Catalyst at BLP-2025 Task 1: Transformer Ensembles and Multi-task Learning Approaches for Bangla Hate Speech Detection

Nahid Hasan
Department of Computer Science and Engineering
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh
nahidhasan2003131@gmail.com

Abstract

We present a compact, cost-efficient system for the BLP-2025 Bangla Multi-task Hate Speech Identification Task 1, which requires fine-grained predictions across three dimensions: type, target, and severity. Our method pairs strong multilingual transformer encoders with two lightweight strategies: task-appropriate ensembling to stabilize decisions across seeds and backbones, and a multi-task head that shares representations while tailoring outputs to each subtask. As Catalyst, we ranked **7th** on Subtask 1A with micro-F1 **73.05**, **8th** on Subtask 1B with **72.79**, and **10th** on Subtask 1C with **72.40**. Despite minimal engineering, careful model selection and straightforward combination rules yield competitive performance and more reliable behavior on minority labels. Ablations show consistent robustness gains from ensembling, while the multi-task head reduces cross-dimension inconsistencies. Error analysis highlights persistent challenges with code-mixed slang, implicit hate, and target ambiguity, motivating domain-adaptive pretraining and improved normalization.

1 Introduction

The fast increase in the number of social media platforms has resulted in a growing concern regarding the harmful online content, especially hate speech in under-resourced languages such as Bangla. Informal language, code-mixing, and culturally sensitive hate utterances make it difficult to detect such content. Although the results of multilingual transformers and domain-adaptive pretraining are promising, they are vulnerable to major code-mixing and distributional drift (Sharif et al., 2021; Caselli et al., 2021).

The Bangla Multi-task Hate Speech Identification shared task (Hasan et al., 2025b) attempts to resolve these issues by proposing a complex framework that does not just deal in binary classification. It lays stress on the practical moderation requirements and promotes systems, which manage delicate phenomena by making organized forecasts along associated dimensions. Our work fills the gaps left by traditional hate speech detection which typically uses binary classification that does not provide the subtlety needed to do content moderation effectively. The common task helps develop the field because it involves systems conducting fine-grained analysis on three dimensions: type of hate, target, and level of severity. In the case of BLP-2025 Task 1, we trained a system that is a combination of transformer-based models and ensemble methods and multi-task learning. We experimented with several multilingual language models and developed simple yet efficient strategies for each subtask, leveraging both cross-lingual transfer and effective ensembling techniques.. Our method shows that basic combinations and good models can be highly effective with simple feature engineering, which confirms results that well-planned model combinations can be more reliable than multicomponent models with more advanced feature engineering designs (Plaza-Del-Arco et al., 2021; Saha et al., 2021).

The main contributions of our work include:

- An extensive analysis of four multilingual hate speech transformer models in Bangla.
- Good ensemble strategies that enhance performance by means of model combination.
- A multi-task learning system that deals with several classification goals at once.
- Competitive outcomes in all three subtasks that included practical and efficient solutions.

Our findings indicate that carefully designed transformer architectures, paired with suitable training strategies, have the ability to deal with the challenges of hate speech detection in Bangla social

Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 468–475

media whilst retaining enough simplicity to be deployable in practice.

2 Related Work

Transformer-based encoders have now become the standard method of abusive and hate speech detection, particularly in Indic code-mixed environments where multilingual models are significantly more effective than classical ML and RNN models of abusive speech detection (Sharif et al., 2021). Ensembling also increases stability and precision, and genetic-algorithm-weighted mixtures of transformer runs achieve the top scores in Dravidian-LangTech 2021 and similar analogous score increases of similar magnitude are reported for Arabic (Saha et al., 2021; de Paula et al., 2025). Multi-task learning that simultaneously predicts hate or aggression and affective cues like sentiment and emotion has the added advantage of sharing a transformer backbone but having task-specific heads (Plaza-Del-Arco et al., 2021). Domain-adaptive pretraining is also significant: HateBERT, which is additionally trained on abuse-rich data on Reddit, performs better and has stronger cross dataset portability than vanilla BERT (Caselli et al., 2021). Finally, multilingual transformers provide strong cross-lingual transfer, motivating the use of few-shot generalization and transfer-aware training to better handle Bangla and code-mixed text (Ni et al., 2020).

