CUET_Sntx_Srfrs at BLP-2025 Task 1: Combining Hierarchical Classification and Ensemble Learning for Bengali Hate Speech Detection

Hafsa Hoque Tripty, Laiba Tabassum, Hasan Mesbaul Ali Taher, Kawsar Ahmed, Mohammed Moshiul Hoque
Department of Computer Science and Engineering,
Chittagong University of Engineering & Technology, Chittagong 4349, Bangladesh
{u2204108,u2204077,u1804038,u1804017}@student.cuet.ac.bd
moshiul_240@cuet.ac.bd

Abstract

Detecting hate speech in Bengali social media content presents considerable challenges, primarily due to the prevalence of informal language and the limited availability of annotated datasets. This study investigates the identification of hate speech in Bengali YouTube comments, focusing on classifying the type, severity, and target group. Multiple machine learning baselines and voting ensemble techniques are evaluated to address these tasks. The methodology involves text preprocessing, feature extraction using TF-IDF and Count vectors, and aggregating predictions from several models. Hierarchical classification with TF-IDF features and majority voting improves the detection of less frequent hate speech categories while maintaining robust overall performance, resulting in an 18^{th} place ranking and a micro F1 score of 68.42%. Furthermore, ablation studies assess the impact of preprocessing steps and n-gram selection, providing reproducible baselines for Bengali hate speech detection. All codes and resources are publicly available at GitHub[1].

1 Introduction

Hate speech detection refers to the automated identification of language that targets individuals or groups with hostility or discrimination based on attributes such as gender, religion, or political affiliation. Detecting hate speech is essential for maintaining safe online environments and is widely used in content moderation, policy enforcement, and research on social dynamics. Detecting hate speech is challenging, especially in less-resourced languages like Bengali, where annotated datasets and robust methods are limited despite growing online activity. Recently, several methods have shown promise for low-resource hate speech detection (Sanoussi et al., 2022; Alaoui et al., 2022). However, most of these focused on binary classification of hate speech, overlooking hate type, severity, or target group (Chiril et al., 2019; Alam et al., 2024; Hossain et al., 2023). This restricts their ability to capture the multi-dimensional nature of hate speech in Bengali (Das et al., 2022). The Bengali multi-task hate speech identification shared task (Hasan et al., 2025b) addresses this with three subtasks: (1A) hate type, (1B) target group, and (1C) multi-task classification of type, severity, and target.

This study investigates machine learning methods for multi-task Bengali hate speech detection in the context of significant data imbalance. Our approach integrates preprocessing, n-gram vectorization, multiple machine learning models, hierarchical classification, and ensemble learning. The main contributions are:

- Developed a framework for Bengali hate speech detection that uses hierarchical classification and majority voting within machine learning ensembles.
- Conducted a systematic evaluation of 16 models across three subtasks, including comprehensive error analysis to assess model effectiveness.

2 Related Work

Hate speech detection has been widely explored in high-resource languages such as English, Arabic, and other European languages. Early studies primarily used machine learning (ML) with handcrafted features. Sanoussi et al. (2022) reported

[1] https://github.com/Hasan-Mesbaul-Ali-Taher/BLP_25_Task_1

Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 523–530
December 23, 2025 ©2025 Association for Computational Linguistics

an accuracy of 95.45% using an SVM classifier, while Alaoui et al. (2022) achieved an accuracy value of 87.23% in the first dataset, and the second dataset attained a value of 93.06%. Al-Hassan and Al-Dossari (2022) explored various ML and DL models, and they achieved the maximum F1 score of 73% on Arabic tweets with the ensemble method (CNN+LSTM). Saleh et al. (2023) utilized BiLSTM for hate speech detection in English and achieved an F1 score of 93%. Khan et al. (2022a) introduced BiCHAT, which leverages BiLSTM and deep CNNs, achieving an F1 score of 0.84 on a Twitter dataset and surpassing benchmarks (Khan et al., 2022b). In contrast, Research on low-resource languages has received relatively limited attention in hate speech detection due to data scarcity and linguistic diversity. Bade et al. (2024) focused on Dravidian languages, employing a Random Forest (RF) model with TF-IDF features and obtained an F1 score of 0.492. (Ibrohim and Budi, 2019) utilized RF+DT for multi-label hate speech detection in Indonesian languages, achieving an accuracy of 77.36%. At the same time, Nasir et al. (2024) proposed a two-level LR technique for Roman-Urdu code-mixed data, which gained 87%accuracy. Bilal et al. (2023) further applied the BERT model to detect hate speech from Roman-Urdu code-mixed text and achieved 97.89% accuracy.

