.090	71.779	83.540	86.131	84.816
8	Panda	73.808	54.430	85.572	66.538	85.655	57.302	68.667	81.870	91.058	86.220
9	EmptyMind	73.797	52.266	86.070	65.038	82.130	63.282	71.485	83.554	86.223	84.868
10	Mavericks	73.699	55.932	82.090	66.532	82.863	61.196	70.400	80.840	87.774	84.164
11	LowResourceNLU	73.468	54.574	86.070	66.795	85.983	57.163	68.672	80.590	89.781	84.937
12	VacLM	72.656	50.286	87.562	63.884	80.536	62.726	70.524	83.183	83.942	83.560
13	LexicalMinds	72.551	51.562	82.090	63.340	83.080	60.779	70.201	81.453	86.953	84.113
14	Score_IsAll_You_Need	72.376	55.805	74.129	63.675	82.163	60.223	69.502	79.624	88.777	83.952
15	winging_it	71.207	45.316	89.055	60.067	83.622	60.362	70.113	83.212	83.668	83.439
16	Semantic_Savants	71.179	51.235	82.587	63.238	82.200	57.163	67.432	79.530	86.496	82.867
–	**Baseline**	**71.073**	**46.690**	**84.081**	**60.033**	**79.680**	**62.732**	**70.194**	**83.271**	**82.663**	**82.970**
17	BpHigh	70.978	53.741	78.607	63.838	80.639	56.189	66.230	78.624	87.591	82.866
18	SUST_Black Box	70.680	47.500	85.075	60.963	83.128	56.189	67.054	81.368	86.861	84.025
19	Team_Syrax	70.450	56.226	74.129	63.948	84.703	51.599	64.131	76.390	91.515	83.271
20	Blue	70.012	45.938	81.592	58.781	82.927	56.745	67.382	81.320	86.588	83.871
21	Team CentreBack	69.390	50.530	71.144	59.091	78.435	57.163	66.130	79.074	87.226	82.950
22	UFAL-ULD	69.009	47.447	78.607	59.176	75.215	60.779	67.231	80.399	80.839	80.619
23	BanglaNLP	68.110	53.650	73.134	61.895	78.602	51.599	62.301	74.646	86.496	80.135
24	KUET_NLP	60.332	36.557	77.114	49.600	75.204	38.387	50.829	76.327	85.310	80.569
25	Shibli_CL	38.427	37.727	41.294	39.430	68.421	01.808	03.523	58.469	94.799	72.329
26	Team Error Point	31.913	08.150	18.408	11.298	31.959	08.623	13.582	63.816	79.653	70.860
27	lixn	31.426	36.000	17.910	23.920	25.000	00.139	00.277	55.126	96.168	70.080

Table 3: The table shows the performance of each team along with the best-performing baseline model (BanglaBERT-base). It contains precision (P), recall (R), and F1 scores of individual categories, and finally a macro F1 score across all categories for final judgment.

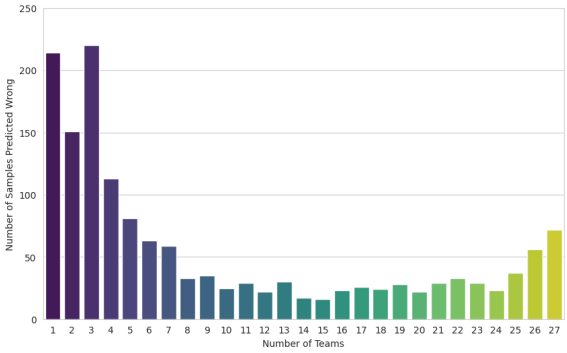

Figure 2: The bar plot shows the number of unique samples (Y-axis) that were predicted wrong by exactly N number of teams (X-axis) out of total 27 teams.

teams. A detailed visualization of these errors can be seen in Figure 2. In summary, a total of 1,510 samples were predicted incorrectly by one or more teams.

For the *Direct Violence* category, out of 201 test instances, 75.05% were predicted accurately by all the teams, while 6.80% were mistakenly identified as *Passive Violence*, and 18.15% were misclassified as *Non-Violence*. The *Passive Violence* test set comprises of 719 samples. Of those, 53.37% were correctly classified by all the teams, while 13.45% were categorized incorrectly as *Direct Violence*, and the rest samples were erroneously categorized as *Non-Violence*. For the *Non-Violence* category, which had 1,096 samples in the test set, an impressive 87.19% were correctly categorized by all the teams. Only 5.54% of those samples were incorrectly identified as *Direct Violence*, and the remaining 7.27% were misclassified as *Passive Violence* (see Figure 3).

