LowResourceNLU at BLP-2023 Task 1 & 2: Enhancing Sentiment Classification and Violence Incitement Detection in Bangla Through Aggregated Language Models

Hariram Veeramani
Department of Electrical and Computer Engineering, UCLA, USA
hariram@ucla.edu

Surendrabikram Thapa
Department of Computer Science, Virginia Tech, Blacksburg, USA
sbt@vt.edu

Usman Naseem
College of Science and Engineering, James Cook University, Australia
usman.naseem@jcu.edu.au

Abstract

Violence incitement detection and sentiment analysis hold significant importance in the field of natural language processing. However, in the case of the Bangla language, there are unique challenges due to its low-resource nature. In this paper, we address these challenges by presenting an innovative approach that leverages aggregated BERT models for two tasks at the BLP workshop in EMNLP 2023, specifically tailored for Bangla. Task 1 focuses on violence-inciting text detection, while task 2 centers on sentiment analysis. Our approach combines fine-tuning with textual entailment (utilizing BanglaBERT), Masked Language Model (MLM) training (making use of BanglaBERT), and the use of standalone Multilingual BERT. This comprehensive framework significantly enhances the accuracy of sentiment classification and violence incitement detection in Bangla text. Our method achieved the 11th rank in task 1 with an F1-score of 73.47 and the 4th rank in task 2 with an F1-score of 71.73. This paper provides a detailed system description along with an analysis of the impact of each component of our framework.

1 Introduction

Natural Language Processing (NLP) has witnessed remarkable advancements in recent years, transforming the way we interact with and understand textual data (Khurana et al., 2023). From chatbots and machine translation to information retrieval and sentiment analysis, NLP has become an indispensable tool for extracting meaning from the vast sea of human-generated text (Sun et al., 2022). Among the diverse array of NLP tasks, sentiment analysis, and violence incitement detection stand out as pivotal areas with far-reaching implications for societal well-being and communication (Khalafat et al., 2021; Castorena et al., 2021).

THIS PAPER CONTAINS EXAMPLES OF VIOLENT TEXT.

Sentiment analysis, also known as opinion mining, is a fundamental NLP task focused on identifying emotional tones and polarities within the text (Cui et al., 2023). It plays a crucial role in various applications, including gauging public opinion, analyzing consumer feedback, monitoring social media, and managing brand perception. By providing insights into sentiment, it empowers informed decision-making, personalized communication, and more effective response strategies (Wankhade et al., 2022). Similarly, in an increasingly digital world, the spread of harmful content, including violence-inciting text, poses significant challenges (Parihar et al., 2021). Violence incitement detection is a critical aspect of content moderation, ensuring online platforms remain safe and free from content that promotes harm, hatred, or illegal activities. Early identification of such content is vital in mitigating potential harm, preserving online discourse, and upholding ethical standards in digital communication.

While the significance of sentiment analysis and violence incitement detection is widely recognized, applying these techniques to low-resource languages presents unique hurdles (Sen et al., 2022). The Bangla language, with its rich linguistic diversity, is a prime example. Despite its extensive speaker base, Bangla remains underrepresented in NLP research, often lacking the comprehensive language resources available for widely spoken languages (Kowsher et al., 2022). This scarcity of resources hinders the development of effective sentiment analysis and violence incitement detection tools for Bangla.

We address the aforementioned problems by presenting a novel approach based on the aggregation of BERT-based models. In this paper, we provide detailed descriptions of our systems for two tasks at the BLP workshop. Our contributions include:

- Our method encompasses three unique

Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), pages 230–235

Text	Translation	Label
ঢাকা কলেজে আগুন লাগিয়ে এই কুলাঙ্গার ছাত্রদের পুরিয়ে মারা উচিৎ,, এরাই এখন গলার কাটা	These Kulanga students should be killed by setting fire to Dhaka College, they are now cut throat	Direct Violence
শয়তান মেরে হাসবে না তো কাঁদবে!!	The devil will not laugh but cry!!	Passive Violence
যে মারা গেল তার ক্ষতিপূরনের ব্যবস্থা করে দেওয়া হোক।	Compensation should be paid to the person who died.	Non-Violence

Table 1: Examples of text used in task 1 (Violence Inciting Text Detection)

approaches: simultaneous fine-tuning of BanglaBERT for MLM and classification tasks, straightforward utilization of Multilingual BERT (mBERT), and a multi-head training strategy addressing two distinct topics (entailment and classification), collectively enhancing performance in natural language processing tasks.