Bengali, as a low-resource language in NLP, lacks a standard benchmark dataset for this task, making consistent evaluation difficult. Due to these constraints, very few research activities have been carried out in this area of Bangla Language Processing (BLP), which are mostly related to hate pr offensive content from social platforms. Momin and Sarker (2025) explored various machine learning and deep learning techniques for hate speech detection and achieved 95% accuracy using a Bi-LSTM with FastText embeddings, while Saha et al. (2024) employed a hybrid BERT-CNN model and achieved an F1 score of 94.44%. Several transformer-based models and LLMs have been explored by Haider et al. (2025) for hate speech detection in transliterated Bengali. This work achieved the highest F1 score of 77.36 with the mBERT model. A recent study utilized BanglaBERT and achieved an accuracy of 74%. Another study employed a hierarchical BERT model (Das et al., 2023) and obtained an F1 score of 0.73797. Faruqe et al. (2023) utilized GRU techniques and gained the highest accuracy of 98.87%. Most previous studies on Bengali hate speech detection have relied on small, imbalanced, and domain-specific datasets, making robust generalization difficult. These works primarily target hate speech in social media posts, neglecting other critical contexts, such as online news comments and discussion forums. To the best of our knowledge, there is still no unified approach for Bengali hate speech detection. This work aims to fill this gap by introducing a more comprehensive approach to detecting hate speech in Bengali.

3 Task and Dataset Description

The Bangla Multi-task hate speech identification shared task (Hasan et al., 2025b) comprises three interrelated subtasks for comprehensive analysis of hate speech on Bangla social media.

- **Subtask 1A:** Classify texts into *Abusive (Ab), Sexism (Sism), Religious Hate (RHate), Political Hate (PHate), Profane (Pfane)*, or *None (Non).*
- **Subtask 1B:** Identify the target as *Individual (In), Organization (Org), Community (Co), Society (So)*, or *None (Non).*
- **Subtask 1C:** Joint classification of hate type, severity (*Little to None (Non), Mild (Mild), Severe (Severe)*), and target group.

3.1 Datasets

The **BanglaMultiHate** corpus (Hasan et al., 2025a) comprises public YouTube comments on politics, sports, and international topics, preprocessed to remove non-Bangla literals and punctuation. Tables 1–3 summarize the distribution of classes across the training, development, and test splits for each subtask. These distributions highlight the imbalance among labels, which poses challenges for model training and evaluation.

Each of the three datasets exhibits strong class imbalance. For Subtask 1A, the *None* class accounts for more than 56% of the training set, whereas minority labels such as *Sexism* (0.34%)

Label	**Train**	**Dev**	**Test**
None	19,954	1,451	5,751
Abusive	8,212	564	2,312
Political Hate	4,227	291	1,220
Profane	2,331	157	709
Religious Hate	676	38	179
Sexism	122	11	29
Total	**35,522**	**2,512**	**10,200**

Table 1: Dataset statistics for hate type.

Label	**Train**	**Dev**	**Test**
None	21,190	1,536	6,093
Individuals	5,646	364	1,571
Organizations	3,846	292	1,152
Communities	2,635	179	759
Society	2,205	141	625
Total	**35,522**	**2,512**	**10,200**

Table 2: Dataset statistics for hate target.

Label	**Train**	**Dev**	**Test**
Little to None	23,489	1,703	6,737
Mild	6,853	483	2,001
Severe	5,180	326	1,462
Total	**35,522**	**2,512**	**10,200**

Table 3: Dataset statistics for hate severity.

and *Religious Hate* (1.9%) are extremely underrepresented. Subtask 1B follows a similar trend, with *None* comprising nearly 60% of all instances, while *Community* and *Society* each account for less than 7%. In Subtask 1C, *Little to None* dominates two-thirds of the dataset, and the combined proportion of *Mild* and *Severe* remains below 34%.

4 Methodology

This work leverages classical machine learning to establish strong baselines for Bangla hate speech detection under resource constraints, using careful preprocessing, hierarchical classification, and ensemble learning. Figure 1 illustrates the abstract process of hate speech detection.

