UFAL-ULD at BLP-2023 Task 1: Violence Detection in Bangla Text

Sourabrata Mukherjee[1], Atul Kr. Ojha[2], Ondřej Dušek[1]
[1]Charles University, Faculty of Mathematics and Physics, Prague, Czechia
[2]Insight SFI Centre for Data Analytics, DSI, University of Galway, Ireland
{mukherjee,odusek}@ufal.mff.cuni.cz
atulkumar.ojha@insight-centre.org

Abstract

In this paper, we present UFAL-ULD team's system, desinged as a part of the BLP Shared Task 1: Violence Inciting Text Detection (VITD). This task aims to classify text, with a particular challenge of identifying incitement to violence into Direct, Indirect or Non-violence levels. We experimented with several pre-trained sequence classification models, including XLM-RoBERTa, BanglaBERT, Bangla BERT Base, and Multilingual BERT. Our best-performing model was based on the XLM-RoBERTa-base architecture, which outperformed the baseline models. Our system was ranked 20th among the 27 teams that participated in the task.

1 Introduction

The rapid proliferation of social media platforms has revolutionized the way we communicate, share information, and engage with diverse communities online. However, with this newfound connectivity and freedom of expression, we have also witnessed a troubling trend – the weaponization of social media for the incitement of violence. The Bengal region, comprising Bangladesh and West Bengal, India, has not remained untouched by this unsettling phenomenon. Online platforms, once hailed as vehicles for progress and connection, are now grappling with the disturbing spread of violence-inciting language, leading to communal discord, destruction, and loss of life.

In this digital age, where the boundaries between the virtual and the real world blur, it becomes imperative to address the multifaceted manifestations of communal violence, particularly in regions like Bengal. The Violence Inciting Text Detection (VITD) shared task emerges as a beacon of hope and a clarion call for the natural language processing (NLP) community to confront this pressing issue head-on.

The VITD shared task centres on the precise categorization and discernment of violence-inciting text within social media comments, echoing the broader challenge of understanding the dark underbelly of online discourse. The violence we seek to detect and categorize transcends mere words on a screen; it has the potential to manifest as explicit threats, divisive propaganda, and derogatory language that can irreparably harm individuals and communities.

This paper discusses our team's system, built as a part of the BLP Shared Task 1: Violence Inciting Text Detection (VITD) (Saha et al., 2023b,a). In this work, we experimented with several pre-trained sequence classification models with the provided data only where we contributed to data augmentation, sampling stratgies, fine-tuning and hyper-parameter tuning to optimize the performance of these models.[1] Our system was ranked 20th among the 27 teams that participated in the task.

2 Related work

The proliferation of hate-speech, verbal threats, aggression, cyberbullying, trolling, abuse, offensive and misogny content are experiencing rapid growth on social media. A considerable number of researchers have been actively involved in investigating the automated detection of offensive and hate speech content as well as many shared tasks were organising (Waseem and Hovy, 2016; Kumar et al., 2018; Mandl et al., 2019; Zampieri et al., 2020; Davani et al., 2023). However, there is considerably less research on violence detection specifically. A few works are as follows: Cano Basave et al. (2013) present the Violence Detection Model (VDM), a probabilistic framework for identifying violent content and extracting violence-related topics from social media without requiring labeled data. VDM uses word prior knowledge derived from relative entropy to cap-

[1]Our code is available at https://github.com/souro/classification_tasks_bangla

Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), pages 220–224

ture word violence indicators, outperforming information gain methods in topic identification and violence classification. Chang et al. (2018) address the detection of aggression and loss in social media, particularly among gang-involved youth. Their system incorporates contextual representations and domain-specific resources, improving the Convolutional Neural Network's performance for detecting aggression and loss. Jahan et al. (2022) introduce BanglaHateBERT, a retrained BERT model for abusive language detection in Bangla. It outperforms generic pre-trained models on various datasets and includes a 15K Bangla hate speech dataset for research use. Zandam et al. (2023) explore the expression of threatening themes in the Hausa language on social media, developing a classification system using machine learning algorithms. XGBoost achieves the highest accuracy of 72% in classifying threatening content. Abercrombie et al. (2023) conduct a systematic review of resources for automated identification of online Gender-Based Violence (GBV), highlighting limitations in existing datasets, such as a lack of theoretical grounding and stakeholder input. The study recommends future resources grounded in sociological expertise and involving GBV experts and those with lived GBV experience.