- We conduct ablation studies to analyze the individual effects of each component in our proposed methodology, shedding light on their respective contributions.

2 Task Descriptions

Task 1: This task focuses on violence incitement text classification (Saha et al., 2023b). The primary objective is to identify and classify Bangla text comments that contain threats associated with violence, which have the potential to incite further acts of violence. Participants were required to categorize the comments into three distinct categories: "Direct Violence", "Passive Violence", and "Non-Violence".

Task 2: It addresses sentiment analysis, aiming to detect the sentiment expressed within a given Bangla text (Hasan et al., 2023a). It constitutes a multi-class classification challenge where participants are tasked with determining whether the sentiment in the Bangla text is "Positive", "Negative", or "Neutral".

3 Dataset

For task 1, participants are presented with a Bangla dataset comprising YouTube comments related to the top 9 violent incidents that have occurred in the Bengal region (comprising Bangladesh and West Bengal) over the past decade, with comments up to 600 words long (Saha et al., 2023a). The training set (2700 samples) comprises approximately 15% direct violence, 34% passive violence, and 51% non-violent instances. In the development set (1330 samples), a similar distribution is observed: 15% direct violence, 31% passive violence, and 54% non-violence. Table 1 shows examples of texts used in task 1.

For task 2, the given dataset combines two primary sources: the MUltiplatform BAngla SEntiment (MUBASE) (Hasan et al., 2023b) and SentNob (Islam et al., 2021) datasets. Thus, this dataset includes public comments on news and videos across 13 domains, and multiplatform content such as Tweets and Facebook posts, all manually annotated for sentiment polarity as shown in Table 2.

Text	Translation	Sentiment
বিবিসি মানের বাবাহীন সন্তান	BBC Standard Fatherless Child.	Negative
আমি আপনার সাথে সম্পূর্ণ একমত । ।	I totally agree with you. .	Neutral
শেখ রেহানা : এক সংগ্রামী জীবনের প্রতিচ্ছবি	Sheikh Rehana: A reflection of a struggling life.	Positive

Table 2: Examples of text used in task 2 (sentiment analysis)

4 System Description

In our methodology, we aggregate three BERT-based language models in order to tackle both classification tasks. The proposed methodology of our system is as shown in Figure 1.

Model A: In this model configuration, we incorporate two heads within the BanglaBERT-large (Bhattacharjee et al., 2022) framework. Our choice of incorporating two heads in the BanglaBERT-large architecture, one for Masked Language Modeling (MLM) and the other for classification, is driven by a thoughtful rationale and strong motivation. Firstly, this dual-headed approach enables us to retain the invaluable language understanding capabilities embedded in pre-

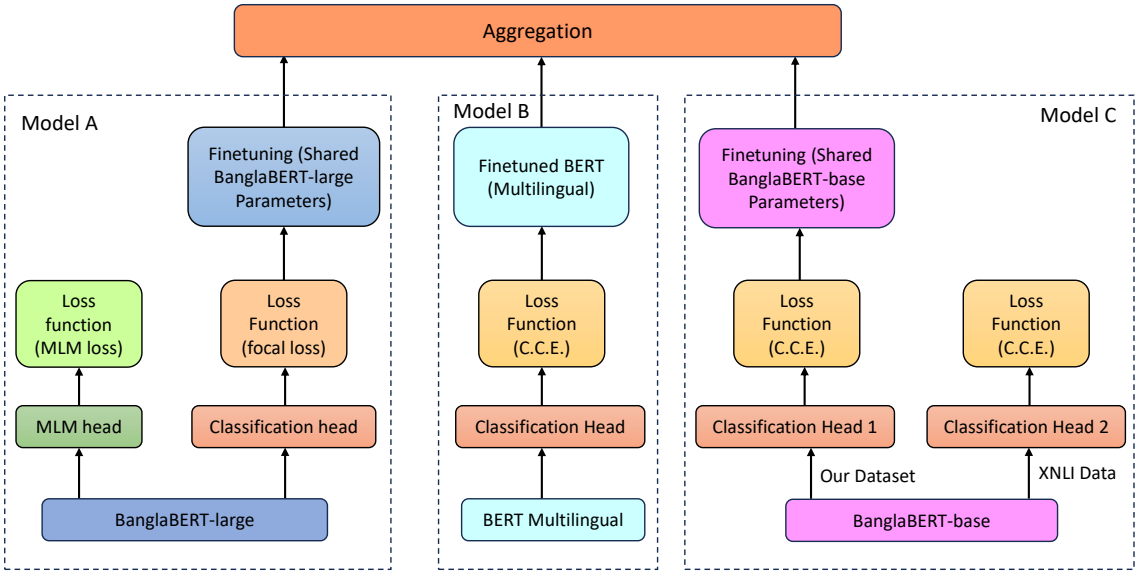

Figure 1: Proposed methodology for our system. Our methodology combines three BERT-based models capitalizing on the strength of each of them.