- **Preprocessing:** To effectively manage noisy Bangla social media text, we performed number removal, stopword elimination using a curated Bangla list, and lemmatization to reduce words to their base forms. The stopword lists and the lemmatizer used are mentioned in the GitHub link. The impact of these preprocessing steps is analyzed in the Ablation Study (Section 5.1).

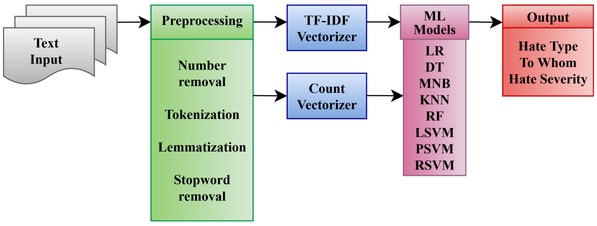


Figure 1: Abstract process of hate speech classification.

- **Tokenization and Vectorization:** Texts were tokenized using whitespace, and features were encoded with Unigram–Bigram–Trigram Count and TF-IDF vectorizers (min_df = 5). The effect of varying n-gram configurations is examined in the Ablation Study (Section 5.1).

- **Classification Models:** Logistic Regression, Decision Tree, Multinomial Naive Bayes, Support Vector Classifier (linear, polynomial, radial), Random Forest, and k-Nearest Neighbors were paired with both vectorizers, forming 16 model–vectorizer combinations.

- **Simple vs. Hierarchical Classification:** In standard classification tasks, models directly predict the target labels. In contrast, hierarchical classification (Figure 2) involves an initial separation between *None* and *Hate* (or *Little to None* and *Hate*). Only Hate instances are subsequently processed by a second classifier, which determines the specific hate category. This method facilitates more detailed identification and categorization of hate-related content. The first model produces $f_1(x) \in \{\text{None}, \text{Hate}\}$, and if $f_1(x) = \text{Hate}$, the second model assigns $f_2(x)$ from the hate-related classes. The final prediction is determined as follows:

$$F(x) = \begin{cases} \text{None}, & \text{if } f_1(x) = \text{None}, \\ f_2(x), & \text{if } f_1(x) = \text{Hate}. \end{cases}$$

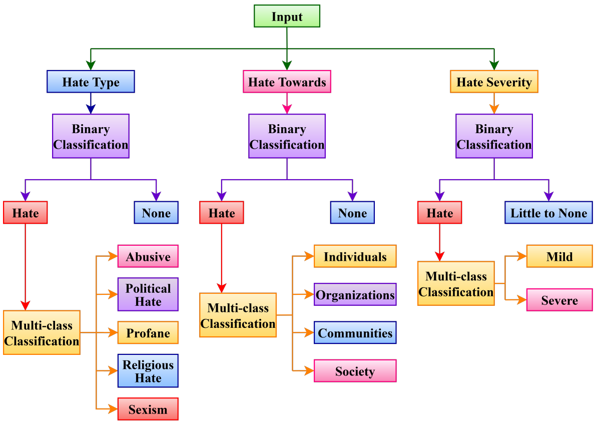


Figure 2: Hierarchical strategy for hate speech classification.

- **Ensemble:** A majority-voting ensemble was applied using the top three models (LR-Count-Hierarchical, SVC-lin-Count-Hierarchical, SVC-rbf-Count-Hierarchical). For each instance, the label agreed upon by at least two models was chosen; if all disagreed, the best-performing model's prediction was used (Algorithm 1).

Algorithm 1 Ensemble Voting for Hate Classification

```
Input: Predictions $P_1, P_2, P_3$ from top-3 models; best-performing model $M^*$
Output: Final predicted label $L$
for each instance $x$ do
    Collect predictions $\{p_1, p_2, p_3\}$
    if two or more predictions agree then
        Assign $L$ to the majority label
    else
        Assign $L$ to the prediction of $M^*$
    end if
end for
Return all $L$
```

4.1 Hyperparameter Settings

Table 4 lists the key hyperparameters used for each machine learning model in our experiments. These settings were selected based on preliminary tuning to balance performance and computational efficiency across the three subtasks.

Model	Key Hyperparameters
LR	C = 1, solver = liblinear
DT	criterion = entropy, min_samples_leaf = 2, random_state = 0
MNB	alpha = 0.5
KNN	n_neighbors = 7, weights = distance
RF	class_weight = balanced, criterion = entropy, max_features = log2, min_samples_leaf = 2, n_estimators = 10, random_state = 0
SVC-l	C = 0.2, kernel = linear, probability = True
SVC-p	C = 10, kernel = poly, gamma = auto
SVC-r	C = 1000, kernel = rbf, gamma = 0.00015, probability = True

Table 4: Hyperparameters used for classification models.

