Computational Story Lab at BLP-2025 Task 1: HateSense: A Multi-Task Learning Framework for Comprehensive Hate Speech Identification using LLMs

Tabia Tanzin Prama[1,2,3,5], **Christopher M. Danforth**[1,2,3,4], **Peter Sheridan Dodds**[1,2,3,5,6]
[1]Computational Story Lab, [2]Vermont Complex Systems Institute,
[3]Vermont Advanced Computing Center, [4]Department of Mathematics and Statistics,
[5]Department of Computer Science, University of Vermont, Burlington, VT 05405, USA
[6]Santa Fe Institute, 1399 Hyde Park Rd, Santa Fe, NM 87501, USA

Abstract

This paper describes HateSense, our multi-task learning framework for the BLP 2025 shared task 1 on Bangla hate speech identification. The task requires not only detecting hate speech but also classifying its type, target, and severity. HateSense integrates binary and multi-label classifiers using both encoder- and decoder-based large language models (LLMs). We experimented with pre-trained encoder models (Bert based models), and decoder models like GPT-4.0, LLaMA 3.1 8B, and Gemma-2 9B. To address challenges such as class imbalance and the linguistic complexity of Bangla, we employed techniques like focal loss and odds ratio preference optimization (ORPO). Experimental results demonstrated that the pre-trained encoders (BanglaBert) achieved state-of-the-art performance. Among different prompting strategies, chain-of-thought (CoT) combined with few-shot prompting proved most effective. Following the HateSense framework, our system attained competitive micro-F1 scores: 0.741 (Task 1A), 0.724 (Task 1B), and 0.7233 (Task 1C). These findings affirm the effectiveness of transformer-based architectures for Bangla hate speech detection and suggest promising avenues for multi-task learning in low-resource languages.

Warning: this paper contains content that may be offensive or upsetting

1 Introduction

The ever-expanding digital landscape, while promising to foster global connectivity and social cohesion, has simultaneously emerged as a breeding ground for hate speech (Castaño-Pulgarín et al., 2021). Hate speech is broadly defined as language that targets, attacks, or incites implicit or explicit hatred or violence against individuals or groups based on specific attributes such as physical appearance, religion, ethnic origin, or gender identity (Papcunová et al., 2021). Its pervasive presence poses severe risks, including the promotion of social division, deterioration of mental health, and escalation of violence (Sahoo et al., 2024). Inadequate moderation of such content further cultivates intolerance, amplifying its negative societal impacts (Hangartner et al., 2021). Addressing online hate speech requires moving beyond a simple "toxic vs. non-toxic" label toward a nuanced analysis that captures its type, severity, and target, enabling deeper insights into motives and potential consequences.

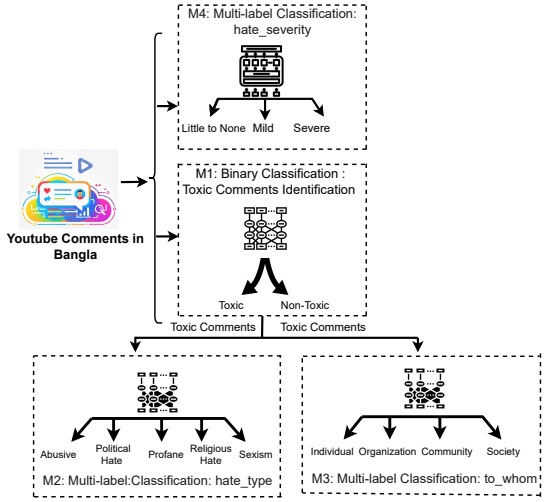


Figure 1: The proposed HateSense framework. It begins with M1 (binary hate speech detection), followed by M2 (multi-label classification of hate type, Task 1A) and M3 (multi-label classification of target, Task 1B). A separate model, M4, classifies hate speech severity. The combined outputs of M1–M4 form the results for Task 1C.

The shared task 1 focuses on Bangla multi-task hate speech identification. Our team, under the name Computational StoryLab and username ttprama, participated in the multi-task hate speech

Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 430–442

identification track. For this task, we proposed HateSense shows in Figure 1, a multi-task learning framework designed to classify Bangla texts into predefined categories of hate type, severity, and targeted group. Transformer-based architectures (Vaswani et al., 2017), such as BERT (Devlin et al., 2019), have revolutionized NLP tasks and consistently achieved state-of-the-art (SOTA) performance across benchmarks (Lan et al., 2019). Meanwhile, the recent surge in LLMs has established them as strong candidates for hate speech detection (Liu et al., 2019), particularly in zero-shot settings. However, while significant progress has been made for English and other high-resource languages, research on Bangla—a low-resource language—remains limited. Prior studies have explored Bangla hate speech detection (Jahan et al., 2019; Prama et al., 2025) and LLM-based methods (Shibli et al., 2022), but no prior work has addressed their intersection within a multi-task learning setting.

To address this research gap, our key contributions are as follows:

- We propose a multi-task learning framework, HateSense, which goes beyond binary detection to jointly predict the type, severity, and targeted group of hate speech.
- We establish several encoder-based baselines for each subtask, with encoders further fine-tuned on Bangla achieving state-of-the-art performance on our dataset.
- We investigate zero-shot(ZS), few-shot and COT prompting with state-of-the-art LLMs and introduce a novel translation-based prompting strategy, which outperforms existing methods on our dataset.

