Team Error Point at BLP-2023 Task 1: A Comprehensive Approach for Violence Inciting Text Detection using Deep Learning and Traditional Machine Learning Algorithm

Rajesh Kumar Das, Jannatul Maowa, Moshfiqur Rahman Ajmain,
Kabid Yeiad, Mirajul Islam, Sharun Akter Khushbu
Department of Computer Science and Engineering
Daffodil International University, Dhaka, Bangladesh
{rajesh15-13032, jannatul15-14095, moshfiqur15-14090, yeiad15-14440,
merajul15-9627, sharun.cse}@diu.edu.bd

Abstract

In the modern digital landscape, social media platforms have the dual role of fostering unprecedented connectivity and harboring a dark underbelly in the form of widespread violence-inciting content. Pioneering research in Bangla social media aims to provide a groundbreaking solution to this issue. This study thoroughly investigates violence-inciting text classification using a diverse range of machine learning and deep learning models, offering insights into content moderation and strategies for enhancing online safety. Situated at the intersection of technology and social responsibility, the aim is to empower platforms and communities to combat online violence. By providing insights into model selection and methodology, this work makes a significant contribution to the ongoing dialogue about the challenges posed by the darker aspects of the digital era. Our system scored 31.913 and ranked 26 among the participants.

1 Introduction

There is a great need for robust detection and classification algorithms in today's digital environment since violent incitement material is spreading so rapidly. This is especially essential for languages like Bangla, where regional context and little changes in language play a large role in determining how violent content operates. The EMNLP BLP shared task on “Violence Inciting Text Detection” serves as a strong appeal to address this topic directly. One of our goals is to make a system that can handle the complicated language of Bangla. This will make it easier and more accurate to find material that encourages violence.The idea for our study came from the important work of (Saha et al., 2023b) and the creation of the Vio-Lens dataset (Saha et al., 2023a). The fundamental purpose of VITD is to detect and classify texts that contain components of incitement to violence. Vio-Lens, a unique annotated collection of over 10,000 Bangla social media posts, marks a significant advancement in detecting and addressing violence-inciting language. With this resource, we aim to push the boundaries of threat assessment in Bangla narratives, including those up to 600 words, seeking to not only identify evident risks but also redefine detection parameters. This research makes a valuable contribution to the wider effort to promote secure digital environments.Several study subjects that have been discussed in the literature are location-independent machine learning approaches for early fake news detection (Liu, 2019), combining audio and text elements to find violent incidents (Anwar, 2022), and the creation of new methods like feature-based Twitter sentiment analysis with enhanced denial handling (Gupta and Joshi, 2021). There is also an investigation into the possible use of a memristive LSTM network for sentiment analysis (Wen et al., 2021). The method used in this study is based on the political security threat prediction framework, which is a mix of a lexicon-based approach and machine learning methods (Razali et al., 2023). Additionally, the system has a racism detection model that leverages a stacked ensemble GCR-NN architecture (Lee et al., 2022). These initiatives demonstrate the applicability of mood analysis in several domains pertaining to security and social justice. To get further details on our research, refer to the publication titled "Sentiment Analysis of Tweets using Heterogeneous Multi-layer Network Representation and Embedding" (Gyanendro Singh et al., 2020). Moreover, a significant advancement is shown in the MC-BERT4HATE model's ability to detect hate speech across many languages and translations (Sohn and Lee, 2019). Even though a lot of work has been made, these improvements also show how hard it is to understand Bangla language. Sometimes, traditional models have trouble understanding all the details in this language.Our proposed methodology employs a diverse range of machine learning models to address the issues

Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), pages 236–240
December 7, 2023 ©2023 Association for Computational Linguistics

mentioned above. The algorithms included in this set are Logistic Regression, Decision Tree, Random Forest, Multi-Naive Bayes, K-Nearest Neighbour (KNN), Support Vector Machine (SVM), and Stochastic Gradient Descent (SGD). This is in addition to using deep learning architectures like Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), and LSTM-CNN hybrids to adapt to the unique features of the Bangla spoken form. The inclusion of all individuals under this methodology facilitates the identification of and categorization of potential hazards, hence streamlining the process. [1] final implementation with an anonymous GitHub link[2]..

