NSU_MILab at BLP-2025 Task 1: Decoding Bangla Hate Speech: Fine-Grained Type and Target Detection via Transformer Ensembles

Md. Mohibur Rahman Nabil[1], Muhammad Rafsan Kabir[1], Rakibul Islam[2], Fuad Rahman[3], Nabeel Mohammed[1], Shafin Rahman[1]

[1]Machine Intelligence Lab (MILab), North South University, Dhaka, Bangladesh
[2]Visa Inc., Atlanta, GA, USA
[3]Apurba Technologies, Sunnyvale, CA, USA

Abstract

This paper describes our participation in Task 1A and Task 1B of the BLP Workshop[1], focused on Bangla Multi-task Hatespeech Identification. Our approach involves systematic evaluation of four transformer models: BanglaBERT, XLM-RoBERTa, IndicBERT, and Bengali-Abusive-MuRIL. To enhance performance, we implemented an ensemble strategy that averages output probabilities from these transformer models, which consistently outperformed individual models across both tasks. The baseline classical methods demonstrated limitations in capturing complex linguistic cues, underscoring the superiority of transformer-based approaches for low-resource hate speech detection. Our solution initially achieved F1 scores of 0.7235 (ranked 12th) for Task 1A and 0.6981 (ranked 17th) for Task 1B among participating teams. Through post-competition refinements, we improved our Task 1B performance to 0.7331, demonstrating the effectiveness of ensemble methods in Bangla hate speech detection.

1 Introduction

In recent years, online communication has become a primary medium for individuals to express opinions and emotions. With the growing use of digital platforms, the prevalence of hate speech has also increased rapidly. Hate speech refers to language that spreads hostility or discrimination against individuals or groups based on attributes such as appearance, religion, ethnicity, or gender (Papcunová et al., 2023b). Such content not only fuels social conflict but can also damage international relations and, in extreme cases, contribute to violent outcomes, including wars (Sahoo et al., 2024). While significant progress has been made in detecting hate speech in high-resource languages such as English (MacAvaney et al., 2019; Kearns et al., 2023), the challenge remains particularly acute for underrepresented languages like Bangla, where datasets, resources, and detection systems are still scarce.

Existing approaches to hate speech detection often rely on classical machine learning algorithms (Mullah and Zainon, 2021; Subramanian et al., 2023), which struggle to capture the linguistic nuances present in hateful texts, especially in low-resource languages. Moreover, most prior studies focus on binary classification (Subramanian et al., 2023), distinguishing hate from non-hate, without addressing the finer-grained categorization of hate into types such as abusive, religious, political, or sexist. Equally overlooked is the identification of the target of hate, whether directed toward individuals, organizations, communities, or society. These limitations highlight a major gap in comprehensive Bangla hate speech identification.

To address these gaps, this study advances beyond binary hate speech detection and tackles two key tasks: ***(a)*** fine-grained classification of hate speech types and ***(b)*** identification of the targeted group. As a baseline, we experimented with classical machine learning classifiers using sentence embeddings from a pretrained sentence transformer, but their limitations in capturing complex linguistic cues underscored the need for transformer-based approaches. We therefore employed four Bangla-specific models, BanglaBERT (Hasan et al., 2020), XLM-RoBERTa (Conneau et al., 2020), IndicBERT (Kakwani et al., 2020), and Bengali-Abusive-MuRIL (Das et al., 2022), and further enhanced performance through an ensemble strategy that averages their output probabilities. The ensemble consistently outperformed individual models across both tasks, demonstrating its effectiveness for Bangla hate speech detection.