5 Results

All experiments were implemented in Python using scikit-learn and executed on Google Colab using a standard CPU, ensuring reproducibility across all subtasks. The overall pipeline was designed as a modular framework that covers preprocessing, feature extraction, training, and evaluation, enabling consistent comparisons of models across identical settings. Table 5 summarizes the performance of all model-vectorizer combinations under both simple and hierarchical settings.

Model-Vectorizer	**F1-micro (Simple)**	**F1-micro (Hierarchical)**
LR-Count	68.12	**68.29**
LR-TF-IDF	67.90	68.10
DT-Count	62.80	63.00
DT-TF-IDF	59.00	59.20
MNB-Count	64.10	64.30
MNB-TF-IDF	66.30	66.50
KNN-Count	63.10	63.30
KNN-TF-IDF	61.60	61.80
RF-Count	56.60	56.80
RF-TF-IDF	55.60	55.80
SVC-lin-Count	68.10	**68.32**
SVC-lin-TF-IDF	66.20	66.40
SVC-poly-Count	60.60	60.80
SVC-poly-TF-IDF	60.60	60.80
SVC-rbf-Count	68.10	**68.30**
SVC-rbf-TF-IDF	67.00	67.20
Majority Voting	–	**68.42**

Table 5: Performance of various models.

The majority voting ensemble achieved the highest F1-micro (**68.42%**), demonstrating robustness across setups. Logistic Regression and SVM consistently outperformed tree- and distance-based models, while hierarchical classification improved detection of minority classes by separating hate from non-hate content. TF-IDF favored linear models, whereas CountVectorizer slightly benefited tree-based methods. Overall, hierarchical classification with TF-IDF and ensemble voting provided the best balance between minority-class sensitivity and global accuracy.

5.1 Ablation Study

This subsection presents an ablation study analyzing the impact of preprocessing techniques and n-gram choices on model performance, along with token statistics at each preprocessing stage.

Impact of Text Preprocessing: We evaluated different preprocessing pipelines on Task 1C using Logistic Regression with trigram features (simple classification). The preprocessing variants are summarized in Table 6.

Preprocessing	**F1-Micro (%)**
text	66.60
nr_text	66.64
l_nr_text	67.31
sr_nr_text	67.12
sr_l_nr_text	67.90

Table 6: Effect of preprocessing on F1-micro. Preprocessing variations: `text` = raw text, `nr_text` = numbers removed, `l_nr_text` = lemmatized + numbers removed, `sr_nr_text` = stopwords removed + numbers removed, `sr_l_nr_text` = stopwords removed + lemmatized + numbers removed.

Impact of N-gram Choice: We tested unigram, unigram+bigram, and unigram+bigram+trigram TF-IDF vectorization with the best preprocessing pipeline on Task 1C (`sr_l_nr_text`) using Logistic Regression in the simple classification setting. The effect of n-gram choice on model performance is shown in Table 7.

Token Counts at Each Preprocessing Stage: Table 8 summarizes total and unique token counts for train, dev, and test sets in Task 1C.

N-gram	**F1-Micro (%)**
Unigram	67.82
Unigram+Bigram	67.75
Unigram+Bigram+Trigram	67.90

Table 7: Effect of n-gram choice on F1-micro.

Dataset	**Preprocessing**	**Total Tokens**	**Unique Tokens**
train_1C_df	text	488,944	51,719
	nr_text	486,001	50,879
	l_nr_text	486,005	41,023
	sr_nr_text	308,212	50,214
	sr_l_nr_text	268,138	40,632
dev_1C_df	text	35,975	9,495
	nr_text	35,772	9,367
	l_nr_text	35,772	6,924
	sr_nr_text	22,537	8,851
	sr_l_nr_text	19,657	6,607
test_1C_df	text	140,677	23,301
	nr_text	139,823	22,972
	l_nr_text	139,825	17,484
	sr_nr_text	89,168	22,356
	sr_l_nr_text	77,634	17,124

Table 8: Total and unique token counts for Task 1C at each preprocessing stage.