2 Literature Review

Due to violence-inciting content, social media is both connecting and alarming. We found a new answer to this essential issue, giving hope. Traditional machine learning and deep learning models classify violence-inciting literature in this study. This study built on natural language processing and hate speech identification research. The NLP survey on hate speech identification is useful for its problem formulations and methods (Schmidt and Wiegand, 2017). The transformative deep bidirectional transformer model BERT by (Devlin et al., 2019) has changed natural language comprehension research. (Van Hee et al., 2015) Cyberbullying detection and classification work shows that online safety awareness has enhanced cyberbullying detection beyond hate speech. (Zhou et al., 2019) and (Zampieri et al., 2019) participated in SemEval-2019 Task 6, which identified and categorized social media offensive language. (Wu et al., 2019) from BNU-HKBU UIC NLP Team 2 employed a BERT model to detect foul language, enriching this field. These studies show the importance of identifying and regulating offensive digital content. Study social media bullying traces and their prognostic potential for online safety (Xu et al., 2012). The necessity of studying protected traits has helped Burnap and Williams improve Twitter cyber hate detection (Burnap and Williams, 2016). Comment embeddings for hate speech identification advance the field and demonstrate their efficacy (Djuric et al., 2015). Mehdad and Tetreault illuminated character-level abusive encounters, improving our comprehension of abusive language (Mehdad and Tetreault, 2016). Due to variances in methods and datasets, these research' results vary in accuracy despite their importance. This comprehensive review uses multiple methodologies and data augmentation to fill this critical gap in our knowledge. We want to improve Bangla sentiment analysis and offensive language identification datasets and models. Our research will illuminate content filtering and internet safety in underrepresented languages.

3 Data and Methodology

In this section, we present the data sources and preprocessing steps, along with the methodology encompassing machine learning and deep learning models.

3.1 Dataset Description

The dataset utilized in our research was sourced from BLP Shared Task 1: Violence Inciting Text Detection (VITD), a valuable resource consisting of two key columns: "text" and "label." The "text" column encompasses textual content harvested from diverse social media platforms. For clarity and reference, we introduce "Label Definition" in Table 1, elucidating the categories assigned to each label within our dataset. Furthermore, Figure 1 illustrates a compelling word cloud visualization, spotlighting the most frequently occurring words in our datasets.

Table 1: Label Definition for BLP Shared Task 1

Label	Category	Total
Direct Violence	2	389
Passive Violence	1	922
Non-Violence	0	1389

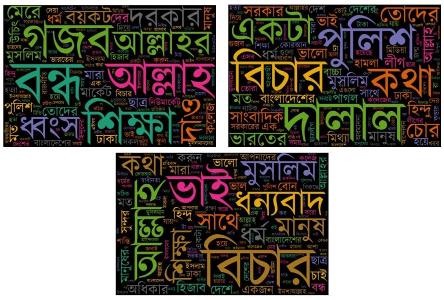

Figure 1: Word Cloud Visualization for Three Label (Non-Violence, Passive Violence, Direct Violence)

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

[2] https://anonymous.4open.science/r/EMNLP_2023_BLP_Workshop_Task1-0FB2

3.2 Preprocessing

The dataset was collected from BLP Shared Task 1: Violence Inciting Text Detection (VITD), which is a shared task in the context of violence inciting text detection. The dataset encompasses a multitude of elements including symbols, URLs, and concealed characters. It also incorporates non-standard characters, Unicode control characters, emoticons, emojis, variations in whitespace, special formatting elements, non-alphanumeric characters, instances of duplicated or reiterated characters, and escape sequences, among others. Hence, we have executed multiple preprocessing procedures to eliminate the noise from the data. We also executed the following actions: elimination of short conversations, exclusion of lengthy conversations, removal of non-Bangla characters, filtering out Stopwords and non-Bangla characters, and Finally we apply stemming. To address the initial label imbalance in our dataset, we employed Up-sampling specifically for the "Direct Violence" category. Table 2 illustrates a comparison between the values before and after the pre-processing phase.