The main contributions are as follows: ***(i)*** Development of a Bangla hate speech detection framework that extends beyond binary classification to perform fine-grained hate categorization and target

[1]https://multihate.github.io/

Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 461–467

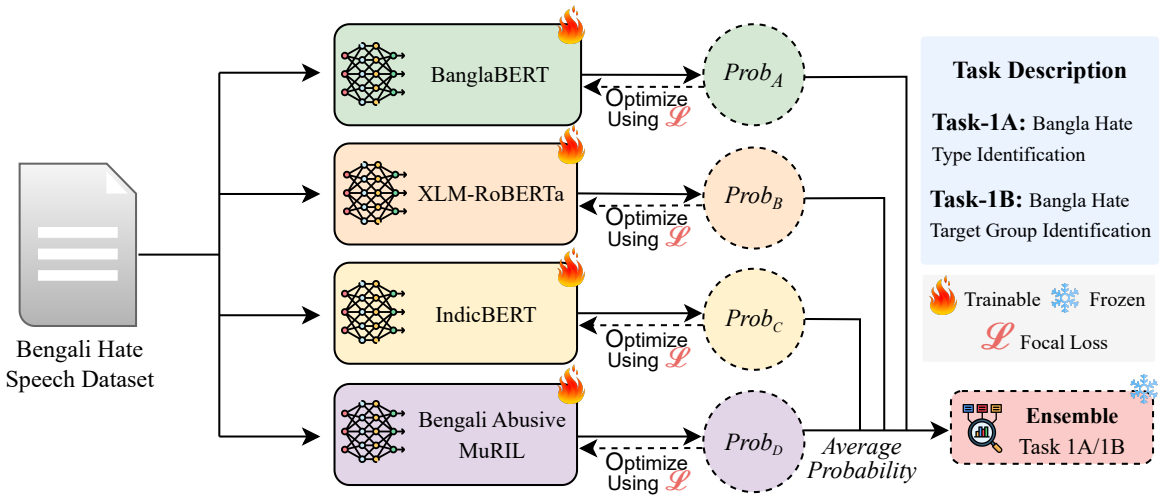

Figure 1: Overview of the proposed Bangla hate speech type identification and Bangla Hate speech target group identification framework. Bangla text inputs are passed through four transformer-based models (BanglaBERT, XLM-RoBERTa, IndicBert, and Bengali-Abusive-MuRIL). A focal loss function is applied to address class imbalance, and the averaged outputs are used for final classification and target group identification.

group identification. ***(ii)*** Systematic evaluation of baseline and advanced approaches, where machine learning classifiers are compared against Bangla-specific transformer models, highlighting the limitations of traditional methods. ***(iii)*** Introduction of a transformer model ensemble that achieves superior performance across both tasks, demonstrating the effectiveness of ensembling in a low-resource language setting.

2 Related Works

Hate speech detection has been widely studied in recent years, with early surveys highlighting challenges such as linguistic subtlety, implicit hate, and limited annotated resources (MacAvaney et al., 2019). While most research has focused on high-resource languages, recent studies emphasize the urgent need for progress in low-resource languages, where annotated corpora and robust models are scarce (Das et al., 2024). For Bangla, initial resources such as BD-SHS (Romim et al., 2022) and BanglaHateBERT (Jahan et al., 2022) have facilitated the development of benchmark systems. Transformer-based approaches have proven effective for hate speech detection (Chakravorty et al., 2024). Monolingual models such as MahaBERT and BanglaBERT often outperform multilingual baselines in capturing language-specific nuances, while multilingual models like MuRIL and XLM-RoBERTa demonstrate strong cross-lingual transfer (Ghosh and Senapati, 2022, 2025). Recent studies further demonstrate that multilingual and multi-task learning can improve generalization across domains and targets of hate (Yuan and Rizoiu, 2025). Generative large language models (LLMs) also show promising results, surpassing traditional transformer baselines in Bangla hate detection tasks (Faria et al., 2024). Finally, datasets such as IndicCONAN (Sahoo et al., 2024) support counter-narrative generation, broadening the scope of hate speech research in Indic languages. Together, these works highlight the evolution from classical methods (Kearns et al., 2023; Papcunová et al., 2023a) to transformer and LLM-driven approaches, underscoring the importance of developing robust systems for Bangla and other low-resource languages. While prior works highlight individual model strengths, we show that an ensemble of state-of-the-art transformers yields more robust and accurate fine-grained Bangla hate speech detection.

