CoU-CU-DSG at BLP-2025 Task 1: Leveraging Weighted Probabilistic Fusion of Language Models for Bangla Hate Speech Detection

Ashraful Alam[1], **Abdul Aziz**[2], **and Abu Nowshed Chy**[3]
Department of Information and Communication Technology
[1]Comilla University, Cumilla-3506, Bangladesh
Department of Computer Science and Engineering
[2]International Islamic University Chittagong, Chattogram-4318, Bangladesh
[3]University of Chittagong, Chattogram-4331, Bangladesh
ashrafulalam.cou.ict@gmail.com, aziz.abdul.cu@gmail.com, nowshed@cu.ac.bd

Abstract

The upsurge of social media and open source platforms has created new avenues for the rapid, global spread of negativity and obscenities targeting individuals and organizations. The process to identify hate speech is critical for the lexical and regional variation as well as the morphological complexity of the texts, especially in low-resource languages, e.g. Bangla. This paper presents our participation in the Hate Speech Detection task at the second workshop on Bangla Language Processing. The objective of this task is not only to detect whether the content is hateful, but also to identify the type of hate, the target group, and its severity. We proposed a Transformer-based weighted probabilistic fusion model to detect the presence of hate speech in Bangla texts. We independently fine-tuned three pre-trained Transformer models, BanglaBERT, XLM-RoBERTa, and MuRIL, to capture diverse linguistic representations. The probability distributions obtained from each model were combined using a weighted fusion strategy, allowing the system to leverage the strengths of all models simultaneously. This fused representation was then used to predict the final labels for the given instances. The experimental results showed that our proposed method obtained competitive performance, ranking 10th in subtask 1A and 15th in subtask 1B among the participants.

1 Introduction

The rapid growth of social networks and online platforms has facilitated communication and information sharing on an unprecedented scale. However, this has also led to the proliferation of harmful content, including hate speech, which can incite violence, discrimination, and social unrest (Roy et al., 2022; Mahajan et al., 2024). Online abuse and the spread of negativity are common practices and an important social problem that is highly correlated with the emergence of social media platforms (Antypas and Camacho-Collados, 2023). Detecting hate speech in Bangla is especially challenging due to the informal language, spelling variations, and the use of slang in comment sections. While most existing hate speech detection systems have been developed for English or other high-resource languages (Toraman et al., 2022; Nozza, 2021), research in Bangla hate speech detection, particularly in YouTube comments, remains limited.

To address this gap, we participated in both Subtask 1A and Subtask 1B of the shared task (Hasan et al., 2025b). Subtask 1A and Subtask 1B are multiclass text classification problems in which each comment is categorized into one of the hate speech classes or labeled as None. We propose a Transformer-based fusion model where we fine-tuned XLM-RoBERTa, BanglaBERT (Bhattacharjee et al., 2022), and MuRIL (Khanuja et al.), under various hyperparameter settings. To handle the strong class imbalance inherent in the data, we integrate a weighted loss function during training and evaluate performance using the micro f1 metric. Our approach achieves competitive performance across both subtasks, demonstrating its effectiveness in identifying harmful Bangla content. These findings highlight the potential of Transformer-based fusion models for low-resource languages and contribute to safer online interactions for Bangla-speaking users.

2 Related Work

Hate speech detection is a growing area in research, particularly with the uprising of social media (Toraman et al., 2022; Salles et al., 2025). While existing work has been done in well-resourced languages like English (Lee et al., 2022), there remains a substantial gap in research for low-resource languages, especially those with complex linguistic structures and script adaptation challenges (Nozza, 2021), such as Bangla.

