BElite at BLP-2025 Task 1: Leveraging Ensemble for Multi Task Hate Speech Detection in Bangla

Zannatul Fardaush Tripty[1,*], Ibnul Mohammad Adib[2,*], Nafiz Fahad[3]
Muhammad Tanjib Hussain[3] Md Kishor Morol[3,†]
[1] Chittagong University of Engineering and Technology
[2]American International University of Bangladesh, [3]Elite Research Lab LLC
Correspondence: kishormorol@ieee.org

Abstract

The widespread use of the internet has made sharing information on social media more convenient. At the same time, it provides a platform for individuals with malicious intent to easily spread hateful content. Since many users prefer to communicate in their native language, detecting hate speech in Bengali poses a significant challenge. This study aims to identify Bengali hate speech on social media platforms. A shared task on Bengali hate speech detection is organized by the Second Bangla Language Processing Workshop (BLP). To tackle this task, we implement five traditional machine learning models (LR, SVM, RF, NB, XGB), three deep learning models (CNN, BiLSTM, CNN+BiLSTM), and three transformer-based models (BanglaBERT, m-BERT, XLM-R). Among all models, a weighted ensemble of transformer models achieves the best performance. Our approach ranks ***3rd*** in Subtask 1A with a micro-*F1* score of 0.734, ***6th*** in Subtask 1B with 0.7315, and, after post-competition experiments, ***4th*** in Subtask 1C with 0.735.

1 Introduction

The rise of social media has enabled billions to share opinions but has also fueled online hate speech, defined as speech inciting hatred against groups based on ethnicity, religion, disability, or sexual identity (American Library Association, 2017). With that comes the caveat of manual content moderation, requiring the development of state-of-the-art content moderation by leveraging artificial intelligence and natural language processing (Amin, 2024).

Most research has focused on high-resource languages like English, while Bangla, despite being the sixth most spoken language globally, remains under-resourced for NLP tasks (Hosain and Morol, 2025; Salam et al., 2016). It poses unique linguistic sociocultural challenges, especially for hate speech detection, including code-switching, spelling variation, and dialect variation. To address this, a shared task on Bangla hate speech classification is organized at BLP (Hasan et al., 2025b), providing a labeled dataset and dividing the task into three subtasks: hate type, severity, and target group, reflecting the complexity of understanding and mitigating hate speech (Hossan et al., 2025; Islam et al., 2024). In this work, we develop an ensemble model for Bangla hate speech classification and conduct thorough experiments across all three subtasks. Our approach demonstrates improved accuracy compared to baseline methods, addressing the gap in NLP research in Bangla.

2 Related Works

Early works in hate speech detection introduced the first annotated dataset, where a GRU with word2vec embeddings outperformed several machine learning models, demonstrating deep learning's superiority (Ishmam and Sharmin, 2019; Hosain et al., 2025a). Subsequent studies moved toward context-aware neural architectures. A two-part encoder–decoder framework with 1D CNNs, BiRNNs, and attention layers was proposed (Das et al., 2021), showing that attention mechanisms outperformed standalone traditional deep learning (Zerine et al., 2020).

Later on, the landscape expanded with BD-SHS, a large benchmark dataset for binary and multi-label classification, which introduced informal fastText embeddings tailored for noisy social media text, highlighting the role of domain-specific representation learning (Romim et al., 2022).

Furthermore, domain-specific embeddings were shown to capture hateful vocabulary better than general-purpose embeddings; therefore, even lighter models were able to rival transformers—an

*Equal contribution.
†Corresponding author.

Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 421–429

important insight for resource-constrained environments (Saleh et al., 2023; Tariquzzaman et al., 2023).

The rise of transformers advanced Bangla hate detection, with monolingual BanglaBERT often outperforming multilingual encoders such as XLM-R and mBERT (Ghosh and Senapati, 2022). Transformer models were further tested with Romanized Bangla compared to standard Bangla, showing that MuRIL excelled in cross-lingual few-shot settings (Das et al., 2022).

Domain-specific transformer models such as BanglaHateBERT (Jahan et al., 2022) yielded consistent gains, while hybrid architectures such as G-BERT (Keya et al., 2023) combined BanglaBERT with a GRU classifier, further improving performance. DeepHateExplainer (Karim et al., 2021) was another pioneering effort with an ensemble of various BERT-based models (Hosain et al., 2025b). Furthermore, it used layer-wise propagation and sensitivity analysis to provide explanations and ensure that the model's decisions were made based on reasonable features; however, it also highlighted the need for more contextual information, especially for labels such as political hate speech.