2 Task and Data

The dataset consists of Bangla YouTube comments (Hasan et al., 2025a), which serve as the input for all subtasks in Task 1 (Hasan et al., 2025b). The BLP Workshop offers three subtasks: Task 1A: categorizing the type of hate speech (abusive, sexism, religious hate, political hate, profane, or none), Task 1B: identifying the targeted group (individuals, organizations, communities, or society), and Task 1C: is a multi-task setup classifying the severity of hate speech (Little to None, Mild, or Severe), type of hate (Task 1A) and targeted group (Task 1B) in Bangla commentd. The label distributions for each subtask are presented in Tables 1, 2, and 3 respectively.

Label	Train	Dev	Test	Total
None	19954	2898	5751	28603
Abusive	8212	1113	2312	11637
Political Hate	4227	574	1220	6021
Profane	2331	342	709	3382
Religious Hate	676	78	179	933
Sexism	122	19	29	170

Table 1: Label distribution of type of hate speech.

Label	Train	Dev	Test	Total
None	21190	3064	6093	30347
Individual	5646	755	1571	7972
Organization	3846	584	1152	5582
Community	2635	338	759	3732
Society	2205	283	625	3113

Table 2: Label distribution of target group of hate speech.

Label	Train	Dev	Test	Total
Little to None	23489	3417	6737	33643
Mild	6853	909	2001	9763
Severe	5180	698	1462	7340

Table 3: Label distribution of severity of hate speech.

3 Methodology

As Task 1 follows a multi-task setup, we adopted the proposed HateSense framework (Figure 1). Tasks 1A and 1B were performed in two stages: first, binary classification of hate speech (M1), followed by multi-label classification of the predicted hate speech instances (M2 for Task 1A and M3 for Task 1B). Task 1C was addressed using a dedicated multi-label classification model (M4), combined with the outputs of M1, M2, and M3. Our baselines fall into two categories: (1) fine-tuning pre-trained language models (LMs) on our dataset, and (2) prompting LLMs with zero-shot (ZS), few-shot (FS), and chain-of-thought (COT) strategies. During evaluation, baseline models were trained on the training set, with the development set used to select the best-performing models. The selected models

were then re-trained on the combined training and development sets, and their predictions were submitted for final test set evaluation

3.1 LM Fine-tuning

For fine-tuning, we experiment with several Transformer encoder-based models, including BanglaBERT (Bhattacharjee et al., 2021), BanglaHateBERT (Jahan et al., 2022) , IndicBERT (Bhattacharyya et al., 2023), XLM-Roberta (Conneau et al., 2019), mDistilBERT (Sanh et al., 2019), and mBERT (Devlin et al., 2019). Each subtask is formulated as a multi-label classification problem by adding a classification head on top of the encoder. Given an input sentence

$$S = (w_1, w_2, \ldots, w_n),$$

it is tokenized and passed through a Transformer encoder f_θ, producing contextual representations:

$$H = \{h_1^l, h_2^l, \ldots, h_n^l\}, \quad l \in \{1, \ldots, L\}, \; h_i^l \in \mathbb{R}^d.$$

From the final layer, the hidden state of the special [CLS] token serves as a sentence-level representation:

$$h_{\text{CLS}} = h_1^L \in \mathbb{R}^d.$$

A dropout layer is applied:

$$\tilde{h}_{\text{CLS}} = \text{Dropout}(h_{\text{CLS}}, p = 0.3)$$

and the result is passed through a linear classifier:

$$z = W\tilde{h}_{\text{CLS}} + b, \quad W \in \mathbb{R}^{k \times d}, \; b \in \mathbb{R}^k,$$

where k is the number of labels. Probabilities are obtained via a sigmoid:

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}}.$$

The model is trained using Binary Cross-Entropy loss:

$$\mathcal{L} = -\frac{1}{k} \sum_{j=1}^{k} \Big[y_j \log(\hat{y}_j) + (1 - y_j) \log(1 - \hat{y}_j) \Big],$$

where $y_j \in \{0, 1\}$. We optimized with AdamW 4×10^{-5} learning rate) for 5 epochs using weighted loss to handle class imbalance, selecting the best checkpoint by validation performance.

3.2 Prompting Strategy

For decoder-only models, each subtask is formulated as a text generation problem, where the model is prompted to produce exactly one label from the predefined set of choices. We experiment with GPT-4.0 (Achiam et al., 2023), LLaMA 3.1 8B (Dubey et al., 2024), and Gemma-2 9B (Riviere et al., 2024). We design base prompts separately for binary classification and multi-label classification. We explore several prompting strategies like Zero-shot prompting (Radford et al., 2019), Few-shot prompting (Brown et al., 2020), Chain-of-Thought (CoT) prompting (Wei et al., 2022)("Let's think step by step" is appended to encourage structured reasoning) and CoT + Few-shot prompting. Appendix A.1, Table 8 , 9, 10 and 11 shows the four prompt strategy we used for this analysis.

3.3 Evaluation Metrics

We evaluate models with micro-F1, aggregating counts over all classes. Let TP_c, FP_c, and FN_c denote true positives, false positives, and false negatives for class $c \in \mathcal{C}$. The micro-F1 is their harmonic mean of micro-precision and micro-recall:

$$\text{F1}_\mu = \frac{2 \sum_c \text{TP}_c}{2 \sum_c \text{TP}_c + \sum_c \text{FP}_c + \sum_c \text{FN}_c}.$$

It weights each instance equally, providing a overall score under class imbalance and across heterogeneous label frequencies.