3 Task Description

This study introduces the Bangla Multi-Task Hate Speech Identification shared task[1], which represents a substantial step beyond the traditional binary-classification paradigm toward a more fine-grained, three-dimensional detection framework. In this shared task, there are three subtasks.

- **Subtask 1A:** Given a Bangla text collected from YouTube comments, classify the text as abusive, sexism, religious hate, political hate, profane, or none.
- **Subtask 1B:** Using a Bangla text gathered from the YouTube comments, classify the hate directed towards individuals, organizations, communities, or society.
- **Subtask 1C:** It is a multi-task arrangement. Based on a Bangla text acquired in YouTube comments, classify it by type of hate, severity, and targeted group.

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

Subtask	Label	Example
1A	Political Hate	আওয়ামী লীগের সন্ত্রাসী কবে দরবেন এই সাহস আপনাদের নাই
1A	Abusive	শালায় এক নাম্বার বাটপার
1B	Organization	খানকির পোলারা হেডার নিউজ দেস নিউজের মাঝখানে দুইবার দেস এড
1B	None	আমার মতো কমেন্ট পরতে ভালোবাসো কারা
1C	Political Hate, Severe, Organization	আপনারা জুতা খাওয়া পাটি

Table 1: Examples by subtask with labels.

3.1 Dataset Description

The dataset (Hasan et al., 2025a) utilized to carry out this task consists of YouTube comments about socially sensitive matters within the Bangla-speaking region. The remarks are in Bangla and show the informal, noisy nature of user-generated content, such as spelling variability, code-mixing and colloquialism. All samples are labeled with the three classification objectives that are related to the subtasks. In the real world hate speech is messy. This is represented in the dataset by its non-uniform categories and a thin line between aggressive political expression and actual hate, and it represents a real test of the subtlety and expertise of a model. Table 1 demonstrates representative examples from the dataset, while Table 2 provides statistical overview.

Subtask	Train	Dev-Test	Test
1A: Hate Type	35,522	2,512	10,200
1B: Target	35,522	2,512	10,200
1C: Multi-task	35,522	2,512	10,200

Table 2: Dataset statistics across all three subtasks.

4 System Description

The proposed system of the hate speech identification shared task utilizes both single-task classification with transformer-based ensembles and multi-task learning with joint prediction. Our approach is a fine-tuning process, in which multilingual transformers trained on general data are fine-tuned to the specific data and ensembled using ensemble techniques. Figure 1 shows the overall system architecture.

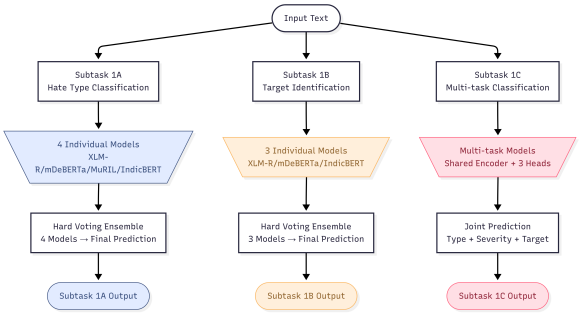


Figure 1: Overall system architecture showing separate pipelines for each subtask with ensemble strategies for 1A/1B and multi-task learning for 1C.

4.1 Models

We employed four cross-lingual Transformer encoders to take advantage of cross-lingual transfer to Bangla and with coverage of complementary architectural options and pretraining corpora.