Beyond Bangla, multi-task learning with user metadata and inter-user or intra-user features improved English hate detection, suggesting that a similar approach could benefit Bangla (Kapil and Ekbal, 2024). Recent explorations with large and small language models (LLMs) also signaled a paradigm shift. GPT-3.5 Turbo with chain-of-thought prompting was shown to outperform BERT baselines on English hate speech; however, performance varied across different languages (Guo et al., 2023; Sakib et al., 2025). TinyLLMs (Phi-2, TinyLlama) fine-tuned with LoRA were also shown to rival large language models in efficiency and accuracy (Sen et al., 2024).

Hate speech detection has remained challenging because datasets vary in annotation quality, domain, and class distribution. Furthermore, creating reliable resources is costly, as it requires strong agreement among annotators to reduce bias (Vasker et al., 2024; Gupta et al., 2025).

3 Task and Dataset Description

The primary objective of this task is to detect hate speech in a Bengali corpus by developing systems capable of accurately classifying text, using the datasets (Hasan et al., 2025a) provided by the organizers of the shared task.[1] A significant class imbalance is observed across all three subtasks of the BLP 2025 dataset, as reflected in Tables 1, 2, and 3. For Subtask 1A (Hate Type Classification), shown in Table 1, the *None* class overwhelmingly dominates the dataset, while several hate categories contain very few examples. In particular, *Sexism* and *Religious Hate* account for only a small portion of the training and evaluation splits, making them the most underrepresented classes. For Subtask

Classes	Train	Dev	Test
None	19954	1451	5751
Abusive	8212	564	2312
Political Hate	4227	291	1220
Profane	2331	157	709
Religious Hate	676	38	179
Sexism	122	11	29
Total	35522	2512	10200

Table 1: Dataset distribution across classes, splits, and word counts for Subtask 1A (hate type)

1B (Target Classification: To Whom), as presented in Table 2, the *None* category remains the most frequent, similar to Subtask 1A. In contrast, the minority classes—especially *Community* and *Society*—have far fewer samples. For Subtask 1C,

Class	Train	Dev	Test
None	21190	1536	6093
Individual	5646	364	1571
Organization	3846	292	1152
Community	2635	179	759
Society	2205	141	625
Total	35522	2512	10200

Table 2: Dataset distribution across classes, splits, and word counts for Subtask 1B (to whom)

which focuses on multi-task classification, an additional column representing *hate severity* is included. The goal of this subtask is to perform multi-task classification, assigning each text a hate type, a target (to whom), and a severity level. The distribution of hate severity across the dataset is shown in Table 3.The majority of instances fall under the *Little to None* severity level, while the *Severe* category appears sparsely across all splits.

[1] https://github.com/AridHasan/blp25_task1

Hate Severity	Train	Dev	Test
Little to None	23489	1703	6737
Mild	6853	483	2001
Severe	5180	326	1462
Total	35522	2512	10200

Table 3: Dataset distribution across hate severity classes for train, dev, and test splits.

4 Methods

To evaluate the performance of Bengali hate speech classification, we implements five machine learning methods (LR, RF, NB, SVM, and XGB), three deep learning techniques (CNN, BiLSTM, and CNN+BiLSTM), and three transformer-based models (mBERT, Bangla-BERT, XLM-R, along with an ensemble strategy). An abstract overview of the system is illustrated in Figure 1.