3 Methods

Problem Formulation: Given a dataset $\mathcal{D} = \{(T_i, y_i)\}_{i=1}^{N}$ of Bangla text samples T_i and their corresponding labels y_i, our objectives are: ***(i)*** *Hate speech type identification*, where $y_i \in \{1, \ldots, 6\}$ denotes one of six predefined hate speech categories, and ***(ii)*** *Target group identification*, where $y_i \in \{1, \ldots, 5\}$ denotes the specific target group five possible categories.

Solution Strategy: Bangla hate speech detection is particularly challenging due to limited resources and significant class imbalance. To address these issues, we adopt the following strategy: ***(a) Multi-***

Model Ensemble: Four transformer-based models are fine-tuned on dataset $\mathcal{D}$ to capture diverse linguistic features. ***(b) Focal Loss Optimization:*** A focal loss function is employed during training to mitigate class imbalance by emphasizing harder, misclassified samples. ***(c) Prediction Aggregation:*** The output probabilities of the four models are averaged to obtain robust final predictions for both hate category and target group identification.

3.1 Revisiting Transformer Models

Bangla hate speech detection is challenging due to linguistic variations, dialect diversity, and code-mixing. As baselines, we fine-tune four distinct transformer-based models on the dataset. **BanglaBERT** is a monolingual model trained on a large Bangla corpus. **XLM-RoBERTa** is a multilingual model that captures cross-lingual representations. **IndicBERT** is trained on multiple Indic languages, including Bangla, and leverages shared linguistic features. **Bengali-Abusive-MuRIL** is pretrained with a focus on abusive content, making it relevant for hate speech detection.

3.2 Ensemble Approach

As Bangla hate speech detection is a complex task due to its low-resource nature, individual models often fail to capture all nuances. Relying on a single model often fails to capture these nuances. To overcome this, we employ an ensemble of four transformer-based models: Bangla BERT, XLM-RoBERTa, IndicBERT, and Bengali-Abusive-MuRIL. Each model is fine-tuned individually on the dataset to learn task-specific features while preserving its pretraining knowledge. As illustrated in Figure 1, the input text is processed in parallel through the four models, and their predictions are later combined. The ensemble strategy leverages complementary strengths of different pretrained models, thereby improving generalization and robustness compared to any single model.

Focal Loss Optimization: The dataset (Hasan et al., 2025b) is highly imbalanced shown in Appendix A.1, with hateful instances being significantly underrepresented. Training with standard cross-entropy loss leads to models biased toward the majority (non-hateful) class. To mitigate this, we adopt focal loss, defined as:

$$\mathcal{L}_{\text{focal}}(p_t) = -\alpha(1-p_t)^{\gamma}\log(p_t)$$

where p_t is the predicted probability for the true class, α is a balancing factor, and γ is a focusing parameter. The modulating term $(1-p_t)^{\gamma}$ reduces the relative loss for well-classified examples, thus placing more emphasis on harder, misclassified samples. This helps the model pay more attention to minority classes and subtle hate speech instances.

Prediction Aggregation: After training, the outputs of each model are combined into a single prediction. We adopt a simple yet effective strategy of averaging all predicted probabilities:

$$\hat{y} = \frac{1}{M}\sum_{m=1}^{M}\hat{y}_m$$

where M is the number of models and $\hat{y}_m$ is the output probability from the m-th model. Averaging stabilizes predictions, reduces variance, and avoids overfitting that may arise from a single model.