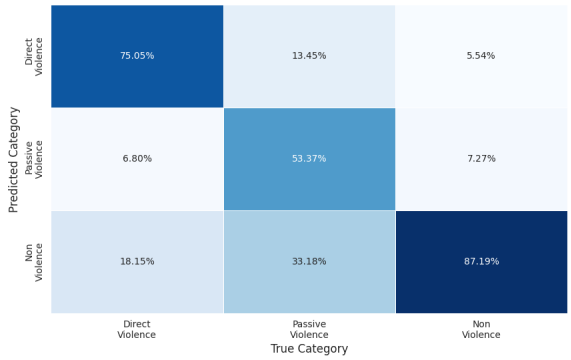

Figure 3: Confusion matrix illustrating category distribution among 27 teams.

We present a few examples from each of the categories, that were predicted wrong by all the teams (see Table 4). For the *Non-Violence* category, no

teams misclassified the same samples, indicating that the category may be easier to predict than the rest.

Example	Category
তোদের মত দাঙ্গাবাজ কুকুরদের বিচার আমি করব (I will judge riotous dogs like you)	Direct Violence
দেখা হবে ভাই মাঠে ময়দানে কোরআন নিয়ে উলটা পালটা কিছু বল্লে (See you in the field if you say something bad about Qur'an)	Direct Violence
ইসলামী আইন অনুযায়ী এই মেয়ের ঘরের মধ্যে আবদ্ধ থাকা উচিত,এ বাইরে কেনো। ইসলামে তো নারীদের যৌন দাসী হিসাবে ব্যবহার করে,এ ঘরের বাইরে গেলে তো ইসলামের অবমাননা করা হয়। (According to Islamic law, this girl should be confined inside the house, why she outside? In Islam, women are used as sex slaves, if she goes out of this house, Islam is insulted.)	Passive Violence
ধর্ম মানেই পাগলামি। সংঘাত, গালাগালি, মারামারি, খুন, ধর্ষন। (Religion means madness, conflict, abuse, fighting, murder, rape.)	Passive Violence

Table 4: This table presents some samples that all the teams predicted wrongfully. It is also to be noted that such wrong predictions were only observed either for *Direct or Passive Violence* categories.

5 Participants System Description

In this section, we present a comprehensive summary of each submitted system for the shared task.

AAmbela (Fahim, 2023) stood second in the competition with an overall Macro-F1 score of 76.040 for the test set. They propose an instruction-finetuned csebuetnlp-BanglaBERT (Bhattacharjee et al., 2022) with three classification heads. As BanglaBERT's vocabulary does not fully cover the tokens in the data, the team added them as special tokens that were learned during the training phase. They also observe the significance of emojis in the dataset, and removing them often leads to a minor result. On the other hand, converting emojis to text and normalizing the text leads to a better result. They experimented with various approaches such as traditional classifiers (SVM, Random Forest, XG-Boost) with Tf-IDF embeddings, Deep learning models (LSTM), and transformer-based architectures (mBERT-case, mDeBerta-v3 base (He et al., 2021a,b), XLM-Roberta base, SagorSarker-BanglaBERT (Sarker, 2020), BanglaBERT (Bhattacharjee et al., 2022). Finally, BanglaBERT trained on three epochs with a batch size of 16 came out on the top.

NLP_CUET (Hossain et al., 2023) achieved 3rd rank in this task with an overall Macro-F1 score of 74.587. They preprocessed data by removing unwanted characters and employed feature extraction methods like TF-IDF and Word2Vec. After investigating several machine learning, deep learning, and transformer-based models, they propose a hybrid method using GAN (Goodfellow et al., 2020) and Bangla-ELECTRA. Here, they considered both labeled data and unlabeled data for model training. The generator and discriminator are both multilayer perceptrons with a single hidden layer of 512 neurons. The generator input is a randomly generated vector of 100 dimensions, and it outputs a fake transformer embedding vector for a single token. The transformer-based model processed the input text, generating a contextualized embedding vector for the CLS token. These embedding vectors from the transformer and generator were then input into the discriminator. The output of the discriminator is extended to K+1 classes where k is the number of classes in this classification task, and the extra class is "REAL." In this approach, they focused on determining whether the embedding produced by the transformer-based architecture is real or fake. During the testing phase, they discarded the generator and used the BERT and discriminator model to classify the input data. They masked the prediction output for the 'REAL' class during testing.