Early foundational work in hate speech detection

Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 492–497
December 23, 2025 ©2025 Association for Computational Linguistics

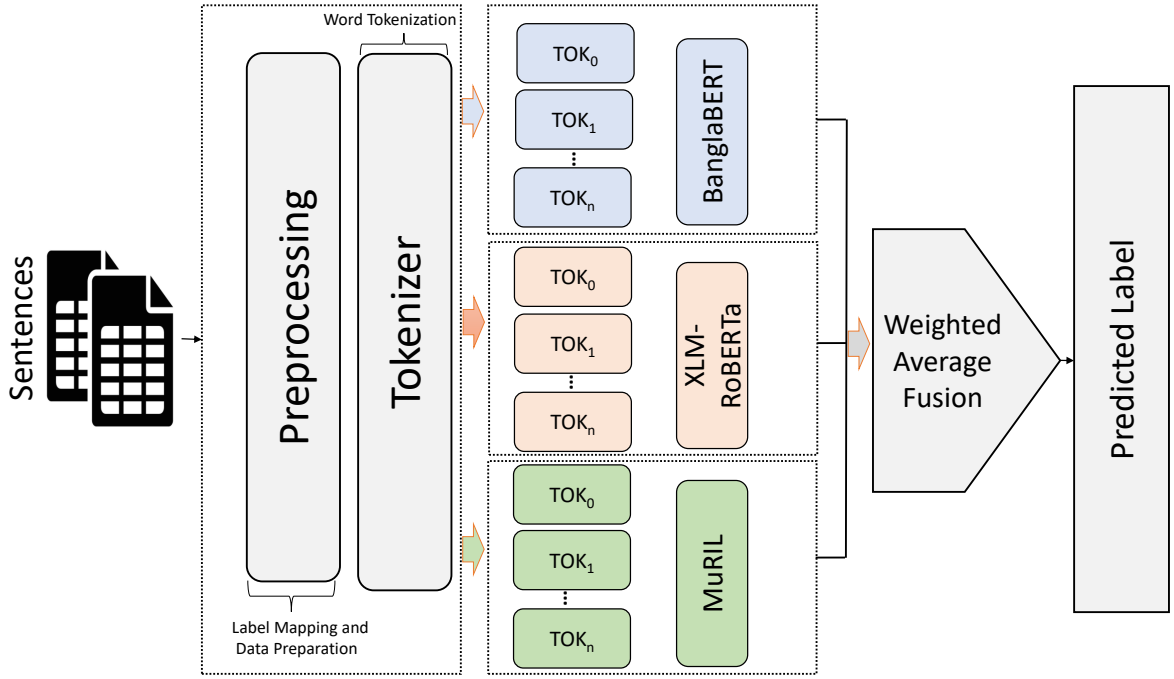

Figure 1: Our proposed model for hate speech detection.

on social media platforms like Twitter has been seen in (Talat and Hovy, 2016). Their research primarily focused on analyzing the impact of extralinguistic features, such as gender. They aimed to enhance the detection of hate speech (Singh and Thakur, 2024; Lee et al., 2022).

In the context of Bangla text, deep learning methods such as recurrent neural networks and long-short-term memory are used to handle hate speech detection in multilabel texts (Das et al., 2022). A key contribution was made by (Bhattacharjee et al., 2022), who developed an annotated dataset of 10,000 Bangla tweets, including both actual Bangla and Romanized Bangla. They implemented models like MuRIL and got excellent performance. More recently, a significant advancement in multiclass classification was made by (Bhattacharjee et al., 2022). They fine-tuned BanglaBERT on their training data. In contrast, we propose a fusion model leveraging three Transformer-based models, including BanglaBERT, XLM-RoBERTa, and MuRIL, to exploit the diverse contextual dimension of Bengali hate speech.

3 Methodology

In this section, we describe our proposed approach for the hate speech detection task. The overview of our framework is depicted in Figure 1.

Given an input text, we first map and assign weights to the labels to address class imbalance, then we employ three transformer models, including BanglaBERT (Bhattacharjee et al., 2022), XLM-RoBERTa (Antypas and Camacho-Collados, 2023), and MuRIL, to detect hate speech. Finally, for the effective fusion of the scores, we take the weighted arithmetic mean of the prediction scores of these models.