6 Conclusion

This work explored various machine learning models to detect hate speech in Bengali. The experiments demonstrate that LR and SVM perform well for the downstream task. Hierarchical classification combined with TF-IDF vectorization improves minority-class detection, while majority voting ensembles enhance robustness and overall F1-micro, achieving **68.42%**. Error analysis shows that subtle, context-dependent, and overlapping classes remain challenging, highlighting the value of hierarchical and ensemble strategies in handling nuanced hate speech. In the future, we plan to investigate deep learning and transformer architectures to more effectively identify contextual, implicit, and nuanced hate expressions that traditional methods often miss. We will also leverage sarcasm-aware, context-sensitive models to enhance the detection of subtle and overlapping hate categories.

Limitations

This section summarizes the main limitations of our study and the steps taken to address them:

- **Class Imbalance:** Rare classes like *Sexism* and *Severe* are under-represented, affecting generalization.

- **Contextual and Sarcastic Language:** Subtle or sarcastic expressions are often misclassified due to reliance on lexical features.
- **Domain Limitation:** Models trained only on YouTube comments may not generalize to other Bangla social media.
- **Sparse Vector Dependence:** Models trained on sparse representations (such as TF-IDF) struggle with complex linguistic variations and long-range dependencies.

Acknowledgments

This work is supported by the NLP Lab, Chittagong University of Engineering & Technology (CUET), Chittagong, Bangladesh. We also thank the organizers of BLP 2025 for conducting this shared task.

References

Areej Al-Hassan and Hmood Al-Dossari. 2022. Detection of hate speech in arabic tweets using deep learning. *Multimedia systems*, 28(6):1963–1974.

Md Alam, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. CUET_NLP_Manning@LT-EDI 2024: Transformer-based approach on caste and migration hate speech detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 238–243, St. Julian's, Malta. Association for Computational Linguistics.

Safae Sossi Alaoui, Yousef Farhaoui, and Brahim Aksasse. 2022. Hate speech detection using text mining and machine learning. *International Journal of Decision Support System Technology (IJDSST)*, 14(1):1–20.

Girma Bade, Olga Kolesnikova, Grigori Sidorov, and José Oropeza. 2024. Social media hate and offensive speech detection using machine learning method. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 240–244.

Muhammad Bilal, Atif Khan, Salman Jan, Shahrulniza Musa, and Shaukat Ali. 2023. Roman urdu hate speech detection using transformer-based model for cyber security applications. *Sensors*, 23(8):3909.

Patricia Chiril, Farah Benamara Zitoune, Véronique Moriceau, and Abhishek Kumar. 2019. The binary trio at SemEval-2019 task 5: Multitarget hate speech detection in tweets. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 489–493, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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, Online only. Association for Computational Linguistics.

Udoy Das, Karnis Fatema, Md Ayon Mia, Mahshar Yahan, Md Sajidul Mowla, Md Fayez Ullah, Arpita Sarker, and Hasan Murad. 2023. EmptyMind at BLP-2023 task 1: A transformer-based hierarchical-BERT model for Bangla violence-inciting text detection. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 174–178, Singapore. Association for Computational Linguistics.

Omar Faruqe, Mubassir Jahan, Md Faisal, Md Shahidul Islam, and Riasat Khan. 2023. Bangla hate speech detection system using transformer-based nlp and deep learning techniques. In *2023 3rd Asian Conference on Innovation in Technology (ASIANCON)*, pages 1–6. IEEE.

Fabiha Haider, Fariha Tanjim Shifat, Md Farhan Ishmam, Md Sakib Ul Rahman Sourove, Deeparghya Dutta Barua, Md Fahim, and Md Farhad Alam Bhuiyan. 2025. BanTH: A multi-label hate speech detection dataset for transliterated Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7217–7236, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

Jawad Hossain, Hasan Mesbaul Ali Taher, Avishek Das, and Mohammed Moshiul Hoque. 2023. NLP_CUET at BLP-2023 task 1: Fine-grained categorization of violence inciting text using transformer-based approach. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 241–246, Singapore. Association for Computational Linguistics.

Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57, Florence, Italy. Association for Computational Linguistics.

Shakir Khan, Mohd Fazil, Vineet Kumar Sejwal, Mohammed Ali Alshara, Reemiah Muneer Alotaibi, Ashraf Kamal, and Abdul Rauf Baig. 2022a. Bichat: Bilstm with deep cnn and hierarchical attention for hate speech detection. *Journal of King Saud University-Computer and Information Sciences*, 34(7):4335–4344.