Seamntic Squad (Dey et al., 2023) received the fifth rank with an overall Macro-F1 score of 74.413. They applied a preprocessing step of removing punctuation, lemmatization, and oversampling/undersampling. Afterward, they used different transformer-based models such as XLM-Roberta (base and large), BanglaBERT (Bhattacharjee et al., 2022) (base and large), and mBERT. Among the approaches, BanglaBERT-base achieved the highest result.

nlpBDpatriots (Raihan et al., 2023) received sixth in the competition with a macro f1 score of 74.313. They applied a rigorous data augmentation process, including translation and back-translation to make the dataset 7 times larger. They applied Statistical machine learning models (Linear Regression, Support Vector Machine), GPT-3.5, and various transformer-based approaches. Their two-step approach first classified violence and non-violence with MuRIL(Khanuja et al., 2021), and later XLM-RoBERTa to classify violence and non-violence on the larger dataset performed best.

the_linguists (Tariquzzaman et al., 2023) achieved 7th rank in this task with an overall

Macro-F1 score of 73.978. Firstly they collected 6.8 million data samples from Facebook and YouTube. Then they applied some preprocessing steps which resulted in a refined dataset containing 3.8 million samples. After that, they applied a semi-supervised methodology for training where the training of the informal FastText word embedding model was done by making use of the preprocessed unlabeled data. These embeddings were then integrated into the LR, SVM, LSTM, BiLSTM, and GRU models which were fine-tuned using the labeled data. And they got the best result from BiLSTM.

EmpytyMind (Das et al., 2023b) achieved 9th rank in this task with an overall Macro-F1 score of 73.797. They first preprocessed the dataset and then normalized the text. After that, they applied statistical machine learning-based approaches (Random Forest and Support Vector Machine, XG-Boost), deep learning-based approaches (one three bidirectional LSTM layers and the other four LSTM layers), and transformer-based approaches using a two-step hierarchical approach. In the hierarchical approach, they first classified the text into violence and non-violence categories, then further classified the violence category into direct violence and passive violence to combat the imbalance dataset, and it yielded the best performance.

Mavricks (Page et al., 2023) received 10th place in the competition with an overall Macro-F1 score of 73.699. They applied different transformer-based models (BanglaBERT, BanglishBERT, MuRIL, XLM-Roberta, and BengaliBERT) and ensembled them. They applied different ensembling methods among which hard voting came out on top.

LowResourceNLU (Veeramani et al., 2023) achieved 11th rank in this task with an overall Macro-F1 score of 73.468. Here, they aggregate three BERT-based language models. They configured the first model by incorporating two heads, one for Masked Language Modeling (MLM) and the other for classification, within the BanglaBERT-*large* framework. They used mBERT as their second model. As their third model, they used BanglaBERT-*base* by incorporating two classification heads. The first head focuses on the Bangla version of the XNLI dataset (Conneau et al., 2018). The second head is dedicated to the dataset. Initially, they extracted individual predictions from each model using the argmax function, selecting the class with the highest confidence score for each model. Then they applied another argmax operation, this time on the maximum logit values obtained from each model. Because of the incorporation of MLM in the first model, the F1 score is enhanced by a substantial margin. Similarly, the joint pretraining with XNLI significantly increased the performance of the third model. The combination of three models exhibits superior performance as compared to the use of a single model alone.

VacLM (Chatterjee et al., 2023) ranked 12th on the competition with an overall Macro-F1 score of 72.656. They introduced external information by incorporating data from Karim et al. (2020) and manually annotating them. They observed augmenting data from external sources in this way actually hampers the performance in the 3-way classification task but generally performs better for the violence and non-violence classification task.

Score_Is_All_You_Need (Ahmed et al., 2023) received 14th place in the competition with an overall Macro-F1 score of 72.376. They applied a two-step approach to first classify violence and Non-Violence. Afterward, from the violence category, they classify direct and passive violence using transformer-based approaches. They applied BanglaBERT, M-BERT, and XLM-RoBERTa using an exhaustive hyperparameter search to fit the model.

