LLM-Based Multi-Task Bangla Hate Speech Detection: Type, Severity, and Target

Md Arid Hasan[1], **Firoj Alam**[2], **Md Fahad Hossain**[3], **Usman Naseem**[4], **Syed Ishtiaque Ahmed**[1]
[1]University of Toronto, Canada, [2]Qatar Computing Research Institute, Qatar
[3]Daffodil International University, Bangladesh, [4]Macquarie University, Australia
{arid, ishtiaque}@cs.toronto.edu, fialam@hbku.edu.qa,
fahadhossain.swe@diu.edu.bd, usman.naseem@mq.edu.au

WARNING: This paper contains examples which may be disturbing to the reader

Abstract

Online social media platforms have become central to communication and information exchange, however, they also serve as fertile ground for hate speech, offensive language, and bullying targeting individuals and communities. Such content undermines online safety and inclusion, underscoring the need for reliable detection systems—especially in low-resource languages with limited moderation tools. For Bangla, existing work provides valuable resources and models, however, they are mostly single-task (e.g., binary hate/offense) with narrow coverage of key dimensions such as type, severity, and target. We address these gaps by introducing *the first multi-task* Bangla hate-speech dataset, *BanglaMultiHate*, one of the largest manually annotated dataset to date. Using this resource, we performed a comparative study across different baselines, monolingual pretrained models, and LLMs under zero-shot, few-shot, and LoRA fine-tuning settings. Our findings show that while LoRA-tuned LLMs rival BanglaBERT, culturally grounded pretraining remains crucial for robust performance. Overall, *BanglaMultiHate* establishes a stronger benchmark for hate speech detection in low-resource contexts. All data[1] and scripts[2] are released for reproducibility.

1 Introduction

The rise of social media has increased the spread of harmful online content (Walther, 2022), with hate speech emerging as a critical societal issue given its potential to perpetuate discrimination (Gelber, 2021), harassment, and violence. Given the large volume of user-generated content, manual moderation is neither scalable nor consistent, highlighting the urgent need for reliable and scalable automated hate speech detection systems. Although

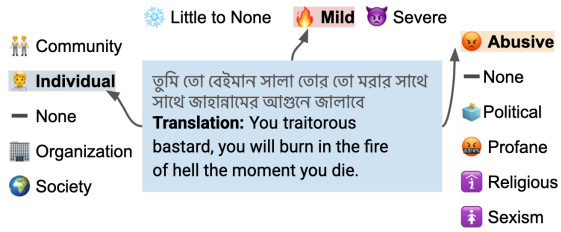

Figure 1: An example of hateful comment with its English translation showing type, severity and target of hate.

substantial progress has been achieved in high-resource languages such as English (Albladi et al., 2025), research effort for low-resource languages like Bangla are relatively limited (Sharma et al., 2025; Das et al., 2022a; Haider et al., 2025; Romim et al., 2022).

Identifying hate speech in Bangla imposes unique challenges due to its rich morphology, free word order, and code-switching with English and other regional dialects, making it difficult for models trained on other languages to generalize effectively. Furthermore, the scarcity of annotated datasets and the lack of high-quality pretrained resources exacerbate the difficulty of building accurate classification systems (Al Maruf et al., 2024). Existing studies often rely on classical machine learning models (Kiela et al., 2020; Mridha et al., 2021; Romim et al., 2022), deep learning models (Romim et al., 2022; Keya et al., 2023), and adapt pretrained models designed primarily for English (Mridha et al., 2021). However, these approaches often fail to capture the cultural, social, and linguistic nuances that shape how hate is expressed in Bangla (Al Maruf et al., 2024), such as context-dependent slurs, metaphorical insults, or region-specific idiomatic usage (Jahan et al., 2022). In addition, most of the studies are limited to single task. Addressing these challenges requires not only improved datasets and resources but also approaches that are sensitive to the sociolinguistic realities of Bangla discourse, ensuring that models move be-

[1]HuggingFace
[2]GitHub

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 33962–33980

yond surface-level understanding and engage with the deeper structures of the language.

In recent years, the rapid advancement of large language models (LLMs) such as GPT-5 (Singh et al., 2025), Claude, Gemini (Comanici et al., 2025), Llama (Dubey et al., 2024), and Qwen (Yang et al., 2025) has shown remarkable success across a variety of downstream NLP tasks, often demonstrating strong generalization abilities in zero-shot or few-shot scenarios (Hasan et al., 2024; Abdelali et al., 2024; Alam et al., 2024b). This has raised important questions about their applicability in sensitive domains such as hate speech detection (Albladi et al., 2025; Alam et al., 2024a), particularly for underrepresented languages. However, the zero-shot performance of LLMs in low-resource contexts is often limited and their inability to capture context-dependent information (Zahid et al., 2025). Moreover, hate speech is highly context-dependent and culturally nuanced, making it difficult for LLMs pretrained on high-resource languages to detect, demonstrating the need for targeted adaptation strategies in low-resource and sensitive tasks.

To address these challenges, we developed the first multi-tasks hate speech dataset named *Bangla-MultiHate* for Bangla. This dataset is specifically designed to support a variety of classification tasks: *(i)* identifying different types of hate speech, *(ii)* the severity of hate, and *(iii)* determining the target of hate. An example of a hateful comment with *type*, *severity* and *target* is demonstrated in Figure 1. This study also conducts a comprehensive evaluation of hate speech detection in Bangla across three tasks, utilizing SVM, BanglaBERT, and zero-shot, few-shot, and LoRA fine-tuning of LLMs. Our contributions can be summarized as follows:

- We developed the first multi-task hate speech dataset for Bangla and one of the largest manually annotated hate speech datasets, which includes type of hate, severity and target.
- We provide comprehensive comparisons of classical, monolingual pretrained, and zero-shot, few-shot, and LoRA fine-tuned approaches using LLMs for Bangla hate speech detection.
- We assess the effectiveness of zero-shot and few-shot inference and LoRA fine-tuning for LLMs, offering insights into their adaptability in low-resource tasks.
- We highlight key limitations and trends, demonstrating that while fine-tuned LLMs are comparable to the performance of BanglaBERT, emphasizing the continued importance of culturally and linguistically grounded pretraining for combating online hate speech in low-resource languages.

Dataset	Size	Type	Labels / Tasks	Source
BD-SHS (Romim et al., 2022)	50,281	Comments	3-level: HS vs. non-HS; target; HS type	SM
Bengali Tweets (Das et al., 2022a)	10,000	Code-mixed	Hate/offense detection	X
TB-OLID (Raihan et al., 2023)	5,000	Transliterated, code-mixed	Offensive vs. Not; target (Indiv./ Group/Untargeted)	FB
BanTH (Haider et al., 2025)	37,350	Transliterated	Multi-label target; HS	YT
BIDWESH (Fayaz et al., 2025)	9,183	Dialectal	Hate vs. non-hate; ~13 types; 7 targets	SM
BanglaMultiHate (Ours)	50,746	Comments	Type, severity, target	YT

Table 1: Overview of existing datasets and ours. SM: Social media. YT: Youtube, FB: Facebook

Our findings are summarized as follows:

- Fine-tuned monolingual BanglaBERT yields superior performance. Moreover, both fine-tuned language-specific models yielded better performance than multilingual models.
- In the zero-shot setting, open-source models underperformed relative to the majority baseline and SVM, while closed-source models achieved stronger results than these baselines.
- SVM performs comparatively better than fine-tuned Qwen on all tasks, and exceeds Llama3 on the target task.
- In zero-shot settings, Gemini outperforms its few-shot configuration, while GPT-5 shows improved performance under few-shot prompting compared to zero-shot.
- Model performance varies significantly with task complexity, with more complex tasks leading to greater performance divergence across models.

2 Related Work

The identification of offensive language and hate speech has become increasingly important due to the extensive use of social media, which has created an environment in which harmful content can spread rapidly (Jiang and Zubiaga, 2024). Research on hate speech identification has progressed rapidly over the past decade (Fortuna and Nunes, 2018), moving from lexicon-based classifiers to transformer models and, more recently LLMs (Albladi et al., 2025).

2.1 Existing Hate Speech Datasets

There has been effort to develop datasets in the past. Gupta et al. (2022) introduced a 150K-comment dataset for abusive speech detection in

five Indic languages, while Sharif et al. (2021) studied offensive language detection in multilingual code-mixed text. These work establish important baselines for future research on code-mixed offensive text detection in Dravidian languages (Saumya et al., 2021b; Chakravarthi et al., 2022).

Some of the notable resources for hate and abusive content on Bangla include 10,178 tweets labeled as hate/offensive/normal (Das et al., 2022b), a 30K comments dataset with 10K hate speech examples (Romim et al.), 3K transliterated Bangla-English abusive comments (Sazzed, 2021), 50K offensive comments from online social networking (Romim et al., 2022), and 10K Bangla posts consisting of 5K actual and 5K Romanized Bengali tweets (Das et al., 2022a). Moreover, a multi-label transliterated Bangla hate speech dataset has been developed by Haider et al. (2024) utilizing a translation-based LLM prompting approach.

