Retriv at BLP-2025 Task 1: A Transformer Ensemble and Multi-Task Learning Approach for Bangla Hate Speech Identification

Sourav Saha, K M Nafi Asib, and Mohammed Moshiul Hoque
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
{sahasourav1170, nafi.asib}@gmail.com
moshiul_240@cuet.ac.bd

Abstract

This paper addresses the problem of Bangla hate speech identification, a socially impactful yet linguistically challenging task. As part of the "Bangla Multi-task Hate Speech Identification" shared task at the BLP Workshop, IJCNLP–AACL 2025, our team "Retriv" participated in all three subtasks: (1A) hate type classification, (1B) target group identification, and (1C) joint detection of type, severity, and target. For subtasks 1A and 1B, we employed a soft-voting ensemble of transformer models (BanglaBERT, MuRIL, IndicBERTv2). For subtask 1C, we trained three multitask variants and aggregated their predictions through a weighted voting ensemble. Our systems achieved micro-f_1 scores of 72.75% (1A) and 72.69% (1B), and a weighted micro-f_1 score of 72.62% (1C). On the shared task leaderboard, these corresponded to 9th, 10th, and 7th positions, respectively. These results highlight the promise of transformer ensembles and weighted multitask frameworks for advancing Bangla hate speech detection in low-resource contexts. We made experimental scripts publicly available for the community.[1]

1 Introduction

With the rapid growth of social media platforms, harmful content such as hate speech and offensive language has become a pressing concern, requiring effective strategies to prevent its spread. Automated detection methods have seen substantial progress in high-resource languages, aided by large datasets and transformer-based models. However, in low-resource languages like Bangla hate speech detection remains challenging due to limited annotated resources, dialectal variation, and frequent code-mixing. Most existing work has focused on binary classification (hate vs. non-hate) or coarse multi-class labeling, leaving fine-grained dimensions such as *type*, *severity*, and *target* underexplored. To address this gap, the BLP Workshop[2] at IJCNLP-AACL 2025 (Hasan et al., 2025b) introduced a shared task comprising three subtasks, including a multitask setup, to advance fine-grained hate speech modeling in Bangla. This paper advances current research by presenting our systems developed for the shared task. The key contributions of this work are illustrated in the following:

- Proposed efficient yet competitive ensemble methods of Bangla-capable transformer models, achieving strong performance for fine-grained hate speech classification (Subtasks 1A and 1B).
- Introduced a weighted voting ensemble within a multitask learning (MTL) framework, enabling joint prediction of hate *type*, *severity*, and *target group*, and demonstrating the viability of MTL for Bangla hate speech.
- Provided a comprehensive empirical study of deep learning and transformer-based approaches, including detailed performance comparisons and error analyses, offering insights for future research in low-resource hate speech detection.

2 Related Work

Research on Bangla hate speech detection has expanded in recent years with several new datasets and modeling approaches. Das et al. (2022) developed a corpus of Bangla and Romanized Bangla posts for hate and offensive language detection, demonstrating strong results with multilingual transformers such as XLM-R and MuRIL. Saha et al. (2023) introduced Vio-lens, covering social media posts linked to communal violence,

[1] https://github.com/sahasourav17/Retriv-BLP25-Task-1

[2] https://blp-workshop.github.io/

while Haider et al. (2025) proposed BanTH, a multi-label dataset for transliterated Bangla that captures multiple target categories and reflects the complexity of real-world hate speech. Hasan et al. (2025a) further introduced BanglaMultiHate, the first multi-task Bangla hate speech dataset jointly modeling *type*, *severity*, and *target*, with extensive experiments using LLMs under zero-shot and LoRA fine-tuning. Beyond datasets, Raza and Chatrath (2024) presented HarmonyNet, an ensemble framework that improves robustness in hate speech identification, and Hossain et al. (2024) proposed a multimodal approach aligning visual and textual features for hateful content detection. Despite these advances, most Bangla work remains focused on binary or coarse multi-class classification.

Multi-task learning (MTL) has been explored as a way to capture complementary signals in hate speech detection. For example, Awal et al. (2021) introduced AngryBERT, which jointly learns *target* and *emotion* alongside hate classification, while in the Bangla context, Saha et al. (2024) proposed MulTSeA, a multitask framework for aspect-based sentiment analysis. These studies suggest that MTL is well-suited for complex tasks like hate speech, where dimensions such as type, severity, and target are interdependent.