3 Dataset

The VITD Shared Task 1 dataset (Saha et al., 2023b) was provided by the task organisers. Individual samples in the dataset are labeled as Direct Violence, Indirect Violence, and Non-Violence, which are represented numerically by 2, 1 and 0 respectively (see Saha et al., 2023b for further details).

The dataset is divided into training, development and test sets, consisting of 2,700, 1,330 and 2,016 samples respectively.[2]

4 Experiments

This section discusses an extensive account of the system we designed for the VITD and Sentiment Analysis of Bangla Social Media Posts tasks. Our strategy encompasses several stages, such as data preprocessing, model choice, hyperparameter adjustment, and advanced methods, all aimed at attaining commendable outcomes.

[2] https://github.com/blp-workshop/blp_task1/tree/main/dataset

4.1 Data Preprocessing

At the outset, a thorough data preprocessing and cleaning phase was performed for our system, which established a robust basis for subsequent operations. We harnessed the tools offered by the Bangla Natural Language Processing (BNLP) toolkit (Sarker, 2021). In addition to basic text processing, we implemented crucial transformations like setting `fix_unicode=True`, `unicode_norm=True`, and `unicode_norm_form="NFKC"`. These steps ensured consistent and standardized text representations, enhancing the quality of our dataset.

4.2 Model Selection

Our system employed a range of pre-trained sequence classification models to tackle the classification tasks effectively. Notable models we experimented with include `XLMRobertaForSequenceClassification`, `BertForSequenceClassification`, and their variants. Specifically, we explored the following models: XLM-RoBERTa (base and large versions) (Conneau et al., 2019), BanglaBERT '' (Bhattacharjee et al., 2022), Bangla BERT Base (Sarker, 2020) and BERT-base-multilingual-cased (Devlin et al., 2018). [3] After thorough evaluation, we found the XLM-RoBERTa-base model to perform best on this task.

4.3 Hyperparameter Tuning

Based on hyperparameter search on the development data, we chose the following hyperparameter settings: batch size of 5,learning rate (lr) 1e-5, using the AdamW optimizer (Loshchilov and Hutter, 2019), training for 15 epochs, setting gradient clipping to 1.0, a weight decay of 0.01, and a dropout rate of 0.1.

4.4 Sampling Strategies

Class imbalance arises when certain classes have notably fewer samples than others, potentially leading to bias in favour of the majority class within the model. This is the case in tasks such as violence detection, where violent texts are in the minority. To

[3] We use the models from HuggingFace: https://huggingface.co/xlm-roberta-base, https://huggingface.co/xlm-roberta-large, https://huggingface.co/csebuetnlp/banglabert, https://huggingface.co/sagorsarker/bangla-bert-base, https://huggingface.co/bert-base-multilingual-cased.

address class imbalance issues, we experimented with both oversampling and undersampling techniques. Although the outcomes were promising, our best-performing model ultimately adopted an alternative approach – focal loss.

Focal Loss (Lin et al., 2017) was incorporated as a specialized loss function to combat the class imbalance issues present in our classification tasks. Focal Loss (Lin et al., 2017) works by significantly reducing the loss for correctly classified examples with high confidence, effectively handling easy instances. Simultaneously, it provides a smaller reduction in loss for difficult-to-classify or misclassified examples, ensuring that the model concentrates on learning from problematic cases. The key idea behind Focal Loss is to give more attention to hard-to-classify examples while reducing the impact of well-classified examples. This is achieved through two essential parameters: `alpha` and `gamma`.