3.1 Transformer Models

Transformer models are adept at capturing long-term dependencies by leveraging multi-head attention and positioned embedding mechanisms. This approach facilitates a robust understanding of the relationships between words, which is essential for obtaining a richer, contextualized representation of the argument's context (Aziz et al., 2023). In this study on hate speech detection, we select three state-of-the-art multilingual and language-specific Transformer models as our foundation (Kim et al., 2022), including BanglaBERT, XLM-RoBERTa, and MuRIL. These models were chosen to effectively handle the lexical diversity present in our dataset, serving as the base architectures upon which our fine-tuning, Transformer-based approach is applied.

3.2 Fusion of Transformer Models

In the field of natural language processing (NLP), combining the strengths of multiple models is a

standard technique to boost performance beyond the capability of any single architecture and mitigate individual model limitations. Our proposed framework adopts a fusion strategy to synthesize the capabilities of BanglaBERT, XLM-RoBERTa, and MuRIL (Romim et al., 2022).

We achieve this synergy by estimating a single, unified probability score for each classification class. This score is derived from fusing the prediction scores generated independently by each of the three fine-tuned Transformer models. Specifically, we employ the weighted probabilistic mean of these three probability scores for the fusion process. By applying weights, we can prioritize the output of the model that shows the highest reliability or relevance for a given prediction. The final label for the input text is then determined by selecting the class associated with the resulting highest fused probability score.

4 Experiments and Results

4.1 Dataset and Preprocessing

We used the official datasets (Hasan et al., 2025a) provided by the Bangla Multi-task Hate Speech Identification Shared Tasks organizers (Hasan et al., 2025b). As shown in Table 1, the training, development, and test sets contain Bangla text instances annotated with six categories for subtask 1A. On the other hand, Table 2 shows the data distribution of subtask 1B that contains 5 categories of labels, such as None, Individual, Society, Community, and Organization.

Labels	Counts
None	19954
Profane	2331
Abusive	8212
Sexism	122
Political Hate	4227
Religious Hate	676

Table 1: Label counts of training dataset (subtask 1A).

The dataset used in this study exhibited a significant class imbalance, where some classes were underrepresented compared to others. To mitigate this issue during training, we employed a weighted cross-entropy loss (Vázquez-Osorio et al., 2024).

$$\mathcal{L}_{\text{weighted}} = -\sum_{i=1}^{C} w_i \, y_i \, \log \hat{y}_i$$

Labels	Counts
None	21190
Individual	5646
Society	2205
Community	2635
Organization	3846

Table 2: Label counts of training dataset (subtask 1B).

In this formulation, C denotes the total number of classes in the dataset. The term y_i represents the ground-truth label for class i expressed in a one-hot encoded format, while $\hat{y}_i$ corresponds to the predicted probability assigned by the model to class i. The coefficient w_i is a class-specific weight that determines the relative importance of each class in the loss calculation, assigning larger penalties to misclassifications from minority classes and smaller penalties to the majority classes. The class weights w_i were derived from the distribution of the training data using the following expression:

$$w_i = \frac{N}{C \, n_i} \tag{1}$$

Here, N is the total number of training samples, n_i denotes the number of samples belonging to class i, and C again is the total number of classes. This formulation ensures that classes with fewer instances are assigned proportionally higher weights, thereby balancing the contribution of each class to the overall loss and reducing the bias towards majority classes. (Al Maruf et al., 2024). We tokenized texts separately for each model using their respective models' tokenizers, BanglaBERT, XLM-RoBERTa, and MuRIL, splitting into subword units, padding or truncating to 256 tokens, and converting them to PyTorch tensors for training, validation, and testing.

