Shahjalal University of Science and Technology

Department of Computer Science and Engineering

Bangla Violence Inciting Text Detection

Hrithik Majumdar Shibu
Reg. No.: 2018331052
4th year, 1st Semester

MD. Sumon Miah
Reg. No.:2018331062
4th year, 1st Semester

Department of Computer Science and Engineering

Machine Learning Lab Project

Mohammad Shahidur Rahman, PhD
Professor
Department of Computer Science and Engineering

28th August, 2023

Abstract

With the rapid proliferation of online communication, there has been a surge in the dissemination of provocative content that fuels real-world violence. This study delves into the development of advanced natural language processing (NLP) techniques tailored to the linguistic nuances of the Bengali language, aiming to identify and classify violent incitement within digital textual content effectively. Leveraging a curated dataset of locally relevant content, the research employs a combination of machine learning algorithms along with deep learning algorithms and semantic analysis to enhance accuracy and adaptability. The findings contribute to the growing field of NLP-driven content moderation and hold crucial implications for maintaining social harmony and preventing the escalation of violence incited by online discourse in Bangladesh. In this work, we are introducing a Bengali dataset consisting of 2700 violent inciting texts labeled as Direct Violence, Passive Violence, and Normal from different kinds of sources including Facebook and YouTube posts, comments, news portals, etc. Besides, we have trained an MLM(Masked Language Model) to learn the context of the various similar domains resulting into a large-scale dataset. Finally, we have fine-tuned a transformer-based Neural Network named Electra-BERT(Bidirectional Encoders Representation from Transformers) to check the performance of our model with respect to our validation dataset and trained MLM model. We have achieved 82 percent of accuracy and about 80 percent of macro F1-score for all our three output classes of the dataset including Direct Violence, Passive Violence and Normal.

Keywords: MLM, Violence Inciting, NLP, Semantic Analysis, Transformer, BERT, ELECTRA

Contents

Chapter 1

Motivation

Unfortunately, despite being the seventh most widely spoken language globally, Bengali is considered one of the notable resource-constrained languages. Statistics reveal that more than 45 million users on Facebook and YouTube are using Bengali daily. Most of these users commonly interact on social media via the textual form. Many textual interactions contain hostile contents that cause the significant rise of hate, violence, and aggression on social media. Thus, to ensure the quality of textual conversation and reduce unlawful activities on these platforms, developing an automated Bengali language system that can identify these violent activities is mandatory. Such a system will flag posts/comments that convey any violence that might threaten national security, try to break communal harmony, and publicize distorted propaganda.

However, developing a system to detect violent textual conversation in a resource-constrained language like Bengali is challenging. The scarcity of benchmark dataset and deficiency of language processing tools are the key barriers to develop such a system in Bengali.

Complicated morphological structure, presence of ambiguous words, diversities in different dialects and rich variations in the constituent parts of a sentence have made the task more complicated. Bengali has a rich vocabulary and unique writing script which has no overlap with other resource-high languages. Moreover, multilingual code-mixing in social media texts has added a new challenge.

Therefore, the key motivations behind doing this project are-

- How can we successfully develop an violence annotated dataset in the Bengali language?

- How can we effectively identify potential violent texts and categorize them into predefined violent categories?

This work develops a Bengali Violence Inciting Text dataset by analysing direct violent, passive violent and normal texts' properties to address the above research questions. We have also deeply analyzed the results and error of the models and shed light on the reasons behind some of the errors and provide a few directions that might help to mitigate the system's deficiency.

Chapter 2

Objectives

The objectives of our "Violence Inciting Text Detection" project outline what we intend to achieve through our work. Here are some potential project objectives:

- Dataset Creation and Annotation
- Dataset Pre-processing
- Model Selection
- Model Training and Optimization
- Performance Evaluation
- Ethical Considerations
- Multilingual Adaptability
- Inspiration for Similar Projects

These objectives provide a clear roadmap for our project and help define its scope, tasks, and expected outcomes. They reflect the various stages of the project, from data collection to model deployment and community engagement.