Building on these efforts, Table 1 provides an overview of existing resources. Our contribution extends this line of work by introducing a larger dataset that not only supports multiple tasks but also incorporates a richer topical hierarchy, spanning 19 topics and 120 sub-topics.

2.2 Existing Approaches

Various classical models (such as logistic regression (LR), SVM, and random forest), deep learning models (e.g., LSTM), and transformer-based models (e.g., BERT, XLM-R, MuRIL, AraBERT, etc.) have been studied in the literature. Sharif et al. (2021) demonstrated that transformer-based pre-trained language models (e.g., Indic-BERT, XLM-R, mBERT) outperform classical models (e.g., LR, SVM). The multi-task learning approach using AraBERT has been studied in Arabic for the identification of offensive language and hate speech (Djandji et al., 2020), while random forest, k-nearest neighbors, and MLP classifiers have been studied for offensive language identification from Dravidian code-mixed texts (B and A, 2021). Pelicon et al. (2021) employed mBERT and LASER models for zero-shot cross-lingual transfer learning, demonstrating promising results in languages such as German, Spanish, Indonesian, and Arabic. Similarly, (Saumya et al., 2021a) explores the impact of cross-cultural transfer learning, showing how biases across cultures affect model performance, examining the impact of cross-cultural transfer learning.

Kiela et al. (2020) utilizes SVM, CNN, and LSTM models to evaluate performance on hateful content. SVM, naive bayes, and random forest, along with transformer models have been studied for multi-label hate speech identification (Ibrohim and Budi, 2019). Mridha et al. (2021) employed L-Boost, a modified AdaBoost algorithm combining BERT embeddings with LSTM models, to identify offensive texts in Bangla and Banglish social media content. SVM, LSTM, and Bi-LSTM models have also been analyzed by Romim et al. on Bangla YouTube and Facebook comments, with results showing that SVM outperforms LSTM and Bi-LSTM. Furthermore, combining BERT and GRU architectures for hate speech detection has been explored in Bengali social media texts (Keya et al., 2023). Explainable hate speech identification has recently attracted attention in the literature (Yang et al., 2023; Piot and Parapar, 2025; Sariyanto et al., 2025).

3 Dataset

3.1 Data Collection

We collected public comments from YouTube videos using the YouTube API[3], primarily from Somoy TV, which is a popular Bangla News channel. The comments belong to 19 different categories, including *Business, Celebrities, Disaster, Entertainment, Fashion, Geopolitics, Health, History, International, Lifestyle, Literature, Miscellaneous, National, Opinion, Politics, Religion, Science, Sports*, and *Technology*, as well as 120 subcategories. In total, we collected approximately 55,000 comments associated with various Bangla news videos. We then removed all entries containing only emojis and URLs, as well as duplicate entries. Additionally, we excluded all Banglish comments (Bangla text written using the English alphabet) from the initial dataset. After applying these filtering and duplicate-removal steps, the dataset contained 50,746 entries. The category-wise data distribution is presented in Figure 2, with more than 90% of the comments falling into five categories.

3.2 Data Annotation

3.2.1 Annotation Guidelines

We developed an annotation guideline to facilitate the data annotation process. Our annotation setup

[3] YouTube API

Figure 2: Distribution of the MultiHate dataset across different categories.

was a multitask annotation. Therefore, each instance could be assigned multiple labels to capture the overlapping and nuanced nature of the content, such as identifying the type of hate, the severity of hate, and the target of hate simultaneously. The guidelines provided clear definitions, decision criteria, and illustrative examples to ensure consistency across annotators. Below, we briefly discuss the annotation guidelines for each annotation task and more detail can be found in Appendix A.

Type of Hate: The purpose of this task is to identify the type of hate from YouTube comments. The annotators classified whether the comments are *Abusive, Sexism, Religious Hate, Political Hate, Profane,* or *None* based on the criteria discussed in the Appendix A. Depending on the label selection, the annotator proceeds with different subsequent labeling tasks. If a comment is labeled as *None*, annotators automatically assign the labels *Little to None* for the severity of hate and *None* for the target of hate; otherwise, they proceed with the regular annotation process.

Severity of Hate: This task aims to assess the degree of hate expressed in a given comment. Annotators evaluate whether the comment reflects *Little to None, Mild,* or *Severe* forms of hate, taking into account factors such as the intensity of derogatory expressions, the use of slurs, and the presence of threats or incitement to violence. The objective is to capture not only the presence of hateful content but also its strength and potential impact.

Target of Hate: This task focuses on identifying the specific *Individuals, Organizations, Communities, Society,* or *None* that is the target of hateful expression. Annotators classify whether the hate is directed toward protected characteristics such as organizations, communities, or society, or if it is aimed at individuals without reference to group identity. In cases where no explicit target is present, annotators assign the label *None*. The goal of this task is to capture the social dimension of hateful language, enabling analysis not only of the presence of hate but also of who or what is being targeted.

3.2.2 Manual Annotation

The annotation team comprised 35 native Bangla-speaking undergraduate students, including both male and female annotators, who contributed to the task voluntarily. The team was trained and supervised by expert annotators to ensure reliability and consistency in the labeling process. Each comment was independently annotated by three annotators, and periodic quality checks of randomly selected samples were conducted, followed by feedback sessions to maintain annotation standards. The final label for each comment was determined by majority agreement among the annotators. In instances of persistent disagreement, consensus meetings were organized to resolve discrepancies and establish the final annotation.

Pair	Cramér's V	NMI
Type vs. Severity	0.63	0.43
Type vs. Target	0.50	0.49
Severity vs. Target	0.56	0.33

Table 2: Cramér's V and normalized mutual information for inter-class correlation.

Annotation Agreement We evaluated the inter-annotator agreement (IAA) of the manual annotations using Fleiss' Kappa coefficient (κ) to assess the reliability of the annotation process across all tasks. The obtained κ scores were 0.71, 0.84, and 0.79 for the type of hate, severity of hate, and target of hate tasks, respectively, indicating substantial to perfect agreement.[4] We also observed that an increase in the number of annotation classes introduces greater challenges for annotators, which is reflected in lower IAA scores for more fine-grained tasks. We also computed *Cramér's V* and *Normalized Mutual Information (NMI)* in Table 2 to show that the annotated dimensions are related (as facets of harm) but not redundant. These scores consistently show moderate dependence but far from duplication and align with our qualitative

[4]According to Landis and Koch (1977), values of κ between 0.61–0.80 represent substantial agreement, while values between 0.81–1.0 represent almost perfect agreement.

analysis. Severity is not fixed by type, and targets are not uniquely tied to either type or severity; the same type can appear with multiple severity levels, and the same target can appear across different types and severities. This is why the task is multi-dimensional rather than a single-label taxonomy.

Class	Train	Dev	Test	Total
Type of Hate				
Abusive	8,212	1,113	2,312	**11,637**
Political Hate	4,227	574	1,220	**6,021**
Profane	2,331	342	709	**3,382**
Religious Hate	676	78	179	**933**
Sexism	122	19	29	**170**
None	19,954	2,898	5,751	**28,603**
Total	**35,522**	**5,024**	**10,200**	**50,746**
Severity of Hate				
Severe	5,180	698	1,462	**7340**
Mild	6,853	909	2,001	**9763**
Little to None	23,489	3,417	6,737	**33,643**
Total	**35,522**	**5,024**	**10,200**	**50,746**
Target of Hate				
Community	2,635	338	759	**3,732**
Individual	5,646	755	1,571	**7,972**
Organization	3,846	584	1,152	**5,582**
Society	2,205	283	625	**3,113**
None	21,190	3,064	6,093	**30,347**
Total	**35,522**	**5,024**	**10,200**	**50,746**

Table 3: Class label distribution across three tasks of the *BanglaMultiHate* dataset.

3.3 Data Split

The dataset was partitioned into training, development, and test sets, comprising 70%, 10%, and 20% of data, respectively, for our experiments. We applied stratified sampling (Sechidis et al., 2011) to ensure a balanced class label distribution across all splits. We provide the detailed data distribution across the splits in Table 3. As shown, the dataset is highly imbalanced across all three annotation dimensions. For the *type of hate* task, the majority of samples fall under the *none* class, while categories such as *sexism* and *religious hate* are comparatively underrepresented. A similar pattern is observed in the *severity of hate* task, where most comments are labeled as *little to none*, followed by *mild* and *severe*. For the *target of hate* task, the *none* class again dominates, whereas labels such as *society* and *community* appear far less frequently. This imbalance highlights the inherent challenges of training reliable models on underrepresented classes and demonstrate the importance of stratification for fair evaluation.