- **XLM-RoBERTa-large** (Conneau et al., 2020) is chosen due to its pretraining on 100 languages allowing a high cross-lingual transfer and an adequate coverage of the subwords in Bangla.
- **microsoft/mdeberta-v3-base** (He et al., 2023) is selected because it implements disentangled attention in an efficient backbone, providing an appropriate trade-off between accuracy and efficiency in our context.
- **google/muril-base-cased** (Khanuja et al., 2021) is included since it is specifically trained for Indian languages and is therefore well suited to regional scripts, code-mixing, and transliteration phenomena.
- **ai4bharat/IndicBERTv2-MLM-only** (Doddapaneni et al., 2023) is included since its tokenizer and vocabulary is optimized to Indic scripts, aiding in the capturing of South Asian morphological and orthographic aspects.

All fine-tuning experiments for the subtasks were conducted on Google Colab using single NVIDIA T4 GPU. The detailed hyperparameters and training configurations are provided in Appendix A, and the complete source code is available in our public repository[2].

[2]https://github.com/nahid2003131/blp25_task1_catalyst

4.2 Subtask 1A: Hate Type Classification

Subtask 1A required assigning each Bangla YouTube comment to one of six categories, *abusive*, *sexism*, *religious hate*, *political hate*, *profane*, or *none*. We fine-tuned four multilingual encoders, *IndicBERTv2*, *XLM-RoBERTa-large*, *mDeBERTa-v3-base*, and *MuRIL-base*, and aggregated predictions via hard voting to stabilize decisions across architectures and seeds. For an instance x_i with model set $\mathcal{M}$ and label set $\mathcal{Y}$, the ensemble prediction was

$$\hat{y}_i = \arg\max_{y \in \mathcal{Y}} \sum_{m \in \mathcal{M}} \mathbf{1}\left\{ y = \hat{y}_i^{(m)} \right\}.$$

This rule yielded a consistent improvement over the strongest single model (Table 3), indicating complementary inductive biases among encoders and increased robustness to label ambiguity.

4.3 Subtask 1B: Target Identification

Subtask 1B identified the target of the hateful content.We fine-tuned three encoders, *XLM-RoBERTa-large*, *mDeBERTa-v3-base*, and *IndicBERTv2*, and applied the same hard voting rule as in Subtask 1A, with $\mathcal{M}$ restricted to these three models. The ensemble outperformed the strongest individual model (Table 3), improving reliability on distinctions among *individual*, *organization*, *community*, and *society* targets.

4.4 Subtask 1C: Multi-task Classification

Our multi-task architecture used a shared encoder with dedicated classification heads for Hate Type, Severity, and Target prediction.We evaluated *IndicBERTv2*, *XLM-RoBERTa-base*, and *mDeBERTa-v3-base* as the shared backbone. Figure 2 illustrates the shared-encoder setup and the three task-specific heads. Each subtask produced an independent cross-entropy loss $\mathcal{L}$, and the overall objective was

$$\mathcal{L}_{total} = \mathcal{L}_{type} + \mathcal{L}_{severity} + \mathcal{L}_{target}.$$

During backpropagation, gradients from all subtasks jointly updated the shared encoder, enabling the model to learn common linguistic and semantic representations across hate-related dimensions. We formed the sentence representation from the encoder's [CLS] token with dropout before the task heads. No explicit loss weighting was applied. Equal contributions from all subtasks yielded stable convergence without noticeable task interfer-

ence. The *IndicBERTv2* variant performed best and was selected as the final submission.

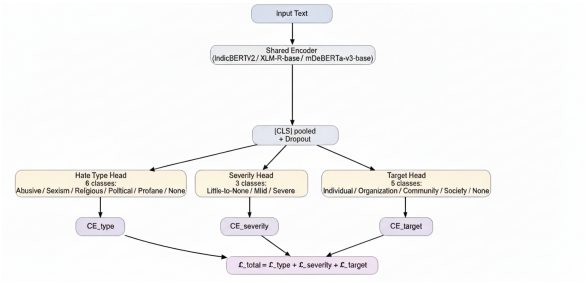


Figure 2: Multi-task learning architecture for Subtask 1C.