Chapter 3

Related Works

There are some related works, but they use a private dataset so it's not easier to work with that dataset for future betterment. Over the last few years, a significant amount of work has been carried out to identify and categorize unwanted texts on various online platforms such as twitter, facebook, reddit and so on. Works included aggression classification, hate speech detection, abuse detection, toxicity classification, misogyny, trolling identification, and offensive text classification have been done previously. But most of them were done for English Language. Based on our literature, there is no available dataset on **Bangla Violence Inciting Text Detection** work. But, kind of similar works have been done for English Language.

Omar et al [1] offered a brand-new, two-level annotated dataset of violent Bengali writing, dubbed "BAD." 14158 texts in level-A have labels indicating whether they are aggressive or not. In level-B, 6807 aggressive texts are divided into groups for verbal, political, religious, and gendered aggression, with 2217, 2085, 2043, and 462 texts in each class, respectively. To identify and categorise the aggressive texts in Bengali, this work suggests a weighted ensemble strategy using m-BERT, distil-BERT, Bangla-BERT, and XLM-Roberta as the basic classifiers.

Tanvirul et al[2] have fine-tuned multilingual transformer models for Bangla text classification tasks in different domains, including sentiment analysis, emotion detection, news categorization, and authorship attribution. We obtain the state of the art results on six benchmark datasets, improving upon the previous results by 5-29 percent accuracy across different tasks.

Chapter 4

Methodology

In this section, we outline the methodologies employed for the development and training of the Bangla Violence Inciting Text Detection model. The process encompasses data preparation, model architecture design, training setup, and evaluation metrics.

4.1 Data Pre-processing

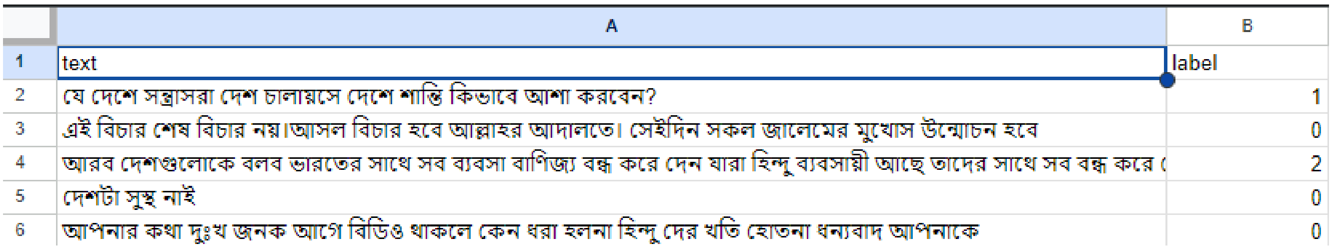
4.1.1 Dataset Creation

A comprehensive dataset containing of Violence Inciting Texts in Bangla Language was curated for this project. Violent texts were collected from multiple social platforms, including Facebook, Twitter and Youtube posts and comments. We have collected a total number of 4000 Bangla violence inciting texts from which 2700 data have been kept for training purpose and 1300 are kept for validation purpose.

4.1.2 Data Labeling

Each text in the dataset was meticulously labelled with the corresponding text classes among Direct Violence(2), Passive Violence(1), Normal text(0), resulting in a dataset suitable for supervised machine learning procedure. Data labelling has done by quite professionals Bangla language practitioners to get less errors in our process.