Our work extends Bangla hate speech identification to a fine-grained, multi-dimensional setting. Unlike prior studies focused on binary or coarse classification, we employ transformer-based ensembles and a multitask framework to jointly model *type*, *severity*, and *target group*, addressing a key gap in Bangla hate speech research.

3 Task and Dataset Descriptions

The primary aim of this shared task (Hasan et al., 2025b) is to conduct fine-grained hate speech identification in Bangla social media content, moving beyond binary detection toward multi-dimensional classification. The task was organized into multiple related subtasks:

- **Subtask 1A (Hate Type)**: Classify a text as *Abusive*, *Sexism*, *Religious Hate*, *Political Hate*, *Profane*, or *None*.
- **Subtask 1B (Target Group)**: Identify whether the hate is targeted at *Individuals*, *Organizations*, *Communities*, or *Society*.
- **Subtask 1C (Multitask)**: Jointly predict the *hate type*, *severity* (*Little to None*, *Mild*, *Severe*), and *target group*.

All three subtasks share the same dataset splits with 35,522 samples for training, 2,512 for development, and 10,200 for testing. On average, the texts contain about 78 tokens across all splits. While the split sizes are identical, the label distributions differ across subtasks. Detailed label distributions for each subtask are provided in the Appendix A

4 System Description

Our approach leverages different architectures tailored to the specific characteristics of each subtask. For subtasks 1A and 1B, we employ a soft ensemble of three pre-trained transformer models, while for subtask 1C, we adopt a multitask learning framework.

4.1 Text Preprocessing

We apply minimal preprocessing to preserve the authentic nature of social media content. The preprocessing pipeline consists of: (1) removal of Bangla digits, and (2) standard tokenization using each model's respective tokenizer. We observe that the provided dataset appears to be well-curated, requiring minimal additional cleaning steps. All input sequences are truncated or padded to a maximum length of 128 tokens.

4.2 Base Models

We utilize three complementary pre-trained transformer models across all subtasks:

- **BanglaBERT** (`csebuetnlp/banglabert`): A monolingual BERT model specifically pre-trained on Bangla text, providing strong language-specific representations. (Bhattacharjee et al., 2022)
- **MuRIL** (`google/muril-base-cased`): A multilingual model covering 17 Indian languages, including Bangla, offering cross-lingual contextual understanding. (Khanuja et al., 2021)
- **IndicBERTv2** (`ai4bharat/IndicBERTv2-MLM-only`): A model trained on 12 major Indian languages with enhanced tokenization for Indic scripts. (Doddapaneni et al., 2023)

4.3 Task-Specific Architectures

Soft Voting Ensemble Approach For hate type classification (1A) and target group identification (1B), each base model is fine-tuned independently with task-specific classification heads. The final prediction is obtained through soft voting by averaging the prediction probabilities from all three models:

$$P_{\text{ensemble}}(y|x) = \frac{1}{3}\sum_{i=1}^{3} P_i(y|x) \quad (1)$$

where $P_i(y|x)$ represents the probability distribution from model i for input x.

Weighted Voting Multitask Approach For the joint prediction task, each base model is fine-tuned independently as a multitask learner with three classification heads: hate type (6 classes), severity (3 classes), and target group (4 classes). The individual multitask objective function combines cross-entropy losses from all tasks:

$$\mathcal{L}_{\text{mtl}} = \alpha\mathcal{L}_{\text{type}} + \beta\mathcal{L}_{\text{severity}} + \gamma\mathcal{L}_{\text{target}} \quad (2)$$

The final ensemble prediction uses weighted voting based on individual development set performance:

$$\begin{aligned} P_{\text{final}}(y \mid x) = {} & 0.5\, P_{\text{MuRIL}}(y \mid x) \\ & + 0.3\, P_{\text{BanglaBERT}}(y \mid x) \\ & + 0.2\, P_{\text{IndicBERTv2}}(y \mid x) \end{aligned} \quad (3)$$

where the weights reflect each model's individual performance ranking on the development set.

