

BANTH: A Multi-label Hate Speech Detection Dataset for Transliterated Bangla

Fabiha Haider^{1*}, Fariha Tanjim Shifat^{1*}, Md Farhan Ishmam^{1,2*}, Deeparghya Dutta Barua¹, Md Sakib Ul Rahman Sourovo¹, Md Fahim^{1,3}, Md Farhad Alam¹

¹Research and Development, Penta Global Limited, Bangladesh

²Islamic University of Technology, Bangladesh

³CCDS Lab, Independent University, Bangladesh

pdcsedu@gmail.com, fahimcse381@gmail.com

Abstract

The proliferation of transliterated texts in digital spaces has emphasized the need for detecting and classifying hate speech in languages beyond English, particularly in low-resource languages. As online discourse can perpetuate discrimination based on target groups, e.g. gender, religion, and origin, multi-label classification of hateful content can help in comprehending hate motivation and enhance content moderation. While previous efforts have focused on monolingual or binary hate classification tasks, no work has yet addressed the challenge of multi-label hate speech classification in transliterated Bangla. We introduce BANTH, the first multi-label transliterated Bangla hate speech dataset comprising 37.3k samples. The samples are sourced from YouTube comments, where each instance is labeled with one or more target groups, reflecting the regional demographic. We establish novel transformer encoder-based baselines by further pre-training on transliterated Bangla corpus. We also propose a novel translation-based LLM prompting strategy for transliterated text. Experiments reveal our further pre-trained encoders achieving state-of-the-art performance on the BANTH dataset while our translation-based prompting outperforms other strategies in the zero-shot setting. The introduction of BANTH not only fills a critical gap in hate speech research for Bangla but also sets the stage for future exploration into code-mixed and multi-label classification challenges in underrepresented languages.

Content Warning: This article contains examples of hateful content.

Note: Throughout the work, we use the term *Bangla* to refer to both *Bangla* and the endonym *Bengali*, both denoting the same language primarily spoken by people of the West Bengal region of India and the vast majority of Bangladesh.

*Equal Contribution

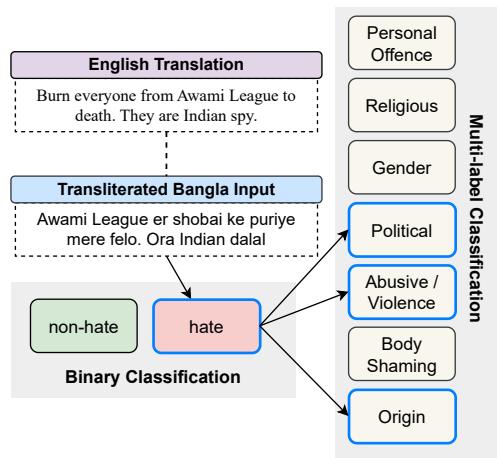


Figure 1: Binary and Multi-label Classification on an example Transliterated Bangla sentence from the BANTH dataset along with its corresponding English translation.

1 Introduction

The ever-expanding digital landscape that promises to improve social cohesion has been a breeding ground for hate speech. Hate speech is defined as any form of language that targets, attacks, or incites implicit or explicit forms of hatred or violence against groups, based on specific characteristics, e.g. physical appearance, religion, national or ethnic origin, and gender identity. Hate speech can potentially inflict societal harm by promoting division among communities, exacerbating mental health problems, and inciting violence. The lack of hate speech moderation cultivates an environment of intolerance and magnifies the negative impact on the target communities (Hangartner et al., 2021). Furthermore, categorizing hate speech requires a multi-faceted approach to capture overlapping hate categories and gain a granular understanding of the underlying motives behind the hateful discourse.

A prevalent informal form of online communication for non-English languages uses transliterated

texts, i.e. writing one language in the script of another. Transliteration is predominant in Bangla, where Latin characters are used to produce a colloquial form of Bangla texts that does not strictly adhere to the original linguistic rules. Despite boasting over a quarter of a billion speakers worldwide, Bangla remains a low-resource language in terms of Natural Language Processing (NLP) tasks. The challenges are compounded when working with the even more under-resourced transliterated Bangla, which dominates online spaces due to the widespread familiarity of users with English keyboard layouts.

Although transliteration facilitates easier typing and cross-linguistic communication, it also complicates the task of automated hate speech detection due to inconsistent spelling and structure, absence of grammar, mixing with English, and the loss of script-specific features (Jahan et al., 2019a). While standard practices have employed transformer-based encoders (Devlin et al., 2019) for automated hate speech detection, the recent surge in popularity of Large Language Models (LLMs) has positioned them as a viable option in hate speech NLP, particularly in the zero-shot setting. While most of the advancements have been centered on English and other high-resource languages, transliterated Bangla has very limited research on hate or hate-like speech detection (Jahan et al., 2019b; Raihan et al., 2023) and LLM-based methods (Shibli et al., 2023), with no work addressing the intersection of these areas.

Addressing the aforementioned research gap, our contribution can be summarized as follows:

- We propose BANTH, the first multi-label hate speech detection dataset on transliterated Bangla with 37,350 samples.
- We establish several encoder-based baselines along with our novel transliterated Bangla further-pretrained encoders, achieving state-of-the-art accuracy on our dataset.
- We explore zero-shot and few-shot prompting techniques using state-of-the-art Large Language Models and introduce a novel translation-based prompting strategy that outperforms existing methods on our dataset.

2 BANTH Dataset

The BANTH dataset comprises 37350 samples, each initially classified into binary labels of *Hate* or

Source Category	#Hate	#Non-hate
News & Politics	8745	23695
People & Blogs	1125	2889
Entertainment	412	584
Total	10282	27167
Dataset Statistics		
Min Character Count		7
Max Character Count		1951
Average Length		58.09
Standard Deviation		65.13
Total Words		379411
Average Words		10.16
Min Word Count		3
Max Word Count		368
Unique Words		75977
BANTH split		
- Train		29879
- Val		3736
- Test		3735
Total		37350

Table 1: Dataset statistics of the BANTH dataset. Min Length and Max Length are the minimum and maximum sentence lengths in terms of the number of characters, respectively. The Std Length indicates the standard deviation of the number of characters per sentence.