System	Micro F1 (Dev)	Micro F1 (Test)	Rank
1A: Hate Type			
XLM-R-large	73.77	71.99	—
MuRIL	73.89	71.25	—
mDeBERTa	72.77	71.58	—
IndicBERTv2	72.73	71.11	—
Ensemble above all	**75.72**	**73.05**	**7/37**
1B: Target			
IndicBERTv2	73.37	72.05	—
mDeBERTa	73.01	71.89	—
XLM-R-large	73.01	71.75	—
Ensemble above all	**74.56**	**72.79**	**8/24**
1C: Multi-task			
XLM-R-base	72.50	71.14	—
mDeBERTa	67.81	69.09	—
IndicBERTv2	**74.59**	**72.40**	**10/21**

Table 3: Performance across subtasks measured using official evaluation metrics (Micro-F1). Bold indicates final submission.

5 Results and Findings

The micro-F1 scores on development-test and official test sets are summarized in Table 3 with respect to all three subtasks. We compare individual transformer model performances with our ensemble approaches for Subtasks 1A and 1B, and evaluate multi-task learning models for Subtask 1C.The final column displays our official rankings on the leaderboard which consisted of 37 teams in Subtask 1A, 24 teams in Subtask 1B, and 21 teams in Subtask 1C.

5.1 Key Observations

- Ensemble methods have a systematic positive effect on single-task performance. Hard-voting ensemble strategy showed considerable improvement on both single-task subtasks. In Subtask 1A, when four models were combined, the resulting test micro-F1 was 73.05, which is a 1.06-point improvement on the highest-performing single model (XLM-RoBERTa-large: 71.99). Likewise, in Subtask 1B, the three-model ensemble scored 72.79 micro-F1, which was 0.74 points higher than the highest single model (IndicBERTv2: 72.05).

- IndicBERTv2 is an efficient multi-task learner, surpassing peers in a wide variety of goals. IndicBERTv2 was the top performer in the multi-task context of Subtask 1C with 72.40 micro-F1, which is significantly higher than the performance of XLM-RoBERTa-base(71.14) and mDeBERTa-v3-base(69.09).

- XLM-RoBERTa-large demonstrates good single-model performance. Although not explicitly trained on Indic languages, XLM-RoBERTa-large achieved competitive results as a general model, with a micro-F1 of 71.99 on Subtask 1A and 71.75 on Subtask 1B. This illustrates the high cross-lingual transfer of large-scale multilingual models.

- Transformer-based methods are far better than their traditional baselines. All the transformer models significantly surpassed the performance of n-gram baselines across the subtasks, and the usefulness of pre-trained contextual representations to detect hate speech in Bangla.

5.2 Error Analysis

Analysis of IndicBERTv2 multi-task model indicates that there are specific patterns of errors:

- **Hate Type Confusion**: Severe overlap between *Abusive* and *Political Hate* indicating the possible overlap of linguistic indicators in the Bangla social media discourse.

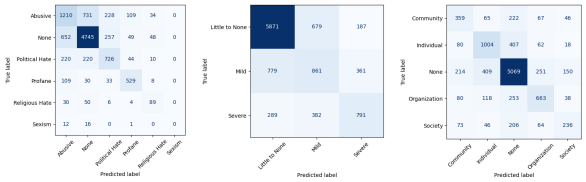
- **Severity Calibration Challenges**: Most frequent mistake (26.25%) with regular *Mild* and *Little to None* confusion which shows that intensity measurement is subjective.

- **Target Ambiguity**: Notable confusion between *Organization*, *Individual*, and *Community* targets, reflecting complex entity references in informal language.

Figure 3 provides quantitative evidence for these error patterns, with detailed analysis revealing specific model weaknesses.