4 Results and Discussion

4.1 Evaluation Phase

Model	M1	M2	M3	M4
BanglaBERT	**0.865**	**0.741**	**0.724**	**0.754**
BanglaHate BERT	0.809	0.724	0.688	0.682
IndicBERT	0.845	0.694	0.705	0.724
XLM-Roberta	0.843	0.728	0.708	0.736
mDistilBERT	**0.865**	0.709	0.675	0.675
mBERT	0.849	0.704	0.708	0.735
VAC-BERT	0.841	0.687	0.698	0.695

Table 4: Performance (micro F1 scores) of fine-tuned models across hate speech detection (M1), hate speech type (M2), target (M3), and severity (M4).

During the evaluation phase on the development set, we assessed models on Sub-tasks 1A, 1B, and

Model	M1	M2	M3	M4
GPT-4o$_{ZS}$	0.701	0.361	0.489	0.453
LLaMA-3.1$_{ZS}$	0.568	0.241	0.336	0.281
Gemma-2$_{ZS}$	0.620	0.226	0.398	0.340
GPT-4o$_{FS}$	0.713	**0.395**	0.478	0.453
LLaMA-3.1$_{FS}$	0.555	0.239	0.319	**0.487**
Gemma-2$_{FS}$	0.607	0.250	0.370	0.326
GPT-4o$_{COT}$	0.736	0.360	0.501	0.476
LLaMA-3.1$_{COT}$	0.579	0.256	0.346	0.298
Gemma-2$_{COT}$	0.628	0.260	0.395	0.357
GPT-4o$_{COT+FS}$	**0.745**	0.365	**0.510**	0.463
LLaMA$_{COT+FS}$	0.593	0.284	0.342	0.301
Gemma-2$_{COT+FS}$	0.638	0.247	0.401	0.349

Table 5: Performance (micro F1 scores) of GPT-4o, LLaMA-3.1, and Gemma-2 under different prompting strategies (ZS = Zero-Shot, FS = Few-Shot, COT = Chain-of-Thought) across hate speech detection (M1), type (M2), target (M3), and severity (M4) classification.

1C. For Sub-tasks 1A and 1B, we employed a two-stage pipeline consisting of a binary toxic comment detector (M1) followed by multi-label classifiers (M2 for hate type and M3 for target). For Sub-task 1C, a separate multi-label classifier (M4) was used in combination with the outputs of M1, M2, and M3.

In toxic comment detection (M1), BanglaBERT and mBERT were the top performers with identical F1 scores of 0.865. Other models, including recent LLMs like GPT-4o$_{COT+FS}$, also demonstrated competitive performance, indicating that both fine-tuned BERT models and large generative models are effective for this task. For Sub-task 1A, performance varied more due to the task's complexity. BanglaBERT achieved the highest F1 score of 0.741, outperforming other fine-tuned models like BanglaHateBERT (0.724) and XLM-Roberta (0.728). Decoder-only LLMs generally struggled with the expanded label space. In Sub-task 1B, BanglaBERT again secured the best performance with an F1 score of 0.7247. The difficulty of this task was evident as only GPT-4o surpassed an F1 score of 0.6 among LLMs. For severity classifier(M4), BanglaBERT was the top fine-tuned model (F1 = 0.7577), while GPT-4o$_{COT+FS}$ led among decoder-only models. Across all subtasks, BanglaBERT consistently delivered the strongest and most reliable performance. Among LLMs, CoT with few-shot prompting proved most effective. Following the HateSense framework, the combined multi-task evaluation for Task 1C (classifying type of hate, severity, and targeted group) in the development phase achieved F1 = 0.7233 (accuracy = 0.7233, precision = 0.7165, recall = 0.7233).

4.2 Testing Phase

Evaluation score	M1	M2	M3	M4
F1 score	0.833	0.704	0.703	0.745
precision	0.836	0.701	0.70	0.741
recall	0.832	0.717	0.708	0.765

Table 6: Performance (F1 micro scores) in testing phase of BanglaBERT across hate speech detection (M1), type, target (M3) and severity (M4) classification.

During testing, we retrained our best-performing model from the evaluation phase, BanglaBERT, using the combined training and development sets to obtain a more generalizable classifier. Following the proposed HateSense framework, we trained separate models for each subtask. The toxic comment detector (M1), formulated as a binary classification task (toxic vs. non-toxic), achieved a micro-F1 score of 0.833. Using M1's predictions as an additional input signal, we then trained a multiclass hate-type classifier (M2) for Subtask 1A, which attained a micro-F1 of 0.704. For Subtask 1B, the target-group multiclass classifier (M3) achieved a micro-F1 of 0.703, while the hate severity multiclass classifier (M4) reached a micro-F1 of 0.745. These results are summarized in Table 6. Aggregating across all subtasks in the test set, our HateSense framework attains a micro-F1 of 0.717 for Task 1C (accuracy = 0.717, precision = 0.718, recall = 0.717).