3.4 Analysis and Statistics

We present the detailed class label distribution in Table 7. The table reports the frequency of labels across the three tasks, for the training, development, and test splits. This breakdown highlights the inherent class imbalance across tasks, with *None* being the most frequent label, whereas categories such as *Sexism* or *Religious Hate* are underrepresented. Such skewed distributions demonstrate the challenge of building robust models capable of handling rare but socially significant cases of hate speech. Moreover, Table 6 presents the distribution of class labels across word length bins for the three tasks. The majority of samples in all splits fall within the ≤20, indicating that most instances of hate speech in Bangla are expressed concisely.

In Figures 3 and 4, we present the relationship between hate type and severity, and between hate type and target, respectively. The *None* category was excluded from the hate type for a concise visual representation. Figure 3 shows that *Abusive* content is most prevalent, peaking at mild severity with a notable severe presence, while *Political Hate* follows a smaller but similar trend. *Profane* content stands out, concentrated in the severe category, suggesting profanity as a marker of high severity. *Religious Hate* appears at low frequency across all severities, and *Sexism* is rare overall. Mild severity emerges as the dominant category across types. Figure 4 highlights that individuals and organizations are the main targets, with abusive expressions disproportionately directed at individuals.

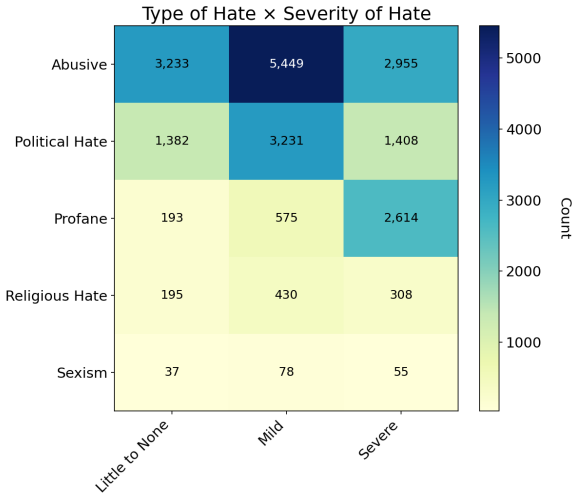

Figure 3: Heatmap demonstrating the relationship between *type* of hate and *severity*.

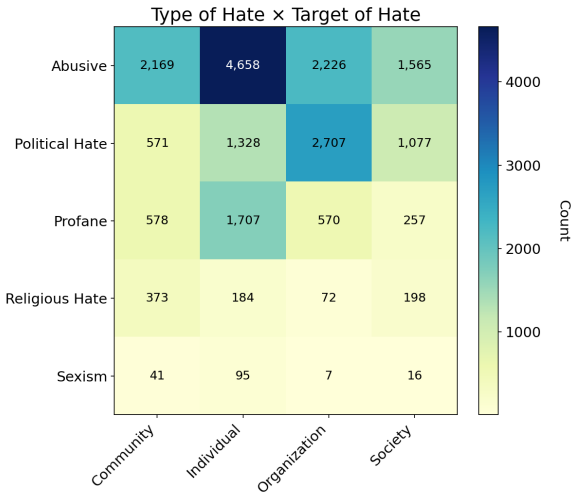

Figure 4: Heatmap demonstrating the relationship between *type* of hate and *target*.

4 Methodology

4.1 Models

We experiment with classical model such as SVM, monolingual pretrained language model such as BanglaBERT (Bhattacharjee et al., 2022), and large language models such as BanglaLLM,[5] Llama-3.2-3B-Instruct,[6] and Qwen3-4B-Instruct-2507.[7] We choose models from different model families to provide extensive evaluation with this dataset.

Baseline. We used a majority-class baseline that always predicts the class with the highest frequency in the training data and a random approach. These methods have been widely used as a baseline technique in numerous prior studies (e.g., (Rosenthal et al., 2017)).

Classical models. We employed SVM (Platt, 1998) with TF-IDF representation which has been extensively utilized in prior research and remains prevalent in low-resource production settings. Our setup employed 1–5 n-grams with TF-IDF weighting and a regularization parameter of $C = 1$.

Pretrained Language Model (PLM). Given that PLMs have demonstrated significant success in the past years and are also computationally reasonable choices for many downstream NLP tasks, we fine-tuned the monolingual BanglaBERT model (Wolf et al., 2020). Following the procedure of Devlin et al. (2019), we trained each model with default hyperparameters (learning rate of $2e^{-5}$, batch size of 16, maximum sequence length of 512, and AdamW optimizer parameters β_1=0.9, β_2=0.999, ϵ=$1e^{-8}$) for 3 epochs. To mitigate training instability, we performed ten runs with different random seeds and selected the best model based on development set performance. All experiments were conducted independently for each task.

LLMs. Recent advances in LLMs have received significant attention from researchers to evaluate the performance of these models, especially for low-resource languages in various downstream NLP tasks. We experiment with Gemini-2.5-pro, GPT-5, BanglaLLM, Llama-3.2-3B-Instruct (Dubey et al., 2024), and Qwen3-4B-Instruct (Yang et al., 2025). We adopt a zero-shot learning setup for all models. To ensure reproducibility, we apply a consistent prompt, response format, output token limit, and decoding configuration (such as temperature set to 0) across models. The prompts were designed using concise instructions, as detailed in Appendix B. We also demonstrate the efficacy of *BanglaMultiHate* dataset by fine-tuning both Llama and Qwen models. We choose PEFT using LoRA (Hu et al., 2022) to reduce the computational cost. We trained the model in full precision (FP16) using the Adam optimizer. The learning rate was set to 2×10^{-4}, with LoRA parameters $\alpha = 16$ and $r = 64$. The maximum sequence length was fixed at 512, and training was performed with a batch size of 8. Fine-tuning was conducted for three epochs without additional hyperparameter tuning. Moreover, both zero-shot and fine-tuned approaches were studied in a multi-task setup due to computational resource constraints.

4.2 Instructions Dataset

We employed a template-based approach to generate diverse English instructions, obtaining 10 hate speech classification task templates per language from GPT-4.1 and Claude-3.5 Sonnet.[8] During fine-tuning and inference, one template was randomly selected and combined with the comment. We report examples of prompts in Appendix B.

4.3 Evaluation Measures

Across all experimental settings, we evaluate performance using accuracy, micro-F1 score, as well as weighted precision and recall, with the weighted metrics chosen to account for class imbalance.

[5] BanglaLLM
[6] Llama-3.2-3B-Instruct
[7] Qwen3-4B-Instruct
[8] claude-3-5-sonnet

5 Results and Discussion

In Table 4, we report model performance across all three tasks in terms of accuracy, precision, recall, and micro-F1.

5.1 Comparison with Baselines

Across all three tasks, both the SVM and pretrained language models substantially outperform the majority and random baselines. Although the majority baseline achieves relatively high accuracy in the hate severity task, this is largely due to label imbalance. Zero-shot results consistently surpass the random baseline across all tasks, yet they still fall behind the majority baseline, demonstrating their limitations without task-specific adaptation. In contrast, PEFT with LoRA yields notable gains. In particular, fine-tuned Llama3 outperforms both baselines across all three tasks, demonstrating the effectiveness of fine-tuning the model. Qwen3 with LoRA shows modest improvements, performing slightly above the majority baseline.

5.2 Classical Model and PLMs

SVM trained with lexical features such as n-grams and TF-IDF, provides consistent but modest improvements over both baselines. For instance, in *type of hate* task, the SVM achieves 0.609 micro-F1 compared to 0.564 micro-F1 from the majority classifier. Similar improvements are seen in the *target of hate* task. These results suggest that traditional supervised methods can exploit word-level patterns that naive baselines miss, making them moderately effective for detecting stereotypical hate expressions. However, the performance changes when used with narrow or context-dependent forms of hate speech, such as implicit derogatory references. This limitation stems from their reliance on surface-level features without deeper semantic understanding.

Leveraging pretraining on large-scale Bangla corpora, BanglaBERT consistently outperforms all models across all tasks. These gains highlight the importance of contextual embeddings from language-specific pretrained models, which enable the system to capture semantic nuances, idiomatic expressions, and subtle markers of abusive tone.

5.3 Zero- and Few-shot Learning

Zero-shot Learning Across all three tasks, BanglaLLM, Gemini-2.5-pro, GPT-5, Llama3, and Qwen3 show mixed performance, while BanglaLLM does not perform better in all three subtasks. For *type of hate*, Llama3 achieves micro-F1 score of 0.275, which is only marginally better than the random baseline (0.164), highlighting the difficulty of zero-shot classification in datasets with imbalanced labels. Qwen3 performs substantially better in the same task with an accuracy and F1 of 0.520, performing better than Llama3 and approaching the majority baseline (0.564), demonstrating that model size significantly influences zero-shot performance. Moreover, Gemini and GPT-5 outperformed both baselines and achieved strong results among LLMs.

