Team CentreBack at BLP-2023 Task 1: Analyzing performance of different machine-learning based methods for detecting violence-inciting texts in Bangla

Refaat Mohammad Alamgir
Independent Researcher
refaat.alamgir@gmail.com

Amira Haque
Independent Researcher
amirahaque1998@gmail.com

Abstract

Like all other things in the world, rapid growth of social media comes with its own merits and demerits. While it is providing a platform for the world to easily communicate with each other, on the other hand the room it has opened for hate speech has led to a significant impact on the well-being of the users. These types of texts have the potential to result in violence as people with similar sentiments may be inspired to commit violent acts after coming across such comments. Hence, the need for a system to detect and filter such texts is increasing drastically with time. This paper summarizes our experimental results and findings for the shared task on The First Bangla Language Processing Workshop at EMNLP 2023 - Singapore. We participated in the shared task 1 : Violence Inciting Text Detection (VITD). The objective was to build a system that classifies the given comments as either non-violence, passive violence or direct violence. We tried out different techniques, such as fine-tuning language models, few-shot learning with SBERT and a 2 stage training where we performed binary violence/non-violence classification first, then did a fine-grained classification of direct/passive violence. We found that the best macro-F1 score of 69.39 was yielded by fine-tuning the BanglaBERT language model and we attained a position of 21 among 27 teams in the final leaderboard. After the competition ended, we found that with some preprocessing of the dataset, we can get the score up to 71.68.

1 Introduction

With the rise of the Internet, it has become easy to post and comment on multiple social media platforms. Ease of access means that people have the power to influence others to commit violent acts. Early detection and removal of these type of content is necessary to avoid regrettable events such as killing, rape or mass murder. To this end, the Violence Inciting Text Detection (VITD) shared task (Saha et al., 2023a) was introduced at the Bangla Language Processing Workshop at EMNLP 2023. To the best of our knowledge, the shared task and its accompanying dataset (Saha et al., 2023b) are the first of its kind. While previous work explored similar tasks such as hate speech detection in Bangla (Ishmam and Sharmin, 2019), (Romim et al., 2021), this task is the first of its kind to call for systems that can classify a given text as likely to incite violence or not. It has a further fine-grained classification label for violence-inciting texts - namely passive and direct.

In this paper, we discuss our submitted systems to the shared task. We present our analysis related to the dataset and also the models that were submitted.[1]. The paper is organized as follows : first, we analyze the dataset, then we describe the experiments performed both before and after the competition ended. Finally, we analyze the systems submitted and discuss their shortcomings, along with possible directions for future work. The accompanying code for our experiments and analysis is publicly made available.[2]

2 Dataset Overview

The dataset was created compiling YouTube comments in Bangla associated with the top 9 violent incidents that have occurred in the Bengal region (Bangladesh and West Bengal) within the past 10 years.

Comments which stated facts or referred to any kind of social discussion were classified under the Non-Violence category. For comments which contained opinions in a derogatory language and statements which attempted to justify violence were classified under the Passive Violence category. Fi-

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

[2] https://github.com/refaat31/team-centreback-blp-task-1

Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), pages 168–173

Text	Label	Category
অসংখ্য ধন্যবাদ জানাই ঢাকা কলেজ শিক্ষার্থীদের	0	Non-Violence
ভোট চুরি করলে তো খুন হবেই	1	Passive Violence
তোদের উপর আল্লাহর গজব পড়বে	2	Direct Violence

Table 1: Example comments from each category

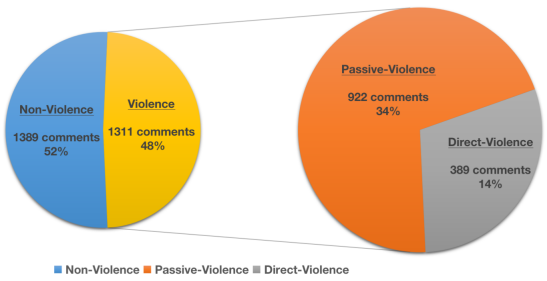

Figure 1: Extensive dataset view showing actual percentages of comments under each category

nally, the comments which presented an unjustified demand, order, actions and any kind of threats were classified under the Direct Violence category. Example comments for each category and an outline of the dataset is given in Table 1.

3 Dataset Analysis

The comments in the dataset have a maximum comment length of up to 600 words. It covers a wide range by considering comments from over the past 10 years.

However, it lacks equality in the amount of data it has for the three different categories. The ratio of Non-Violence to Violence does seem like a roughly equal distribution from a high level. In spite of that, if we dive deep we see that the ratio of Passive Violence to Direct Violence is roughly 2:1. The complete and extensive view of the dataset showing the actual percentages of the comments under each category is given in Figure 1.