4.4 Training Configuration

All models are trained using the `AdamW` optimizer with identical hyperparameters across all subtasks: learning rate of 2×10^{-5}, batch size of 16, and 3 training epochs. For subtasks 1A and 1B, each model in the soft voting ensemble is trained independently with task-specific objectives. For subtask 1C, each model is trained independently as a multitask learner before combining predictions through weighted voting.

5 Results and Analysis

5.1 Performance Against Baselines

Table 1 reports system performance across all three subtasks, evaluated with the official metrics (micro-f_1 for 1A and 1B, and weighted micro-f_1 for 1C). The organizers also released three baselines-Random, Majority, and n-gram, which obtained 16.38, 56.38, and 60.20% on 1A; 20.43, 59.74, and 62.09% on 1B; and 23.04, 60.72, and 63.05% on 1C, respectively.

Model	Micro-f_1		W-Micro-f_1
	1A	**1B**	**1C**
BiLSTM (GV)	69.39	64.49	–
BiLSTM (FT)	68.67	62.74	–
BiGRU (GV)	69.35	68.75	–
BiGRU (FT)	66.92	68.75	–
MRL	74.00	74.60	74.79
BNB	71.00	73.61	73.35
INB	74.00	73.17	71.22
SV (MRL+BNB+INB)	**75.72**	**74.96**	74.08
HV (MRL+BNB+INB)	72.53	74.16	73.42
WV (MRL+BNB+INB)	74.16	74.56	**75.12**

Table 1: Performance of employed models across all subtasks. A dash (–) denotes that the model was not evaluated for the corresponding subtask. Abbreviations: GV=GloVe, FT=FastText, MRL=MuRIL, BNB=BanglaBERT, INB=IndicBERTv2, SV=Soft Voting, HV=Hard Voting, WV=Weighted Voting.

Our systems substantially outperform these baselines. For example, BanglaBERT attains 71.00% on subtask 1A, more than 10 points higher than the n-gram baseline. The best-performing systems are ensembles: soft voting achieves 75.72% (1A) and 74.96% (1B), while weighted voting performs best on 1C (75.12%).

5.2 RNNs vs. Transformers vs. Ensembles

On subtasks 1A and 1B, BiLSTM and BiGRU models with static embeddings (GloVe, FastText) yield scores in the mid-60s, significantly lower than transformer-based models. Based on these results, we did not extend RNNs to Subtask 1C.

Among transformers, MuRIL and IndicBERTv2 perform comparably and generally surpass BanglaBERT, reflecting their larger multilingual training. However, ensemble methods consistently outperform individual models. Soft voting is most effective in 1A and 1B, whereas weighted voting excels in 1C, suggesting that the optimal ensemble strategy depends on task complexity.

5.3 Error Analysis

Confusion Patterns. In subtask 1A (hate type), the system frequently confuses *Abusive* with *None*, and *Political Hate* with *Profane*. Minority categories such as *Sexism* and *Religious Hate* suffer

from very low recall due to class imbalance. In subtask 1B (target group), *Organization* is often predicted as *None*, and *Society* is confused with *Individual*. Subtask 1C inherits these trends, with additional difficulty distinguishing between *Mild* and *Severe* hate severity. Confusion matrices and error examples are provided in Appendix B.

Qualitative Errors. Representative examples highlight typical misclassifications. For instance,

- **Implicit or sarcastic hate:** ভাইয়া আপনি অভিনেতা হইয়েন না না হলে সবাই বাচ্চা চাইবে *(Brother, don't act like an actor, otherwise everyone will demand children from you)* was annotated as *Abusive, Individual*, but all systems predicted *None*.
- **Subtask complementarity:** এটা রে আওয়ামী লীগ ভোট দিবে কেনমাথায় আসে না *(Why would anyone vote for Awami League, I cannot imagine)* was misclassified in Subtask 1B (*Organization*) but correctly resolved in the multitask model.
- **Trade-offs in multitask learning:** হারামি মোসাদ্দেক দেখি এইখানে *(That bastard Mosaddek, I see him here)* was correctly identified as *Profane, Individual* in single-task models, but misclassified as *Abusive, Individual* in multitask.