Shakir Khan, Ashraf Kamal, Mohd Fazil, Mohammed Ali Alshara, Vineet Kumar Sejwal, Reemiah Muneer Alotaibi, Abdul Rauf Baig, and Salihah Alqahtani. 2022b. Hcovbi-caps: Hate speech detection using convolutional and bi-directional gated recurrent unit with capsule network. *IEEE Access*, 10:7881–7894.

Md. Emnul Momin and Hasan Sarker. 2025. Enhancing hate speech detection in bengali language using machine learning and deep learning algorithms. In *2025 2nd International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM)*, pages 1–6, Los Alamitos, CA, USA. IEEE Computer Society.

Sarah Nasir, Ayesha Seerat, and Muhammad Wasim. 2024. Hate speech detection in roman urdu using machine learning techniques. In *2024 5th International Conference on Advancements in Computational Sciences (ICACS)*, pages 1–7. IEEE.

Sagor Kumar Saha, Afrina Akter Mim, Sanzida Akter, Md. Mehraz Hosen, Arman Habib Shihab, and Md Humaion Kabir Mehedi. 2024. Bengalihatecb: A hybrid deep learning model to identify bengali hate speech detection from online platform. In *2024 6th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, pages 439–444. IEEE.

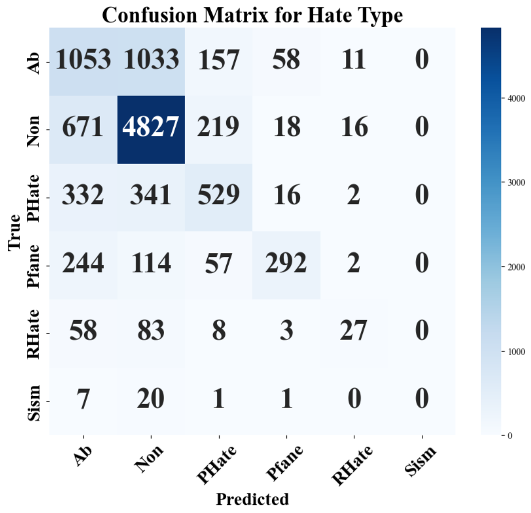
Hind Saleh, Areej Alhothali, and Kawthar Moria. 2023. Detection of hate speech using bert and hate speech word embedding with deep model. *Applied Artificial Intelligence*, 37(1):2166719.

Mahamat Saleh Adoum Sanoussi, Chen Xiaohua, George K Agordzo, Mahamed Lamine Guindo, Abdullah MMA Al Omari, and Boukhari Mahamat Issa. 2022. Detection of hate speech texts using machine learning algorithm. In *2022 IEEE 12th annual computing and communication workshop and conference (CCWC)*, pages 0266–0273. IEEE.

A Error Analysis

To assess the model's efficacy in various tasks, a detailed error analysis is conducted.

- **Quantitative Analysis:** Confusion matrices are used to conduct the quantitative analysis that highlights misclassification patterns. For **hate type**, the majority class *None* was well predicted (4827), but minority classes like *Sexism* were often confused with *None* or *Abusive* (Figure 3).



	Ab	Non	PHate	Pfane	RHate	Sism
Ab	1053	1033	157	58	11	0
Non	671	4827	219	18	16	0
PHate	332	341	529	16	2	0
Pfane	244	114	57	292	2	0
RHate	58	83	8	3	27	0
Sism	7	20	1	1	0	0

Figure 3: Majority vote ensemble Confusion matrix for Hate Type.

Hate severity showed reliable detection for *Little to None* (6325), while *Mild* and *Severe* were underestimated (Figure 4).

Target group predictions were strong for *None* (5293), but overlapping categories (e.g., *Community* vs. *Individual*) were frequently misclassified (Figure 5).