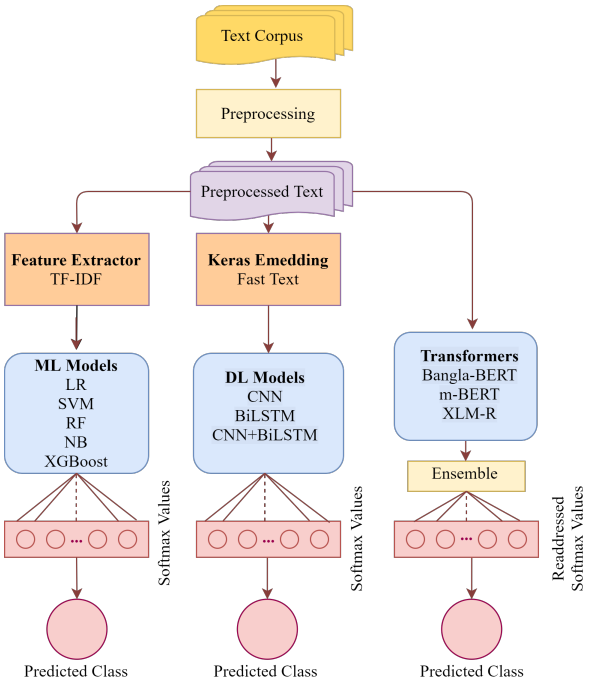


Figure 1: Abstract process diagram of Bengali hate speech detection system.

4.1 Preprocessing

As the dataset is relatively clean, only minimal preprocessing was applied. Text normalization is performed using the Bangla Normalizer tool (Hasan et al., 2020), which standardizes spacing, punctuation, and character representations for consistency.

4.2 Machine Learning Models

All five ML models use unigram features extracted via TF-IDF. For instance, Logistic Regression (LR) employs the `liblinear` solver with $l2$ regularization ($C = 5.0$), 500 iterations, and balanced class weights. Meanwhile, Random Forest (RF) uses 300 trees with the `gini` criterion, considers all features, and requires at least two samples to split a node. In contrast, Naive Bayes (NB) applies additive smoothing ($\alpha = 0.5$) to optimize probability estimates. Similarly, the Support Vector Machine (SVM) uses a linear kernel with $C = 2$, balanced class weights, and 500 iterations to ensure convergence. Finally, XGBoost (XGB) is configured for multi-class classification with 500 trees and a learning rate of 0.1. These settings are carefully chosen to maximize accuracy and stability across all classifiers.

4.3 Deep Learning Models

We implement three deep learning models: CNN, BiLSTM, and hybrid CNN+BiLSTM. All models use pretrained FastText embeddings (Joulin et al., 2016) to obtain dense word representations. The CNN comprises two convolutional blocks, with 128 filters of size 5 and 64 filters of size 3, each followed by max-pooling layers of sizes 5 and 3, respectively. The features are then flattened and passed through a dense layer with 128 *ReLU*-activated neurons, followed by a dropout layer with rate 0.5 and a softmax output layer. The BiLSTM uses a bidirectional LSTM with 200 units and dropout 0.2, followed by a dense layer with 128 *ReLU*-activated neurons, dropout 0.5, and a softmax output. The hybrid CNN+BiLSTM first applies the CNN convolutional and pooling layers, then feeds the resulting features into a bidirectional LSTM with 200 units and dropout 0.2. The output is flattened, passed through a dense layer with 128 *ReLU* neurons, a dropout layer with rate 0.5, and a softmax layer for classification.

4.4 Transformer Models

Past studies show that transformer models trained in monolingual, multilingual, or cross-lingual settings achieve state-of-the-art performance in hate speech classification (Mazari et al., 2024; Saleh et al., 2023). In this work, we select Bangla-BERT, XLM-R, and mBERT for our ensemble because they represent complementary architectures that have demonstrated strong performance in Bangla

NLP tasks. Bangla-BERT (Bhattacharjee et al., 2022) is a monolingual model that effectively captures language-specific morphology and culturally grounded expressions, making it particularly suitable for Bangla hate speech detection. XLM-R (Conneau et al., 2020) is a cross-lingual model that provides robust representations across multiple languages and handles noisy or code-mixed social media text effectively. mBERT (Devlin et al., 2019), although trained on a smaller multilingual corpus, has shown strong generalization across low-resource languages, including Bangla. All three models are pre-trained transformers that we fine-tune on the shared-task dataset, accessed via the Hugging Face library[2] .

The models are fine-tuned on the dataset using the Hugging Face Trainer API[3] for 3 epochs, with a batch size of 8 for both training and evaluation. A learning rate of $2e^{-5}$ and a weight decay of 0.01 are applied. Evaluation and model checkpointing are performed every 500 steps, and the best-performing model on the validation set is automatically loaded at the end of training.