Non-Hate. To effectively capture the target group, the hate samples are further multi-labeled as *Political*, *Religious*, *Gender*, *Personal Offense*, *Abusive/Violence*, *Origin*, and *Body Shaming*. A detailed description of the target group is provided in Appendix C.1. The flowchart of the dataset creation is given in Fig. 2.

2.1 Data Sourcing

We construct the dataset by scraping user comments from public YouTube videos using the YouTube API¹. The scraped comments include three categories of videos: “News & Politics”, “People & Blogs”, and “Entertainment”, totaling 26 different YouTube channels, outlined in Table 1. The contents often cover a wide range of topics, including political analysis, interviews, talk shows, product reviews, and travel videos, regionally relevant to West Bengal of India and Bangladesh.

¹<https://developers.google.com/youtube/v3/getting-started>

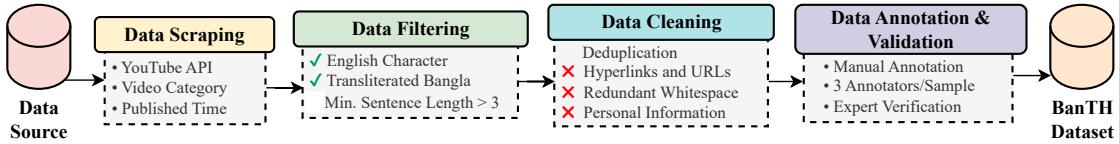


Figure 2: Overview of the BANTH dataset creation pipeline.

The scraped comments were collected from January 2020 to July 2024, a time frame that captures significant recent events, e.g. the COVID-19 pandemic and the Bangladesh quota reform movement². Such events led to increased social media activity. The variety of content appeals to a diverse audience with a wide range of interests and perspectives. The label-wise distribution of video categories is elaborated in Appendix C.2.

2.2 Data Filtering

We primarily filtered comments containing English characters. During annotation, non-Bangla and non-English transliterated comments, such as comments in Hindi, were discarded. The scraped comments were filtered through English character filtering and were further filtered using dictionary-based language identification. Original English and Bangla and code-mixed texts were filtered out using the dictionary filtering approach. Transliterated texts in any other language such as Hindi, were dropped during annotation. Only the transliterated Bangla texts with lengths greater than or equal to three were kept in the dataset.

2.3 Data Cleaning

During data cleaning, we removed URLs using regular expressions as they were not relevant to our datasets. Comments that were verbatim duplicates of others were removed. Personal information like addresses and contact numbers were also removed during annotation.

2.4 Data Annotation

We hired four annotators and two domain experts for the data annotation process. All the annotators were high school graduates and native Bangla speakers. They had extensive exposure to social media content and actively used transliterated Bangla text. The annotators were compensated fairly for their work, with a rate of BDT 1 per sample. The domain experts were remunerated

hourly. The annotation process was supervised by domain experts with experience in working with transliterated Bangla text, who provided guidance and ensured clear and proper annotation guidelines, outlined in Appendix B.

Each data sample was annotated by three annotators to ensure consistency and capture a shared perspective. Each data sample was first instructed to be classified as hate or non-hate and each resulting hate sample was then instructed to be multi-labeled according to the target group descriptions given above. In cases where annotators encountered difficulties in reaching a consensus or understanding the text, these issues were addressed and instantly resolved by the domain experts to avoid further confusion.

Fleiss' Kappa	Expert-Annotator	Inter-Annotator
Binary Labeling	0.75	0.71
Political	0.62	0.72
Religious	0.68	0.66
Gender	0.64	0.68
Personal Offense	0.70	0.75
Abusive/Violence	0.72	0.71
Origin	0.68	0.72
Body Shaming	0.64	0.66

Table 2: Fleiss' Kappa Score of annotators and domain experts across categories.

2.5 Data Validation

Following the annotation process, the dataset underwent a thorough review by the domain experts to ensure labeling accuracy. The experts verified whether any samples remained unlabeled, whether any samples were mislabeled, or if there were any instances where a sample was labeled as hate without the corresponding multi-label, and vice versa. Additionally, the dataset was examined by the domain experts to ensure that all samples were in transliterated Bangla, as samples in other languages were considered beyond the scope of the study and were subsequently removed. The agreement between the annotators and the domain experts, and the inter-annotator was measured using the Fleiss'

²https://en.wikipedia.org/wiki/2024_Bangladesh_quota_reform_movement

Kappa metric and the result is shown the Table 2. An inter-annotator agreement score of 0.53 indicates a moderate agreement across the dataset (Islam et al., 2021).

2.6 Dataset Statistics

Following data validation, the dataset is split into train, test, and val datasets maintaining the standard 80:10:10 ratio. During the split, a stratification approach was followed to maintain a uniform distribution of hate and non-hate labels and further multi-label categories across all the splits. Further data statistics are given in Table 1.

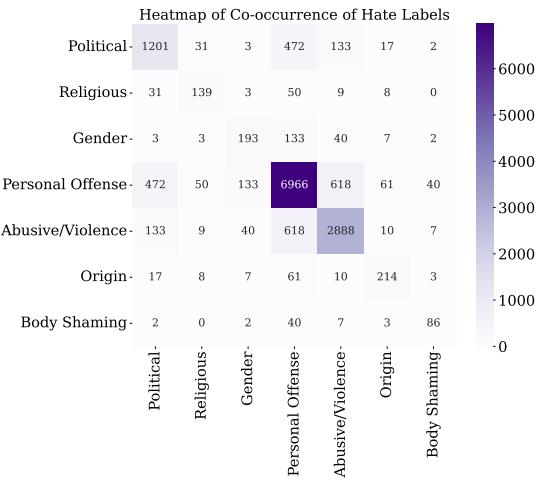


Figure 3: Heatmap of co-occurrence of multi-label Hate in the BANTH dataset.