Effect of Label Imbalance. The dataset distribution reveals severe class imbalance across subtasks (see Appendix A). In Subtask 1A, categories such as *Sexism* (only 122 training instances) and *Religious Hate* (676 instances) are underrepresented, explaining their very low recall in our experiments. Similarly, in Subtask 1B, minority classes like *Community* and *Society* are frequently misclassified as *None* or *Individual*. For Subtask 1C, the dominance of the *Little to None* category in severity prediction makes it challenging for models to correctly identify *Mild* and *Severe* hate. These imbalances highlight the need for data augmentation and re-weighting strategies in future work.

5.4 Summary of Findings

Our analysis shows that (i) transformer ensembles consistently outperform single models and RNN baselines, (ii) multitask learning captures complementary signals across hate type, severity, and target group, though sometimes introducing inconsistencies, and (iii) errors stem primarily from subtle linguistic cues, overlapping class boundaries, and severe class imbalance. Together, these findings validate the complementary strengths of ensemble and multitask strategies for Bangla hate speech identification.

Official Shared Task Results. On the blind test set used for leaderboard evaluation, our submissions achieved 72.75% Micro-f1 in Subtask 1A (9th), 72.69% in Subtask 1B (10th), and 72.62% Weighted Micro-f1 in Subtask 1C (7th). These results confirm the competitiveness of our ensemble and multitask strategies in a challenging shared-task setting.

6 Conclusion

In this paper, we presented our systems for the BLP 2025 Shared Task 1 on Bangla hate speech identification. We explored transformer-based ensembles for subtasks 1A and 1B, and designed an efficient yet competitive multitask learning framework for subtask 1C. Our models achieved competitive performance across all subtasks, demonstrating the viability of both ensemble strategies and multitask learning in a low-resource setting like Bangla. The analysis further revealed challenges such as class imbalance and the difficulty of modeling underrepresented categories (e.g., *Sexism*, *Religious Hate*). For future work, we aim to explore data augmentation, cross-lingual transfer, and more robust multitask architectures to improve fine-grained hate speech detection in Bangla and extend these approaches to other low-resource languages.

7 Limitations

While our system demonstrates competitive performance on fine-grained hate speech detection, we acknowledge certain limitations. Our ensemble approaches require training and inference across multiple transformer models, increasing computational overhead compared to single-model solutions. The fixed weighting strategy for subtask 1C, while empirically determined from development performance, may benefit from more sophisticated dynamic weighting mechanisms. Additionally, our evaluation focuses on the shared task dataset, and broader cross-domain validation would strengthen the generalizability claims of our ensemble strategies.

References

Md Rabiul Awal, Rui Cao, Roy Ka-Wei Lee, and Sandra Mitrović. 2021. Angrybert: Joint learning target and emotion for hate speech detection. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 701–713. Springer.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States. Association for Computational Linguistics.

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.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

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.

Eftekhar Hossain, Omar Sharif, Mohammed Moshiul Hoque, and Sarah Masud Preum. 2024. Align before attend: Aligning visual and textual features for multimodal hateful content detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 162–174.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, and 1 others. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Shaina Raza and Veronica Chatrath. 2024. Harmonynet: navigating hate speech detection. *Natural Language Processing Journal*, 8:100098.

Sourav Saha, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. Multsea: Aspect-based sentiment analysis using multitask learning from bengali texts. In *2024 27th International Conference on Computer and Information Technology (ICCIT)*, pages 25–31. IEEE.

Sourav Saha, Jahedul Alam Junaed, Maryam Saleki, Arnab Sen Sharma, Mohammad Rashidujjaman Rifat, Mohamed Rahouti, Syed Ishtiaque Ahmed, Nabeel Mohammed, and Mohammad Ruhul Amin. 2023. Vio-lens: A novel dataset of annotated social network posts leading to different forms of communal violence and its evaluation. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 72–84.

A Datasets Analysis

A.1 Label-wise Distribution

Subtask	Label	Train	Dev	Test
1A (Type)	None	19,954	1,447	5,751
	Abusive	8,212	549	2,312
	Political Hate	4,227	283	1,220
	Profane	2,331	185	709
	Religious Hate	676	40	179
	Sexism	122	8	29
1B (Target)	None	21,190	1,528	6,093
	Individual	5,646	391	1,571
	Organization	3,846	292	1,152
	Community	2,635	159	759
	Society	2,205	142	625
1C (Type)	None	19,954	1,447	5,751
	Abusive	8,212	549	2,312
	Political Hate	4,227	283	1,220
	Profane	2,331	185	709
	Religious Hate	676	40	179
	Sexism	122	8	29
1C (Severity)	Little to None	23,489	1,714	6,737
	Mild	6,853	426	2,001
	Severe	5,180	372	1,462
1C (Target)	None	21,190	1,528	6,093
	Individual	5,646	391	1,571
	Organization	3,846	292	1,152
	Community	2,635	159	759
	Society	2,205	142	625

Table 2: Label distributions across Train, Dev, and Test splits for all subtasks.