4.5 Proposed Ensemble Model

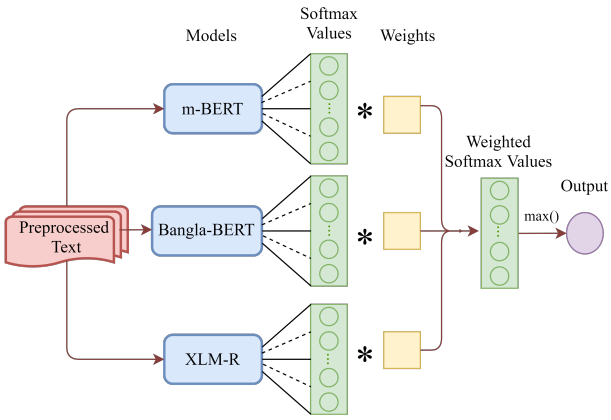


Figure 2: Proposed ensemble method.

Recent studies (Karim et al., 2021; Singh et al., 2023) have demonstrated that ensembles of transformer models can substantially improve the effectiveness of classification tasks. Ensemble learning leverages the complementary strengths of individual models to enhance overall predictive accuracy. In this work, three pretrained transformer models—Bangla-BERT, XLM-R, and m-BERT—are fine-tuned on their respective datasets and subsequently combined using both averaging *(A-ensemble)* and weighted *(W-ensemble)* approaches. The average ensemble computes the mean of the softmax probabilities generated by the participating models and assigns the class with the highest probability as the final prediction. In contrast, weighted ensemble combines the predictions of multiple models by assigning different importance (weights) to each model based on their prior performance. In this experiment, the weighted ensemble assigns weights based on micro-*F1* scores from the evaluation dataset. These weights are normalized and combined with the softmax probabilities generated by the fine-tuned BERT models, thereby allowing models with superior prior performance to exert greater influence on the final prediction. The overall process of the weighted ensemble is illustrated in Figure 2.

Let $M = \{M_1, M_2, \ldots, M_L\}$ represent the set of fine-tuned models, where $L = 3$ in our case. For a given instance, let $p_i(c)$ denote the softmax probability predicted by model M_i for class c, and let f_i be the micro-*F1* score of model M_i on the evaluation dataset.

We first compute normalized weights for each model based on their micro-*F1* scores:

$$w_i = \frac{f_i}{\sum_{j=1}^{L} f_j}, \quad i = 1, 2, \ldots, L \tag{1}$$

The weighted ensemble probability for class c is then computed as:

$$P(c) = \sum_{i=1}^{L} w_i \cdot p_i(c) \tag{2}$$

Finally, the predicted class $\hat{y}$ is determined as the class with the highest weighted probability:

$$\hat{y} = \arg\max_{c} P(c) \tag{3}$$

This approach ensures that models with higher prior performance (as measured by micro-*F1*) contribute more to the final prediction, while still leveraging the complementary strengths of all models in the ensemble.

5 Experiments and Results

The evaluation results of individual models on the test set are presented in Table 4. The results indicate that among the machine learning approaches, XGB with TF-IDF features achieved the highest micro *F1*-scores, recording 0.66, 0.67, and 0.68 for

[2] https://huggingface.co/

[3] https://huggingface.co/docs/transformers/main_classes/trainer

	Subtask 1A			Subtask 1B			Subtask 1C		
Models	P	R	F1	P	R	F1	P	R	F1
TF-IDF+LR	0.65	0.66	0.65	0.65	0.65	0.65	0.66	0.67	0.67
TF-IDF+SVM	0.63	0.64	0.64	0.64	0.63	0.63	0.65	0.65	0.65
TF-IDF+RF	0.64	0.65	0.65	0.63	0.66	0.66	0.64	0.67	0.67
TF-IDF+NB	0.60	0.61	0.61	0.57	0.63	0.63	0.60	0.64	0.64
TF-IDF+XGB	0.65	0.67	0.66	0.64	0.67	0.67	0.65	0.68	0.68
FT+CNN	0.65	0.68	0.68	0.63	0.67	0.67	0.65	0.69	0.69
FT+BiLSTM	0.67	0.69	0.69	0.66	0.69	0.69	0.69	0.70	0.70
FT+C+B	0.66	0.68	0.68	0.64	0.68	0.68	0.66	0.69	0.69
m-BERT	0.68	0.70	0.70	0.69	0.70	0.70	0.67	0.69	0.69
XLM-R	0.70	0.69	0.69	0.70	0.71	0.71	0.70	0.71	0.71
Bangla-BERT	0.72	0.72	0.72	0.72	0.72	0.72	0.72	**0.74**	0.73
W-Ensemble	**0.72**	**0.73**	**0.734**	**0.72**	**0.73**	**0.731**	**0.72**	0.72	**0.735**
A-Ensemble	0.71	0.71	0.71	0.70	0.71	0.71	0.70	0.71	0.72

Table 4: Performance of various models on the Subtask 1A and Subtask 1B test sets where P, R, and F denote precision, recall, and micro F1-score, respectively. Here, C+B represents the CNN+BiLSTM model and FT represents FastText.