SUST_Black_Box (Shibu et al., 2023) ranked 18th in the competition with an overall Macro-F1 score of 70.680. They applied to incorporate data from similar sentiment and hate speech-related datasets for data augmentation. They used different transformer-based techniques such as SagorSarker-BanglaBERT(Sarker, 2020), M-BERT, and RoBERTa on the augmented dataset. Finally, they applied different ensembling methods to the augmented dataset.

Team_Syrax (Riyad et al., 2023) received 19th in the competition with an overall Macro-F1 score of 70.450. They applied traditional preprocessing steps such as emoji and punctuation removal. Then, they applied data augmentation from the Bengali hate speech detection dataset (BAD, BD-SHS). They applied different ensemble methods such as bagging and hard majority voting for the classification.

Team CentreBack (Alamgir and Haque, 2023)

ranked 21st in the competition with an overall F1 score of 69.390 in the test set. They applied several approaches using transformer-based architectures (BanglaBERT and XLM-Roberta) and a two-stage approach where they first classified violence and non-violence and then further classified the violence into direct and indirect violence. They also applied a few-shot approach with SBERT but it ultimately resulted in a poor performance. Among those approaches, BanglaBERT (20 epochs received the highest approach with the stage approach closely behind.

UFAL-ULD (Mukherjee et al., 2023) ranked 22nd in the competition with an overall Macro-F1 score of macro 69.009 for the test set. They applied different transformers-based models: XLM-Roberta-base, XLM-Roberta-large, BanglaBERT-Sagor, BanglaBERT-BUET and BanglaBERT-BUET-large. They used focal loss to handle the issue of class imbalance and applied simple data augmentation techniques like synonym replacement, insertion, deletion, swap, and shuffle.

BanglaNLP (Saha and Nanda, 2023) ranked 23rd in the competition with an overall Macro-F1 score of 68.110 for the test set. They used a general paraphrasing technique for data augmentation. In addition using general classification techniques such as logistic regression, SGD classifier, and multinomial naive bayes with ensembling techniques such as majority voting and stacking. They finally used BanglaBERT (Sarker) (Sarker, 2020) and Multilingual-E5-base as transformer-based model, with the later ultimately provided the best performance.

Team Error Point (Das et al., 2023a) ranked 26th with an overall Macro-F1 score of 31.913. They applied different traditional machine learning classifiers along with CNN and LSTM. Their combination of LSTM and CNN achieved the highest performance.

6 Discussion

6.1 Popular Architechture

The large majority of the participants (14 teams) employed transformer-based methods. They used mBERT, mDeBerta-v3 base, XLM-Roberta (*base* and *large*), SagorSarker-BanglaBERT, BanglaBERT (base and large), MuRIL, etc. Notably, variants of BanglaBERT consistently outperformed other models. Several submissions explored statistical machine learning methods leveraging FastText and Word2Vec for word-embeddings and subsequently used SVM, Logistic Regression, and XGBoost for classification. Another popular technique used by some teams is the two-steps approach to first classify the violence and non-violence and then subsequently classify them into *Direct and Passive Violence*. NLP_CUET used a GAN-based architecture. Please see Table 5 for details.

6.2 Popular Methods

Ensembling of different classifiers and transformers is the most prominent method used by the participants. Among the ensembling methods, hard voting gave the best results. Some teams used a two-step approach to classify the violence category and then the direct and passive violence from that category. Some teams tended to add more data to the dataset. They primarily adopted two approaches: One of the approaches included operations on the dataset such as insert, substitution, deletion, translation, and back-translation. The other approaches included datasets from similar datasets such as the Bangla Hate Dataset (Romim et al., 2021), and XNLI Dataset (Conneau et al., 2018), etc.

6.3 Insights

Generally, most of the successful process has been monolingual pre-trained language model modified with various task-specific process. Specially BanglaBERT (Bhattacharjee et al., 2022) has been the most impactful monolingual model. Emojis played a crucial role in the dataset build-up process and played a crucial role in the annotation. So, removing those has a negative impact on the prediction (Fahim, 2023). Also, statistical machine learning methods such as SVM, and XGBoost embedded after Fastext or Word2Vec don't capture the complex context of the dataset and fall short in the prediction. Deep Learning methods such as RNN, LSTM, and Bi-LSTM generally perform better than the statistical machine especially Das et al. (2023b) showed a significant score using a combination of lstm and bi-lstm with a two-step approach. Ultimately BanglaBERT (Bhattacharjee et al., 2022) was the most prominent for all the teams having a vast amount of pretrained knowledge of Bangla at its disposal.