Type	Text	Actual	Pred.
HT	বিদ্যুৎ জ্বালানি খাতে আওয়ামী লীগের আমলে সবচেয়ে বেশি দুর্নীতি...	Abusive	Political Hate
HT	ভারতীয় দালাল সময় টিভিকে বয়কট করুন...	Political Hate	Abusive
HT	কত সুন্দর করে বড় ভাই বলছে শাট টা কিনে নিছ ভাইয়া হালার পু হাল...	Abusive	Profane
SC	ইজরায়েলের বিচার হওয়া উচিৎ...	Mild	Little to None
SC	হেন কাপ পুলিশের মারে... বিচার হবে কি...	Little to None	Mild
SC	কিছুই হবে না বদমাশ ইসরাইল...	Severe	Mild
TA	হাসিনা তাহলে বি এন পিকে ভয় পেয়েছে...	Organization	Individual
TA	আমলিগরা হচ্ছে ছুর... তিস্তা নদির পান...	Organization	Community
TA	আমি হালায় টাকার অভাবে বিয়ে করতে পারতেছি না...	Individual	Community

Table 4: Qualitative error examples with actual vs. predicted labels. HT = Hate Type, SC = Severity Calibration, TA = Target Ambiguity.



(a) Hate Type (b) Severity (c) Target

Figure 3: Confusion matrices on the test set.

Hate Type Confusion: The most frequent confusion occurs between *Abusive* and *Political Hate* categories, with **448** total misclassifications (228 Abusive→Political, 220 Political→Abusive). Additionally, substantial non-*None* content is misclassified as *None*, including **731** Abusive, **220** Political Hate, and **30** Profane instances.

Severity Calibration: Errors concentrate at the low-moderate severity boundary, with **679** *Little-to-None*→*Mild* and **779** *Mild*→*Little-to-None* misclassifications, indicating significant threshold uncertainty.

Target Ambiguity: Notable confusion exists among *Organization*, *Individual*, and *Community* targets. Many targeted cases are incorrectly predicted as *None*, affecting **407** Individual, **253** Organization, **222** Community, and **206** Society instances.

These systematic errors highlight key challenges in fine-grained hate speech analysis for low-resource languages and outline clear priorities for future model refinement.

5.3 Cost Analysis

In real-world scenarios, particularly for low-resource languages, the computational cost and environmental impact of a model are as critical as its predictive performance. While billion-parameter models such as **mT5-Large** (Xue et al., 2021), **mT5-XL** (Xue et al., 2021), and **BLOOMZ-1B7** (Muennighoff et al., 2022) achieve strong results, their prohibitive resource demands limit accessibility and scalability. Our work instead emphasizes parameter efficiency, showing that carefully selected compact encoders can offer competitive performance at a fraction of the cost. Table 5 presents the verified parameter counts of our encoders compared to widely used large multilingual models.

Model	Parameters
mT5-Large	1.2B
mT5-XL	3.7B
BLOOMZ-1B7	1.7B
XLM-RoBERTa-large	559M
mDeBERTa-v3-base	276M
MuRIL-base-cased	177M
IndicBERTv2-MLM-only	278M

Table 5: Parameter count comparison highlighting the large gap between our compact encoders (bottom group) and billion-scale models (top group).

Our largest encoder, XLM-RoBERTa-large (559M), is under half the size of the smallest billion-scale model, while others are an order smaller (under 280M). Following scaling trends (Kaplan et al., 2020; Tay et al., 2020), these models use roughly 3–5x less GPU memory and train several times faster, offering a strong balance between accuracy, efficiency, and sustainability for multilingual hate speech detection.

6 Conclusion

The paper introduces a low-cost method to detect hate speech in Bangla with transformer ensembles and multi-task learning. Our findings indicate that by employing well-crafted ensemble techniques to multilingual transformers we can achieve competitive results relative to large individual models, at the same time being computationally efficient. The multi-task learning paradigm proved to be especially successful in capitalizing on common representations among related classification problems. We believe that an increase in the size and diversity of the dataset will result in better performance of our approach in many respects, thus expanding the range of potential uses to other text classification tasks in Bangla.