Subtasks 1A, 1B, and 1C, respectively. This performance is superior to LR(0.65,0.66,0.65), SVM (0.64, 0.63, 0.65), RF (0.65, 0.66, 0.67), and NB (0.61, 0.63, 0.64). Within deep learning based methods, BiLSTM with FastText embeddings provides the most consistent results, achieving 0.69, 0.69, and 0.70 across Subtasks 1A, 1B, and 1C. This slightly exceeds the performance of CNN (0.68, 0.67, 0.69) and CNN+BiLSTM (0.68, 0.68, 0.69). Nevertheless, all DL models still fall short compared to transformer-based approaches. Among transformers, Bangla-BERT delivers the best results, with micro *F1*-scores of 0.72, 0.72, and 0.73, outperforming m-BERT (0.70, 0.70, 0.69) and XLM-R (0.69, 0.70, 0.71). Finally, the ensemble strategies proved most effective overall. The Weighted Ensemble achieves the highest scores of 0.734, 0.7315, and 0.735, surpassing both individual models and the Averaging Ensemble (0.71, 0.71, 0.72). A key finding of this study is the effectiveness of ensemble strategies. These results show that the weighted ensemble outperforms standalone ML, DL, and transformer models.

6 Error Analysis

Error analysis is performed using both quantitative and qualitative approaches. Detailed results are provided in Appendices A and B. Quantitative analysis identifies systematic misclassifications via confusion matrices, while qualitative analysis explores underlying causes such as class imbalance, contextual subtleties, and overlapping linguistic cues.

7 Conclusion

In this paper, we propose a weighted ensemble approach for multi-task hate speech detection in a Bengali corpus, leveraging the complementary strengths of multiple transformer-based models. By combining fine-tuned Bangla-BERT, m-BERT, and XLM-R, the ensemble captured both language-specific nuances and cross-lingual semantic information, outperforming individual models across all subtasks. It achieved micro-*F1* scores of 0.734, 0.7315, and 0.735 on Subtask 1A, 1B, and 1C, respectively.

To validate our approach, we conduct extensive experiments with traditional machine learning and deep learning models using various feature extraction and embedding strategies. The results highlight the effectiveness of the ensemble and its potential for low-resource languages, where limited annotated data and linguistic complexity pose significant challenges for automated text classification.

Limitations

Our approach has several potential limitations. Notably, the dataset exhibits substantial class imbalance, which can cause models to overfit the ma-

jority classes while underperforming on underrepresented ones. Addressing this limitation requires effective data augmentation techniques, such as SMOTE (Chawla et al., 2002), ADASYN (He et al., 2008), or other strategies to increase the amount of data, which can help balance the dataset and improve model generalization.

References

American Library Association. 2017. Hate speech and hate crime. https://www.ala.org/advocacy/intfreedom/hate. Accessed September 22, 2025.

MM Amin. 2024. Ai-powered personalized marketing: A deep dive into customer segmentation and targeting. *INTERNATIONAL JOURNAL*, 11(12).

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.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

Amit Kumar Das, Abdullah Al Asif, Anik Paul, and Md Nur Hossain. 2021. Bangla hate speech detection on social media using attention-based recurrent neural network. *Journal of Intelligent Systems*, 30(1):578–591.

Mithun Das, Somnath Banerjee, Punyajoy Saha, and Animesh Mukherjee. 2022. Hate speech and offensive language detection in bengali. *arXiv preprint arXiv:2210.03479*.

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.

Koyel Ghosh and Apurbalal Senapati. 2022. Hate speech detection: a comparison of mono and multilingual transformer model with cross-language evaluation. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 853–865.

Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. 2023. An investigation of large language models for real-world hate speech detection. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1568–1573.

Rajan Das Gupta, Md Kishor Morol, Nafiz Fahad, Md Tanzib Hosain, Sumaya Binte Zilani Choya, and Md Jakir Hossen. 2025. Brains: A retrieval-augmented system for alzheimer's detection and monitoring. *arXiv preprint arXiv:2511.02490*.