The heatmap in Figure 3 provides a detailed visualization of multi-label co-occurrence within the BANTH dataset, illustrating the frequency with which pairs of labels appear together in the text data. Each cell in the heatmap represents the count of co-occurrence between two specific labels, with darker shades of purple indicating higher levels of co-occurrence. The diagonal cells depict the total occurrences of each label independently, with *Personal Offense* emerging as the most frequent category, appearing in 6966 samples. Conversely, *Body Shaming* is the least frequent label in the dataset. Significant co-occurrence patterns are observed between *Personal Offense* and *Abusive/Violence* with 618 samples, as well as between *Personal Offense* and *Political* with 472 samples. In contrast, labels such as *Religious* and *Body Shaming* demonstrate minimal co-occurrence with other categories, reflecting their relative isolation in the dataset. The heatmap underscores the prominence of *Personal Offense* in the BANTH dataset, indicating that a

substantial portion of the text data is classified under this label.

3 Methodology

We use various baseline approaches to assess the quality of BANTH, establishing the benchmark performance of it. The experiments are carried out in two stages: Binary Classification of the dataset, followed by Multi-label classification of the hate speech predicted in the first stage. For evaluation, we follow two approaches: one involves finetuning the pre-trained LMs on our corpus, BANTH, and the other involves prompting state-of-the-art large GPT-based LLMs. The following sections discuss the experiments in detail.

3.1 Further Pretraining

Pretrained LMs such as BERT (Devlin et al., 2019) and BanglaBERT (Bhattacharjee et al., 2022) are typically trained on general and formal texts. However, these models often lack transliterated samples in their pretraining datasets. To enhance the performance of these models, we employ the technique of *Further Pretraining (FPT)* (Qiu et al., 2021; Gururangan et al., 2020), utilizing the Masked Language Modeling (MLM) loss (Devlin et al., 2019; Zhuang et al., 2021) as our pretraining objective. This approach aims to improve the models’ effectiveness on our BANTH dataset.

Specifically, for a given sentence S , we represent the input tokens as $X = (x_1, x_2, \dots, x_n)$. We randomly mask 15% of these tokens with the *[MASK]* token, resulting in a masked sequence denoted as \tilde{X} . Let $M \subseteq \{1, 2, \dots, n\}$ represent the set of indices of the masked tokens. We then input \tilde{X} into the language model f_{LM} . For each masked token x_i (where $i \in M$), the model generates predicted probabilities $P(x_i|\tilde{X})$. The loss for each masked token is calculated as $L_i = -\log P(x_i|\tilde{X})$. The overall loss across all masked tokens is given by:

$$L_{MLM} = -\frac{1}{|M|} \sum_{i \in M} \log P(x_i|\tilde{X})$$

For our FPT process, we utilize the BanglaTLit-PT corpus³, which contains 243k unlabeled Bangla transliterated texts. After further pretraining the models on BanglaTLit-PT, we build Bangla transliteration-enhanced encoder models namely TB-Encoders (Transliterated Bangla Encoders).

³<https://www.kaggle.com/datasets/farihatanjimshifat1/bangla-transliteration-further-pretraining-dataset>

3.2 LM Finetuning

In our fine-tuning experiments, we utilize a pre-trained language model f_{LM} that has been specifically trained on our dataset. We input a sentence S into f_{LM} to obtain layer-wise contextual representations $\mathbf{H} = \{\mathbf{h}_i^l\}_{l=1}^L$, where $\mathbf{H} = f_{LM}(S)$ and L = no. of layers in f_{LM} . We use the representation from the last layer's [CLS] token, h_{CLS}^L , as the sentence representation, which is then passed through a Multi-Layer Perceptron (MLP) for classification. The transformation is defined as follows:

$$z = W_2 \cdot (\text{ReLU}(W_1 \cdot h_{CLS}^L + b_1)) + b_2$$

The resulting representation z is then employed for calculating the loss. In the case of TB-Encoders, we also similarly fine-tune those models after doing FPT on the BanglaTLit-PT dataset.

3.3 Prompting Strategy

We adopt one of the following strategies: Why Positive/Why Negative (Wang et al., 2023), HARE (Yang et al., 2023), non-explanatory, explanation-based, CoT based or translation based. An overview of the prompts has been reported in Tab. 3. We consider GPT 4o and GPT-3.5 in the prompting experiments and have designed separate base prompts for binary and multi-label classification. Each of these base prompts is then extended to one of the prompting strategies. We consider both *zero-shot* and *few-shot* prompting techniques, reported in App. E. Overall, all the variations of the prompts can be categorized as follows:

Non-Explanatory. The non-explanatory prompting strategy forms the base prompts for other prompting strategies, where we only ask the LLM to do binary or multi-label classification. The prompt also includes the hate speech definition, the target groups, and their definition in multi-label classification, key indicators of hate, and labeling instructions.

Explanation Based. In the explanation-based prompting approach, we include the question *Explain why* with the base prompt and ask the LLM to return its reasoning.

HARE. In HARE-based prompting strategy, we insert the line *Let's explain step by step*, as adapted in (Yang et al., 2023), along with the base-prompt and asked the LLM to return its reason for the

corresponding binary or multi-label.

Why Positive Why Negative. As explained in (Wang et al., 2023), we adopt the approach by adding the question *why the given text is positive or negative* appending to the base prompt in this variation.

CoT-Based. CoT-based prompting strategy is implemented by adding the question *Let's think step by step* along with the base prompt.

Translation Based. LLMs are trained primarily on extensive formal text, which may limit their performance and explanatory capabilities when handling transliterated text. To address this issue, we propose a simple yet effective prompting strategy called the *Translation-Based* prompting strategy. This approach requires LLMs to first convert Bengali transliterated text into standard Bangla or English before proceeding with the usual classification. We test this proposed prompting method using Non-Explanatory, Explanatory, and CoT-based prompting techniques.

All the prompts are provided in Appendix E.

4 Experimental Result

We conduct experiments on fine-tuning LMs along with utilizing prompt-based techniques. Our experiment setup for finetuning LMs is described in Appendix D and the prompts are reported in Appendix E. Table 5 shows the benchmarking of various models on our dataset for both binary and multi-label classification tasks.