4.3 Error Analysis

Effect of Class Imbalance. Class-wise accuracies (Appendix A.2, Figure 2) and confusion matrices (Figure 3) reveal consistent patterns across subtasks. Due to class imbalance, performance drops notably for minority and nuanced classes like Sexism (Task 1A), Society (Task 1B), and Mild (Task 1C). While the model handles clear categories well, it struggles with subtle hate. Task 1A accuracy is high for None (0.89) and Profane (0.68) but drops to 0.00 for Sexism. Task 1B favors

Comment	Gold Label	Model Prediction	Subtask
তোমার মতন মেয়ে বেঁচে থাকার চেয়ে মরে যাওয়া অনেক ভালো তোমরা হচ্ছ পাপী অন্য ছেলের সাথে ফটো দাও কেন *"For a girl like you, it's much better to die than to live. You are sinners – why do you post photos with other men?"*	Sexism	Abusive	1A
আগে ইসরাইল বদ করতে হবে সবাই একসাথে *"We must destroy Israel first, all together."*	Abusive	Political Hate	1A
ইহুদির বাচ্চা ইহুদী ই হবে *"A Jew's child will be a Jew."*	Community	Society	1B
ভারতীয় দালাল সময় টিভিকে বয়কট করুন *"Boycott Somoy TV, the Indian stooges."*	Organization	Community	1B
ইজরায়েলের বিচার হওয়া উচিৎ *"Israel should be brought to justice."*	Abusive, Mild, Society	Abusive, Little-to-None, Society	1C
আল্লাহ্ এসব জানোয়ারদের শেষ করে দাও *"Allah, wipe out these animals."*	Profane, Severe, Community	Abusive, Little-to-None, Community	1C

Table 7: Examples of comments, gold labels, and model predictions under the HateSense framework for subtasks 1A (hate type), 1B (target), and 1C (severity). Incorrect model predictions are highlighted in red.

None (0.85) over Society (0.34), and Task 1C excels on Little-to-None (0.95) while failing on Mild (0.36). Confusion matrices confirm these trends: M1 effectively separates toxic from non-toxic, whereas M2–M4 bias toward dominant labels (e.g., None/Little-to-None). Consequently, nuanced hate often misclassifies as neutral; Sexism is never identified (0 correct), and Mild/Severe cases are frequently predicted as Little-to-None.

Qualitative Error Analysis. Table 7 illustrates characteristic failures across subtasks. In Subtask 1A (type classification), we observe frequent confusion between Abusive and Profane, alongside under-predictions of subtle Political Hate, often stemming from short texts or figurative language. For Subtask 1B (target identification), the primary challenge lies in distinguishing Organization versus Community, particularly when targets are implied or indirect. Despite multitask modeling capturing interdependencies, systematic errors persist, including specificity loss, target mismatches, and severity underestimation:

- **Sexism → Generic Abuse:** Explicit gendered hate (e.g., তোমার মতন মেয়ে বেঁচে..) is reduced to generic *Abusive*, missing honor-policing and sexism.
- **Target Granularity Confusion:** The model detects group hate but confuses labels (e.g., Jews → *Society*, TV channels → *Community*), failing to distinguish Community / Society / Organization.
- **Severity Underestimation:** Violent rhetoric (e.g., "আল্লাহ্ এসব জানোয়ারদের..") is gold-labeled *Severe* but predicted as *Little-to-None*, showing the model downgrades threat intensity.

5 Conclusion

We introduced HateSense, a multi-task framework for Bangla hate speech detection that jointly models hate type, target, and severity. Leveraging encoder–decoder transformers with focal loss, Odds Ratio Preference Optimization (ORPO), and CoT + few-shot prompting, our system achieved strong performance across all subtasks in BLP 2025, demonstrating the effectiveness of transformer-based approaches for low-resource, morphologically rich languages. At the same time, our analysis exposed persistent challenges, particularly class imbalance and the difficulty of modeling underrepresented categories such as Sexism and Religious Hate. In future work, we plan to explore data augmentation, cross-lingual transfer, and more robust multitask architectures to improve fine-grained Bangla hate speech detection and extend these methods to other low-resource languages. Our code is available for reproducibility at: https://github.com/HateSense.

6 Limitations

While our framework achieved strong performance, several limitations remain. First, fine-tuning decoder-based models such as GPT, LLaMA, or Gemma on the shared task dataset could further improve performance; however, computational resource constraints prevented us from exploring this option. Second, the dataset suffers from class imbalance. For example, hate speech type Sexism, target category Society, and severity level Severe are underrepresented, leading to reduced model performance in these classes. Although we employed focal loss and Odds Ratio Preference Optimization (ORPO) to address imbalance, the models still struggle with fine-grained distinctions in ambiguous or borderline cases.

Moreover, as Bangla is a low-resource language, language models face challenges in capturing cultural-specific hate speech phenomena, which limits their ability to generalize beyond surface-level patterns. Fine-tuning on more culturally nuanced datasets could enhance detection accuracy. Another constraint is the limited availability of pre-trained models specifically focused on Bangla, which restricts opportunities for leveraging domain-specific linguistic features. Finally, the overall dataset size for training and evaluation remains relatively small, which reduces the robustness and generalizability of our models to real-world applications.

References

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, and 260 others. 2023. Gpt-4 technical report.

Abhik Bhattacharjee, Tahmid Hasan, Kazi Samin Mubasshir, Md. Saiful Islam, Wasi Uddin Ahmad, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2021. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. In *NAACL-HLT*.