trained BanglaBERT. The MLM head, through cross-entropy loss (label vs. predicted), maintains and refines BanglaBERT's grasp of linguistic nuances, ensuring that it remains adept at capturing contextual word relationships. To enhance the MLM's effectiveness, we employ a balanced masking strategy. Specifically, within the MLM head, we utilize a 50% deterministic and 50% random masking approach. In random masking, we mask random words in the input text, while in deterministic masking, we append a mask token to our text. This dual approach enhances MLM's robustness in capturing contextual language information.

Secondly, the classification head, leveraging a specialized loss function like focal loss, empowers BanglaBERT to adapt swiftly and effectively to specific downstream tasks i.e. classification. This dynamic adaptability is crucial, as it enables BanglaBERT to excel in diverse applications, such as sentiment analysis, textual entailment, or any classification task at hand. Simultaneous training with shared parameters efficiently fuses the strength of both heads, resulting in a compact and versatile model that excels in various natural language processing tasks (Veeramani et al., 2023b,f,a), particularly classification.

Model B: It is Multilingual BERT (mBERT) (Devlin et al., 2019), a versatile architecture designed to handle multiple languages (Kassner et al., 2021; Xu et al., 2021; Veeramani et al., 2023c,e,d), including Bangla. Leveraging mBERT's rich multilingual knowledge, our methodology gains valuable linguistic insights, enhancing our understanding of Bangla text.

Model C: This model introduces a multi-head training strategy, simultaneously addressing two distinct yet interrelated tasks. The first head within BanglaBERT-base focuses on the XNLI dataset (Conneau et al., 2018), specifically targeting the task of textual entailment. This choice is motivated by the rationale of knowledge fusion, aiming to merge insights and linguistic patterns from both textual entailment and classification domains. XNLI has languages like Hindi, Urdu, and Swahili whose dialects and cultural nuances are similar to Bangla. We hypothesize that this helps the model to acquire better parameters. By sharing parameters across heads, the model seeks to develop a deeper and more comprehensive understanding of Bangla language nuances. The second head is dedicated to our data, which is centered around a classification problem. This dual-task approach not only boosts efficiency but also contributes to achieving our primary objective: solving the classification problem. The inclusion of the textual entailment task acts as an auxiliary training signal, facilitating the acquisition of versatile and adaptable language representations. This, in turn, aids in achieving superior performance in our core clas-

sification task, making Model C a powerful and efficient component of our methodology.

For all models A, B, and C, we made trials with focal loss and cross-entropy loss and used the loss function which gave the optimal performance. We also made trials with BanglaBERT-large and BanglaBERT-base and selected the most optimal framework as shown in Figure 1. All models have objective function as classification.

Aggregation: Our aggregation technique employs a multi-step process to effectively combine predictions from multiple models. Initially, we extract individual predictions from each model using the argmax function (Davani et al., 2022; Kanagasabai et al., 2023), selecting the class with the highest confidence score for each model. Subsequently, to consolidate these individual predictions, we apply another argmax operation, this time on the maximum logit values obtained from each model. This step ensures that we capture the most confident prediction across all models. If two labels have equal highest probabilities, we select the majority sample class.

5 Results

Performance on the task 1 and task 2 were evaluated on the basis of macro and micro F1-score respectively. Our team ranks 11th in task 1 with F1-score of 73.47. Similarly, our team ranks 4th in task 2 with macro F1-score of 71.72. Table 3 provides a comprehensive analysis of the impact of various models within our architecture, presenting macro-averaged F1-scores, precision, and recall for both tasks. In our analysis, we meticulously evaluate the impact of all models, focusing on a detailed assessment of Model A and Model C. We specifically delve into the effects of two crucial aspects: the integration of MLM (Masked Language Model) in Model A and the influence of joint pretraining with the XNLI dataset in Model C. Our Task 1 results demonstrate that Model A enhances the F1-score by a substantial margin, surpassing a 3.3-point improvement through the incorporation of MLM. Similarly, the joint pretraining with XNLI significantly enhances the performance of Model C by approximately 2.1 points. Model B alone gives an F1-score of 69.45. The combination of all components (Model A + B + C) exhibit superior performance as compared to use of single model alone.