These patterns result from several challenges. Extreme label imbalance leads models to favor the majority classes, making small but key categories such as *Sexism*, *Severe*, and *Community* more likely to be misclassified. Many comments express hate implicitly, indi-

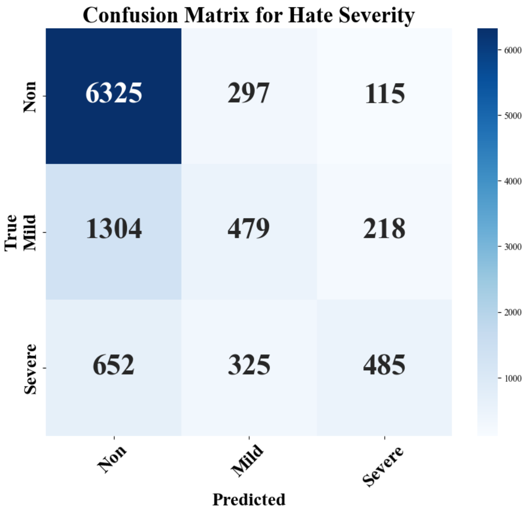


Figure 4: Majority vote ensemble Confusion matrix for Hate Severity

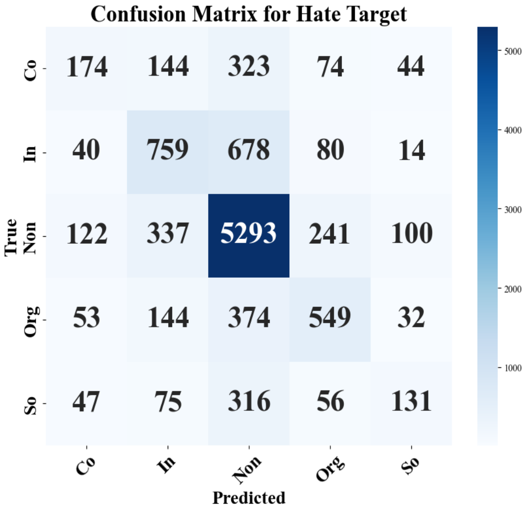
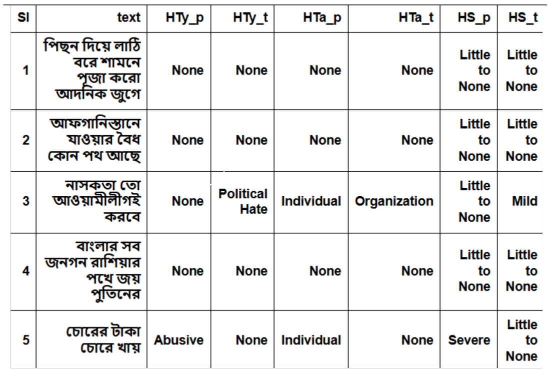


Figure 5: Majority vote ensemble Confusion matrix for Hate Target

rectly, or through sarcasm, which sparse TF-IDF features fail to capture, leading to confusion between labels such as *Mild*, *Little*, and *None*. Target group boundaries often blur in real user comments, leading to ambiguity and errors, such as predicting *Community* as *Individual*. These insights show the limits of classical models in handling nuanced, context-dependent hate speech and reinforce the value of the hierarchical and ensemble strategies used in this study.

- **Qualitative Analysis:** Examining predictions across **HTy, HTa, HS** (Figure 6) reveals systematic patterns.



SI	text	HTy_p	HTy_t	HTa_p	HTa_t	HS_p	HS_t
1	পিছন দিয়ে লাঠি বরে শামনে পূজা করো আদনিক জুগে	None	None	None	None	Little to None	Little to None
2	আফগানিস্তানে যাওয়ার বৈধ কোন পথ আছে	None	None	None	None	Little to None	Little to None
3	নাসকতা তো আওয়ামীলীগই করবে	None	Political Hate	Individual	Organization	Little to None	Mild
4	বাংলার সব জনগন রাশিয়ার পখে জয় পুতিনের	None	None	None	None	Little to None	Little to None
5	চোরের টাকা চোরে খায়	Abusive	None	Individual	None	Severe	Little to None

Figure 6: Sample predictions from the Majority Vote Ensemble. HTy = Hate Type, HTa = Hate Target, HS = Hate Severity; suffix p = predicted, suffix t = true label, SI = Sample Index.

Political and abusive content was occasionally misclassified (e.g., SI 3: *Political Hate* predicted as *None*). Implicit targets caused errors (SI 3: *Organization* predicted as *Individual*). Severity distinctions were often missed (SI 3: *Mild* predicted as *Little to None*, SI 5: *Little to None* predicted as *Severe*). Neutral examples were accurately classified (SI 1, 2, 4), but subtle, context-dependent hate remained challenging. These qualitative insights complement quantitative findings, emphasizing the utility of hierarchical and ensemble methods for nuanced class detection.