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.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M Sohel Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for bengali-english machine translation. *arXiv preprint arXiv:2009.09359*.

Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. 2008. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. Ieee.

Md Tanzib Hosain, Rajan Das Gupta, and Md Kishor Morol. 2025a. Multilingual question answering in low-resource settings: A dzongkha-english benchmark for foundation models. *arXiv preprint arXiv:2505.18638*.

Md Tanzib Hosain and Md Kishor Morol. 2025. B-reaso: A multi-level multi-faceted bengali evaluation suite for foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9260–9274.

Md Tanzib Hosain, Md Kishor Morol, and Md Jakir Hossen. 2025b. A hybrid self attentive linearized phrase structured transformer based rnn for financial sentence analysis with sentence level explainability. *Scientific Reports*, 15(1):23893.

Tanvir Hossan, BM Taslimul Haque, Md Shihabun Sakib, Niladry Chowdhury, and Md Minhajul Amin. 2025. Ethical challenges in business analytics: Balancing data privacy and profit. *Open Access Library Journal*, 12(2):1–12.

Alvi Md Ishmam and Sadia Sharmin. 2019. Hateful speech detection in public facebook pages for the bengali language. In *2019 18th IEEE international conference on machine learning and applications (ICMLA)*, pages 555–560. IEEE.

Md Mainul Islam, Ismoth Zerine, Md Arifur Rahman, Md Saiful Islam, and Md Yousuf Ahmed. 2024. Ai-driven fraud detection in financial transactions-using machine learning and deep learning to detect anomalies and fraudulent activities in banking and e-commerce transactions. *Available at SSRN 5287281*.

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.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Prashant Kapil and Asif Ekbal. 2024. A unified multi-task learning architecture for hate detection leveraging user-based information. *arXiv preprint arXiv:2411.06855*.

Md Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Sagor Sarker, Mehadi Hasan Menon, Kabir Hossain, Md Azam Hossain, and Stefan Decker. 2021. Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. In *2021 IEEE 8th international conference on data science and advanced analytics (DSAA)*, pages 1–10. IEEE.

Ashfia Jannat Keya, Md Mohsin Kabir, Nusrat Jahan Shammey, Md Rashedul Islam, and Yutaka Watanobe. 2023. G-bert: an efficient method for identifying hate speech in bengali texts on social media. *IEEE Access*, 11:79697–79709.

Ahmed Cherif Mazari, Nesrine Boudoukhani, and Abdelhamid Djeffal. 2024. Bert-based ensemble learning for multi-aspect hate speech detection. *Cluster Computing*, 27(1):325–339.

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. *arXiv preprint arXiv:2206.00372*.

Tanjil Hasan Sakib, Md Tanzib Hosain, and Md Kishor Morol. 2025. Small language models: Architectures, techniques, evaluation, problems and future adaptation. *arXiv preprint arXiv:2505.19529*.

Abdus Salam, Ishtiaq Mohammed Chowdhury, Mohammad Masum Sadeque, and BM Taslimul. 2016. Save time for public transport users in a developing country. *International Journal of Education and Management Engineering*, 11(8):27–33.

Hind Saleh, Areej Alhothali, and Kawthar Moria. 2023. Detection of hate speech using bert and hate speech word embedding with deep model. *Applied Artificial Intelligence*, 37(1):2166719.

Tanmay Sen, Ansuman Das, and Mrinmay Sen. 2024. Hatetinyllm : Hate speech detection using tiny large language models. *Preprint*, arXiv:2405.01577.

Kartik Singh, Meenakshi Tripathi, Basant Agarwal, and Abhay Kumar Sain. 2023. Ensemble of transformer based approach for hate speech detection on twitter data. In *2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, volume 10, pages 894–899. IEEE.