4 Experiments

4.1 Setup

Dataset: We conducted our experiments on the BLP Shared Task 1A and Task 1B datasets (Hasan et al., 2025a,b) for Bangla hate speech. Each dataset consists of 35,522 training samples, 2,512 validation samples, 2,512 dev-test samples, and 10,200 test samples. Task 1A focuses on classifying the type of hate speech across six categories: None, Abusive, Political Hate, Religious Hate, Sexism, and Profane. Task 1B, on the other hand, involves identifying the target group with five classes: None, Individual, Organization, Community, and Society. The detailed class-wise distributions across all dataset splits are provided in Appendix A.1.

Evaluation Metrics: Performance was assessed using Precision (P), Recall (R), and F1-score, reported on both the dev-test and test datasets for both Task 1A and Task 1B.

4.2 Main Results

Table 1 presents the performance of individual models and the ensemble system on both subtasks: Task 1A (Bangla hate Speech Type Identification) and Task 1B (Bangla hate Speech Target Group Identification) for the dev-test and test datasets.

For **Task 1A**, BanglaBERT and Bengali-Abusive-MuRIL achieved strong results, with F1-scores of 0.7312 and 0.7075 on the dev-test set, respectively. IndicBERT performed moderately with an F1 score of 0.6455, while XLM-RoBERTa lagged behind, achieving only 0.4211 F1 on the

Model	Task 1A:Bangla Hate Speech Type Identification						Task 1B:Bangla Hate Speech Target Group Identification					
	Dev-Test			Test			Dev-Test			Test		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
BanglaBERT (Hasan et al., 2020)	0.7385	0.7269	0.7312	0.7178	0.7046	0.7092	0.7399	0.7186	0.7259	0.7320	0.7123	0.7187
XLM-RoBERTa (Conneau et al., 2020)	0.3318	0.5760	0.4211	0.3179	0.5638	0.4066	0.3700	0.6083	0.4601	0.3568	0.5974	0.4468
IndicBERT (Kakwani et al., 2020)	0.6446	0.6525	0.6455	0.6364	0.6481	0.6393	0.6619	0.6712	0.6644	0.6491	0.6612	0.6526
Bengali-Abusive-MuRIL (Das et al., 2022)	0.7058	0.7109	0.7075	0.6978	0.7019	0.6996	0.7288	0.7277	0.7275	0.7068	0.7081	0.7070
Ensemble of All (with focal loss)	**0.7358**	**0.7448**	**0.7396**	**0.7211**	**0.7271**	**0.7235**	**0.7447**	**0.7464**	**0.7444**	**0.7320**	**0.7350**	**0.7331**

Table 1: Performance of individual models and the ensemble on Task 1A and Task 1B. Results are reported on dev-test and test sets in terms of Precision (P), Recall (R), and F1-score. Best results are in **bold**, and the second-best are underlined. More detailed ensembling results are presented in Appendix A.2.

dev-test. The ensemble of the four models with focal loss outperformed all individual models, achieving the best F1-scores of 0.7396 on the dev-test and 0.7235 on the test set. For **Task 1B**, a similar trend was observed. BanglaBERT and Bengali-Abusive-MuRIL were competitive baselines, with F1-scores of 0.7259 and 0.7275 on the dev-test, respectively. IndicBERT performed slightly lower, and XLM-RoBERTa again underperformed compared to monolingual and multilingual models tailored for Indic languages. Our proposed ensemble achieved the highest overall performance, reaching an F1 score of 0.7444 on the dev-test and 0.7331 on the test set.

These results demonstrate that while Bangla-specific and Indic-focused models are strong baselines for hate speech detection, the ensemble strategy with focal loss provides consistent gains across both subtasks, highlighting the benefits of combining diverse model predictions.

4.3 Error Analysis

Figure 2 shows error patterns across both tasks, closely tied to the dataset distribution (Table 2 in Appendix A.1). In Task 1A, most errors occur in Abusive (1,088 errors) and Political Hate (533 errors), the largest hate-related categories, where greater lexical diversity makes classification harder. In contrast, Religious Hate (28 errors) and Sexism (89 errors) appear better handled, though this is partly due to their small sample sizes (676 and 122 samples), which limit variability rather than stronger generalization. Profane shows moderate difficulty with 163 errors.