Limitations

This study is subject to several limitations. The primary challenge was the scarcity of high-quality, diverse Bangla hate speech data, which is exacerbated by the sensitive nature of the content and the limited number of publicly available sources. While the dataset exhibits class imbalance, preliminary experiments with class weighting and data sampling techniques did not yield significant performance improvements. Consequently, no explicit imbalance-handling strategy was implemented in the final models, which may adversely affect performance on minority classes. Additionally, the subjective nature of hate speech annotation, particularly for severity calibration, likely introduced labeling inconsistencies that are not fully captured by our error analysis.

References

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Angel Felipe Magnossão de Paula, Imene Bensalem, Paolo Rosso, and Wajdi Zaghouani. 2025. Transformers and ensemble methods: A solution for hate speech detection in arabic languages. *Preprint*, arXiv:2303.09823.

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.

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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. MuRIL: Multilingual representations for indian languages. *Preprint*, arXiv:2103.10730.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, and 1 others. 2022. Crosslingual generalization through multitask finetuning.

Jian Ni, Taesun Moon, Parul Awasthy, and Radu Florian. 2020. Cross-lingual relation extraction with transformers. *Preprint*, arXiv:2010.08652.

Flor Miriam Plaza-Del-Arco, M. Dolores Molina-González, L. Alfonso Ureña-López, and María Teresa Martín-Valdivia. 2021. A multi-task learning approach to hate speech detection leveraging sentiment analysis. *IEEE Access*, 9:112478–112489.

Debjoy Saha, Naman Paharia, Debajit Chakraborty, Punyajoy Saha, and Animesh Mukherjee. 2021. Hate-alert@DravidianLangTech-EACL2021: Ensembling strategies for transformer-based offensive language detection. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 270–276, Kyiv. Association for Computational Linguistics.

Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2021. NLP-CUET@DravidianLangTech-EACL2021: Offensive language detection from multilingual code-mixed text using transformers. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 255–261, Kyiv. Association for Computational Linguistics.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2020. Efficient transformers: A survey. *arXiv preprint arXiv:2009.06732*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A Experimental Setup and Hyperparameters

All fine-tuning experiments were conducted on Google Colab using a single NVIDIA T4 GPU (16 GB VRAM). To ensure reproducibility, a fixed random seed of 42 was used across all runs. Common training configurations included mixed-precision training (fp16), a maximum sequence length of 512, and the AdamW optimizer. Model-specific hyperparameters, determined through a limited search on a held-out validation set, are detailed in the following subsections.

Subtask 1A: Hate Type Classification

For this subtask, all models shared a common base configuration: they were trained for 4 epochs with a linear learning rate scheduler, a weight decay of 0.01, a warmup ratio of 0.01, and a maximum sequence length of 512. The effective batch size was standardized to 16, achieved either directly or through gradient accumulation. The distinct learning rates and gradient accumulation steps for each model are provided in Table 6.

Model	Learning Rate	BS	Grad. Ac.
XLM-RoBERTa-large	1.5×10^{-5}	4	4
mDeBERTa-v3-base	2×10^{-5}	16	1
IndicBERTv2-MLM-only	2×10^{-5}	16	1
MuRIL-base	2×10^{-5}	16	1

Table 6: Hyperparameters for Subtask 1A. Abbreviations: BS = Batch Size, Grad. Accum. = Gradient Accumulation Steps.

Subtask 1B: Target Identification

The setup for Subtask 1B was similar, employing a linear scheduler, weight decay of 0.01, and an effective batch size of 16. However, the number of epochs and warmup ratio were tuned specifically for this task. The model-specific learning rates, epochs, and warmup ratios are summarized in Table 7.