4.2 Experimental Settings

We now present the details of the experimental setup, including the specific hyperparameter configurations and fine-tuning strategies utilized to develop our proposed system. We performed the fine-tuning process using the following key configurations, which were defined within the Training Arguments class. The models were trained for three epochs. We utilized a training batch size of 16, and set the learning rate to the standard pre-trained optimization value of 2e-5. Additionally, a weight

decay of 0.01 was applied to mitigate overfitting. Training logs were recorded every 50 steps. To optimize resources, we disabled immediate evaluation during training and set the model saving frequency to zero. In our weighted probabilistic fusion, we consider weights of 0.5, 0.3, and 0.2 for BanglaBERT, XLM-RoBERTa, and MuRIL, respectively, based on their individual performance.

4.3 Results and Analysis

To evaluate the performance of the participant's system at the BLP25 hate speech detection shared task (Hasan et al., 2025b), the micro f1 score is considered as the main evaluation matrics for subtask 1A and subtask 1B.

Team	Position	Score
shifat_islam	1st	0.7362
SyntaxMind	2nd	0.7345
nahidhasan	7th	0.7305
CoU-CU-DSG	10th	0.7273
pritampal98	19th	0.7057
programophile	21st	0.7013
intfloat	33rd	0.6634

Table 3: Comparative results with other selected participants (Subtask 1A).

Team	Position	Score
mahim_ju	1st	0.7356
shifat_islam	2nd	0.7335
nahidhasan	8th	0.7279
CoU-CU-DSG	15th	0.7114
pritampal98	19th	0.6974
lamiaa	24th	0.2848

Table 4: Comparative results with other selected participants (Subtask 1B).

The performance of our proposed system in the BLP25 hate speech detection shared task is analyzed across Subtask 1A and Subtask 1B in this section. Table 3 and Table 4 present the comparative results for Subtask 1A and Subtask 1B, respectively, contrasting our system with other top entries. At first, we presented the performance of our proposed system. We also presented the performance of top-ranked participating systems and the baseline used in subtask 1A and subtask 1B. Here, we see that our proposed method obtained a good score in terms of the primary evaluation metric micro f1 score.

Method	Dev set	Test set
BanglaBERT	0.7356	0.7156
XLM-R	0.6937	0.6735
MuRIL	0.6846	0.6628
BanglaBERT+XLM-R	0.7308	0.7242
BanglaBERT+MuRIL	0.7348	0.7211
XLM-R+MuRIL	0.7225	0.7078
Proposed Fusion Model	0.7456	0.7273

Table 5: Ablation study of our proposed model (Subtask 1A). XLM-R represents the XLM-RoBERTa model.

In our proposed system, we perform the effective fusion of three Transformer models. However, to validate the performance of our fusion strategy, we conduct evaluate the performance of each model used in our proposed system. The results of the ablation study of our proposed model are articulated in Table 5. From the results, it is observed that BanglaBERT performed better compared to other models when considering individual model performances. However, combining the three models' prediction scores by using a weighted average improved the performance. It shows that the fusion strategy improve the $\sim$2% performance compared to the BanglaBERT model and improves the $\sim$6% performance compared to the other two models in terms of the evaluation measure micro f1 score. This validates the importance of our fusion strategy.

5 Conclusion and Future Directions

In this paper, we present an approach to detect hate speech from Bangla texts leveraging three BERT variants, including BanglaBERT, XLM-RoBERTa, and MuRIL, with an effective weighted fusion strategy. Experimental results demonstrate the efficiency of our fusion model, which helped us to obtain a competitive position in both subtask 1A and subtask 1B.

In the future, we intend to explore feature engineering, training strategies, and other pretrained models for further improvement. We also have a plan to explore the external knowledge of other similar domains, as well as the strength of large language models (LLMs) in this task.

Limitations

Although our proposed weighted probabilistic fusion approach demonstrates promising improvements for Bangla hate speech detection, some limitations remain. First, the performance of the fused models is still bounded by the representational capacity and pretraining data of the individual language models. For example, BanglaBERT is trained primarily on curated Bangla corpora, while XLM-RoBERTa and MuRIL include multilingual data, which may introduce noise or under-represent Bangla-specific linguistic phenomena such as code-mixing, dialectal variations, or colloquial expressions. Second, despite our fusion strategy improving the performance, more adaptive or robust fusion mechanisms could potentially yield stronger results. Finally, although our method reduces variance compared to relying on a single model, it increases computational cost during inference by requiring predictions from multiple language models. This may limit practical deployment in low-resource or real-time settings.