Pramit Bhattacharyya, Joydeep Mondal, Subhadip Maji, and Arnab Bhattacharya. 2023. Vacaspati: A diverse corpus of bangla literature. *ArXiv*, abs/2307.05083.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Sergio Andrés Castaño-Pulgarín, Natalia Suárez-Betancur, Luz Magnolia Tilano Vega, and Harvey Mauricio Herrera López. 2021. Internet, social media and online hate speech. systematic review. *Aggression and Violent Behavior*, 58:101608.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *ArXiv*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 510 others. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

Dominik Hangartner, Gloria Gennaro, Sary Alasiri, Nicholas Bahrich, Alexandra Bornhoft, Joseph Boucher, Buket Buse Demirci, Lauren M. Derksen, Aldo Hall, Matthias Jochum, María M. Muñoz, Marc Richter, Franziska Vogel, Salome Wittwer, Felix Wüthrich, Fabrizio Gilardi, and Karsten Donnay. 2021. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences of the United States of America*, 118.

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.

Maliha Jahan, Istiak Ahamed, Md. Rayanuzzaman Bishwas, and Swakkhar Shatabda. 2019. Abusive comments detection in bangla-english code-mixed and transliterated text. *2019 2nd International Conference on Innovation in Engineering and Technology (ICIET)*, pages 1–6.

Md Saroar Jahan, Mainul Haque, Nabil Arhab, and Mourad Oussalah. 2022. Banglahatebert: Bert for abusive language detection in bengali. In *RESTUP*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Jana Papcunová, Marcel Martonik, Denisa Fedáková, Michal Kento, Miroslava Bozogáňová, Ivan Srba, Róbert Móro, Matú Pikuliak, Marián Simko, and Matú Adamkovi. 2021. Hate speech operationalization: a preliminary examination of hate speech indicators and their structure. *Complex & Intelligent Systems*, 9:2827–2842.

Tabia Tanzin Prama, Jannatul Ferdaws Amrin, Md. Mushfique Anwar, and Iqbal H. Sarker. 2025. Ai enabled user-specific cyberbullying severity detection with explainability. *ArXiv*, abs/2503.10650.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Gemma Team Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, L'eonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ram'e, Johan Ferret, Peter Liu, Pouya Dehghani Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, and 176 others. 2024. Gemma 2: Improving open language models at a practical size. *ArXiv*, abs/2408.00118.

Nihar Ranjan Sahoo, Gyana Prakash Beria, and Pushpak Bhattacharyya. 2024. Indicconan: A multilingual dataset for combating hate speech in indian context. In *AAAI Conference on Artificial Intelligence*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

G. M. Shahariar Shibli, Md. Tanvir Rouf Shawon, Anik Hassan Nibir, Md. Zabed Miandad, and Nibir Chandra Mandal. 2022. Automatic back transliteration of romanized bengali (banglish) to bengali. *Iran Journal of Computer Science*, 6:69–80.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

A Appendix

A.1 Prompts

Task	Zero-Shot Prompt
Binary Hate Speech Detection	You are an expert at detecting hate speech in Bangla text. Your task is to classify each input text as either: - true (if it contains hate speech) - false (if it does not contain hate speech) Text: "<INSERT_TEXT_HERE>" Answer:
Hate Speech Type Classification	You are an expert at detecting hate speech in Bangla text. Your task is to classify each input text into one of the following categories: - Abusive - Sexism - Religious Hate - Political Hate - Profane - Non-hate Text: "<INSERT_TEXT_HERE>" Answer:
Target of Hate Speech	You are an expert at analyzing hate speech in Bangla text. Your task is to identify the primary target of hate speech in each input text. The possible targets are: - Individuals - Organizations - Communities - Society - Non-hate Text: "<INSERT_TEXT_HERE>" Answer:
Hate Speech Severity	You are an expert at detecting hate speech in Bangla text. Your task is to classify the severity of the text into one of the following categories: - Little to None - Mild - Severe Text: "<INSERT_TEXT_HERE>" Answer:

Table 8: Zero-shot prompts for all four hate speech classification subtasks (Hate Speech Detection (M1), Hate Speech Type Classification (M2), Target of Hate Speech (M3) and Hate Speech Severity (M4).