Alpha Parameter (`alpha`): In our system, we set `alpha` to 1. This value signifies that we assigned equal weight to all classes. By doing so, we aimed to ensure that our model did not exhibit bias towards any specific class. However, adjusting `alpha` allows for a flexible weighting scheme, where higher values give more importance to minority classes.

Gamma Parameter (`gamma`): We chose a `gamma` value of 2. This parameter regulates the rate at which the loss decreases as the predicted probability for the correct class increases. A higher `gamma` value, as in our case, slows down the loss reduction for well-classified examples. This design decision helped our model focus on challenging or misclassified instances, potentially leading to improved overall performance.

In summary, Focal Loss played a crucial role in enhancing the performance of our system, especially in scenarios with imbalanced class distributions. Our choice of `alpha` and `gamma` parameters aligns with standard practices for effectively leveraging Focal Loss to tackle classification challenges.

4.5 Data Augmentation

The diversity and robustness of our model was enhanced through data augmentation. A data augmentation strategy with a probability of 0.5 was introduced on the original data (Saha et al., 2023b).

Model	macro-F1
BanglaBERT Baseline	0.7879
XLM-RoBERTa base Baseline	0.7292
BERT multilingual base (cased) Baseline	0.6819
BLP Shared Task 1 winning system	76.044
Our system	69.009

Table 1: UFAL-ULD team and baseline systems results

The techniques employed included synonym replacement, insertion, deletion, swap, and shuffling (cf. Mukherjee and Dusek, 2023). Through a collective application of these techniques, a diverse set of augmented data was generated that proved vital to the performance of our best-reported model.

In summary, a systematic approach for data preprocessing, model selection, hyperparameter tuning, class imbalance handling, the integration of advanced loss functions, and data augmentation was employed to achieve competitive results for the VITD task.

5 Results

The macro-F1 metric has been used for evaluation measure in the BLP Shared Task 1 (Saha et al., 2023a), with comparisons made against the ground truth labels. This metric signifies the comprehensive effectiveness of our system in accurately categorizing text that incites violence into the specified classifications: Direct Violence, Passive Violence, and Non-Violence. The macro F1 score is a resilient measurement that considers precision and recall across all categories, making it particularly suitable for tasks with imbalanced class distributions. Our system achieved a macro F1 score of 69.01 on the test set (see Table 1), outperforming baselines. Our system was ranked 20th among the 27 teams that participated in the task.[4]

6 Conclusion

In this shared task on Violence Inciting Text Detection (VITD), we have presented our system's approach and results, emphasizing the significance of addressing the challenging problem of identifying and categorizing violence-inciting text in the Bangla language. Our system, equipped with a comprehensive set of natural language processing techniques, achieved a competitive macro F1 score of 69.009 on the test set. Our system was ranked 20th among the 27 teams that participated in the task.

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

We remain committed to further refining our system and exploring innovative approaches to contribute to the ongoing efforts in violence detection and prevention.

Limitations

While our system performed well in the VITD shared task, it is essential to acknowledge certain limitations:

Data Availability: Our system's performance heavily relies on the quality and quantity of training data. The availability of more extensive and diverse annotated datasets in Bangla could further enhance our system's capabilities.

Ethical Considerations: As with any content analysis task, there is the potential for bias and sensitivity in handling violent or offensive text. Ensuring ethical considerations and responsible AI practices are crucial in the development and deployment of such systems.

Ethics Statement

In developing our system for the Violence Inciting Text Detection task, we adhered to ethical principles and guidelines for responsible AI. We are committed to the following ethical considerations:

Data Privacy: We respect data privacy and ensure that any data used in our experiments are anonymized and do not contain personally identifiable information.

Bias Mitigation: We took measures to mitigate bias in our system, both in terms of model performance and the potential impact of our work on society. We recognize the importance of fairness and impartiality in automated content analysis.

Transparency: We are committed to transparency in our research and have provided a detailed system description, including preprocessing steps, model selection, and evaluation metrics.