In Task 2, which focuses on sentiment analysis,

Models	F1-score	Precision	Recall
Model A only	73.41	73.65	77.64
Model B only	69.45	70.28	70.87
Model C only	73.42	73.91	77.73
Model A w/o MLM	70.10	72.06	73.51
Model C w/o XNLI	71.34	73.17	76.00
Proposed (Model A + B + C)	**73.47**	**74.1**	**77.92**

Table 3: Results for Task 1 (Violence Incitement Text Detection). The F1-score, precision and recall are macro-averaged.

Table 4 provides a detailed performance analysis of various models. Model A without the inclusion of the Masked Language Model (MLM) component achieves an F1-micro score of 71.03, while Model C, operating without joint pretraining using the XNLI dataset, achieves an F1-micro score of 71.06. When evaluated independently, Model A attains an F1-micro score of 71.71, and Model C achieves a slightly higher F1-micro score of 71.72. Model B, on the other hand, was able to score an micro F1-score of 69.47. However, our proposed framework, which combines all three models (Model A, Model B, and Model C), outperforms these individual models. It achieves the highest F1-micro score of 71.73, highlighting the substantial improvement gained through the synergy of all models. Additionally, the framework excels in macro-averaged precision, recall, and F1-score, with values of 71.08, 71.73, and 71.36, respectively. These results underscore the effectiveness of our integrated approach in sentiment analysis, showcasing the value of combining multiple models for superior accuracy and performance.

Models	$F1_{mic}$	Pre_{mac}	Rec_{mac}	$F1_{mac}$
Model A only	71.71	70.43	71.72	70.67
Model B only	69.47	68.32	70.85	68.50
Model C only	71.72	71.06	71.70	71.34
Model A w/o MLM	71.03	68.95	71.00	69.00
Model C w/o XNLI	71.06	69.39	71.03	69.20
Proposed (Model A + B + C)	**71.73**	**71.08**	**71.73**	**71.36**

Table 4: Results for task 2 (sentiment analysis). The $F1_{mic}$ stands for micro-averaged F1-score. Similarly, Pre_{mac}, Rec_{mac}, and $F1_{mac}$ represents macro-averaged precision, recall and F1-score.

6 Conclusion

In conclusion, our methodology presents a detailed and novel approach to addressing the challenges of sentiment analysis and violence detection in Bangla text. By aggregating insights from three different language models, we achieve a high performance in both tasks. Through a detailed ablation analysis, we have analyzed the impact of each component, demonstrating the efficiency of our proposed approach. While our primary focus lies in sentiment analysis and violence detection, the consistently high performance across both tasks underscores the potential versatility of our method in various other text analysis applications in Bangla. In the future, more research can be done on bias mitigation, ensuring responsible and equitable deployment of our framework in a real-world context.

Limitations

We proposed a methodology primarily focused on sentiment analysis and violence incitement detection. In this process, we might be potentially overlooking other aspects of text analysis. The adaptability to different domains may require further fine-tuning, and the scalability of our approach could be challenged with very large datasets.

Ethics Statement

The framework may potentially generate biased interpretations, a critical aspect that requires thorough investigation before considering the deployment of our model in real-world applications. It is essential to note that we did not undertake a comprehensive bias analysis within the scope of this work, highlighting the need for future research to meticulously examine and mitigate any biases that might arise in practical implementations of our methodology.

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, Seattle, United States. Association for Computational Linguistics.

Carlos M Castorena, Itzel M Abundez, Roberto Alejo, Everardo E Granda-Gutiérrez, Eréndira Rendón, and Octavio Villegas. 2021. Deep neural network for gender-based violence detection on twitter messages. *Mathematics*, 9(8):807.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. 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*. Association for Computational Linguistics.

Jingfeng Cui, Zhaoxia Wang, Seng-Beng Ho, and Erik Cambria. 2023. Survey on sentiment analysis: evolution of research methods and topics. *Artificial Intelligence Review*, pages 1–42.