Team Name	Embedding			Statistical Machine Learning							Deep Learning					Transformer Based Approach								GAN
	FastText	Word2Vec	Tf-Idf	LR	SVM	KNN	DT	RF	MNB	SGD	XGBoost	RNN	LSTM	Bi-LSTM	GRU	CNN	BanglaBERT (Sagor)	BanglaBERT (csebuetnlp)	MBERT	XLM-RoBERTa	MuRIL	Indic Bert	mDeBerta -v3 base	
AAmbela			✓		✓			✓			✓		✓				✓	✓	✓				✓	
NLP_CUET		✓	✓	✓	✓		✓	✓	✓	✓				✓		✓		✓						✓
Seamntic Squad																	✓	✓	✓	✓				
nlpBDpatriots				✓	✓														✓	✓	✓			
the_linguists	✓			✓	✓								✓	✓	✓									
EmpytyMind		✓			✓			✓			✓		✓	✓				✓						
Mavricks																		✓		✓	✓			
LowResourceNLU																		✓	✓					
VacLM																		✓			✓	✓		
Score_Is_All_You_Need																		✓	✓	✓				
SUST_Black_Box																	✓	✓	✓	✓				
Team_Syrax																		✓	✓	✓				
Team CentreBack																		✓		✓				
UFAL-ULD																	✓	✓	✓	✓				
BanglaNLP			✓	✓					✓	✓							✓		✓					
Team Error Point			✓	✓	✓		✓	✓	✓	✓			✓			✓								

Table 5: This table shows the most popular techniques used by different teams.

7 Limitations

Quantitative Limitations: The main limitation of the shared task arises from the dataset. First of all, the dataset is small in size, with only 4k data points for the training and validation sets, and around 2k data points for the test set. This often creates problems in terms of over-fitting on large models. Additionally, the dataset is highly imbalanced with only a minor fraction of the data for direct violence creating a challenge for class detection which is also reflective of the participant's results.

Qualititive Limitations: Emojis play a crucial role in sentence classification, so removing any emoji during preprocessing leads to a loss of context. The dataset consists of data from Bangladesh and West Bengal, comprising only in Bengali language. Therefore, the nature of violence-inciting text's nature may differ based on culture and language. Finally, the dataset requires a hectic process to annotate and validate thus expanding the correct data is much more difficult.

Procedural Limitations: The dataset is fully annotated by Bangladeshi residents, all undergraduate students, with an expert resolving the dispute. The annotation is done based on previous literature, personal observations, and a strict framework for annotators to rely on. Then relying on a majority vote and expert adjudication to produce is used to reach a gold standard label. Several previous studies reveal that annotator identity is a critical determinant of data annotation patterns (Sap et al., 2019; Larimore et al., 2021; Waseem, 2016) and so majority voting doesn't always capture the subjective nature of the annotation (Davani et al., 2022). Nonetheless, the definition of violence and its subcategories in taxonomy and how the authors' builders built their dataset and the annotators applied their best judgment are based on societal papers primarily from Galtung (1969, 1990), does not take FRS (Faith, Religion and Societal Impact) into account. Therefore, any dataset and corresponding systems will have the mentioned limitations. Thus, others with different cultural, societal, or religious backgrounds may disagree with some of the annotations.

8 Conclusion

In this paper, we have presented an overview of the shared task on the Violence Inciting Text Detection (VITD) dataset. The contest fostered submissions from 27 teams with 16 teams outperforming the highest baseline system BanglaBERT (Bhattacharjee et al., 2021), and 17 teams submitted the system paper. The system description and subsequent analysis and limitations discussion demonstrate the successful completion of the task.

The task has some vast scope for improvement. As mentioned in Saha et al. (2023), there is significant unlabelled data ready for further improvement of the systems to invoke larger systems without over-fitting the larger models. A potential scope for improvement is adding more data from huge unlabelled data. A future version of the task may be arranged with the challenge of more data from dif-

ferent sources, languages, and regions. Also, real-time violence detection models can be the next step of the task.