The first 5 rows of our dataset is shown below from a csv file:



	A	B
1	text	label
2	যে দেশে সন্ত্রাসরা দেশ চালায়সে দেশে শান্তি কিভাবে আশা করবেন?	1
3	এই বিচার শেষ বিচার নয়।আসল বিচার হবে আল্লাহর আদালতে। সেইদিন সকল জালেমের মুখোস উন্মোচন হবে	0
4	আরব দেশগুলোকে বলব ভারতের সাথে সব ব্যবসা বাণিজ্য বন্ধ করে দেন যারা হিন্দু ব্যবসায়ী আছে তাদের সাথে সব বন্ধ করে	2
5	দেশটা সুস্থ নাই	0
6	আপনার কথা দুঃখ জনক আগে বিডিও থাকলে কেন ধরা হলনা হিন্দু দের খতি হোতনা ধন্যবাদ আপনাকে	0

4.2 Model Architecture Design

4.2.1 Masked Language Model

By masking off words and guessing their meaning based on the context, the Masked Language Model (MLM) is a neural network architecture used in natural language processing (NLP) to anticipate missing words in phrases. MLM models function by using words that are hidden, reducing the discrepancy between forecasts and actual words, and taking into account bidirectional context. They can be honed for specialised tasks like text classification, sentiment analysis, and question answering after receiving pretraining on big corpora of text data to learn broad language patterns. NLP tasks including sentence completion, text generation, and language translation have been changed by MLM models. We have fine tuned the MLM on the large scale corpora to get a better performance of our detection process

4.2.2 Experimental Models

We have experimented on the Mask Language Model of 3 different transformers based model including BERT, XLM-Roberta and Electra For Masked LM. Among these models Electra Model has given the highest performance metrics on accuracy and F1-Score.

4.3 Model Compilation and Training

4.3.1 Domain Adaptation

We have used domain adaptation process in MLM model fine-tuning process. Domain adaptation refers to the process of adjusting or adapting a pre-trained MLM to perform well on a particular domain or dataset that may differ from the domain on which the model was initially trained. In other words, it involves modifying the model's expertise to suit the needs of data from a specific area. An MLM that has been pretrained, picks up context and linguistic patterns from a variety of text corpora. Domain adaptation for masked language models(MLM) involves the process of fine-tuning a pre-trained model to perform well in a specific domain by training it on domain-specific data. This helps ensure that the model's linguistic representations are tailored to the nuances of the target domain, leading to improved performance in real-world applications.

4.3.2 Experimental Setup

We have collected all the aggressive data[1], negative emotion/sentiment data [2], hate speech data[3] and others available negative meaning Bangla text contents from various domains to train a Masked Language Model(MLM). We have used 2700 data with the domain adapted huge corpora of total 93000 text data which we have collected for training purpose and kept 1300 data for validation. We have fine tuned the Learning Rate, Weight Decay, Batch Sizes and also applied freezing to the encoder layers for the experimental models. The primary evaluation metrics chosen were accuracy and macro F1-score.

4.3.3 Training Setup

The model was trained using the generated training and validation data generators. An epoch count of 10 was selected for Masked Language Model training. The training phase consists of 2 consecutive steps. Firstly, we have trained the Masked Language Model combining the created and collected dataset by domain adaptation. Secondly, we have experimented different models as Bangla-BERT, Electra-BERT, XLM-Roberta on the train dataset and validation dataset among which Electra has given us the best performance.

4.3.4 Training Process

The model underwent 50 epochs for the classification purpose with the fine-tuned early stopping rate. The number of steps per epoch and validation steps was set according to the length of the respective generators. In each epoch we have shown all the performance matrices including accuracy, precision, recall and f1-score.

4.4 Evaluation

4.4.1 Metrics

Model performance was evaluated using accuracy and macro F1-score as the primary metrics. Accuracy quantifies the ratio of correctly predicted violence inciting text classes. F1-score is the harmonic mean of precision and recall. In this case, our dataset is imbalanced. That's why we are using Macro F1 score alongside accuracy. The macro F1 score calculates the F1 score for each class independently and then takes the average of those F1 scores.

4.5 Software and Hardware Environment

4.5.1 Software

The project was implemented using Python programming language, utilizing PyTorch and Transformers libraries for deep learning tasks on Google Colab.

4.5.2 Hardware

The training and evaluation were performed on a machine with an 8GB Ram and 4GB GPU environment.