Although we do have a total of around 1350 comments for both Non-Violence and Violence categories, if we only consider the Violence categories then we do not have the Direct and Passive Violence categories equally distributed. On one hand for Direct Violence category we have a deficit of around 300 and on the other hand we have a surplus of 300 for the Passive Violence category. Hence, we can conclude that we have a significantly low amount of data for the direct violence category and overall have an imbalance of data in the dataset. Besides, confusion arises

Text	Label	Our Confusion
ওবায়দুল কাদের সরাত হবে একদিন মনে রাখেন ক্ষমতার জন্য আল্লাহ তায়ালার কোরআন এর সাথে বেয়াদবি করছেন এর বিচার খুবই তারাতাড়ি হবে	2	Here we are not seeing any direct threat or action. Instead this is more of an opinion made about an individual so we think this should be label 1 instead of 2.
আওয়ামীলীগের চেতনা	0	It looks like a sarcastic and derogatory remark about a political party so we think this should be category 1.
আল্লাহ কোন দেশে বসবাস করি।আওয়ামীলীগের উপর গজব নাজিল করো।	2	These sentences are similar to one another. Yet both have different labels. We think both of these comments should have label 2 since here a urge towards wishing misfortune for a group of individuals is potrayed.
আল্লাহ তুমি এদের উপর গজব নাজিল করো।	1	
আমরা সনাতনলম্বীরা একটু শান্তি চাই আমাদের সাথে হওয়ার অন্যায়ইয়ের বিচার চাই আর এই নাস্তিকগুলোর ফাঁসি চাই 🙏🙏🙏😡😡😡😡😡😡😡 সব নাস্তিক দের শাস্তি চাই 🙏🙏🙏🙏😡😡😡😡😡	0	This comment is expressing hate and also calling for punishment of a specific community. We think it should have a label of 2 rather than 0.
ওরে বাটপার 😀😀	0	This can have different meanings depending on context.

Table 2: Confusing comment-label combinations

for multiple comments where according to the description it should have been identified as one category but it was denoted as another category. Examples are shown in Table 2.

Furthermore, throughout the dataset instead of portraying religion in a non-discriminative way a biasness was observed towards Islam or Hinduism. Religion was displayed in more of a negative tone which was further strengthened by the lower number of neutral religious comments across the whole dataset. This in turn can lead to a potential misuse and give rise to some ethical concerns. To add to it, the dataset consists of only Bengal regions (Bangladesh and West Bengal). Also, since the only source for the dataset was YouTube comments, a wide range of other diverse origins like educational sites such as Encyclopedias, Blogs, E-books and factual sites like news channel websites were not taken into account. Thus, considering a worldwide perspective a major portion of variety was missing. To summarize, there was an absence of impartiality in terms of religion and regarding the scope a more narrowed down sense was noticed throughout the whole dataset.

Moreover, lots of repetition and redundancy was seen on all parts of the dataset. Not only are there recurrences of comments but also loads of spelling mistakes at frequent intervals. Example comments which show repetition in comments are shown in Table 3. To add to it, throughout the dataset we have seen multiple comments which had missing spacing between independent words. This might have also contributed to the number of misclassifications that we had for the test set. Example of such comments are shown in Table 4.

Also, comments should normally have a variety in the speaking tone. However, no such steps were

Text	Label
ইসলামের বিরুদ্ধে লিখলে ই বিজ্ঞান মনস্ক হয়ে যাবে নাকি?কেন অভিজিৎ কে মারা হলো তা একবারও বলল না।	1
অভিজিৎকে কিসের জন্য মারা হইছে এটা বলা হইলো না হায়রে সাংবাদিক তোরা ও ইসলামের বিরুদ্ধে কথা বলোছ।	1

Table 3: Examples of repetitive comments

Text	Label
ধর্ম যার যারউৎসব সবার,,,,তাহলে , বিয়ে করবে করেছে আসাদুজ্জামান খান কামাল বউ সবার , পুকুর কামালের ,মাছ সবার , কামালের টাক মাথা ঢোল বাজাবে সবাই, গাড়ি কামালের চড়বে সবাই , মেয়ে কামালের সেই মেয়ের সাথে সেক্স করার অধিকার সবার , ইত্যাদি ইত্যাদি ইত্যাদি ইত্যাদি	1
আজ যদি শেখ মুজিবুর রহমান বেঁচে থাকতেনতাহলে হয়তো উনি বলতেন72 এর যুদ্ধ করেছিলাম বাংলার জন্য...এ আমার সোনার বাংলার মানুষ জন্য	0

Table 4: Example comments showing no spacing between independent words

taken to consider the different tones of speaking to further normalize the dataset as a whole. Examples for these kind of words in comments found in the dataset are shown in Table 5.