4.1 Fine-Tuning LMs Baseline

For the binary classification, among the fine-tuned models, TB-mBERT achieves the highest accuracy (82.57%) and Macro-F1 score (77.36%). The best-performing BERT variant is TB-BanglaBERT, with an accuracy of 81.61% and Macro-F1 of 77.12%. In multi-label classification, for Macro-F1, TB-BERT leads with 30.17%, followed closely by XLM-RoBERTa (28.24%). In subset accuracy, TB-mBERT performs best (54.71%), with TB-BERT close behind (54.19%). Hamming Loss is lowest (best) for TB-BanglaBERT (7.26). From the experiments, our proposed TB-encoders beat the other LMs for both binary and multi-label classification.

Prompt-Strategy	Prompt
Non-Explanatory	Base-prompt
Chain of Thought (CoT)	Base-prompt + "Let's think step by step"
Explanation-based (Exp)	Base-prompt + "Explain why"
HARE	Base-prompt + "Let's explain step by step"
Why [Positive]	Base-prompt + "Explain why the comment is positive"
Why [Negative]	Base-prompt + "Explain why the comment is negative"
Translation Based [BAN]	"Translate the following transliterated text into standard Bangla" + Prompt-Strategy
Translation Based [ENG]	"Translate the following transliterated text into standard English" + Prompt-Strategy

Table 3: The prompting strategies used in our benchmarks. We consider binary/multi-label classification and zero-shot/few-shot approaches along with their corresponding prompts.

Model	Label	Macro-F1 \uparrow	Acc. \uparrow
GPT 3.5	Positive	62.48	69.03
	Negative	62.71	66.77
GPT 4o	Positive	63.33	74.91
	Negative	69.98	75.29

Table 4: Performance of Why prompting on the BanTH test split for binary classification.

4.2 Prompt Based LLM Performance

Among prompt-based LLMs, GPT-4o + Few-shot performed the best for multi-label classification, achieving the highest Macro-F1 (39.53%), Subset Accuracy (26.76%), and lowest Hamming Loss (14.16%). In binary classification, GPT 4o+Exp+Few-shot has the highest Macro-F1 (70.70%). In terms of accuracy, our Translation Prompting approach gains the highest result (74.87%). If we consider *zero-shot* prompting techniques, the proposed approach consistently outperforms all other prompt variations.

Our extensive experiments reveal that fine-tuned models outperform all prompt-based results for both binary and multi-label classification with a good margin (~8% improvement over prompting). The *few-shot* prompting strategy enhances model performance compared to the *zero-shot* approach, although this is not true for all cases. Additionally, GPT-4 consistently outperforms GPT-3.5 across all experiments and prompting variations, yielding improvements of 8-10%. We also find that generating explanations can occasionally boost performance, but it may also lead to worse results in some instances. Our proposed prompting techniques achieve the best performance in *zero-shot* settings. Overall, our proposed TB encoders demonstrate superior performance, surpassing all fine-tuned LMs and prompt-based results.

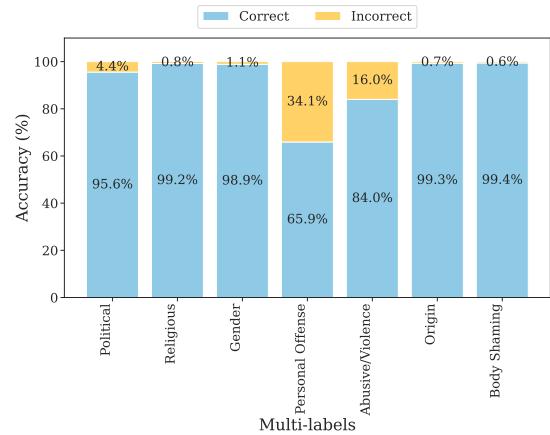


Figure 4: Category-wise classification Macro-F1 score in the BANTH dataset with the best-performing model.

5 Error Analysis

Figure 4 illustrates per-label accuracy with the best performing model, TB-mBert. From the fig, it is evident that the classification accuracy for all the labels except *Personal Offense* label. Figure 3 indicates that most hate speech labels are categorized under *Personal Offense*. Consequently, the subset accuracy of the multi-label classification, as presented in Table 5, is lower than the per-label accuracy. This suggests that the model struggles to effectively classify samples of *Personal Offense* label. As illustrated by Fig. 5, the minimum accuracy for binary and multi-label classification is at least 80% and 50%, respectively, with *Entertainment* videos holding the highest accuracy in both the classifications, showing higher subset accuracy in multi-label classification (79.5%).

It is evident from Tab. 6 that the number of *Personal Offense* in *Entertainment* video type is comparable to that of the other multi-labels of hate, whereas in the other two categories, *News & Politics* and *People & Blogs*, the number of *Personal Offense* hate is noticeably high. The overall im-