Ethical Considerations

We release the dataset and baseline classes and individual systems for specific classes containing violence-inciting texts. We also shared the participants' system descriptions. The malicious actors can use this information to train a generative model and use it for malicious purposes (Kirk et al., 2022). However, we believe that the risk is negligible to the huge potential of such systems in detecting violence-inciting text detection. The annotators were interviewed by the task organizers and they assured that they were given proper mental support and did not face any challenges at the time or after completing the annotation procedure.

References

Kawsar Ahmed, Md Osama, Md. Sirajul Islam, Md Taosiful Islam, Avishek Das, and Mohammed Moshiul Hoque. 2023. Score_isall_you_need at blp-2023 task 1: A hierarchical classification approach to detect violence inciting text using transformers. In *Proceedings of the 1st Workshop on Bangla Language Processing (BLP 2023)*, Singapore. Association for Computational Linguistics.

Refaat Mohammad Alamgir and Amira Haque. 2023. Team centreback at blp-2023 task 1: Analyzing performance of different machine-learning based methods for detecting violence-inciting texts in bangla. In *Proceedings of the 1st Workshop on Bangla Language Processing (BLP 2023)*, Singapore. Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63.

Abhik Bhattacharjee, Tahmid Hasan, Kazi Samin, M Sohel Rahman, Anindya Iqbal, and Rifat Shahriyar. 2021. Banglabert: Combating embedding barrier for low-resource language understanding. *arXiv preprint arXiv:2101.00204*.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad Uddin, Kazi Mubasshir, Md Saiful Islam, Anindya Iqbal, M Sohel Rahman, and Rifat Shahriyar. 2022. Banglabert: Lagnuage model pretraining and benchmarks for low-resource language understanding evaluation in bangla. *Findings of the North American Chapter of the Association for Computational Linguistics: NAACL*.

Cristina Bosco, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, Tesconi Maurizio, et al. 2018. Overview of the evalita 2018 hate speech detection task. In *Ceur workshop proceedings*, volume 2263, pages 1–9. CEUR.

Shilpa Chatterjee, P J Leo Evenss, and Pramit Bhattacharyya. 2023. Vaclm at blp-2023 task 1: Leveraging bert models for violence detection in bangla. In *Proceedings of the 1st Workshop on Bangla Language Processing (BLP 2023)*, Singapore. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Rajesh Kumar Das, Jannatul Maowa, Moshfiqur Rahman Ajmain, Kabid Yeiad, Mirajul Islam, and Sharun Akter Khushbu. 2023a. Team error point at blp-2023 task 1: A comprehensive approach for violence inciting text detection using deep learning and traditional machine learning algorithm. In *Proceedings of the 1st Workshop on Bangla Language Processing (BLP 2023)*, Singapore. Association for Computational Linguistics.

Udoy Das, Karnis Fatema, Md Ayon Mia, Mahshar Yahan, Md Sajidul Mowla, MD Fayez Ullah, Arpita Sarker, and Hasan Murad. 2023b. Emptymind at blp-2023 task 1: A transformer-based hierarchical-bert model for bangla violence-inciting text detection. In *Proceedings of the 1st Workshop on Bangla Language Processing (BLP 2023)*, Singapore. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Krishno Dey, Prerona Tarannum, Md. Arid Hasan, and Francis Palma. 2023. Semantics squad at blp-2023 task 1: Violence inciting bangla text detection with fine-tuned transformer-based models. In *Proceedings of the 1st Workshop on Bangla Language*

Processing (BLP 2023), Singapore. Association for Computational Linguistics.

Md Fahim. 2023. Aambela at blp-2023 task 1: Focus on [unk] tokens: Analyzing violence inciting bangla text with adding dataset specific new word tokens. In *Proceedings of the 1st Workshop on Bangla Language Processing (BLP 2023)*, Singapore. Association for Computational Linguistics.

Caleb Ferguson, Sally C Inglis, Phillip J Newton, Peter JS Cripps, Peter S Macdonald, and Patricia M Davidson. 2014. Social media: a tool to spread information: a case study analysis of twitter conversation at the cardiac society of australia & new zealand 61st annual scientific meeting 2013. *Collegian*, 21(2):89–93.