Table 2 presents the detailed label-wise distributions across train, development, and test sets for all subtasks. These statistics highlight the inherent class imbalance, such as the small number of *Sexism* and *Religious Hate* instances in Subtask 1A, which makes the task more challenging.

B Detailed Error Analysis

We include confusion matrices for all three subtasks to illustrate misclassification patterns, along with representative failure cases that highlight challenges such as class imbalance, and subtle contextual cues.

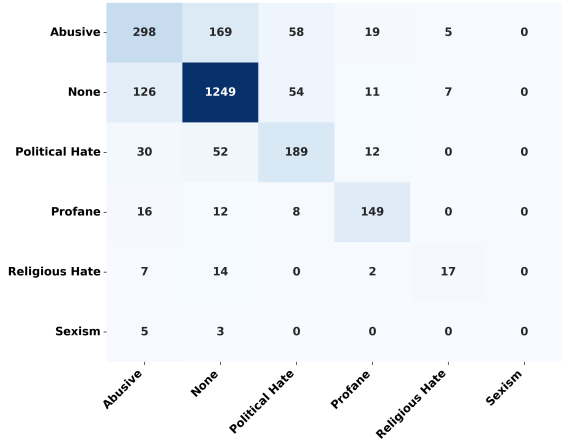


Figure 1: Confusion Matrix for Subtask 1A

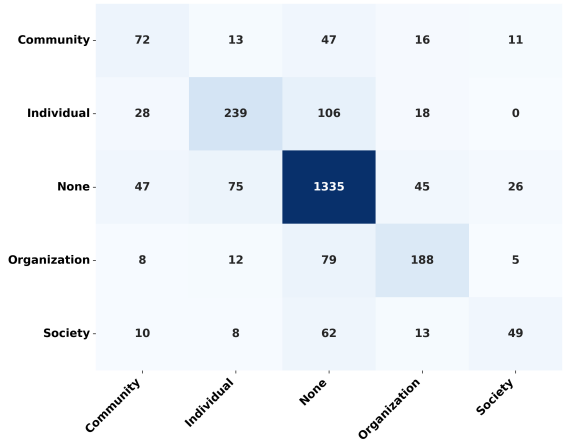
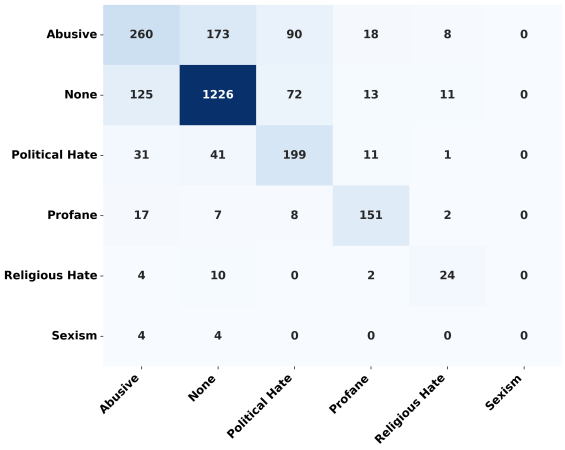
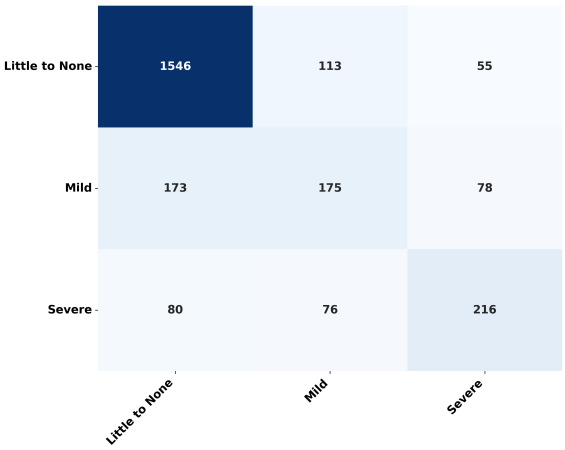