In the *severity of hate* task, Gemini, GPT-5, Llama3, and Qwen3 achieve micro-F1 score of 0.698, 0.651, 0.508, and 0.589, respectively. Both Gemini and GPT-5 models perform better than both baselines, while Llama3 and Qwen3 models perform only better than the random baseline (0.327) but do not perform better than the majority baseline (0.660), suggesting that the inherent structure of the severity task requires a nuanced understanding of language intensity, limiting the effectiveness of zero-shot approaches. For the *target of hate*, the models achieve, 0.510 (Gemini), 0.434 (GPT-5), 0.340 (Llama3), and 0.434 (Qwen3) micro-F1 score, similarly perform better than random but below the majority baseline, indicating that identifying the *target of hate* often requires explicit task-specific knowledge that zero-shot models may not fully possess. Overall, zero-shot learning provides a reasonable starting point, particularly for Qwen3, but generally falls short of the majority baseline, emphasizing the need for task adaptation to achieve high performance.

Few-shot Learning We conducted 3-shot prompting experiments using Gemini-2.5-pro and GPT-5 to assess whether few-shot examples improve LLM performance on the Bangla hate speech tasks. Our results show that few-shot prompting provides no improvement compared to zero-shot performance for Gemini. For GPT-5, we observe modest gains only on the *Target of Hate* task and little on the *Type of Hate* task, while *severity* remains largely unchanged from zero-shot.

5.4 Fine-tuning LLMs

Fine-tuning with LoRA significantly improves performance across all three tasks. We also performed analysis on loss behavior during fine-

Model	Type of Hate					Severity of Hate					Target of Hate				
	Acc.	P.	R.	F1	$F1_m$	Acc.	P.	R.	F1	$F1_m$	Acc.	P.	R.	F1	$F1_m$
Random Baseline	0.164	0.385	0.164	0.164	0.120	0.327	0.486	0.327	0.327	0.293	0.204	0.404	0.204	0.204	0.167
Majority Baseline	0.564	0.318	0.564	0.564	0.120	0.660	0.436	0.660	0.660	0.265	0.597	0.357	0.597	0.597	0.150
SVM	0.609	0.574	0.609	**0.609**	0.320	0.672	0.607	0.672	**0.672**	0.400	0.629	0.568	0.629	**0.629**	0.320
BanglaBERT	0.712	0.716	0.712	**0.712**	0.530	0.722	0.727	0.722	**0.722**	0.608	0.715	0.716	0.715	**0.715**	0.582
Zero-Shot															
LLama3	0.275	0.619	0.275	0.275	0.214	0.508	0.729	0.508	0.508	0.348	0.340	0.465	0.340	0.340	0.075
Qwen3	0.520	0.542	0.520	0.520	0.210	0.589	0.639	0.589	0.589	0.455	0.434	0.508	0.434	0.434	0.075
Gemini-2.5-pro	0.674	0.726	0.674	**0.674**	0.453	0.698	0.770	0.698	**0.698**	0.576	0.510	0.593	0.510	0.510	0.118
GPT-5	0.638	0.710	0.638	**0.638**	0.397	0.651	0.750	0.651	0.651	0.581	0.434	0.546	0.434	0.434	0.324
BanglaLLM	0.099	0.669	0.099	0.099	0.061	0.276	0.712	0.276	0.276	0.194	0.149	0.564	0.149	0.149	0.003
Few-Shot															
Gemini-2.5-pro	0.648	0.698	0.648	**0.648**	0.491	0.643	0.727	0.643	0.643	0.555	0.452	0.652	0.459	0.452	0.136
GPT-5	0.654	0.711	0.654	**0.654**	0.425	0.658	0.746	0.658	0.658	0.583	0.648	0.689	0.648	**0.648**	0.231
Fine-tuned															
LLama3	0.620	0.725	0.620	**0.620**	0.120	0.685	0.682	0.685	**0.685**	0.265	0.610	0.716	0.610	**0.610**	0.150
Qwen3	0.595	0.453	0.595	**0.595**	0.220	0.661	0.436	0.661	**0.661**	0.265	0.598	0.470	0.598	**0.598**	0.150
BanglaLLM	0.693	0.677	0.693	**0.693**	0.281	0.722	0.736	0.722	**0.722**	0.626	0.631	0.683	0.631	**0.631**	0.362

Table 4: Performance of different models on Bangla hate speech detection across three tasks. Acc.: Accuracy, P.: Precision, R.: Recall, F1: micro-F1 score, $F1_m$: macro-F1 score. **Bold** indicates results that surpass both baseline methods for the respective task in terms of micro-F1 score, while Underline denotes the best overall performance across all three tasks.

tuning shown in Appendix C. For *type of hate* task, Llama3 achieves micro-F1 score of 0.620, surpassing both zero-shot Llama3 (0.275) and Qwen3 (0.520), and also performs better than majority baseline. Fine-tuned Qwen3 achieves 0.595 micro-F1, slightly below Llama3; however, still demonstrates improvement over its zero-shot experiment. However, BanglaLLM shows notably larger performance gains, suggesting that language- and domain-specific pretraining offers a clear advantage over general-purpose LLMs. These results show that fine-tuning enables the models to better capture nuanced patterns.

In the *severity of hate* task, Llama3 achieves a micro-F1 score of 0.685, while Qwen3 and BanglaLLM obtain 0.661 and 0.722. All models outperform the majority baseline (0.660) while BanglaLLM shows the best performance along with BanglaBERT, demonstrating the effectiveness of task-specific adaptation in assessing the intensity of hateful content. Similarly, in the *target of hate* task, Llama3 reaches a micro-F1 score of 0.610, with Qwen3 and BanglaLLM achieving 0.598 and 0.631, marking a clear improvement over zero-shot performance. These results demonstrate that LoRA fine-tuning enables the models to capture subtle contextual cues that indicate the intended target of hate. Overall, LoRA fine-tuning effectively transforms pre-trained LLMs into task-aware classifiers, narrowing the performance gap with classical models like SVM and pretrained models such as BanglaBERT, while providing a scalable approach for hate speech detection in low-resource languages.

5.5 Additional Experiments: LoRA with Chain-of-Thought (CoT)

We performed a cross-domain experiment on BD-SHS dataset using BanglaLLM, and details are provided in Appendix D. We have observed that BanglaLLM performs noticeably lower than dataset-trained baselines, particularly on the Target classification task.

To further explore the reasoning capabilities of LLMs in hate speech detection, we conducted additional experiments using CoT prompting and LoRA fine-tuning. We generated CoT using *Gemini-2.5-pro* with detailed definitions (see Listing 5) for each task. We then fine-tuned the Llama and Qwen models using LoRA, using the same hyperparameters; however, the Qwen model was fine-tuned for 3 epochs to see if training longer improves the performance. The results are summarized in Table 5.

Across all tasks, Qwen consistently outperforms Llama, demonstrating stronger baseline reasoning and contextual understanding. Moreover, Qwen fine-tuned for three epochs achieves the better performance compared without using CoT. We provide the prompts in Listing 6.

Models	Acc.	P	R	F1
Type of Hate				
Llama3	0.570	0.600	0.570	0.570
Qwen	0.601	0.619	0.601	0.601
Qwen*	0.634	0.658	0.634	**0.634**
Severity of Hate				
Llama3	0.624	0.648	0.624	0.624
Qwen	0.647	0.664	0.647	0.647
Qwen*	0.665	0.688	0.665	**0.665**
Target of Hate				
Llama3	0.586	0.606	0.586	0.586
Qwen	0.616	0.630	0.616	0.616
Qwen*	0.630	0.656	0.630	**0.630**

Table 5: Performance of fine-tuned LLMs using LoRA with CoT. Acc.: Accuracy, P.: Precision, R.: Recall, F1: micro-F1 score. * indicates model trained to 3 epochs. **Bold** indicates the best results for their respective task.

5.6 Findings

Does language-specific pretraining improve Bangla hate-speech classification across tasks? *BanglaBERT* achieves the highest scores on all three tasks, indicating that pretraining on linguistically and culturally relevant Bangla data is most effective, especially for fine-grained distinctions such as severity and target.

Are zero-shot LLMs sufficient, or is task-specific fine-tuning required? Zero-shot approaches are insufficient for Bangla hate speech. Task-specific fine-tuning substantially improves LLMs performance. Fine-tuned *Llama3* is a promising alternative, whereas *Qwen3* exhibits weaker gains, suggesting differences in pretraining data and alignment.

How do fine-tuned LLMs compare to monolingual PLMs? Fine-tuned LLMs performs reasonably, however, do not surpass *BanglaBERT*. This shows that the continued importance of language-specific pretraining for reliable detection in low-resource settings.

What dataset properties most affect evaluation? Class imbalance inflates baseline performance (e.g., majority-class predictions, particularly for the severity task). Robust evaluation should report macro-F1 and per-class metrics and consider stratified splits.

Does task-specific training help uniformly across tasks? Task-specific training improves performance on all three tasks, with the largest practical benefits for nuanced categories (e.g., severity levels and target groups), where generic zero-shot models struggle.