Elisabetta Fersini, Paolo Rosso, Maria Anzovino, et al. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. *Ibereval@ sepln*, 2150:214–228.

Johan Galtung. 1969. Violence, peace, and peace research. *Journal of peace research*, 6(3):167–191.

Johan Galtung. 1990. Cultural violence. *Journal of peace research*, 27(3):291–305.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Jawad Hossain, Hasan Mesbaul Ali Taher, Avishek Das, and Mohammed Moshiul Hoque. 2023. Nlp_cuet at blp-2023 task 1: Fine-grained categorization of violence inciting text using transformer-based approach. In *Proceedings of the 1st Workshop on Bangla Language Processing (BLP 2023)*, Singapore. Association for Computational Linguistics.

Md Zobaer Hossain, Md Ashraful Rahman, Md Saiful Islam, and Sudipta Kar. 2020. Banfakenews: A dataset for detecting fake news in bangla. *arXiv preprint arXiv:2004.08789*.

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. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 390–399. IEEE.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Hannah Kirk, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. 2022. Handling and presenting harmful text in nlp research. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 497–510.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90.

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. *arXiv preprint arXiv:1907.11692*.

Murat Mengü and Seda Mengü. 2015. Violence and social media. *Athens Journal of Mass Media and Communications*, 1(3):211–227.

Sourabrata Mukherjee, Atul Kr Ojha, and Ondrej Dusek. 2023. Ufal-uld at blp-2023 task 1: Violence detection in bangla text. In *Proceedings of the 1st Workshop on Bangla Language Processing (BLP 2023)*, Singapore. Association for Computational Linguistics.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Saurabh Page, Sudeep Mangalvedhekar, Kshitij Deshpande, Tanmay Chavan, and Sheetal S. Sonawane. 2023. Mavericks at blp-2023 task 1: Ensemble-based approach using language models for violence inciting text detection. In *Proceedings of the 1st Workshop on Bangla Language Processing (BLP 2023)*, Singapore. Association for Computational Linguistics.

Md Nishat Raihan, Dhiman Goswami, Sadiya Sayara Chowdhury Puspo, and Marcos Zampieri. 2023. nlpbdpatriots at blp-2023 task 1: Two-step classification for violence inciting text detection in bangla - leveraging back-translation and multilinguality. In *Proceedings of the 1st Workshop on Bangla Language Processing (BLP 2023)*, Singapore. Association for Computational Linguistics.

Raquel Recuero. 2015. Social media and symbolic violence. *Social media+ society*, 1(1):2056305115580332.

Omar Faruqe Riyad, Trina Chakraborty, and Abhishek Dey. 2023. Team_syrax at blp-2023 task 1: Data augmentation and ensemble based approach for violence inciting text detection in bangla. In *Proceedings of the 1st Workshop on Bangla Language Processing (BLP 2023)*, Singapore. Association for Computational Linguistics.

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.

Saumajit Saha and Albert Aristotle Nanda. 2023. Banglanlp at blp-2023 task 1: Benchmarking different transformer models for violence inciting text detection in bangla. In *Proceedings of the 1st 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. 2023. 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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2019. Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891*.

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

Hrithik Majumdar Shibu, Shrestha Datta, Zhalok Rahman, Shahrab Khan Sami, MD. SUMON MIAH, Raisa Fairooz, and Md Adith Mollah. 2023. Sust_black box at blp-2023 task 1: Detecting communal violence in texts: An exploration of mlm and weighted ensemble techniques. In *Proceedings of the 1st Workshop on Bangla Language Processing (BLP 2023)*, Singapore. Association for Computational Linguistics.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Md. Tariquzzaman, Md Wasif Kader, Audwit Nafi Anam, Naimul Haque, Mohsinul Kabir, Hasan Mahmud, and Md Kamrul Hasan. 2023. the_linguists at blp-2023 task 1: A novel informal bangla fasttext embedding for violence inciting text detection. In *Proceedings of the 1st Workshop on Bangla Language Processing (BLP 2023)*, Singapore. Association for Computational Linguistics.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023. Lowresourcenlu at blp-2023 task 1 2: Enhancing sentiment classification and violence incitement detection in bangla through aggregated language models. In *Proceedings of the 1st Workshop on Bangla Language Processing (BLP 2023)*, Singapore. Association for Computational Linguistics.

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.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.