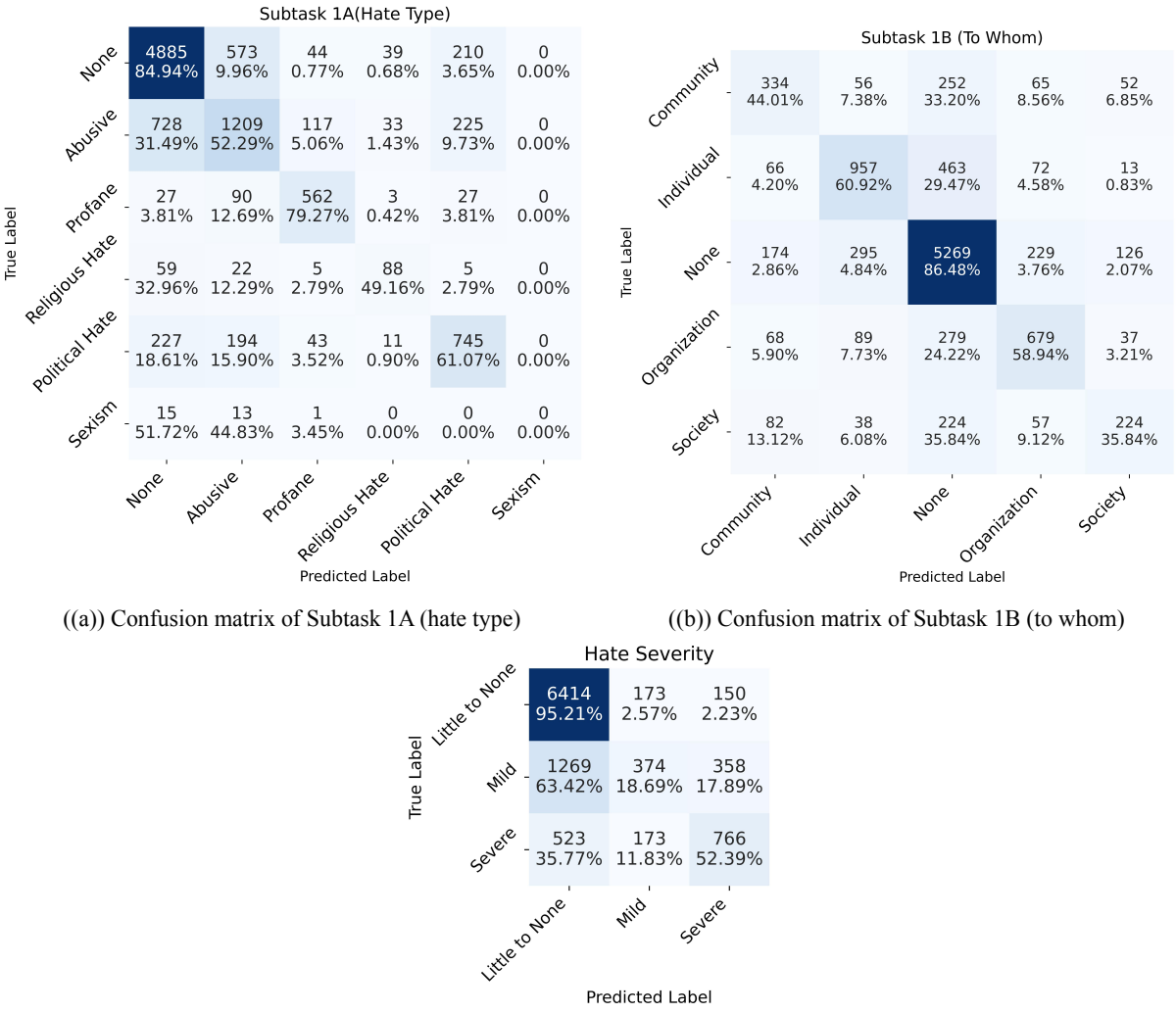
Md Tariquzzaman, Md Wasif Kader, Audwit Anam, Naimul Haque, Mohsinul Kabir, Hasan Mahmud, and Md Kamrul Hasan. 2023. the_linguists at blp-2023 task 1: A novel informal bangla fasttext embedding for violence inciting text detection. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 214–219.

Nishat Vasker, Anika Tabassum Nafisa, MD Arifur Rahman, and Mahamudul Hasan. 2024. Heart disease classification with xai and kernel shap. In *2024 IEEE 3rd International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things (RAAICON)*, pages 82–85. IEEE.

Ismoth Zerine, Md Mainul Islam, Md Saiful Islam, Md Yousuf Ahmad, and Md Arifur Rahman. 2020. Climate risk analytics for us agriculture sustainability: Modeling climate impact on crop yields and supply chain to support federal policies food security and renewable anergy adoption. *Cuestiones de Fisioterapia*, 49(3):241–258.