Model	LR	Epochs	WR
XLM-RoBERTa-large	1.5×10^{-5}	4	0.01
mDeBERTa-v3-base	2×10^{-5}	3	0.1
IndicBERTv2-MLM-only	2×10^{-5}	3	0.1

Table 7: Fine-tuning hyperparameters for Subtask 1B models. Column abbreviations: LR = Learning Rate, WR = Warmup Ratio.

Model	LR	Epochs	WR	Sched.
IndicBERTv2-MLM-only	2×10^{-5}	4	0.01	Cosine
XLM-RoBERTa-base	2×10^{-5}	3	—	Linear
mDeBERTa-v3-base	2×10^{-5}	4	0.1	Cosine

Table 8: Fine-tuning hyperparameters for Subtask 1C models. Column abbreviations: LR = Learning Rate, WR = Warmup Ratio, Sched. = Scheduler.

Subtask 1C: Multi-task Classification

For the multi-task learning setup (Subtask 1C), we explored different learning rate schedulers. All models were trained with an effective batch size of 16 and a fixed random seed of 42. The complete hyperparameter set for each multi-task model is detailed in Table 8.

B Detailed Performance Metrics

This appendix provides complete performance evaluations for our final models across all subtasks. We report per class precision, recall, and F1 scores to supplement the main paper's Micro F1 results.

The official evaluation metric, Micro F1, is derived from global counts across all classes. Micro Precision is calculated as $\frac{\sum TP}{\sum TP + \sum FP}$, Micro Recall as $\frac{\sum TP}{\sum TP + \sum FN}$, and Micro F1 as their harmonic mean.

We also provide Macro F1 scores, computed as the arithmetic mean of per class F1 scores, to ensure balanced performance across all categories. These metrics collectively offer a comprehensive assessment of model effectiveness for multilingual hate speech detection.

Label	Precision	Recall	F1
Abusive	57.34	53.24	55.21
Political Hate	58.76	61.31	60.01
Profane	72.76	81.38	76.83
Religious Hate	48.89	49.16	49.03
Sexism	33.33	3.45	6.25
None	82.80	83.57	83.18
Macro Avg	58.98	55.35	55.09
Micro Avg	**73.05**	**73.05**	**73.05**

Table 9: Per-class precision, recall, and F1-scores for Subtask 1A (Hate Type).

-

Target Type (per-class)	Precision	Recall	F1
Community	44.29	45.98	45.12
Individual	64.57	64.04	64.30
Organization	61.59	59.29	60.42
Society	48.64	43.04	45.67
None	82.66	84.00	83.32
Macro Avg	60.35	59.27	59.77
Micro Avg	**72.79**	**72.79**	**72.79**

Table 10: Per-class precision, recall, and F1-scores for Subtask 1B (Target Type).

Hate Type (per-class)	Precision	Recall	F1
Abusive	54.19	52.34	53.25
None	81.92	82.51	82.21
Political Hate	58.08	59.51	58.79
Profane	71.88	74.61	73.22
Religious Hate	47.09	49.72	48.37
Sexism	0.00	0.00	0.00
Macro Avg	52.19	53.11	52.64
Micro Avg	**71.56**	**71.56**	**71.56**
Hate Severity (per-class)			
Little to None	84.61	87.15	85.86
Mild	44.80	43.03	43.89
Severe	59.07	54.10	56.48
Macro Avg	62.83	61.43	62.08
Micro Avg	**73.75**	**73.75**	**73.75**
To Whom (per-class)			
Community	44.54	47.30	45.88
Individual	61.14	63.91	62.50
None	82.33	83.19	82.76
Organization	59.89	57.55	58.70
Society	48.36	37.76	42.41
Macro Avg	59.25	57.94	58.45
Micro Avg	**71.87**	**71.87**	**71.87**
Overall Averages (across tasks)			
Micro Avg	**72.40**	**72.40**	**72.40**
Macro Avg	58.09	57.49	57.72

Table 11: Per-class results for Subtask 1C (multi-task model) across Hate Type, Hate Severity, and Target