Table 2: Comparison of Data Before and After Preprocessing

Label	Before Preprocessing	After Preprocessing
Non Violence	1389	1336
Passive Violence	922	881
Direct Violence	389	750
Total	2700	2967

3.3 Models

In our study, we employed a diverse set of models, encompassing both deep learning and traditional machine learning approaches. The deep learning models included Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), and a hybrid model, LSTM-CNN, each tailored for text classification. These models excel at capturing sequential information and local features within the text data. Additionally, we leveraged traditional machine learning models such as Logistic Regression, Decision Tree, Random Forest, Multinomial Naive Bayes, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Stochastic Gradient Descent (SGD). We purposefully chose the models for our study based on their distinct advantages and applicability to solving the challenging problem of identifying texts that incite violence. Here are the reasons we chose these models: LSTM chosen for its expertise in capturing sequential information, making it perfect for analyzing the complex language in texts that incite violence. CNN selected for its ability to identify structural patterns and components indicating violent content in text. Combines LSTM and CNN advantages, using local features and sequential information for comprehensive text classification. Traditional machine learning models chosen for their diverse techniques and effectiveness in text categorization.

3.4 Experimental Setup

To initiate the training of our traditional models, we first converted the preprocessed data into TF-IDF vectors. We went a step further by incorporating weighted n-grams, encompassing not only unigrams but also bigrams and trigrams. This strategy allowed us to harness contextual information more effectively, enhancing our model's understanding. We meticulously fine-tuned the model parameters to optimize performance and ensure the robustness of our deep learning-based classification approach, as detailed in Table 3. The dataset is divided into two subsets: "Training set" containing 2373 samples for model training, and "Test set" comprising 594 samples for evaluation.

4 Results and Discussion

In this section, we present the results of our experiments and engage in a comprehensive discussion of the findings. Our study aimed to address the challenge of violence inciting text detection using a combination of machine learning and deep learning models. We used various algorithms and techniques to analyze and classify text data into different categories of violence, namely Direct Violence, Passive Violence, and Non-Violence.

The machine learning models displayed varying degrees of performance in classifying violence inciting text in table 4. Notably, the Random Forest and Support Vector Machine (SVM) models outperformed the others in terms of accuracy and F1 score. These models achieved accuracy levels above 76.09%, demonstrating their effectiveness in distinguishing between different categories of violence.

Table 3: Experimental Setup for Deep Learning Models

Model	Embedding Dimension	Input Length	Vocabulary Size	Number of Classes	Batch Size	Number of Epochs
LSTM	128	300	5000	3	64	50
CNN	128	300	5000	3	64	50
LSTM-CNN Combine	128	300	5000	3	64	50

Table 4: Machine Learning Model Performance

Model Name	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Logistic Regression	73.91	75.26	73.91	72.11
Decision Tree	69.02	69.33	69.02	68.72
Random Forest	76.09	77.60	76.09	74.02
Multi. Naive Bayes	70.54	71.52	70.54	70.13
KNN	61.78	62.93	61.78	61.48
SVM	76.94	76.50	76.94	76.10
SGD	76.94	76.75	76.94	75.64

Table 5: Deep Learning Model Performance

Model	Class	Accuracy	Precision	Recall	F1-Score
LSTM	No-Violence	67.68	82.44	81.20	81.82
	Passive Violence		76.72	60.75	67.81
	Direct Violence		50.22	69.05	58.15
CNN	No-Violence	68.69	73.03	83.46	77.89
	Passive Violence		73.63	68.60	71.02
	Direct Violence		56.80	57.14	56.97
LSTM-CNN	No-Violence	66.50	64.85	80.45	71.81
	Passive Violence		74.60	64.16	68.99
	Direct Violence		56.50	59.52	57.97

Our ensemble of deep learning models, including LSTM, CNN, and LSTM-CNN, displayed strong performance in classifying violence-inciting text listed in table 5. It is evident that the CNN model has the highest accuracy at 68.69%, followed closely by the LSTM model with an accuracy of 67.68%. The LSTM-CNN hybrid model, while still respectable, trails slightly behind with an accuracy of 66.50%.