A Quantitative analysis

In Subtask 1A (hate type), the confusion matrix 3(a) shows that the model performed strongly on the *None* category, achieving an accuracy of 84.94%. However, this dominance contributes to frequent misclassification of the minority hate types. A significant portion of *Abusive* samples are mislabeled as *None* (31.49%, 728 instances) or as *Profane* (5.06%, 117 instances), while *Profane* instances are often predicted as *Abusive* (12.69%, 90 instances). This pattern highlights the fine-grained overlap between abusive language and profanity. Similarly, *Political Hate* is frequently misclassified as *Abusive* (15.90%, 194 instances), reflecting the difficulty of detecting implicit or coded political language. The weakest performance is observed in *Sexism*, where most sexist utterances are mislabeled as *None* (51.72%, 15 instances) or *Abusive* (44.83%, 13 instances), emphasizing the impact of severe class imbalance and subtle linguistic cues.



((a)) Confusion matrix of Subtask 1A (hate type)

((b)) Confusion matrix of Subtask 1B (to whom)

((c)) Confusion matrix of hate severity

Figure 3: Confusion matrix of the proposed ensemble transformer models on task

Text	Hate Type		To Whom		Severity	
	A	P	A	P	A	P
হামাস আর বোম মার (Hamas, bomb them)	Abusive	None	Organization	None	Severe	Mild
সবগুলোই চুর হারামি টাকা খায় (They all take stolen, illicit money)	Profane	Abusive	Community	Individual	Little to None	Severe
আমেরিকা নেটো যুদ্ধ অপরাধী (America and NATO are war criminals)	Political Hate	Abusive	Organization	Society	Little to None	Mild
আজ কাল পবিত্র সংসদ বড়ে গেছে নর-তকি আর বেসসা দিয়ে (These days the sacred parliament has become full with prostitutes and immoral people)	Profane	Abusive	Community	Organization	Severe	Mild
শালা মিথ্যুক দালাল ও মুনাফিক (Damn liar, broker, and hypocrite)	Abusive	Profane	None	Individual	Mild	Severe

Table 5: Illustrative data samples highlighting the diverse behavior of the ensemble model. Here, A denotes the ground-truth label, while P denotes the predicted label.

The confusion matrix of Subtask 1B (To Whom) 3(b), non-targeted content (*None*) is identified with relatively high accuracy (86.48%, 5,269 instances). Nevertheless, the majority class again overshadows the minority categories. For example, labels such as *Community*, *Organization*, and *Society* are

frequently misclassified, likely due to their overlapping meanings and nuances in the Bengali context. Overall, all labels exhibit a tendency to be predicted as *None*, reflecting the dominance of this category in the dataset.

Regarding the confusion matrix of hate severity 3(c) ,the model tends to underestimate intensity, with a large portion of instances being classified as *Little to None*. Specifically, 63.42% of *Mild* cases and 35.77% of *Severe* cases are predicted as *Little to None*, indicating challenges in distinguishing subtle variations in hate severity.

B Qualitative analysis

The model shows systematic biases and confusions that stem primarily from class imbalance, subtle contextual cues, and overreliance on explicit lexical signals as shown in Table 5. For hate type, the model struggles particularly with *Sexism*, *Religious Hate*, often misclassifying them as *None* or *Abusive*. The dataset does not provide enough representative examples for the model to learn their patterns. In contrast, *Profane* speech is detected more reliably, since it is usually tied to explicit keywords. Common misclassifications include *Abusive* vs. Non-abusive, *Profane* vs. *Abusive*, and *Political Hate* vs. *Abusive*, which arise from overlapping language patterns and insufficient contextual representation. For hate severity, the model frequently mispredicts *Mild*, often confusing it with *Little to None* because of subtle gradations of severity. For Target (to whom), the model distinguishes *Individuals* fairly well but struggles with *Community*, *Society*, and *Organization*. These categories are often implicit, underrepresented, or context-dependent, causing the model to confuse them with one another. The model tends to overpredict majority classes like *None* and *Little to None*, reflecting its bias toward heavily represented categories, while underperforming on minority classes. Moreover, contextual nuance is crucial to separate closely related categories such as *Little to None* vs. *Mild* or *Organization* vs. *Society*.

The ensemble approach helps by boosting confidence and compensating for some data gaps, but categories like *Sexism* remain difficult to predict simply because there is not enough training data to establish strong patterns.