Models	Binary		Multi-label		
	Macro-F1 ↑	Acc. ↑	Macro-F1 ↑	Subset Acc. ↑	Hamming Loss ↓
Language Model (LM) Fine-tuning					
Bangla LM					
BanglaBERT	76.50	81.04	20.82	54.08	7.26
BanglighBERT	75.07	80.62	20.61	52.74	7.42
BanglaHateBERT	70.92	77.54	11.34	49.83	7.97
VAC-BERT	74.19	79.76	18.45	52.56	7.50
Indian LM					
MuRIL	75.29	80.83	14.58	53.98	7.55
IndicBERT	74.51	80.48	13.39	51.56	7.88
Multilingual LM					
mBERT	74.97	80.37	28.24	52.19	7.78
XLM-R	77.35	81.37	29.29	53.28	7.23
Character-based LM					
CharBERT	76.61	80.91	19.66	53.21	7.44
Transliterated Bangla (TB) LM (Ours)					
TB-BERT	76.27	79.25	30.17	54.19	7.18
TB-mBERT	77.36	82.57	27.07	54.71	7.28
TB-XLM-R	77.04	81.29	29.04	52.86	7.26
TB-BanglaBERT	77.12	81.61	22.52	53.97	7.37
TB-BanglighBERT	77.12	81.39	21.41	52.79	7.42
Large Language Model (LLM) Prompting					
Non-Explanatory					
GPT 3.5	61.44	64.02	24.12	18.27	19.77
GPT 4o	70.05	74.30	36.07	22.60	16.91
GPT 3.5 + Few-shot	61.85	65.65	23.86	16.85	20.36
GPT 4o + Few-shot	68.77	74.18	39.53	26.76	14.16
Chain of Thought (CoT) Prompting					
GPT 3.5 + CoT	61.87	65.77	25.61	20.34	19.28
GPT 4o + CoT	69.87	74.14	35.97	22.60	16.64
GPT 3.5 + CoT + Few-shot	63.30	67.80	28.35	20.15	19.15
GPT 4o + CoT + Few-shot	69.50	74.94	36.49	21.10	17.36
Explanation-based (Exp) Prompting					
GPT 3.5 + Exp	61.69	65.16	22.61	17.33	19.84
GPT 4o + Exp	69.91	74.16	32.58	19.12	17.94
GPT 3.5 + Exp + Few-shot	62.69	66.60	25.00	18.36	19.57
GPT 4o + Exp + Few-shot	70.70	75.15	35.61	21.33	17.15
HARE Prompting					
GPT 3.5 + HARE	62.64	66.77	25.99	20.90	18.94
GPT 4o + HARE	69.71	74.75	36.67	20.80	17.10
GPT 3.5 + HARE + Few-shot	62.20	66.72	24.68	20.06	19.18
GPT 4o + HARE + Few-shot	69.12	74.75	34.06	21.59	17.37
Translation Prompting (Ours)					
GPT 4o + Translation [BAN]	69.63	72.70	35.72	23.28	16.73
GPT 4o + Translation [ENG]	69.38	72.96	34.28	20.72	16.91
GPT 4o + Translation [BAN] + CoT	69.01	72.66	36.22	21.35	17.36
GPT 4o + Translation [ENG] + CoT	69.74	74.87	35.94	21.19	16.91
GPT 4o + Translation [BAN] + Exp	<u>70.34</u>	73.65	36.20	21.54	16.46
GPT 4o + Translation [ENG] + Exp	70.19	73.84	<u>36.25</u>	23.92	<u>16.23</u>

Table 5: Model Benchmarking on BANTH Dataset on test split for Binary and Multi-label Classification. For prompting LLMS the results show the zero-shot and few-shot approaches for each of the strategies. All the Translation Prompting is designed following a zero-shot approach. **Bold** represents overall the best results whereas Underline represents the best performing results in *zero-shot* settings.

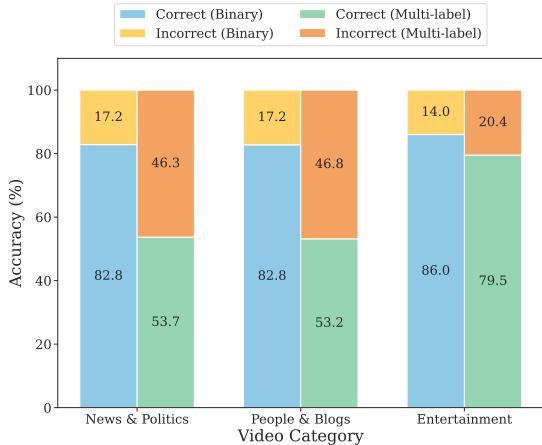


Figure 5: Video category-wise class classification accuracy on the BanTH dataset with the best-performing model. We consider Macro-Accuracy for Binary and Subset-Accuracy for Multi-label Classification.

proved accuracy in *Entertainment* category aligns with the result produced in Fig. 4.

The results in Tab. 5 show that our Translation-based models, like GPT 4o Bangla/English Translation with explanations, perform nearly as well as top prompt-based models. As observed in Tab. 7, LLMs can naturally translate Bangla transliterated text into English, even without explicit instructions. However, when asked to translate, the LLM tends to give more literal translations. We also observe that poor translations often lead to incorrect classifications, especially with regional dialects or terms lacking direct English equivalents. English translations generally outperform Bangla ones. When not prompted to translate, the LLM focuses on identifying hate-related terms, which works for explicit hate speech but struggles with subtler forms of offensive content, leading to misclassification.

6 Related Work

Hate Speech. Natural Language Processing (NLP) tasks on hate speech primarily involve hate speech detection (Schmidt and Wiegand, 2017) which can be binary (Mutanga et al., 2020; Bose and Su, 2022), multi-class (Yigezu et al., 2023; Hashmi and Yayilgan, 2024), or multi-label classification (Mollas et al., 2022). Multi-class and multi-label tasks generally categorize hate based on race, religion, gender, origin, and disability. Hate speech detection has also been explored in multi-modal (Gomez et al., 2020; Kiela et al., 2020) and multi-lingual (Aluru et al., 2020; Ousidhoum et al., 2019) settings.

Bangla Hate Speech. As a low-resource language, Bangla has limited work on hate speech detection (Romim et al., 2022; Das et al., 2022a) but has made significant progress on other hate-related tasks involving abusive content (Hussain et al., 2018; Jahan et al., 2022a), cyberbullying (Emon et al., 2022), sexism (Jahan et al., 2023), and violence (Fahim, 2023). The multi-label variants are generally categorized based on religion, politics, and gender (Sharif et al., 2022). Key areas of multi-labeling tasks include aggression (Hossain et al., 2023), cyberbullying (Saifuddin et al., 2023), hate speech (Shakil, 2022), and toxic content (Belal et al., 2023).

Transliterated Bangla. Research on NLP tasks using transliterated or romanized Bangla is notably scarce, with limited works on back-transliteration (Shibli et al., 2023), abusive content detection (Jahan et al., 2019b; Sazzed, 2021), hate speech detection (Das et al., 2022a), offensive language identification (Raihan et al., 2023), sentiment analysis (Hassan et al., 2016), and cyberbullying detection (Ahmed et al., 2021). While multi-label classification tasks on transliterated Bangla have been explored as sentiment and emotion analysis tasks (Tripto and Ali, 2018), there remains a significant research gap on multi-label hate speech classification for transliterated Bangla.

7 Conclusion

Our novel multi-label hate speech dataset, BANTH, addresses a crucial research gap in the domain of transliterated Bangla. Extensive experimentations on both traditional language models and LLMs highlight their usefulness in their unique settings. We envision our novel prompting strategies to be generalized to the processing of transliterated text in other languages. We believe our dataset can serve as a valuable resource in the creation of safer digital platforms for the future.