Additionally, similar token length values were used for all the categories. However, the value taken was comparatively a very small value. The mean value was around 19.6 and the standard deviation value was around 16.6. Hence, the deviation from the mean was very high. In ideal scenarios we would expect the standard deviation to be as low as possible. Standard deviation and mean token length values are shown in Table 6.

Finally, it is worth noting that the emojis have a significance while denoting the category labels. For example, for the comment shown in Table 7, if we consider that emoji has a significance, then the category (1) which it has been given seems justified. However, leaving aside the emoji, it seems like this is simply a normal statement and does not imply a violent tone. Hence, across the whole dataset the emojis played a vital role while classifying the

Text	Word
এই বার ইন্ডিয়ার বিরুদ্দে কথা বলা দরকার কেন জানেন তারা কি জন্য সেনা পাটাবে দেশে সেটা ক কিচু বলচেন না আপনারা কি জন্য বয় করেন আমরা ত দেকতেচি শুনতেচি তারা কি বাজে মনত্তব করতেচে	বলচেন
জ্বি জাই। মনের কথা বলসেন। এত খারাপ পরিস্থিতিতেও মানুষ শপিং এ যাই? চিন্তা করেন	বলসেন

Table 5: Example words in comments showing different speaking tones

	Non-Violence	Passive Violence	Direct Violence
Mean	18.6	21.2	19.2
Standard Deviation	15.9	17.9	16.0

Table 6: Category wise mean and standard deviation token length values

Text	Category
এটা গুজব নয় এইটা সত্যি সত্যিই ঘটেছে 😡😡😡😡	1

Table 7: Example comment showing significance of emoji

comments.

4 System Overview

For all the experiments, we have used either Nvidia Tesla V100 GPU or T4 GPU provided by Google Colaboratory, depending on the availability. During our submission for the competition we did not consider any preprocessing for the dataset and focused fully on the methodology of the model. After the competition ended, we performed some preprocessing to remove punctuations completely from the dataset.

It is important to note that, in the competition, there were two phases - in the first round, we were provided a test set with the ground truth labels, while in the second phase, we had a hidden test set, whose labels were provided after the competition ended. Thus, we have reported our results (in Table 8) on both the first and second round of the competition, as well as the result obtained from experiments performed after the competition ended.

In this competition, the evaluation metric was the macro-F1, which takes the arithmetic mean of the per-class F1 scores. The F1-score is calculated for each class in the following way [3] - $2 * \frac{precision * recall}{precision + recall}$, where precision tells us what fraction of the positive predictions are correct, and recall tells us what fraction of the positive labels have been correctly identified. Here, positive label means that the comment belongs to the class, for which F1-score is being calculated.

4.1 Fine-tuning Language Models

This task can be thought of as a sequence classification task, since we are assigning each comment a category : non-violence, direct violence or passive

[3] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

Model name	macro-F1 (first round)	macro-F1 (second round)
XLM-ROBERTa (base)	68.97	65.43
DistilBERT (base-multilingual)	64.26	63.50
Few-shot learning with SBERT	38.22	33.87
Two-stage (BanglaBERT + catboost with SBERT embeddings)	71.30	69.20
BanglaBERT (50 epochs)	74.81	68.92
BanglaBERT (20 epochs)	76.50	69.39
BanglaBERT with preprocessing done after the competition ended (10 epochs)	77.38	71.68

Table 8: Macro-F1 for different models on the first and second test set provided by organizers

violence. We have fine-tuned a few language models for this purpose by adding a classification head at the end. The training dataset is composed of 2700 Bangla comments and we have used an 80:20 train-test split for our training.

First, we used BanglaBERT (Bhattacharjee et al., 2022) as it was pre-trained on Bangla text and has been shown to have good scores for sentiment analysis task, which is a form of sequence classification. We also used XLM-ROBERTa (Conneau et al., 2019) which is an advanced version of BERT (Devlin et al., 2019), trained on multilingual data. Finally we used the multilingual version of DistilBERT (Sanh et al., 2019). We have trained each of these models for 50 epochs and have noticed that all of the models overfit quite quickly after a certain number of epochs. This can be attributed to the relatively small amount of training data. Finally, BanglaBERT trained for 20 epochs gave the best macro-F1 score of 76.50 for the first test phase set.

4.2 Two stage approach

This was our second best model. The reason for using different models in the two stages is that the models learn unique things in different ways when put under contrasting scenarios. This would result in the model being more versatile and adaptable to any circumstances.

For this method in stage 1, we have done a 80:20 train-test split on the provided original dataset of 2700 comments. Conversely, for stage 2 we have only considered the violence section (both passive and direct violence classes) of the dataset and hence performed a 80:20 train-test split on only the 1311 violence comments.