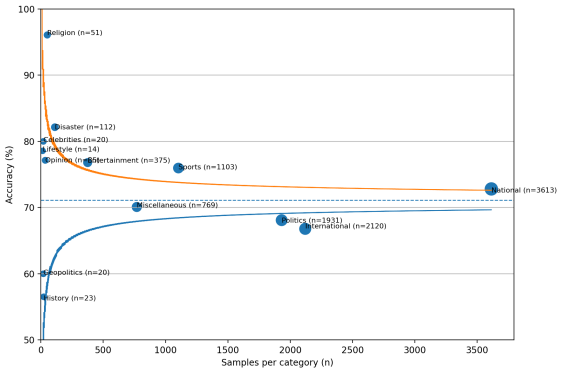

Figure 5: Category-wise accuracy for the *hate type* task using BanglaBERT model. The categories with less than 10 samples are excluded.

Category-wise model accuracy. To examine whether categories are uniform within a task, Figure 5 plots category accuracy (y) *vs.* sample size (x) for the *hate type* task using *BanglaBERT*. The dashed line marks the overall accuracy ($\approx$ 71%); curves show 95% binomial control limits around this mean. Most categories lie within the limits, indicating that residual differences are consistent with sampling variability given n. Among high-support groups, *national* is slightly above the mean, while *international* and *politics* are modestly below; *sports* and *entertainment* are near or slightly above the mean.Categories such as *religion*, *disaster*, *celebrities*, *lifestyle*, *opinion* appear higher but remain inconclusive due to wide uncertainty. Overall, the funnel indicates that the task-specific gains are broadly uniform across major categories. More details of these analyses are reported in Appendix E.

6 Conclusions and Future Work

In this study, we present *BanglaMultiHate*, a Bangla hate-speech dataset that is among the largest manually annotated corpora in a multi-task setting. The dataset comprises approximately 51K instances spanning 19 topics and 120 sub-topics. To demonstrate its utility, we conduct comprehensive experiments comparing classical approaches, PLMs, and LLMs. Our findings demonstrate the importance of language-specific pretraining as well as task-specific fine-tuning for robust performance. As future work, we plan to extend the dataset with reasoning to support task-level interpretability and explanation.

Limitations

This study has several limitations. First, our dataset is collected from YouTube comments, which contain examples that may be disturbing or offensive to readers. During the annotation process, annotators were explicitly cautioned about this content and provided with appropriate warnings. Second, from a modeling perspective, the dataset is highly imbalanced across classes, which may affect both training stability and performance evaluation. Addressing these issues, for example, through data augmentation, re-sampling strategies, or collecting additional underrepresented examples, remains an important direction for future work.

Ethics and Broader Impact

Our dataset consists solely of comments and does not include any personally identifiable user information, thereby posing no direct privacy risks. Nonetheless, it is important to acknowledge that annotation is inherently subjective, which can introduce biases into the dataset. To mitigate this, we designed a clear annotation schema and provided detailed guidelines to annotators, aiming to ensure greater consistency and reliability. However, we encourage researchers and practitioners to remain careful of these limitations when using the dataset for model development or further studies.

Despite these issues, the dataset holds significant potential for positive societal impact. Models trained on *BanglaMultiHate* can support social media platforms in identifying and moderating harmful content, thereby contributing to healthier online discourse.

References

Ahmed Abdelali, Hamdy Mubarak, Shammur Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Samir Abdaljalil, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Youssef Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, and Firoj Alam. 2024. LAraBench: Benchmarking Arabic AI with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 487–520, St. Julian's, Malta. Association for Computational Linguistics.

Abdullah Al Maruf, Ahmad Jainul Abidin, Md Mahmudul Haque, Zakaria Masud Jiyad, Aditi Golder, Raaid Alubady, and Zeyar Aung. 2024. Hate speech detection in the bengali language: a comprehensive survey. *Journal of Big Data*, 11(1):97.

Firoj Alam, Md Rafiul Biswas, Uzair Shah, Wajdi Zaghouani, and Georgios Mikros. 2024a. Propaganda to hate: A multimodal analysis of arabic memes with multi-agent llms. In *International Conference on Web Information Systems Engineering*, pages 380–390. Springer.

Firoj Alam, Shammur Absar Chowdhury, Sabri Boughorbel, and Maram Hasanain. 2024b. LLMs for low resource languages in multilingual, multimodal and dialectal settings. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, St. Julian's, Malta. Association for Computational Linguistics.

Aish Albladi, Minarul Islam, Amit Das, Maryam Bigonah, Zheng Zhang, Fatemeh Jamshidi, Mostafa Rahgouy, Nilanjana Raychawdhary, Daniela Marghitu, and Cheryl Seals. 2025. Hate speech detection using large language models: A comprehensive review. *IEEE Access*.

Bharath B and S. Ajith A. 2021. SSNCSE_NLP@DravidianLangTech-EACL2021: Offensive language identification on multilingual code mixing text. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, DravidianLangTech*, pages 313–318. Association for Computational Linguistics.

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.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P McCrae. 2022. Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. *Language Resources and Evaluation*, 56(3):765–806.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Mithun Das, Somnath Banerjee, Punyajoy Saha, and Animesh Mukherjee. 2022a. 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.

Mithun Das, Somnath Banerjee, Punyajoy Saha, and Animesh Mukherjee. 2022b. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, Minneapolis, Minnesota, USA.

Mouadh Djandji, Freddy Baly, Wissam Antoun, and Hady Hajj. 2020. Multi-task learning using arabert for offensive language detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection (OSACT4)*, pages 97–101. European Language Resource Association.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Azizul Hakim Fayaz, MD. Shorif Uddin, Rayhan Uddin Bhuiyan, Zakia Sultana, Md. Samiul Islam, Bidyarthi Paul, Tashreef Muhammad, and Shahriar Manzoor. 2025. BIDWESH: A bangla regional based hate speech detection dataset.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *Acm Computing Surveys (Csur)*, 51(4):1–30.

Katharine Gelber. 2021. Differentiating hate speech: a systemic discrimination approach. *Critical Review of International Social and Political Philosophy*.

Vikram Gupta, Sumegh Roychowdhury, Mithun Das, Somnath Banerjee, Punyajoy Saha, Binny Mathew, Hastagiri Prakash Vanchinathan, and Animesh Mukherjee. 2022. Macd: Multilingual abusive comment detection at scale for indic languages. In *36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks*.

Fabiha Haider, Fariha Tanjim Shifat, Md Farhan Ishmam, Deeparghya Dutta Barua, Md Sakib Ul Rahman Sourove, Md Fahim, and Md Farhad Alam. 2024. Banth: A multi-label hate speech detection dataset for transliterated bangla. *arXiv preprint arXiv:2410.13281*.

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, Shudipta Das, Afiyat Anjum, Firoj Alam, Anika Anjum, Avijit Sarker, and Sheak Rashed Haider Noori. 2024. Zero-and few-shot prompting with llms: A comparative study with fine-tuned models for bangla sentiment analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17808–17818.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

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

Md Saroar Jahan, Mainul Haque, Nabil Arhab, and Mourad Oussalah. 2022. BanglaHateBERT: BERT for abusive language detection in Bengali. In *Proceedings of the Second International Workshop on Resources and Techniques for User Information in Abusive Language Analysis*, pages 8–15, Marseille, France. European Language Resources Association.

Aiqi Jiang and Arkaitz Zubiaga. 2024. Cross-lingual offensive language detection: A systematic review of datasets, transfer approaches and challenges. *arXiv preprint arXiv:2401.09244*.

A. J. Keya, M. M. Kabir, N. J. Shammey, M. F. Mridha, M. R. Islam, and Y. Watanobe. 2023. G-bert: An efficient method for identifying hate speech in bengali texts on social media. *IEEE Access*, 11:79697–79709.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems (NeurIPS) 33*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1).

Md. Firoj Mridha, Md. Abul Hasnat Wadud, Md. Abdul Hamid, Md. Mostafa Monowar, Md. Abdullah-Al-Wadud, and Atif Alamri. 2021. L-boost: Identifying offensive texts from social media post in bengali. *IEEE Access*, 9:164681–164699.

Anze Pelicon, Raghav Shekhar, Matjaz Martinc, Blaž Škrlj, Matthew Purver, and Simon Pollak. 2021. Zero-shot cross-lingual content filtering: Offensive language and hate speech detection. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 30–34. Association for Computational Linguistics.

Paloma Piot and Javier Parapar. 2025. Towards efficient and explainable hate speech detection via model distillation. In *European Conference on Information Retrieval*, pages 376–392. Springer.

John Platt. 1998. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods - Support Vector Learning*. MIT Press.

Md Nishat Raihan, Umma Tanmoy, Anika Binte Islam, Kai North, Tharindu Ranasinghe, Antonios Anastasopoulos, and Marcos Zampieri. 2023. Offensive language identification in transliterated and code-mixed bangla. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 1–6.

Nauros Romim, Mosahed Ahmed, Md Saiful Islam, Arnab Sen Sharma, Hriteshwar Talukder, and Mohammad Ruhul Amin. 2022. BD-SHS: A benchmark dataset for learning to detect online Bangla hate speech in different social contexts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5153–5162, Marseille, France. European Language Resources Association.

Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam. Hate speech detection in the bengali language: A dataset and its baseline evaluation. *IJCACI 2020*, page 457.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*.

Niloofar Safi Samghabadi, Deepthi Mave, Sudipta Kar, and Thamar Solorio. 2018. Ritual-uh at trac 2018 shared task: Aggression identification. *arXiv preprint arXiv:1807.11712*.

Happy Khairunnisa Sariyanto, Diclehan Ulucan, Oguzhan Ulucan, and Marc Ebner. 2025. Towards explainable hate speech detection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12883–12893.

S. Saumya, A. Kumar, and J. P. Singh. 2021a. Offensive language identification in dravidian code mixed social media text. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, DravidianLangTech*, pages 36–45. Association for Computational Linguistics.

Sunil Saumya, Abhinav Kumar, and Jyoti Prakash Singh. 2021b. Offensive language identification in Dravidian code mixed social media text. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 36–45, Kyiv. Association for Computational Linguistics.

Salim Sazzed. 2021. Abusive content detection in transliterated bengali-english social media corpus. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 125–130, Online. Association for Computational Linguistics.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 145–158. Springer.

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, DravidianLangTech*, pages 255–261, Kyiv. Association for Computational Linguistics.

Deepawali Sharma, Tanusree Nath, Vedika Gupta, and Vivek Kumar Singh. 2025. Hate speech detection research in south asian languages: a survey of tasks, datasets and methods. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(3):1–44.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Zeerak Talat and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Joseph B Walther. 2022. Social media and online hate. *Current Opinion in Psychology*, 45:101298.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. 2023. HARE: Explainable hate speech detection with step-by-step reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5490–5505, Singapore. Association for Computational Linguistics.

Anwar Hossain Zahid, Monoshi Kumar Roy, and Swarna Das. 2025. Evaluation of hate speech detection using large language models and geographical contextualization. *arXiv preprint arXiv:2502.19612*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.

Appendix

A Detailed Annotation Guideline

A.1 Definitions

The primary goal of this annotation task is to categorize YouTube comments in the Bangla language into specific categories based on the nature and severity of hate speech they contain, as well as identifying the target of such speech.

Type of Hate Categorization This annotation task involves having annotators annotate Bangla text samples according to the type of hate expressed. Each text is categorized into one of six classes: *Abusive*, *Sexism*, *Religious Hate*, *Political Hate*, *Profane*, or *None*. The goal is to capture the specific nature of hateful content, enabling models to distinguish between different forms of hate speech and non-hateful content. Abusive vs. Profane follows the distinction between targeted abuse and untargeted profanity used in OffensEval and large-scale abusive language datasets (Zampieri et al., 2019). Sexism/Religious/Political hate are treated as identity-/ideology-directed subtypes in line with prior hate-speech corpora (Talat and Hovy, 2016).

- **Abusive:** Comments that are directly insulting, intending to belittle or harm someone's dignity. For example, তুমি একেবারেই অকেজো (*English: You are completely useless*). This comment degrades someone by calling them utterly useless.
- **Political Hate:** Comments that display hostility towards political beliefs, parties, or figures. For example, সব রাজনৈতিক নেতারা চোর (*English: All political leaders are thieves*).
- **Profane:** Comments that use swear words or vulgar language, intended to shock or offend without targeting anyone specifically. For example, শুয়োরের বাচ্চা তোর সাহস অনেক বড় (*English: Son of a pig, you got guts*). This comment uses profanity to express frustration.
- **Religious Hate:** Comments targeting individuals or groups based on their religion or religious beliefs. For example, এই ধর্মের লোকেরা সব খারাপ (*English: All people in this religion are bad*). This comment generalizes a whole religion as bad.
- **Sexism:** Comments that discriminate or belittle someone based on their gender, often reflecting stereotypes. For example, মেয়েদের কেবল রান্না করা উচিত (*Women should cook only*). This comment reinforces the stereotype expression.
- **None:** Comments that do not exhibit hate or negativity, including neutral or positive comments. For example, আমি আজ মুভিটা দেখলাম (*English: I saw the movie today*). This comment do not reflect any hate content.

Severity of Hate This annotation task involves having annotators label Bangla text samples according to the intensity of hateful content expressed. Severity is defined as an ordinal three-level scale- *Little to None*, *Mild*, *Severe*-aligned with aggression and hate-severity work (e.g., non/-covert/overt aggression; non-violent vs. violent hate), with explicit threats or incitement always coded as Severe (Samghabadi et al., 2018). This classification scheme is designed to capture the varying degrees of harmfulness in hate speech.

- **Severe:** Comments that contain threats, extreme prejudice, or are highly offensive. For example, তুই বাসা থেকে বের হ, তোর মত জানোয়ারকে দেকে নিব (*English: Get out of the house, I'll take care of a beast like you*).
- **Mild:** Comments that are derogatory or mildly offensive but do not contain threats.

For example, তুমি তো বেইমান সালা তোর তো মরার সাথে সাথে জাহান্নামের আগুনে জালাবে (*English: You are a traitor, you will burn in hell as soon as you die*).

- **Little to None:** Comments that are slightly negative, ambiguous, or completely neutral or positive. For example, হানিফ পরিবহণ বন্ধো করে দেওয়া হোক (*English: Hanif transport should be stopped*).

Target of Hate This annotation task involves having annotators label Bangla text samples based on the intended target of the hateful expression. The labels are divided into five categories: *Community*, *Individual*, *Organization*, *Society*, and *None*. Target is defined based on who is attacked Individuals, Organizations, Communities, Society, None, mapping to the Individual/Group/Other structure in OLID/OffensEval and multi-aspect resources, with identity attributes (e.g., religion, gender, political ideology) recorded when applicable. This categorization aims to capture whether the hate speech is directed at a specific person, a collective group, broader societal structures, or institutions, while also accounting for instances where no explicit target is present.

- **Community:** Comments against a specific racial, ethnic, gender, or religious group. For example, এই সম্প্রদায়ের মানুষ বিশ্বাস করার যোগ্য নয় (*English: People from this community are not trustworthy*).

- **Individual:** Comments targeting a specific person, either by name or implication. For example, শেখ হাসিনা তুমি চোর চোরের মায়ের বড় গলা (*English: Sheikh Hasina you are thief, the thief's mother has a loud voice*).

- **Organization:** Comments aimed at specific companies, governmental bodies, or any formal group. For example, সময় টেলিভিশন মনে হয় সরকারের (*English: Somoy television seems to be government's*).

- **Society:** Comments that critique societal norms, values, or general community practices. For example, ইতালির লোক ভাত পায়না আবার জন্ম হার বাড়াবে (*English: Italians do not get food, they will increase birth rate again*).

- **None:** Comments that are ambiguous, or completely neutral or positive. For example, এরকম এই হওয়া উচিত যে কোন খেলার ভিতর রাজনীতি নেয়াটাই দুষ্কর (*English: It should be like this: it's hard to bring politics into any game*).

A.2 Analysis and Statistics

We present the detailed class label distribution in Table 7. Moreover, Table 6 presents the distribution of class labels across word length bins for the three tasks.

B Prompts

We provide the instructions used to generate prompts for all three tasks in Listings 1, 2, and 3. Additionally, the prompts employed for zero-shot learning, fine-tuning, and inference are presented in Listing 4.

```
We are creating an English
instruction-following dataset for Type of
Hate hate speech detection.
Read the given text carefully and choose the
most appropriate label for the task from the
label lists.
For the 'Type of Hate' task, the labels are
'Abusive', 'Sexism', 'Religious Hate',
'Political Hate', 'Profane', and 'None'.
Select only one correct label for each task
based on the information provided in the text
and return your response in the following
json format.
{"type_of_hate": "Abusive"}

Write 10 very diverse and concise English
instructions. Only return the instructions
without additional text. Do not generate
additional text.

Return the instructions in a list format as
follows.
['sent1', 'sent2']
```

Listing 1: Prompt for generating instructions for Type of Hate task.

```
We are creating an English
instruction-following dataset for Hate
Severity hate speech detection. Here is an
example instruction:
Read the given text carefully and choose the
most appropriate label for the task from the
label lists.
For the 'Hate Severity' task, the labels are
'Little to None', 'Mild', and 'Severe'.
Select only one correct label for each task
based on the information provided in the text
and return your response in the following
json format.
{"severity_of_hate": "Mild"}

Write 10 very diverse and concise English
instructions. Only return the instructions
```