For Task 1B, the None class dominates errors (929 errors), reflecting its overwhelming size (21,190 samples) and the challenge of distinguishing non-targeted from subtly targeted text. Other categories, including Individual (555 errors), Organization (470 errors), and Society (359 errors), show comparable difficulty, while Community (389 errors) is relatively more stable despite fewer examples. Overall, class imbalance drives most errors, with frequent misclassifications in majority classes and limited coverage for minority ones.

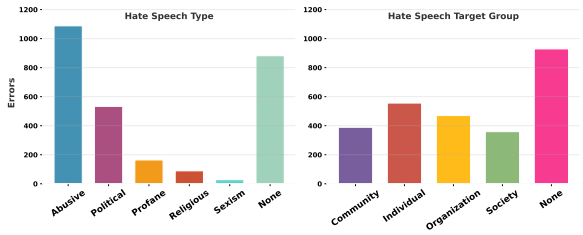

Figure 2: Error distribution across classification tasks. The bar charts show the number of misclassified samples for each class in Task 1A (left) and Task 1B (right).

4.4 Discussion

Our findings show that Bangla-specific models such as BanglaBERT and Bengali-Abusive-MuRIL outperform general multilingual models like XLM-RoBERTa, emphasizing the importance of language-focused pretraining. IndicBERT achieved moderate results, but its multilingual nature limited its effectiveness compared to Bangla-focused models. Across both subtasks, the ensemble approach consistently showed the best scores. These results highlight the value of combining diverse models to enhance generalization in Bangla hate speech detection.

Conclusion

In this study, we presented our proposed method and results on the BLP Shared Task 1A (Bangla hate speech type classification) and Task 1B (target group identification). Our work highlights the importance of tackling the challenging problem of hate speech detection in a low-resource language (Bangla), particularly in identifying both the type of hate speech and its target group. To this end, we employed an ensemble of four transformer-based

models, demonstrating the effectiveness of robust NLP systems in mitigating harmful online content.

Limitations

Although our ensemble framework achieved competitive rankings in the shared task, it faced notable constraints. The reliance on pretrained transformer models introduced high computational costs during fine-tuning, which may not be feasible in resource-limited environments. Furthermore, class imbalance in the dataset, particularly for underrepresented categories such as Sexism and Religious Hate, limited the models ability to generalize across all classes. These challenges contributed to misclassifications observed in the error analysis, highlighting difficulties in handling overlapping or subtle linguistic cues. Another limitation is that the ensemble approach, while effective, increases inference time compared to single-model systems, which could hinder real-time or large-scale deployment. Moreover, our ensembling method used uniform averaging, which may not optimally capture the varying strengths of individual models.

Future Works: Building on the competition results, future work can focus on improving class balance through techniques such as data augmentation, resampling strategies, or more adaptive loss functions. Exploring alternative ensemble strategies beyond simple probability averaging, such as weighted ensembling or stacking, could further enhance performance by leveraging model-specific strengths. Finally, incorporating more efficient fine-tuning methods (e.g., parameter-efficient tuning) may reduce computational demands, enabling broader participation in similar low-resource shared tasks while maintaining strong performance.

References

Debamalya Chakravorty, Arijit Das, and Diganta Saha. 2024. Multilingual hate speech detection using transformer-based deep learning approaches. In *2024 11th International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 126–131.

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, Online. Association for Computational Linguistics.

Mithun Das, Somnath Banerjee, and Animesh Mukherjee. 2022. Data bootstrapping approaches to improve low resource abusive language detection for indic languages. In *Proceedings of the 33rd ACM conference on hypertext and social media*, pages 32–42.