Accountability: We are open to feedback and accountability for our work. We encourage responsible use and scrutiny of AI technologies, and we remain responsive to concerns or issues related to our system's functionality.

By adhering to these principles, we aim to contribute to the responsible development and deployment of AI systems for content analysis, with a focus on promoting online safety and mitigating harm.

Acknowledgements

This research was supported by the European Research Council (Grant agreement No. 101039303 NG-NLG) and by Charles University projects GAUK 392221 and SVV 260575. We acknowledge of the use of resources provided by the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth, and Sports project No. LM2018101).
Atul Kr. Ojha would like to acknowledge the support of the Science Foundation Ireland (SFI) as part of Grant Number SFI/12/RC/2289_P2 Insight_2, Insight SFI Centre for Data Analytics.

References

Gavin Abercrombie, Aiqi Jiang, Poppy Gerrard-abbott, Ioannis Konstas, and Verena Rieser. 2023. Resources for automated identification of online gender-based violence: A systematic review. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 170–186, Toronto, Canada. Association for Computational Linguistics.

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, Seattle, United States. Association for Computational Linguistics.

Amparo Elizabeth Cano Basave, Yulan He, Kang Liu, and Jun Zhao. 2013. A weakly supervised Bayesian model for violence detection in social media. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 109–117, Nagoya, Japan. Asian Federation of Natural Language Processing.

Serina Chang, Ruiqi Zhong, Ethan Adams, Fei-Tzin Lee, Siddharth Varia, Desmond Patton, William Frey, Chris Kedzie, and Kathy McKeown. 2018. Detecting gang-involved escalation on social media using context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 46–56, Brussels, Belgium. Association for Computational Linguistics.

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.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Md Saroar Jahan, Mainul Haque, Nabil Arhab, and Mourad Oussalah. 2022. BanglaHateBERT: BERT for abusive language detection in Bengali. In *Proceedings of the Second International Workshop on Resources and Techniques for User Information in Abusive Language Analysis*, pages 8–15, Marseille, France. European Language Resources Association.

Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. *COLING 2018*, page 1.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandalia, and Aditya Patel. 2019. Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in indo-european languages. In *FIRE '19: Forum for Information Retrieval Evaluation, Kolkata, India, December, 2019*, pages 14–17. ACM.

Sourabrata Mukherjee and Ondrej Dusek. 2023. Leveraging Low-resource Parallel Data for Text Style Transfer. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 388–395, Prague, Czechia. Association for Computational Linguistics.

Sourav Saha, Jahedul Alam Junaed, Nabeel Mohammed, Sudipta Kar, and Mohammad Ruhul Amin. 2023a. Blp-2023 task 1: Violence inciting text detection (vitd). In *Proceedings of the 1st International Workshop on Bangla Language Processing (BLP-2023)*, Singapore. Association for Computational Linguistics.

Sourav Saha, Jahedul Alam Junaed, Maryam Saleki, Arnab Sen Sharma, Mohammad Rashidujjaman Rifat, Mohamed Rahout, Syed Ishtiaque Ahmed, Nabeel Mohammad, and Mohammad Ruhul Amin. 2023b. Vio-lens: A novel dataset of annotated social network posts leading to different forms of communal violence and its evaluation. In *Proceedings of the 1st International Workshop on Bangla Language Processing (BLP-2023)*, Singapore. Association for Computational Linguistics.

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

Sagor Sarker. 2021. BNLP: natural language processing toolkit for bengali language. *CoRR*, abs/2102.00405.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the Student Research Workshop, SRW@HLT-NAACL 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 88–93. The Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of SemEval*.

Abubakar Yakubu Zandam, Fatima Adam Muhammad, and Isa Inuwa-Dutse. 2023. Online threats detection in hausa language. In *Proceedings of the 4th Workshop on African Natural Language Processing, AfricaNLP@ICLR 2023, Kigali, Rwanda, May 1, 2023*.