Split	Task	Label	Word Length Bins					
			<=10	11-20	21-30	31-40	41-50	51+
Train	Type of Hate	Abusive	4374	2438	801	311	115	173
		Political Hate	1614	1415	625	259	139	175
		Profane	1329	624	214	72	45	47
		Religious Hate	281	233	81	42	19	20
		Sexism	57	36	14	9	0	6
		None	12624	4814	1353	508	265	390
	Severity of Hate	Little to None	14433	5840	1699	683	336	498
		Mild	3207	2176	825	307	153	185
		Severe	2639	1544	564	211	94	128
	Target of Hate	Community	1131	888	336	134	62	84
		Individual	3146	1552	539	202	89	118
		Organization	1755	1259	470	175	86	101
		Society	951	694	293	124	58	85
		None	13296	5167	1450	566	288	423
Dev	Type of Hate	Abusive	599	315	121	37	16	25
		Political Hate	212	197	92	33	16	24
		Profane	190	104	26	6	7	9
		Religious Hate	27	35	9	0	4	3
		Sexism	11	7	0	1	0	0
		None	1809	717	185	87	42	58
	Severity of Hate	Little to None	2079	864	239	106	54	75
		Mild	413	300	118	30	23	25
		Severe	356	211	76	28	8	19
	Target of Hate	Community	141	111	49	18	6	13
		Individual	429	203	76	19	12	16
		Organization	261	196	78	23	11	15
		Society	124	92	33	13	11	10
		None	1893	773	197	91	45	65
Test	Type of Hate	Abusive	1224	711	228	68	40	41
		Political Hate	434	431	183	91	23	58
		Profane	371	207	79	32	9	11
		Religious Hate	73	59	28	12	4	3
		Sexism	17	10	1	1	0	0
		None	3645	1390	373	159	59	125
	Severity of Hate	Little to None	4153	1678	484	204	73	145
		Mild	885	680	243	96	38	59
		Severe	726	450	165	63	24	34
	Target of Hate	Community	329	242	106	46	16	20
		Individual	861	468	136	54	25	27
		Organization	494	399	142	59	20	38
		Society	260	206	91	36	11	21
		None	3820	1493	417	168	63	132

Table 6: Detailed class label distribution with word length bin count.

```
without additional text. Do not generate
additional text.

Return the instructions in a list format as
follows.
['sent1', 'sent2']
```

Listing 2: Prompt for generating instructions for Severity of Hate task.

```
We are creating an English
instruction-following dataset for the Target
of Hate task of hate speech detection. Here
is an example instruction:
Read the given text carefully and choose the
most appropriate label for the task from the
label lists."+
For the 'Target of Hate' task, the labels are
'Individuals', 'Organizations',
'Communities', 'Society', and 'None'. Select
only one correct label for the task based on
the information provided in the text and
return your response in the following json
format.
{"type_of_hate": "Society"}
```

Split	Type of Hate	Severity of Hate			Total	Target of Hate					Total
	Class	LN	Mild	Severe		Comm.	Indiv.	Org.	Society	None	
	Abusive	2,254	3,867	2,091	**8,212**	1,521	3,340	1,510	1,118	723	**8,212**
	Political Hate	978	2,237	1,012	**4,227**	404	935	1,896	745	247	**4,227**
	Profane	127	396	1,808	**2,331**	399	1,182	385	183	182	**2,331**
Train	Religious Hate	150	301	225	**676**	283	119	51	147	76	**676**
	Sexism	26	52	44	**122**	28	70	4	12	8	**122**
	None	19,954	–	–	**19,954**	–	–	–	–	19,954	**19,954**
	Total	**23,489**	**6,853**	**5,180**	**35,522**	**2,635**	**5,646**	**3,846**	**2,205**	**21,190**	**35,522**
	Abusive	333	507	273	**1,113**	207	427	251	141	87	**1,113**
	Political Hate	141	292	141	**574**	43	130	270	101	30	**574**
	Profane	25	63	254	**342**	52	171	58	23	38	**342**
Dev	Religious Hate	16	36	26	**78**	28	18	4	17	11	**78**
	Sexism	4	11	4	**19**	8	9	1	1		**19**
	None	2,898	–	–	**2,898**	–	–	–	–	2,898	**2,898**
	Total	**3,417**	**402**	**698**	**5,024**	**338**	**755**	**584**	**283**	**3,064**	**5,024**
	Abusive	646	1,075	591	**2,312**	441	891	465	306	209	**2,312**
	Political Hate	263	702	255	**1,220**	124	263	541	231	61	**1,220**
	Profane	41	116	552	**709**	127	354	127	51	50	**709**
Test	Religious Hate	29	93	57	**179**	62	47	17	34	19	**179**
	Sexism	7	15	7	**29**	5	16	2	3	3	**29**
	None	5,751	–	–	**5,751**	–	–	–	–	5,751	**5,751**
	Total	**6,737**	**2,001**	**1,462**	**10,200**	**759**	**1,571**	**1,152**	**625**	**6,093**	**10,200**

Table 7: Class label distribution of the dataset. LN: Little to None, Comm.: Community, Indiv.: Individual, Org.: Organization

```
Write 10 very diverse and concise English
instructions. Only return the instructions
without additional text. Do not generate
additional text.

Return the instructions in a list format as
follows.
['sent1', 'sent2']
```

Listing 3: Prompt for generating instructions for Target of Hate task.

```
You are a Bangla AI assistant specialized in
the hate speech detection task. Your task is
to identify the correct label for the task.


Read the text and assign the correct labels
for the type of hate, severity of hate, and
target of hate. Return only the answer
without any explanation, justification, or
additional text.
For the 'Type of Hate' task, the labels are
'Abusive', 'Sexism', 'Religious Hate',
'Political Hate', 'Profane', and 'None'.
For the 'Severity of Hate' task, the labels
are 'Little to None', 'Mild', and 'Severe'.
And for the 'Target of Hate' task, the labels
are 'Individuals', 'Organizations',
'Communities', 'Society', and 'None'.
Select only one correct label for each task
based on the information provided in the text
and return your response in the following
JSON format.

{
    "type_of_hate": "Abusive",
    "severity_of_hate": "Mild",
    "target_of_hate": "Society"
}

If you select the 'None' label for the 'Type
of Hate' task, the labels for the 'Severity
of Hate' and 'Target of Hate' tasks would be
'Little to None' and 'None'.
```

Listing 4: Sample Prompt for zero-shot learning, model fine-tuning, and inference.

```
You are a Bangla AI assistant specialized in
the hate speech detection task. Your task is
to identify the correct label for the task.


<think>
### General Instructions
1. Understand the definitions for each task:
type of hate, severity of hate, and target of
hate.
    * Type of Hate: this annotation task
    involves having annotators annotate
    Bangla text samples according to the type
    of hate expressed. Each text is
    categorized into one of six classes:
    Abusive, Sexism, Religious Hate,
    Political Hate, Profane, or None.
```

```
    {Here we put the annotation definition
    from Appendix A.1.}

    * Severity of Hate: this annotation task
    involves having annotators label Bangla
    text samples according to the intensity
    of hateful content expressed. The
    severity is categorized into three
    levels: Severe, Mild, and Little to None.
        {Here we put the annotation definition
        from Appendix A.1.}

    * Target of Hate: this annotation task
    involves having annotators label Bangla
    text samples based on the intended target
    of the hateful expression. The labels are
    divided into five categories: Community,
    Individual, Organization, Society, and
    None.
        {Here we put the annotation definition
        from Appendix A.1.}

###Task Instructions
2. **Analyze the input text and labels**
    * Read the content carefully.
    * Read the provided labels carefully.
    * Generate a synthetic Chain-of-thought
    thinking style dataset in Bangla using
    the given definitions, input text, and
    labels.
    * Do not mention the labels in
    Chain-of-thought.
3. Provide explanations in the following JSON
format, and explanations should be in plain
text:
{
    "bangla_cot": []
}
```

Listing 5: Sample Prompt for CoT generation.

```
[
    {
    "role": "system",
    "content": "You are a Bangla AI assistant
    specialized in the hate speech detection
    task. Your task is to identify the
    correct label for the task."
    },
    {
    "role": "user",
    "content": """Begin by confirming you
    understand the three annotation schemes
    (Type, Severity, Target). For each input
    comment, determine the most appropriate
    single label from each list below:
        * Type: Abusive, Sexism, Religious
        Hate, Political Hate, Profane, None.
        * Severity: Little to None, Mild,
        Severe.
        * Target: Individual, Organization,
        Community, Society, None.

    After labeling, produce a concise
    Chain-of-thought in plain text describing
    the cues in the comment (words, context,
    implied target, intensity) that led to
    your choices. Format and return the
    result as the JSON example below.

    {
        "Chain-of-thought": "",
        "Labels": {
            "type_of_hate": "",
            "severity_of_hate": "",
            "target_of_hate": ""
        }
    }
    Input text: {input_text}"""
    },
    {
    "role": "assistant",
    "content": """{
        "Chain-of-thought": "",
        "Labels": {
            "type_of_hate": "",
            "severity_of_hate": "",
            "target_of_hate": ""
        }
    }"""
    }
]
```

Listing 6: Sample Prompt for CoT model fine-tuning, and inference.