Susmita Das, Arpita Dutta, Kingshuk Roy, Abir Mondal, and Arnab Mukhopadhyay. 2024. A survey on automatic online hate speech detection in low-resource languages. arXiv preprint arXiv:2411.19017.

Fatema Tuj Johora Faria, Laith H. Baniata, and Sangwoo Kang. 2024. Investigating the predominance of large language models in low-resource bangla language over transformer models for hate speech detection: A comparative analysis. *Mathematics*, 12(23):3687.

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 (PACLIC)*, pages 853–865, Manila, Philippines. Association for Computational Linguistics.

Koyel Ghosh and Apurbalal Senapati. 2025. Hate speech detection in low-resourced indian languages: An analysis of transformer-based monolingual and multilingual models with cross-lingual experiments. *Natural Language Processing*, 31(2):393–414.

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. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2612–2623, Online. Association for Computational Linguistics.

M. S. Jahan, M. Haque, N. Arhab, and M. Oussalah. 2022. BanglaHateBERT: Bert for abusive language detection in bengali. In *Proceedings of the Second Workshop on Abusive Language Online (co-located with LREC 2022)*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul NC, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian

languages. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 4948–4961.

Colm Kearns, Gary Sinclair, Jack Black, Mark Doidge, Thomas Fletcher, Daniel Kilvington, Katie Liston, Theo Lynn, and Pierangelo Rosati. 2023. A scoping review of research on online hate and sport. *Communication & Sport*, 11(2):402–430.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Nanlir Sallau Mullah and Wan Mohd Nazmee Wan Zainon. 2021. Advances in machine learning algorithms for hate speech detection in social media: a review. *IEEE access*, 9:88364–88376.

Jana Papcunová, Marcel Martončík, Denisa Fedáková, Michal Kentoš, and Matúš Adamkovič. 2023a. Perception of hate speech by the public and experts: Insights into predictors of the perceived hate speech towards migrants. *Cyberpsychology, Behavior, and Social Networking*, 26:546–553.

Jana Papcunová, Marcel Martončik, Denisa Fedáková, Michal Kentoš, Miroslava Bozogáňová, Ivan Srba, Robert Moro, Matúš Pikuliak, Marián Šimko, and Matúš Adamkovič. 2023b. Hate speech operationalization: a preliminary examination of hate speech indicators and their structure. *Complex & intelligent systems*, 9(3):2827–2842.

N. Romim, M. Ahmed, M. S. Islam, A. S. Sharma, H. Talukder, and M. R. Amin. 2022. Bd-shs: A benchmark dataset for learning to detect online bangla hate speech in different social contexts. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 5153–5162.

Nihar Ranja Sahoo, Gyana Prakash Beria, and Pushpak Bhattacharyya. 2024. Indicconan: A multilingual dataset for combating hate speech in indian context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22313–22321.

Malliga Subramanian, Veerappampalayam Easwaramoorthy Sathiskumar, G Deepalakshmi, Jaehyuk Cho, and G Manikandan. 2023. A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. *Alexandria Engineering Journal*, 80:110–121.

Lanqin Yuan and Marian-Andrei Rizoiu. 2025. Generalizing hate speech detection using multi-task learning: A case study of political public figures. *Computer Speech & Language*, 89:101690.

A Appendix

A.1 Dataset Distribution Details

Table 2 provides a detailed class-wise breakdown of the dataset used in our experiments, covering both Task 1A (hate speech type classification) and Task 1B (target group identification). The dataset was split into four partitions: training, validation, dev-test, and test.