Limitations

The BANTH dataset contains 27167 non-hate samples out of 37350, i.e. a dummy classifier predicting all samples as non-hate achieves an accuracy of 72.73%. This undermines the performance of a few methods, notably, GPT-3.5 performs worse than this baseline. We also recognize that our work does not cover all possible hate classes and may

not adequately represent other regionally relevant hate classes, e.g. hate based on socio-economic conditions.

Ethical Considerations

The use of comments scraped from the YouTube videos complies with the YouTube API’s terms of services⁴. All forms of Personal Identification Information (PII) have been removed from the dataset to prevent privacy violations. We ensured fair distribution of annotation workload. The hired annotators and domain experts were compensated on an hourly basis at a rate above the industry standard.

References

Md Tofael Ahmed, Maqsudur Rahman, Shafayet Nur, Azm Islam, and Dipankar Das. 2021. Deployment of machine learning and deep learning algorithms in detecting cyberbullying in bangla and romanized bangla text: A comparative study. In *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pages 1–10. IEEE.

Tasnim Ahmed, Shahriar Ivan, Mohsinul Kabir, Hasan Mahmud, and Kamrul Hasan. 2022. Performance analysis of transformer-based architectures and their ensembles to detect trait-based cyberbullying. *Social Network Analysis and Mining*, 12(1):99.

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*.

Faisal Yousif Al Anezi. 2022. Arabic hate speech detection using deep recurrent neural networks. *Applied Sciences*, 12(12):6010.

Tanveer Ahmed Belal, GM Shahriar, and Md Hasanul Kabir. 2023. Interpretable multi labeled bengali toxic comments classification using deep learning. In *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pages 1–6. IEEE.

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.

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

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media*, pages 36–41, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Saugata Bose and Dr. Guoxin Su. 2022. Deep one-class hate speech detection model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7040–7048, Marseille, France. European Language Resources Association.

Patricia Chiril, Farah Benamara, and Véronique Moriceau. 2021. “be nice to your wife! the restaurants are closed”: Can gender stereotype detection improve sexism classification? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2833–2844, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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, Punyajoy Saha, Binny Mathew, and Animesh Mukherjee. 2022b. HateCheckHIn: Evaluating Hindi hate speech detection models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5378–5387, Marseille, France. European Language Resources Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

⁴<https://developers.google.com/youtube/terms/api-services-terms-of-service>

Md Imdadul Haque Emon, Khondoker Nazia Iqbal, Md Humaion Kabir Mehedi, Mohammed Julfikar Ali Mahbub, and Annajiat Alim Rasel. 2022. Detection of bangla hate comments and cyberbullying in social media using nlp and transformer models. In *International Conference on Advances in Computing and Data Sciences*, pages 86–96. Springer.

Md Fahim. 2023. [Aambela at BLP-2023 task 1: Focus on UNK tokens: Analyzing violence inciting Bangla text with adding dataset specific new word tokens](#). In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 201–207, Singapore. Association for Computational Linguistics.

Md Fahim, Md Shihab Shahriar, Bangladesh Gazipur, and Mohammad Ruhul Amin. Hateexplain space model: Fusing robustness with explainability in hate speech analysis.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1470–1478.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don’t stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Dominik Hangartner, Gloria Gennaro, Sary Alasiri, Nicholas Bahrich, Alexandra Bornhoft, Joseph Boucher, Buket Buse Demirci, Laurenz Derkßen, Aldo Hall, Matthias Jochum, et al. 2021. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, 118(50):e2116310118.

Ehtesham Hashmi and Sule Yildirim Yayiglan. 2024. Multi-class hate speech detection in the norwegian language using fast-rnn and multilingual fine-tuned transformers. *Complex & Intelligent Systems*, 10(3):4535–4556.

Asif Hassan, Mohammad Rashedul Amin, Abul Kalam Al Azad, and Nabeel Mohammed. 2016. Sentiment analysis on bangla and romanized bangla text using deep recurrent models. In *2016 International Workshop on Computational Intelligence (IWCI)*, pages 51–56. IEEE.

Nimali Hettiarachchi, Ruvan Weerasinghe, and Randil Pushpanda. 2020. Detecting hate speech in social media articles in romanized sinhala. In *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 250–255. IEEE.

Jawad Hossain, Avishek Das, Mohammed Moshiul Hoque, and Nazmul Siddique. 2023. Multilabel aggressive text classification from social media using transformer-based approaches. In *2023 26th International Conference on Computer and Information Technology (ICCIT)*, pages 1–6. IEEE.

Md Gulzar Hussain, Tamim Al Mahmud, and Waheda Akthar. 2018. An approach to detect abusive bangla text. In *2018 International Conference on Innovation in Engineering and Technology (ICIET)*, pages 1–5. IEEE.

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.

Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. 2021. [SentNoB: A dataset for analysing sentiment on noisy Bangla texts](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3265–3271, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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

Md Saroor Jahan, Mainul Haque, Nabil Arhab, and Mourad Oussalah. 2022a. 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.

Md Saroor Jahan, Mainul Haque, Nabil Arhab, and Mourad Oussalah. 2022b. 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.

Sarif Sultan Saruar Jahan, Raqeebir Rab, Peom Dutta, Hossain Muhammad Mahdi Hassan Khan, Muhammad Shahriar Karim Badhon, Sumaiya Binte Hasan, and Ashikur Rahman. 2023. Deep learning based misogynistic bangla text identification from social media. *Computing and Informatics*, 42(4):993–1012.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. [IndicNLP Suite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages](#). In *Findings of EMNLP*.

Tae Young Kang, Eunrang Kwon, Junbum Lee, Youngeun Nam, Junmo Song, and JeongKyu Suh.

2022. Korean online hate speech dataset for multilabel classification: How can social science improve dataset on hate speech? *arXiv preprint arXiv:2204.03262*.

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

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. *Advances in neural information processing systems*, 33:2611–2624.