Figure 2: Confusion Matrix for Subtask 1B

Since all three subtasks were derived from the same dataset, we analyze errors using a shared set of representative examples. This allows us to highlight systematic issues across subtasks in a more consistent manner. Instead of comparing systems directly against each other, we focus on cases where each subtask fails, succeeds, or faces sys-



(a) Hate Type



(b) Hate Severity

(c) To Whom

Figure 3: Confusion matrices for subtask 1C

tematic challenges. Wrong predictions are highlighted in blue.

B.1 Subtask 1A: Hate Type Classification

Table 3 shows cases where Subtask 1A (type classification) fails. We observe frequent confusion between *Abusive* and *Profane*, as well as underprediction of subtle *Political Hate*. These errors often arise from short texts or figurative language.

Text	Gold Annotation	1A Prediction
সয়তানর খালাম্মা কয়কি *(What did the devil's aunt say?)*	Abusive	None
নির্বাচক মণ্ডলীর দের কে খেলানো দরকার হালারা *(The election committee members should be made to play, idiots.)*	Political Hate	Profane

Table 3: Examples where Subtask 1A (type) failed. Wrong predictions are in blue.

B.2 Subtask 1B: Target Group Classification

Table 4 presents errors from Subtask 1B (target group identification). The main challenge lies in distinguishing *Organization* vs. *Community*, and in cases where hate is implied but the target is indirect.

Text	Gold Annotation	1B Prediction
তোমাদের রাজনীতি হাস্যকর *(Your politics is ridiculous.)*	Organization	Community
ভালো থাকলে কাপড় খুলে বের হস ... জাহান্নামে গেলি *(If you're fine, strip off your clothes ... went to hell.)*	Community	Individual

Table 4: Examples where Subtask 1B (target) failed. Wrong predictions are in blue.

B.3 Subtask 1C: Multitask Classification

Table 5 highlights errors in Subtask 1C (multitask). Although multitask modeling captures interdependencies between type, severity, and target, we observe systematic errors such as over-predicting *Severe*, mismatched targets, and type drift.

B.4 Cases Where All Subtasks Fail

Finally, Table 6 shows difficult examples where all systems fail. These include sarcasm, implicit hate, or ambiguous targets, which remain challenging for current transformer models.

Text	Gold Annotation	1C Prediction
গোয়া মারা দিয়ে আছে বাংলাদেশ মাদারচোদ নিউজ করে সালার পুত পাম দেস *(Bangladesh is ruined motherfucker making news and buttering them up.)*	Profane / Mild / Organization	Profane / Severe / Individual
হারামি মোসাদ্দেক দেখি এইখানে *(That bastard Mosaddek, I see him here.)*	Profane / Little to None / Individual	Abusive / Severe / Individual

Table 5: Examples where Subtask 1C (multitask) failed. Wrong predictions are in blue.

Text	Gold Annotation	System Predictions
দনের নিবাচন ফাইজলামি। *(Today's election is a farce.)*	Profane / Mild / None	None / None / None
তোমরা শুধু প্রতিবাদে জানাতে পারবে ... জনগণের টাকা দিয়ে খুজতেছে জনগণকে রক্ষা করার জন্য *(You will only be able to protest ... wasting people's money pretending to protect them.)*	Abusive / Little to None / Community	Political Hate / None / None

Table 6: Challenging examples where all subtasks failed. Wrong predictions are in blue.

C Additional Training Details

The training configuration used in our experiments is summarized in Table 7.

Parameter	Value
Optimizer	AdamW
Learning rate	2×10^{-5}
Batch size	16
Epochs	3
Max sequence length	128

Table 7: Training hyperparameters used across all transformer models.

D Reproducibility Note

All experiments were conducted on a single NVIDIA RTX 3090 GPU with 24 GB VRAM. We used the PyTorch[3] deep learning framework together with the HuggingFace[4] Transformers library. All hyperparameters are listed in Table 7. Random seeds were fixed across runs for consistency, although minor variations in results may occur due to non-deterministic GPU operations.

[3] https://pytorch.org/
[4] https://huggingface.co/