Task 1A: Hate Speech Type				
Class	**Train**	**Val**	**Dev-Test**	**Test**
None	19,954	1,451	1,447	5,751
Abusive	8,212	564	549	2,312
Political Hate	4,227	291	283	1,220
Religious Hate	676	38	40	179
Sexism	122	11	8	29
Profane	2,331	157	185	709
Total	**35,522**	**2,512**	**2,512**	**10,200**

Task 1B: Target Identification				
Class	**Train**	**Val**	**Dev-Test**	**Test**
None	21,190	1,536	1,528	6,093
Individual	5,646	364	391	1,571
Organization	3,846	292	292	1,152
Community	2,635	179	159	759
Society	2,205	141	142	625
Total	**35,522**	**2,512**	**2,512**	**10,200**

Table 2: Dataset distribution showing class-wise breakdown for Task 1A (Bangla hate speech type classification) and Task 1B (Bangla hate speec target group identification) across all splits.

Bangla Hate Speech Type Classification: This task focuses on categorizing each instance into one of six classes: *None*, *Abusive*, *Political Hate*, *Religious Hate*, *Sexism*, and *Profane*. The distribution is highly imbalanced, with the majority of samples belonging to the *None* and *Abusive* categories. Specifically, the training set contains 19,954 instances of *None* and 8,212 instances of *Abusive*, compared to only 122 samples of *Sexism*. Such imbalance poses challenges for model training, as minority classes like *Sexism* and *Religious Hate* are be underrepresented.

Bangla Hate Speech Target Group Identification: This task aims to identify the target group of the hateful expression, categorized into five groups: *None*, *Individual*, *Organization*, *Community*, and *Society*. As shown in Table 2, the distribution is again skewed, with *None* being the dominant category (21,190 training instances), followed by *Individual* (5,646 samples) and *Organization* (3,846 samples). The minority categories, such as *Community* (2,635 samples) and *Society* (2,205 samples),

Model / Ensemble	Task 1A: Bangla Hate Speech Type Identification						Task 1B: Bangla Hate Speech Target Group Identification					
	Dev-Test			Test			Dev-Test			Test		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
2-Model Ensembles												
BanglaBERT + XLM-RoBERTa	0.7202	0.7237	0.7201	0.7049	0.7071	0.7034	0.7464	0.7392	0.7415	0.7005	0.6937	0.6305
BanglaBERT + IndicBERT	0.7022	0.7133	0.7051	0.6874	0.6994	0.6909	0.7334	0.7336	0.7312	0.7203	0.7216	0.7182
BanglaBERT + B-A-MuRIL	0.7166	0.7233	0.7192	0.7054	0.7107	0.7077	0.7417	0.7408	0.7399	0.7210	0.7212	0.7200
XLM-RoBERTa + IndicBERT	0.6318	0.6524	0.6348	0.6310	0.6507	0.6319	0.7259	0.7328	0.7285	0.5786	0.6437	0.5491
XLM-RoBERTa + B-A-MuRIL	0.7065	0.7149	0.7097	0.6958	0.7021	0.6986	0.7391	0.7412	0.7397	0.6795	0.6912	0.6385
IndicBERT + B-A-MuRIL	0.7040	0.7169	0.7082	0.6835	0.6965	0.7088	0.7006	0.7320	0.7261	0.7052	0.7165	0.7087
3-Model Ensembles												
BanglaBERT + XLM-RoBERTa + IndicBERT	0.6865	0.7042	0.6868	0.6774	0.6950	0.6768	0.6979	0.7165	0.6832	0.6929	0.7084	0.6734
BanglaBERT + XLM-RoBERTa + B-A-MuRIL	0.7210	0.7336	0.7256	0.7068	0.7186	0.7109	0.7320	0.7452	0.7308	0.7141	0.7301	0.7138
BanglaBERT + IndicBERT + B-A-MuRIL	0.7201	0.7289	0.7237	0.7159	0.7229	0.7190	0.7385	0.7416	0.7382	0.7243	0.7291	0.7248
XLM-RoBERTa + IndicBERT + B-A-MuRIL	0.6950	0.7125	0.6936	0.6770	0.6977	0.6763	0.7029	0.7221	0.6901	0.6855	0.7059	0.6712
Ensemble (with focal loss)	**0.7358**	**0.7448**	**0.7396**	**0.7211**	**0.7271**	**0.7235**	**0.7447**	**0.7464**	**0.7444**	**0.7320**	**0.7350**	**0.7331**

Table 3: Performance comparison of two-model ensembles, three-model ensembles, and the final four-model ensemble on Task 1A and Task 1B. Results are reported in terms of Precision (P), Recall (R), and F1-score. Best results are in **bold**. B-A-MuRIL denotes Bengali-Abusive-MuRIL.

contain relatively fewer examples.