References

Abdullah Al Maruf, Ahmad Jainul Abidin, Md Mahmudul Haque, Zakaria Masud Jiyad, Aditi Golder, Raaid Alubady, and Zeyar Aung. 2024. Hate speech detection in the bengali language: a comprehensive survey. *Journal of Big Data*, 11(1):97.

Dimosthenis Antypas and Jose Camacho-Collados. 2023. Robust hate speech detection in social media: A cross-dataset empirical evaluation. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 231–242, Toronto, Canada. Association for Computational Linguistics.

Abdul Aziz, Md Akram Hossain, and Abu Nowshed Chy. 2023. Csecu-dsg at semeval-2023 task 4: Fine-tuning deberta transformer model with cross-fold training and multi-sample dropout for human values identification. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1988–1994.

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.

Mithun Das, Somnath Banerjee, Punyajoy Saha, and Animesh Mukherjee. 2022. Hate speech and offensive language detection in bengali. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 286–296.

Md Arid Hasan, Firoj Alam, Md Fahad Hossain, Usman Naseem, and Syed Ishtiaque Ahmed. 2025a. Llm-based multi-task bangla hate speech detection: Type, severity, and target. *arXiv preprint arXiv:2510.01995*.

Md Arid Hasan, Firoj Alam, Md Fahad Hossain, Usman Naseem, and Syed Ishtiaque Ahmed. 2025b. Overview of blp 2025 task 1: Bangla hate speech identification. In *Proceedings of the Second International Workshop on Bangla Language Processing (BLP-2025)*, India. Association for Computational Linguistics.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, and 1 others. Muril: Multilingual representations for indian languages.

Youngwook Kim, Shinwoo Park, and Yo-Sub Han. 2022. Generalizable implicit hate speech detection using contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6667–6679, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jean Lee, Taejun Lim, Heejun Lee, Bogeun Jo, Yangsok Kim, Heegeun Yoon, and Soyeon Caren Han. 2022. K-MHaS: A multi-label hate speech detection dataset in Korean online news comment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3530–3538, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Esshaan Mahajan, Hemaank Mahajan, and Sanjay Kumar. 2024. Ensmulhatecyb: Multilingual hate speech and cyberbully detection in online social media. *Expert systems with applications*, 236:121228.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

Nauros Romim, Mosahed Ahmed, Md Saiful Islam, Arnab Sen Sharma, Hriteshwar Talukder, and Mohammad Ruhul Amin. 2022. 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*, pages 5153–5162.

Pradeep Kumar Roy, Snehaan Bhawal, and Chinnaudayar Navaneethakrishnan Subalalitha. 2022. Hate speech and offensive language detection in dravidian

languages using deep ensemble framework. *Computer Speech & Language*, 75:101386.

Isadora Salles, Francielle Vargas, and Fabrício Benevenuto. 2025. HateBRXplain: A benchmark dataset with human-annotated rationales for explainable hate speech detection in Brazilian Portuguese. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6659–6669, Abu Dhabi, UAE. Association for Computational Linguistics.

Akshay Singh and Rahul Thakur. 2024. Generalizable multilingual hate speech detection on low resource Indian languages using fair selection in federated learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7211–7221, Mexico City, Mexico. Association for Computational Linguistics.

Zeerak Talat and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France. European Language Resources Association.

Jesús Vázquez-Osorio, Gerardo Sierra, Helena Gómez-Adorno, and Gemma Bel-Enguix. 2024. PCICU-NAM at WASSA 2024: Cross-lingual emotion detection task with hierarchical classification and weighted loss functions. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 490–494, Bangkok, Thailand. Association for Computational Linguistics.