Task	Few-Shot Prompt
Binary Hate Speech Detection	You are an expert at detecting hate speech in Bangla text. Classify each input text as either True (if it contains hate speech) or False (if it does not). Examples: Text: ”ইজরায়েলের বিচার হওয়া উচিৎ” Answer: True Text: ”১২ বছরের ভিতর কোনো ভোট দিতে পারিনি” Answer: False Now classify the following text: Text: ”<INSERT_TEXT_HERE>” Answer:
Hate Speech Type Classification	You are an expert at detecting hate speech in Bangla text. Classify each input text into one of the following categories: Abusive, Sexism, Religious Hate, Political Hate, Profane, Non-hate Examples: Text: ”ইজরায়েলের বিচার হওয়া উচিৎ” Answer: Abusive Text: ”শামীম ওসামা বিন হাসিনা” Answer: Non-hate Text: ”ইহুদির বাচ্চা ইহুদী ই হবে” Answer: Religious Hate Text: ”যে রাষ্ট্র আমেরিকার সাথে নিজেদের তুলনা দেয় হঠাৎ কোনও গরিব হয়ে গেলো এখন ঋণ চায় তার মানে কি চাপাবাজ আওয়ামী লীগ” Answer: Political Hate Text: ”সময় টিভি একটা জাউড়া মিডিয়া মিথ্যা তথ্য প্রচার করে বেড়ায়” Answer: Profane Text: ”তুই মারা গেলে ভালো কারণ তার স্ত্রী থাকতে কেন তার সাথে শুইতে গেলি” Answer: Sexism Now classify the following text: Text: ”<INSERT_TEXT_HERE>” Answer:
Target of Hate Speech	You are an expert at detecting the target of hate speech in Bangla text. Classify each input text into one of: Individuals, Organizations, Communities, Society, Non-hate. Examples: Text: ”ইজরায়েলের বিচার হওয়া উচিৎ” Answer: Society Text: ”শামীম ওসামা বিন হাসিনা” Answer: Non-hate Text: ”আল্লাহ্ এসব জানোয়ারদের শেষ করে দাও” Answer: Community Text: ”আলহামদুলিল্লাহ দেশ এগিয়ে যাচ্ছে বিএনপি জামাতীদের জুতা পেটা করতে হবে” Answer: Organization Text: ”চোর চোর ভোট চোর হাসিনা ভোট চোর” Answer: Individual Now classify the following text: Text: ”<INSERT_TEXT_HERE>” Answer:
Hate Speech Severity	You are an expert at assessing the severity of hate speech in Bangla text. Classify each input text into one of: Little to None, Mild, Severe. Examples: Text: ”ইজরায়েলের বিচার হওয়া উচিৎ” Answer: Mild Text: ”শামীম ওসামা বিন হাসিনা” Answer: Little to None Text: ”আল্লাহ্ এসব জানোয়ারদের শেষ করে দাও” Answer: Severe Now classify the following text: Text: ”<INSERT_TEXT_HERE>” Answer:

Table 9: Few-shot prompts for the four hate speech classification subtasks (Hate Speech Detection (M1), Hate Speech Type Classification (M2), Target of Hate Speech (M3) and Hate Speech Severity (M4).

Task	CoT Prompt
Binary Hate Speech Detection	You are an expert at detecting hate speech in Bangla text. Internally follow these reasoning steps: 1. Read the full sentence. 2. Identify offensive or hostile words/phrases. 3. Consider the context to see if the text expresses hate. 4. If hate elements are present → classify as True. Otherwise → False. Important: Do all reasoning internally and return only the final classification. Text: "<INSERT_TEXT_HERE>" Answer (True or False):
Hate Speech Type Classification	You are an expert at detecting hate speech in Bangla text. Internally follow these reasoning steps: 1. Read the sentence carefully. 2. Identify abusive, gender-related, religious, political, or profane words. 3. Consider the context to determine the type of hate. 4. Map the text into one of these categories: - Abusive - Sexism - Religious Hate - Political Hate - Profane - Non-hate Important: Do all reasoning internally and return only the final category. Text: "<INSERT_TEXT_HERE>" Answer:
Target of Hate Speech	You are an expert at analyzing the target of hate speech in Bangla text. Internally follow these reasoning steps: 1. Read the sentence. 2. Identify who/what is being attacked. 3. Determine if the target is an individual, organization, community, or society. 4. If no hate is detected, return Non-hate. Possible categories: - Individual - Organization - Community - Society - Non-hate Important: Do all reasoning internally and return only the final target. Text: "<INSERT_TEXT_HERE>" Answer:
Hate Speech Severity	You are an expert at detecting the severity of hate speech in Bangla text. Internally follow these reasoning steps: 1. Read the sentence. 2. Identify hostile or violent language. 3. Check if the tone is harmless, mildly offensive, or severely hateful/violent. 4. Classify into one of the following categories: - Little to None - Mild - Severe Important: Do all reasoning internally and return only the final severity label. Text: "<INSERT_TEXT_HERE>" Answer:

Table 10: Chain-of-Thought (CoT) prompts for the four hate speech classification subtasks (Hate Speech Detection (M1), Hate Speech Type Classification (M2), Target of Hate Speech (M3) and Hate Speech Severity (M4).

Table 11: Chain-of-Thought + Few-shot prompts for the four hate speech classification subtasks (Hate Speech Detection (M1), Hate Speech Type Classification (M2), Target of Hate Speech (M3) and Hate Speech Severity (M4)