A.2 Detailed Results of Model and Ensemble Experiments

Table 3 presents the complete experimental results for both sub-tasks: **Task 1A** (Bangla Hate Speech Type Identification) and **Task 1B** (Bangla Hate Speech Target Group Identification). We report Precision (P), Recall (R), and F1-score on both the **dev-test** and **test** splits for all two-model ensembles, three-model ensembles, and the final four-model ensemble trained with focal loss.

Two-Model Ensembles: Pairwise ensembles demonstrate varying levels of performance depending on model combinations. For Task 1A, **BanglaBERT + Bengali-Abusive-MuRIL** achieves the strongest results among 2-model ensembles with F1-scores of 0.7192 (dev-test) and 0.7077 (test). The **BanglaBERT + XLM-RoBERTa** combination yields comparable dev-test performance (0.7201 F1) but slightly lower test scores (0.7034 F1). In Task 1B, **BanglaBERT + XLM-RoBERTa** shows strong dev-test performance with an F1 score of 0.7415, though it experiences a notable drop on the test set (0.6305 F1). The most stable 2-model ensemble for Task 1B is **BanglaBERT + Bengali-Abusive-MuRIL**, achieving 0.7399 (dev-test) and 0.7200 (test). Ensembles involving XLM-RoBERTa without BanglaBERT show weaker performance, particularly **XLM-RoBERTa + IndicBERT**, which achieves only 0.6348 F1 (dev-test) and 0.6319 F1 (test) in Task 1A.

Three-Model Ensembles: Three-way ensembles demonstrate improved stability and performance over most pairwise combinations. The **BanglaBERT + IndicBERT + Bengali-Abusive-MuRIL** ensemble achieves the highest performance among this group, with F1-scores of 0.7237 (dev-test) and 0.7190 (test) in Task 1A, and 0.7382 (dev-test) and 0.7248 (test) in Task 1B. The **BanglaBERT + XLM-RoBERTa + Bengali-Abusive-MuRIL** combination also performs strongly, achieving 0.7256 (dev-test) and 0.7109 (test) in Task 1A, and 0.7308 (dev-test) and 0.7138 (test) in Task 1B. These results indicate that combining Bangla-specific pretrained models with multilingual counterparts helps capture diverse linguistic cues while maintaining robustness. Ensembles excluding BanglaBERT consistently underperform, with **XLM-RoBERTa + IndicBERT + Bengali-Abusive-MuRIL** achieving only 0.6936 F1 (dev-test) in Task 1A.

Final Four-Model Ensemble with Focal Loss: The strongest performance is obtained with the final **all-model ensemble** (BanglaBERT, XLM-RoBERTa, IndicBERT, and Bengali-Abusive-MuRIL) trained using focal loss. This configuration achieves **0.7396 F1 (dev-test) and 0.7235 F1 (test) on Task 1A**, and **0.7444 F1 (dev-test) and 0.7331 F1 (test) on Task 1B**, outperforming all other ensemble configurations. Notably, this represents improvements of 0.0159 F1 points (dev-test) and 0.0045 F1 points (test) over the best 3-model ensemble in Task 1A, and 0.0062 F1 points (dev-test) and 0.0083 F1 points (test) in Task 1B. The final ensemble also demonstrates better generalization, with precision of 0.7358 and recall of 0.7448 on the dev-test for Task 1A, indicating a balanced approach to both false positives and false negatives.