5 Conclusion

Our research underscores the critical importance of detecting and classifying violent incitement text within the realm of Natural Language Processing (NLP). Drawing inspiration from the EMNLP BLP shared assignment on Violence Inciting Text Detection and building upon the foundational work, we aimed to redefine the parameters of danger assessment in the context of the Bangla language. This study undertakes a comprehensive evaluation of machine learning and deep learning models to assess their effectiveness in categorizing literature that incites violence. Conventional machine learning algorithms, such as Logistic Regression, Decision Tree, Random Forest, Multi-Naive Bayes, KNN, SVM, and SGD, consistently demonstrate strong and reliable performance. Notably, Support Vector Machines (SVM) and Stochastic Gradient Descent (SGD) stand out for their efficacy in accurately classifying violent content. Deep learning models, including Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), and the hybrid LSTM-CNN, also exhibit significant capabilities. LSTM, in particular, emerges as a standout performer among the deep learning models. This study's limitations include language and dataset specificity, data imbalance, model interpretability, and computational resource requirements. Future research may encompass multilingual expansion, contextual analysis, user-level profiling, ethical considerations, human-in-the-loop approaches, cross-domain application, and real-world deployment of violence-inciting text detection models.

References

A. Anwar. 2022. Deepsafety: Multi-level audio-text feature extraction and fusion approach for violence detection in conversations.

P. Burnap and M. L. Williams. 2016. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data science*, 5:1–15.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pretraining 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.

N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30.

I. Gupta and N. Joshi. 2021. Feature-based twitter sentiment analysis with improved negation handling. *IEEE Transactions on Computational Social Systems*, 8(4):917–927.

L. Gyanendro Singh, A. Mitra, and S. Ranbir Singh. 2020. Sentiment analysis of tweets using heterogeneous multi-layer network representation and embedding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

E. Lee et al. 2022. Racism detection by analyzing differential opinions through sentiment analysis of tweets using stacked ensemble gcrnn model. *IEEE Access*, 10:9717–9728.

H. Liu. 2019. A location independent machine learning approach for early fake news detection. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 4740–4746, Los Angeles, CA, USA.

Y. Mehdad and J. Tetreault. 2016. Do characters abuse more than words? In *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue*, pages 299–303.

N. A. M. Razali et al. 2023. Political security threat prediction framework using hybrid lexicon-based approach and machine learning technique. *IEEE Access*, 11:17151–17164.

S. Saha, J. A. Junaed, A. S. S. Api, N. Mohammad, and M. R. Amin. 2023a. 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.

S. Saha, J. A. Junaed, N. Mohammed, S. Kar, and M. R. Amin. 2023b. 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.

A. Schmidt and M. Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*.

H. Sohn and H. Lee. 2019. Mc-bert4hate: Hate speech detection using multi-channel bert for different languages and translations. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 551–559.

C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, et al. 2015. Detection and fine-grained classification of cyberbullying events. In *Proceedings of the international conference recent advances in natural language processing*, pages 672–680.

S. Wen et al. 2021. Memristive lstm network for sentiment analysis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(3):1794–1804.

Z. Wu, H. Zheng, J. Wang, W. Su, and J. Fong. 2019. Bnu-hkbu uic nlp team 2 at semeval-2019 task 6: Detecting offensive language using bert model. In *Proceedings of the 13th International Workshop on Semantic Evaluation*.

J. M. Xu, K. S. Jun, X. Zhu, and A. Bellmore. 2012. Learning from bullying traces in social media. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 656–666.

M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval).

C. Zhou, J. Wang, and X. Zhang. 2019. Ynu-hpcc at semeval-2019 task 6: Identifying and categorising offensive language on twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*.