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.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Md. Arid Hasan, Firoj Alam, Anika Anjum, Shudipta Das, and Afiyat Anjum. 2023a. Blp-2023 task 2: Sentiment analysis. In *Proceedings of the 1st International Workshop on Bangla Language Processing (BLP-2023)*, Singapore. Association for Computational Linguistics.

Md Arid Hasan, Shudipta Das, Afiyat Anjum, Firoj Alam, Anika Anjum, Avijit Sarker, and Sheak Rashed Haider Noori. 2023b. Zero-and few-shot prompting with llms: A comparative study with fine-tuned models for bangla sentiment analysis. *arXiv preprint arXiv:2308.10783*.

Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. 2021. SentNoB: A dataset for analysing sentiment on noisy Bangla texts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3265–3271, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rajaraman Kanagasabai, Saravanan Rajamanickam, Hariram Veeramani, Adam Westerski, and Kim Jung Jae. 2023. Semantists at ImageArg-2023: Exploring Cross-modal Contrastive and Ensemble Models for Multimodal Stance and Persuasiveness Classification. In *Proceedings of the 10th Workshop on Argument Mining*, Online and in Singapore. Association for Computational Linguistics.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual lama: Investigating knowledge in multilingual pretrained language models. *arXiv preprint arXiv:2102.00894*.

Monther Khalafat, S Alqatawna Jafar, Rizik Al-Sayyed, Mohammad Eshtay, and Thaeer Kobbaey. 2021. Violence detection over online social networks: An arabic sentiment analysis approach. *iJIM*, 15(14):91.

Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2023. Natural language processing: State of the art, current trends and challenges. *Multimedia tools and applications*, 82(3):3713–3744.

Md Kowsher, Abdullah As Sami, Nusrat Jahan Prottasha, Mohammad Shamsul Arefin, Pranab Kumar Dhar, and Takeshi Koshiba. 2022. Bangla-bert: transformer-based efficient model for transfer learning and language understanding. *IEEE Access*, 10:91855–91870.

Anil Singh Parihar, Surendrabikram Thapa, and Sushruti Mishra. 2021. Hate speech detection using natural language processing: Applications and challenges. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1302–1308. IEEE.

Sourav Saha, Jahedul Alam Junaed, Arnab Sen Sharma Api, Nabeel Mohammad, and Mohammad Ruhul 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.

Sourav Saha, Jahedul Alam Junaed, Nabeel Mohammed, Sudipta Kar, and Mohammad Ruhul 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.

Ovishake Sen, Mohtasim Fuad, Md Nazrul Islam, Jakaria Rabbi, Mehedi Masud, Md Kamrul Hasan, Md Abdul Awal, Awal Ahmed Fime, Md Tahmid Hasan Fuad, Delowar Sikder, et al. 2022. Bangla natural language processing: A comprehensive analysis of classical, machine learning, and deep learning-based methods. *IEEE Access*, 10:38999–39044.

Tian-Xiang Sun, Xiang-Yang Liu, Xi-Peng Qiu, and Xuan-Jing Huang. 2022. Paradigm shift in natural language processing. *Machine Intelligence Research*, 19(3):169–183.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023a. Automated Citation Function Classification and Context Extraction in Astrophysics: Leveraging Paraphrasing and Question Answering. In *Proceedings of the Second Workshop on Information Extraction from Scientific Publications*, Online. Association for Computational Linguistics.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023b. DialectNLU at NADI 2023 Shared Task: Transformer Based Multitask Approach Jointly Integrating Dialect and Machine Translation Tasks in Arabic. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023c. Enhancing ESG Impact Type Identification through Early Fusion and Multilingual Models. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing (FinNLP)*. Association for Computational Linguistics.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023d. KnowTellConvince at ArAIEval Shared Task: Disinformation and Persuasion Detection in Arabic using Similar and Contrastive Representation Alignment. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023e. LowResContextQA at Qur'an QA 2023 Shared Task: Temporal and Sequential Representation Augmented Question Answering Span Detection in Arabic. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023f. Temporally Dynamic Session-Keyword Aware Sequential Recommendation system. In *2023 International Conference on Data Mining Workshops (ICDMW)*.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

Haoran Xu, Benjamin Van Durme, and Kenton Murray. 2021. Bert, mbert, or bibert? a study on contextualized embeddings for neural machine translation. *arXiv preprint arXiv:2109.04588*.