C Training vs Validation Loss

We analyze the loss dynamics during fine-tuning to assess training stability and generalization. As shown in Figure 6, both training and validation losses decrease steadily across all three epochs and remain closely aligned throughout, with no observable divergence. This consistent trend indicates stable optimization and provides no evidence of overfitting during fine-tuning.

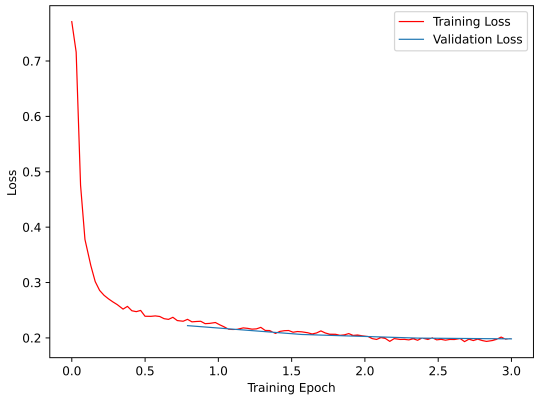

Figure 6: Training vs Validation loss of *Qwen3*.

D Cross-Domain Evaluation

We also performed a cross-domain experiment using BanglaLLM on the BD-SHS dataset. As shown in Table 8, BanglaLLM performs noticeably lower than dataset-trained baselines, particularly on the Target classification task. This reinforces our main claim: LLMs and fine-tuned

Model	Task	Acc.	P.	R.	F1
Baseline	Hate	–	0.901	0.901	0.901
BanglaLLM		0.803	0.804	0.803	0.803
Baseline	Type	–	0.773	0.681	0.721
BanglaLLM		0.630	0.690	0.630	0.630
Baseline	Target	–	0.885	0.857	0.868
BanglaLLM		0.482	0.822	0.482	0.482

Table 8: Cross-Domain performance on BD-SHS dataset using BanglaLLM. Acc.: Accuracy, P.: Precision, R.: Recall, F1: micro-F1 score.

models exhibit domain sensitivity in Bangla hate speech, and transferring from YouTube discourse to social-media corpora remains challenging.

E Detailed Result Analysis

Category-wise Performance We present the category-wise performance of BanglaBERT and Gemini in Figures 7 and 8, respectively. As shown, BanglaBERT demonstrates more consistent and better performance across categories, particularly in linguistically complex or sensitive domains such as National and Religion. In contrast, Gemini exhibits greater variability across domains, performing competitively in general or informal categories like Entertainment and Sports but lagging in culturally nuanced contexts such as National and Politics. These results highlight the advantage of in-language pretraining for capturing domain-specific and culturally grounded expressions of hate speech. Moreover, both models performed poorly in the History category, indicating their limited understanding of historical information.

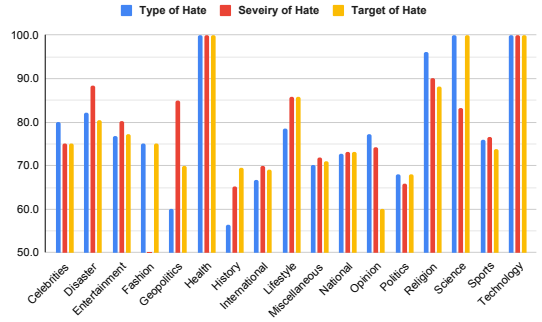

Figure 7: Category-wise result analysis of *BanglaBERT*.

Error Analysis We also conduct a class-wise performance analysis across several models. The results reveal substantial variation in how models handle low-frequency categories, with aggregate metrics such as micro-F1 often masking important class-specific deficiencies. For instance, BanglaBERT fails to correctly predict any instances of the Sexism class, whereas GPT-5 correctly identifies 18 out of 19 Sexism samples, indicating strong sensitivity to this rare category. However, this advantage does not generalize: GPT-5 performs poorly on more frequent yet semantically heterogeneous classes such as Abusive (818/2312 correct) and Profane (206/709 correct), highlighting pronounced inconsistency across categories. We observe similar patterns for the Gemini-2.5-pro and Qwen3 models, which also exhibit uneven class-wise performance—showing sensitivity to certain low-frequency categories while struggling with more frequent, semantically diverse classes. These findings demonstrate that strong performance on rare classes does not necessarily imply robust overall classification. To better illustrate these class-wise error patterns, we present the confusion matrices for Type, Severity, and Target of Hate tasks using BanglaBERT and GPT-5 in Table 9, 10, 11, 13, 12, and 14 respectively. Moreover, we also include representative qualitative examples illustrating common failure patterns for BanglaBERT and GPT-5 in Figure 9. These cases highlight how models often misinterpret po-

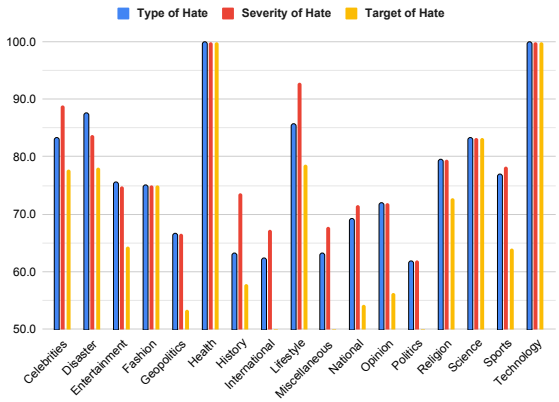

Figure 8: Category-wise result analysis of *Gemini-2.5-pro*.

Class	Abus.	None	Pol. Hate	Prof.	Rel. Hate	Sexism
Abusive	1275	638	245	39	115	0
None	742	4641	269	55	44	0
Political Hate	245	196	727	9	43	0
Profane	129	33	28	516	3	0
Religious Hate	27	40	7	5	100	0
Sexism	14	13	1	0	1	0

Table 9: Confusion matrix of BanglaBERT for the Type of Hate task. Column headers are abbreviated as follows: Abus. = Abusive, Pol. Hate = Political Hate, Prof. = Profane, and Rel. Hate = Religious Hate.

litically charged language as general abuse or incorrectly infer hate in contexts involving sensitive geopolitical actors.

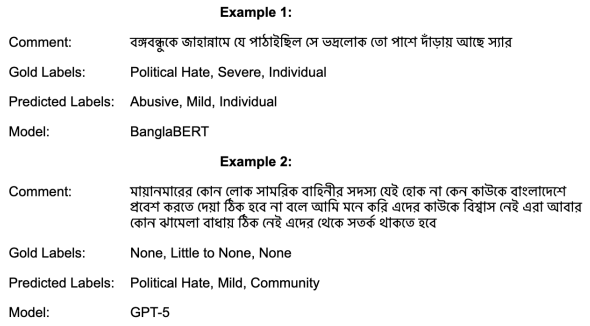

Figure 9: Category-wise result analysis of *BanglaBERT* and *GPT-5*.

Class	Little to None	Mild	Severe
Little to None	5,684	813	240
Mild	685	864	452
Severe	244	401	817

Table 10: Confusion matrix of BanglaBERT for the Severity of Hate task.

Class	Community	Individual	None	Organization	Society
Community	357	73	224	81	24
Individual	84	999	390	77	21
None	230	400	5,023	261	179
Organization	94	103	253	661	41
Society	87	38	181	65	254

Table 11: Confusion matrix of BanglaBERT for the Target of Hate task.

	Little to None	Mild	Severe
Little to None	4,610	1,807	320
Mild	271	1,224	506
Severe	89	572	800

Table 12: Confusion matrix of GPT-5 for the Severity of Hate task.

	Abus.	Comm.	None	Pol. Hate	Prof.	Rel. Hate	Sex.
Abusive	818	3	497	772	32	139	51
Communities	0	0	0	0	0	0	0
None	464	1	4,291	698	25	210	62
Political Hate	31	0	138	1,014	5	31	1
Profane	242	0	26	173	206	38	24
Religious Hate	1	0	6	15	0	157	0
Sexism	4	0	7	0	0	0	18

Table 13: Confusion matrix of GPT-5 for the Type of Hate task. Column header abbreviations are as follows: Abus. = Abusive, Comm. = Communities, Pol. Hate = Political Hate, Prof. = Profane, Rel. Hate = Religious Hate, and Sex. = Sexism.

	Comm.	Ctry.	Indiv.	None	Org.	Pol.	Soc.
Community	267	0	252	79	156	0	5
Countries	0	0	0	0	0	0	0
Individual	89	0	1,113	252	97	0	20
None	462	2	709	4,405	485	1	29
Organization	79	0	155	154	757	1	6
Politics	0	0	0	0	0	0	0
Society	214	8	69	110	198	1	25

Table 14: Confusion matrix of GPT-5 for the Target of Hate task. Column header abbreviations are as follows: Comm. = Community, Ctry. = Countries, Indiv. = Individual, Org. = Organization, Pol. = Politics, and Soc. = Society.

F Data Release

We made *BanglaMultiHate* dataset publicly available under the CC BY-NC-SA 4.0.[9]

[9] https://creativecommons.org/licenses/by-nc-sa/4.0/