Jean Lee, Taejun Lim, Heejun Lee, Bogeun Jo, Yangsok Kim, Heegeun Yoon, and Soyeon Caren Han. 2022. **K-MHaS: A multi-label hate speech detection dataset in Korean online news comment**. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3530–3538, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Wentao Ma, Yiming Cui, Changlei Si, Ting Liu, Shijin Wang, and Guoping Hu. 2020. **Charbert: Character-aware pre-trained language model**. *Preprint*, arXiv:2011.01513.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hateexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigoris Tsoumacas. 2022. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, 8(6):4663–4678.

Aida Mostafazadeh Davani, Ali Omrani, Brendan Kennedy, Mohammad Atari, Xiang Ren, and Morteza Dehghani. 2021. **Improving counterfactual generation for fair hate speech detection**. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 92–101, Online. Association for Computational Linguistics.

Raymond T Mutanga, Nalindren Naicker, and Oluayo O Olugbara. 2020. Hate speech detection in twitter using transformer methods. *International Journal of Advanced Computer Science and Applications*, 11(9).

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. **Multilingual and multi-aspect hate speech analysis**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Hrushikesh Patil, Abhishek Velankar, and Raviraj Joshi. 2022. **L3Cube-MahaHate: A tweet-based Marathi hate speech detection dataset and BERT models**. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 1–9, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Zhengqi Pei, Zhewei Sun, and Yang Xu. 2019. **Slang detection and identification**. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 881–889, Hong Kong, China. Association for Computational Linguistics.

Flor Miriam Plaza-del Arco, M Dolores Molina-González, L Alfonso Urena-López, and M Teresa Martín-Valdivia. 2021. Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, 166:114120.

Rahul Pradhan, Ankur Chaturvedi, Aprna Tripathi, and Dilip Kumar Sharma. 2020. A review on offensive language detection. *Advances in Data and Information Sciences: Proceedings of ICDIS 2019*, pages 433–439.

Yao Qiu, Jinchao Zhang, and Jie Zhou. 2021. Different strokes for different folks: Investigating appropriate further pre-training approaches for diverse dialogue tasks. *arXiv preprint arXiv:2109.06524*.

Md Nishat Raihan, Umma Hani Tanmoy, Anika Binte Islam, Kai North, Tharindu Ranasinghe, Antonios Anastasopoulos, and Marcos Zampieri. 2023. Offensive language identification in transliterated and code-mixed bangla. *arXiv preprint arXiv:2311.15023*.

Tharindu Ranasinghe and Marcos Zampieri. 2020. **Multilingual offensive language identification with cross-lingual embeddings**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online. Association for Computational Linguistics.

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.

Md Saifuddin, Mohiuddin Ahmed, Spandan Basu, and Pritam Acharjee. 2023. Enhancing online safety:

Natural language processing based multi-label cyberbullying classification in bangla. In *2023 26th International Conference on Computer and Information Technology (ICCIT)*, pages 1–6. IEEE.

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.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Mahmudul Hasan Shakil. 2022. *A hybrid deep learning model and explainable AI-based Bengali hate speech multi-label classification and interpretation*. Ph.D. thesis, Brac University.

Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2022. M-bad: A multilabel dataset for detecting aggressive texts and their targets. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 75–85.

GM Shahriar Shibli, Md Tanvir Rouf Shawon, Anik Hassan Nibir, Md Zabed Miandad, and Nibir Chandra Mandal. 2023. Automatic back transliteration of romanized bengali (banglish) to bengali. *Iran Journal of Computer Science*, 6(1):69–80.

Fariha Tanjim Shifat, Fabiha Haider, MSUR Souro, Deeparghya Dutta Barua, Md Farhan Ishmam, Md Fahim, and Farhad Alam Bhuiyan. 2024. Penta-nlp at exist 2024 task 1–3: Sexism identification, source intention, sexism categorization in tweets. *Working Notes of CLEF*.

Manuel Tonneau, Diyi Liu, Samuel Fraiberger, Ralph Schroeder, Scott Hale, and Paul Röttger. 2024. From languages to geographies: Towards evaluating cultural bias in hate speech datasets. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 283–311, Mexico City, Mexico. Association for Computational Linguistics.

Nafis Irtiza Tripto and Mohammed Eunus Ali. 2018. Detecting multilabel sentiment and emotions from bangla youtube comments. In *2018 international conference on Bangla speech and language processing (ICBSLP)*, pages 1–6. IEEE.

Grigoris Tsoumacas and Ioannis Katakis. 2009. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3:1–13.

Kanishk Verma, Tijana Milosevic, Keith Cortis, and Brian Davis. 2022. Benchmarking language models for cyberbullying identification and classification from social-media texts. In *Proceedings of the First Workshop on Language Technology and Resources for a Fair, Inclusive, and Safe Society within the 13th Language Resources and Evaluation Conference*, pages 26–31, Marseille, France. European Language Resources Association.

Han Wang, Ming Shan Hee, Md Rabiul Awal, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2023. Evaluating gpt-3 generated explanations for hateful content moderation. *Preprint*, arXiv:2305.17680.

Mengzhou Xia, Anjolie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, Online. Association for Computational Linguistics.

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. *Preprint*, arXiv:2311.00321.

Mesay Gemedu Yigezu, Olga Kolesnikova, Grigori Sidorov, and Alexander F Gelbukh. 2023. Transformer-based hate speech detection for multi-class and multi-label classification. In *IberLEF@SEPLN*.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Additional Related Work

Hate-like Tasks. Hate speech lacks a consistent definition across the literature and has significant overlap with similar detection tasks involving offensive language (Pradhan et al., 2020; Ranasinghe and Zampieri, 2020), sexism (Chiril et al., 2021; Shifat et al., 2024), abusive content (Nobata et al., 2016), slang (Pei et al., 2019), and cyberbullying (Verma et al., 2022; Ahmed et al., 2022). The studies on hate speech also focused on areas, such as explainability (Mathew et al., 2021; Fahim et al.), fairness (Mostafazadeh Davani et al., 2021), and bias mitigation (Xia et al., 2020).