Task	CoT + Few-Shot Prompt
Binary Hate Speech Detection	You are an expert at detecting hate speech in Bangla text. Internally follow these steps: 1. Read the sentence fully. 2. Identify offensive or hostile words. 3. Consider the context to see if hate is expressed. 4. Decide: True if hate speech, False otherwise. Important: Do all reasoning internally and return only the final classification. Examples: Text: ”ইজরায়েলের বিচার হওয়া উচিৎ” Answer: True Text: ”১২ বছরের ভিতর কোনো ভোট দিতে পারিনি” Answer: False Now classify the following text: Text: ”<INSERT_TEXT_HERE>” Answer:
Hate Speech Type Classification	You are an expert at detecting hate speech in Bangla text. Internally follow these steps: 1. Read the sentence carefully. 2. Identify abusive, gender-related, religious, political, or profane terms. 3. Map them into one category. Categories: Abusive, Sexism, Religious Hate, Political Hate, Profane, Non-hate Important: Do all reasoning internally and return only the final category. Examples: Text: ”ইজরায়েলের বিচার হওয়া উচিৎ” → Abusive Text: ”শামীম ওসামা বিন হাসিনা” → Non-hate Text: ”ইহুদির বাচ্চা ইহুদী ই হবে” → Religious Hate Text: ”যে রাষ্ট্র আমেরিকার সাথে নিজেদের তুলনা দেয়…” → Political Hate Text: ”সময় টিভি একটা জাউড়া মিডিয়া…” → Profane Text: ”তুই মারা গেলে ভালো…” → Sexism Now classify the following text: Text: ”<INSERT_TEXT_HERE>” Answer:
Target of Hate Speech	You are an expert at identifying the target of hate speech in Bangla text. Internally follow these steps: 1. Read the sentence carefully. 2. Identify who/what is being attacked. 3. Map the target into one of the given categories. Categories: Individual, Organization, Community, Society, Non-hate Important: Do all reasoning internally and return only the final target. Examples: Text: ”ইজরায়েলের বিচার হওয়া উচিৎ” → Society Text: ”শামীম ওসামা বিন হাসিনা” → Non-hate Text: ”আল্লাহ্ এসব জানোয়ারদের শেষ করে দাও” → Community Text: ”আলহামদুলিল্লাহ দেশ এগিয়ে যাচ্ছে বিএনপি জামাতীদের জুতা পেটা করতে হবে” → Organization Text: ”চোর চোর ভোট চোর হাসিনা ভোট চোর” → Individual Now classify the following text: Text: ”<INSERT_TEXT_HERE>” Answer:
Hate Speech Severity	You are an expert at detecting the severity of hate speech in Bangla text. Internally follow these steps: 1. Read the sentence. 2. Identify hostile or violent language. 3. Judge whether it is harmless, mildly offensive, or severely hateful. Categories: Little to None, Mild, Severe Important: Do all reasoning internally and return only the final severity label. Examples: Text: ”ইজরায়েলের বিচার হওয়া উচিৎ” → Mild Text: ”শামীম ওসামা বিন হাসিনা” → Little to None Text: ”আল্লাহ্ এসব জানোয়ারদের শেষ করে দাও” → Severe Now classify the following text: Text: ”<INSERT_TEXT_HERE>” Answer:

A.2 Class-wise perfromance analysis

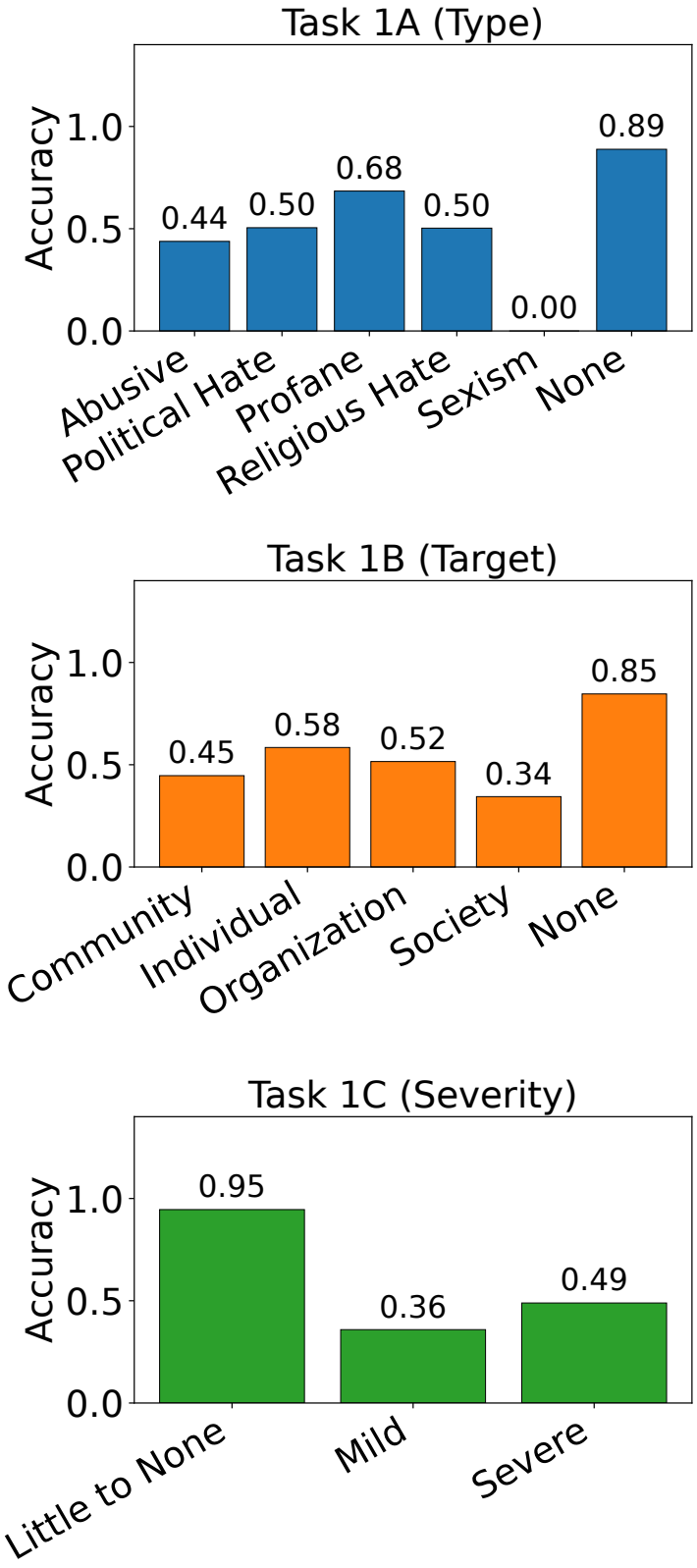
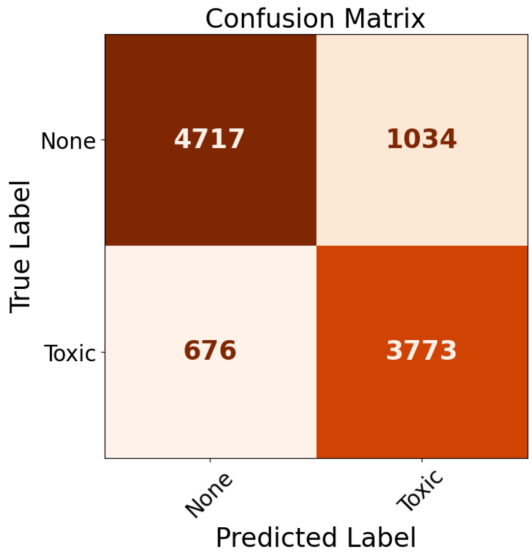
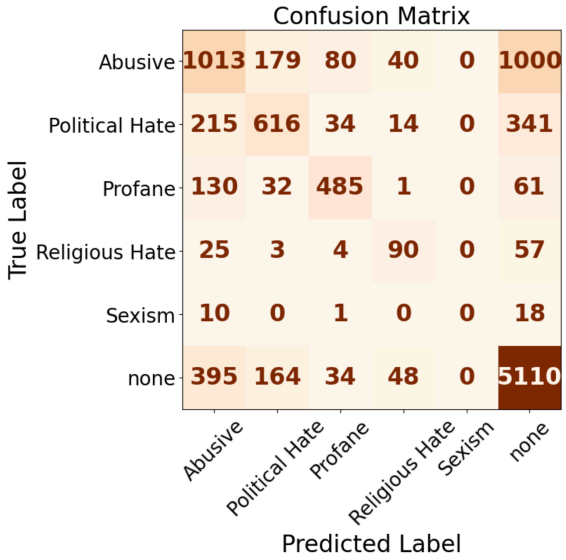