The above methodologies encapsulate the steps taken to create, develop, train, and evaluate Bangla Violence Inciting Text Detection model. This section provides a clear understanding of the strategies employed to achieve the project's objectives and outcomes.

Chapter 5

Results

In this section, we present the results achieved through the training and evaluation of the Bangla Violence Inciting Text Detection model. The model's performance metrics are highlighted, and the implications of the outcomes are discussed.

5.1 Model Performance

The trained detection model was evaluated using a variety of performance metrics to gauge its accuracy and effectiveness in detecting Bangla Violence Inciting Texts. The following key results were obtained:

- Validation Measures: The model achieved an impressive validation accuracy of 82.03 percent and macro F1-score of 80.38 percent respectively which are good in our observation.

5.2 Limitations

It's important to acknowledge the limitations of the project:

- **Dataset Size:** The model's performance could be further improved with a larger and more diverse dataset.
- **Imbalanced Datset:** Our dataset is imbalanced due to the scarcity of bangla violence inciting texts in online or social platforms.

Chapter 6

Discussion

The validation accuracy of 82 percent achieved by the Bangla Violence Inciting Text Detection model is a notable accomplishment. This accuracy indicates that the model has successfully learned to distinguish between 3 different kinds of violent texts based on their linguistic characteristics. The effectiveness of the model in classifying Bangla Violence Inciting Texts is promising, as it opens up opportunities for various applications like social media content moderation, law enforcement support and counter terrorism efforts etc.

6.1 Implications and Applications

The accuracy achieved by the model holds practical implications for various domains:

6.1.1 Conflict Resolution and Peace-building:

Organizations working in conflict zones can use the system to monitor online discussions, identify potential triggers, and intervene to mitigate tensions and promote peaceful dialogue.

6.1.2 Education and awareness

The curated dataset and trained model or this technology can be incorporated into educational programs to raise awareness about the impacts of violent language and to encourage responsible digital communication among students and the general public.

Chapter 7

Conclusion

In conclusion, the creation and application of a system for identifying Bangla texts that incite violence represent an important step towards creating a more secure and peaceful online environment in Bangladesh. This technology has the ability to reduce the spread of violent content and hate speech online by utilising cutting-edge natural language processing techniques customised to the linguistic nuances of the Bengali language. Applications for it can be found in many fields, including journalism, education, and conflict resolution as well as social media moderation and help for police enforcement. As this technology develops, its effects on averting actual violence, developing responsible digital communication, and encouraging a culture of peace all highlight how important it is in a world that is becoming more interconnected. To solve new difficulties, it is crucial to constantly improve the system's accuracy, assure ethical concerns, and work with stakeholders. In the end, the Bangla violence-inciting text detection system serves as a crucial tool for protecting the digital sphere from the negative impacts of online violence and extremism, thereby building a more secure and inclusive society for all. With an accuracy over 82 percent and macro F1-score of 80 percent respectively, this dataset along with the model will really be a good benchmark for the further researches on this topic in future.

Despite the project's outstanding outcomes, there are still areas for improvement and expansion. To address issues like imbalanced dataset, future rounds might investigate data augmentation approaches, fine-tuning tactics, and the expansion of the dataset. By analysing the performance and statistics of the dataset, the evaluation measures can be developed more in future.

References

[1] O. Sharif and M. M. Hoque, “Tackling cyber-aggression: Identification and fine-grained categorization of aggressive texts on social media using weighted ensemble of transformers,” *Neurocomputing*, vol. 490, pp. 462–481, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231221018567

[2] T. Alam, A. Khan, and F. Alam, “Bangla text classification using transformers,” 2020.

[3] N. Romim, M. Ahmed, M. S. Islam, A. Sen Sharma, H. Talukder, and M. R. Amin, “BD-SHS: A benchmark dataset for learning to detect online Bangla hate speech in different social contexts,” in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 5153–5162. [Online]. Available: https://aclanthology.org/2022.lrec-1.552