The process we followed for both the stages are as follows - first we tried to perform one kind of binary classification to categorize comments as either violence or non-violence. Then we did a fine-grained classification of the comments labeled as violent by further classifying them as either direct or passive violence. For the two stages, we used BanglaBERT for the first stage and catboost (Dorogush et al., 2018) with SBERT (Reimers and Gurevych, 2020) embeddings for the second stage. Initially we tried out BanglaBERT for both the stages. In stage 1 it was used for classifying between violence and non-violence and in stage 2 similarly it was used for classifying between direct violence and passive violence. It was seen that it performed better in stage 1 and hence for our two stage approach it was chosen for the first stage. On the other hand, for the second stage both catboost and BanglaBERT gave similar scores hence we thought of going with something non-identical compared to stage 1 and choose catboost as opposed to BanglaBERT. Finally, we believe that two stage training is a possible future direction.

4.3 Few-shot learning with SBERT

As the number of training examples for the category "direct violence" is significantly less compared to the other two categories, we wanted to see if few-shot learning would yield good results. Furthermore, considering the structure of the comments, where they consist of one or more sentences, we encoded them using SBERT. Since SBERT has been shown to perform excellent results on measuring semantic text similarity (STS), we converted our dataset into a suitable format for fine-tuning an STS model. We randomly sampled 100 examples from each class first. Then, if sentence 1 and sentence 2 are of the same class, we gave the sentence pair a label of 1, else we gave it a label of 0.

Finally, for inference, we computed the semantic textual similarity for each sentence embedding as follows: we computed its cosine similarity with every training example, and took the maximum. The class for which we got the highest score, we assigned that class to the test example. The training was done for 50 epochs, and this yielded a poor result as shown in Table 8.

5 Error Analysis

After the competition ended, we performed further analysis to determine the reason for the comparatively low macro-F1 scores. In the test set, we noticed that among the misclassifications done by our best model, a portion of those comments had a lot of repetitive punctuations as shown in Table 9. Furthermore, comments consist of either single sentences or multiple sentences. In order to ensure the

	Predicted	Ground Truth
কারা যেনো বলছিলো ঢাকা কলেজ এর ছাত্ররা নির্দোষ ,,,,,,,, ভাই আপনাদের মুখটা একটু দেখতে চাই	0	1
এই হলো আওয়ামী সংস্কৃতি!!! মন্ডপে কেন প্রতিমার উপর কেন কোরআন শরিফ রাখা হলো??? এটার বিচার আগে করেন কাউয়া কাদের	0	1

Table 9: Example comments showing repetitive punctuations

model does not treat comments of variable length sentences differently, we determined that the complete removal of punctuations was necessary. We achieved this with the bnlp toolkit (Sarker, 2021), which is an excellent library for preprocessing text in Bangla. After this was done, it improved our score from 76.5 to 77.38 for test set 1 and 69.39 to 71.68 for test set 2 . We also noticed that training for 10 epochs seems to give us the best score for the final test set. This is in line with our previous observation that the model overfits quite quickly due to the relatively small amount of training data. Thus, the best model is actually BanglaBERT trained for 10 epochs which gives a score of 71.68 on comments that have punctuations completely removed.

6 Future Works

Although we tried out different methods but our system did not take into account a number of things. Firstly, the spelling mistakes and missing spaces between two independent words in both training and inference stages. Secondly, the significance of emojis was also not taken into consideration. Furthermore, additional knowledge bases for fine-tuning could have also been used to see if it solves the issue with the limited dataset. Lastly, the repetition of similar comments throughout the whole dataset was also not taken into account.

The points mentioned above can be considered for future work for improving violence inciting text detection. In addition, the performance of large language models can also be investigated in this task, as they have been recently shown to perform well on different NLP tasks. (Liu et al., 2023)

7 Conclusion

In this paper, we have presented our experiments and findings for the BLP Shared Task 1 : Violence Inciting Text Detection. Initially, we provide a detailed analysis of the dataset, showing statistics and discussing problems with the dataset. We have found that BanglaBERT fine-tuned for 20 epochs gives us the best macro-F1 score of 69.39. After the competition ended, we analyzed the possible reasons for misclassifications. To further explore and overcome some of those causes, we conducted different experiments that led to a further improvement, taking the macro-F1 score to 71.68. Finally, we discussed the shortcomings of our system and the various possible directions for future work that can improve the detection of violence-inciting texts.

References

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M Sohel Rahman, and Rifat Shahriyar. 2022. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327.

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. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. 2018. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*.

Alvi Md Ishmam and Sadia Sharmin. 2019. Hateful speech detection in public facebook pages for the bengali language. In *2019 18th IEEE international conference on machine learning and applications (ICMLA)*, pages 555–560. IEEE.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Lin Zhao, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. 2023. Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta-Radiology*, 1(2):100017.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

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.

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

Sagor Sarker. 2021. Bnlp: Natural language processing toolkit for bengali language. *arXiv preprint arXiv:2102.00405*.