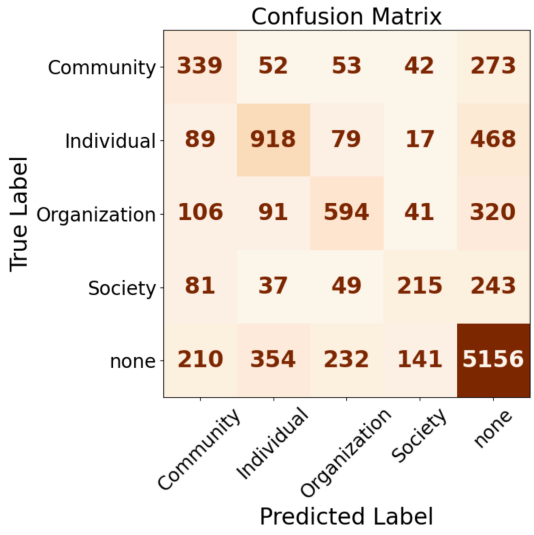
Figure 2: Class-wise accuracy for Task 1A (type), Task 1B (target), and Task 1C (severity).



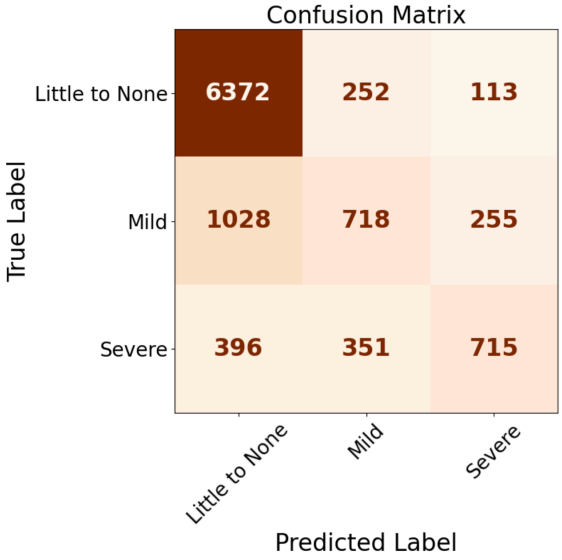
(a) Confusion matrix of hate speech detection (M1)



(b) Confusion matrix of hate speech type classification (M2)



(c) Confusion matrix of hate speech target classification (M3)



(d) Confusion matrix of hate speech severity classification (M4)

Figure 3: Confusion matrices of the HateSense framework across all models: (a) Hate speech detection(M1), (b) Hate type classification (M2), (c) Hate target classification (M3), (d) Hate severity classification (M4) of the test phase.