Non-English Hate Speech. While the majority of the hate speech datasets are in US English (Tonneau et al., 2024), the task has been explored in several non-English languages including Arabic (Anezi, 2022), Hindi (Das et al., 2022b), Indonesian (Ibrohim and Budi, 2019), Korean (Kang et al., 2022; Lee et al., 2022), Marathi (Patil et al., 2022), and Spanish (Plaza-del Arco et al., 2021). Due to the prevalence of English

alphabets in digital spaces, some of these works involve transliterated (Hettiarachchi et al., 2020) and code-mixed texts (Bohra et al., 2018).

B BanTH Dataset Annotation Guidelines

This document outlines the procedures required for accurate annotation of texts within the BanTH dataset. Annotators are expected to follow each step meticulously to ensure consistency and precision. Your role as an annotator is critical to maintaining the integrity and quality of the dataset.

Binary Classification - Hate/Non-Hate: Texts containing offensive, abusive, or harmful language targeting individuals or groups based on their identity, characteristics, or affiliations are to be labeled hate. If the text does not include offensive or harmful content it is to be labeled non-hate.

Steps:

1. Review the text in transliterated Bangla.
2. Classify it as either Hate Speech or Non-Hate Speech.
3. If the comment is Non-Hate, stop here. No further labeling is necessary.

Multi-Label Hate Speech Classification: For texts classified as Hate Speech, assign one or more relevant labels from the following categories. Each text may fit multiple categories, so assign all that apply.

- **Political:** Texts that attack, criticize, or promote hate based on political affiliations, views, or ideologies. This includes language dehumanizing political opponents or encouraging harm.

Example:

[BN] "Oder rajniti desh ta noshto korche."
[EN] "Their politics are destroying the country."

- **Religious:** Texts targeting someone's religious beliefs or lack thereof, or that promote violence or discrimination based on religion.

Example:

[BN] "Ei dhormo manushkei bhag kore."
[EN] "This religion only divides people."

- **Gender:** Texts that reinforce harmful gender stereotypes or discriminate based on gender

identity or expression. Includes sexist or transphobic remarks.

Example:

[BN] "Meyera kaj korte pare na, barir kaji kore."

[EN] "Women can't work, they only do household chores."

- **Personal Offense:** Deeply insulting or offensive remarks directed at an individual, which do not target a specific group but are highly personal.

Example:

[BN] "Tui ekta boka chele."

[EN] "You're a stupid boy."

- **Abusive/Violence:** Texts that contain explicit threats of violence or encourage violent acts against individuals or groups.

Example:

[BN] "Tor matha guriye debo."

[EN] "I'll smash your head."

- **Origin:** Texts targeting someone based on their nationality, ethnicity, or race. This includes racial slurs or calls for exclusion based on origin.

Example:

[BN] "Ei desher manush er kono kotha nai."

[EN] "People from this country don't matter."

- **Body Shaming:** Texts that criticize or mock someone's physical appearance, such as body size or traits, including physical disabilities or conditions.

Example:

[BN] "Tor pet eto boro kano?"

[EN] "Why is your stomach so big?"

Each sample will be annotated by three different annotators. After all of this, domain experts will verify the labels to ensure quality.

Notes: Firstly, carefully read and understand each text before assigning labels. Note that, a sample can receive multiple labels if it fits into more than one category. Please ensure cross-verification by collaborating with fellow annotators. If the problem is not solved even after collaboration then contact domain experts because accuracy is essential for a high-quality dataset.

C Additional Details of BANTH

C.1 Formal Description of Target Groups

The description of the target group with suitable examples is given in the following:

Political: Statements that marginalize, threaten, or incite violence against individuals or groups based on their political affiliations or ideologies. Includes calls for harming political opponents, dehumanizing metaphors applied to political groups, or the deliberate spread of disinformation intended to foment hatred.

Religious: Expressions targeting individuals or groups due to their religious beliefs, practices, or lack thereof. Includes language that demonizes religious communities, advocates for discrimination against religious groups, or incites violence toward religious institutions or adherents.

Gender: Extends beyond sexist remarks and includes language promoting gender-based violence, discrimination based on gender identity, misogynistic or misandrist content, transphobic speech, and language reinforcing harmful gender stereotypes.

Personal Offense: Deeply insulting or offensive language directed at an individual. Includes mocking tragedies, using insulting names, or making derogatory comments.

Abusive/Violence: Explicit threats of violence, detailed descriptions of harm one wishes to inflict on others, or speech that glorifies or encourages violent acts against individuals or groups.

Origin: Expressions that target individuals or groups based on their national, ethnic, or racial background. Includes racial slurs, promotion of racist/nationalist ideologies, calls for racial segregation or deportation, or speech that attributes negative characteristics to groups.

Body Shaming: Derogatory comments on a person's physical appearance, often related to weight, shape, or size. Can extend to mocking people with physical disabilities, skin conditions, or other visible physical traits.

C.2 Category-wise Multi-label Distribution

Multi-labels	#News & Politics	#People & Blogs	#Entertainment
Political	1005	174	21
Religious	109	30	0
Gender	151	38	4
Personal Offense	5883	784	297
Abusive/Violence	2491	276	121
Origin	201	11	2
Body Shaming	56	29	1

Table 6: Distribution of multi-label hate speech among the YouTube video categories of the BANTH dataset.

The distribution of multi-labels within the BanTH dataset in Table 6 reveals notable variances in the prevalence of various types of hate speech across video categories, including News & Politics, People & Blogs, and Entertainment. Table 1 shows that the News & Politics video category holds a major share in the BanTH dataset, so it is less surprising that the number of hate samples of each label will be greater in that. The *Personal Offense* and *Body Shaming* labels are the most and least prominent in the News & Politics category, with 5883 and 56 samples respectively. In fact, the *Personal Offense* label is prominent in every video category.

C.3 Wordcloud

From Figure 6 and the corresponding English translations in Figure 7, we can visualize the mostly mentioned words for each label in the BanTH dataset. For *Political* label, people often referred to Hasina (ex-prime minister of Bangladesh), bnp (a major political party in Bangladesh), league/lig (wing of a political party in Bangladesh), dalal (broker), police (a law enforcement force), etc. As we have covered transliterated Bangla YouTube comments of July 2024, the BanTH dataset includes the public responses related to the political unrest of Bangladesh during that period. In *Religious* labeled sample texts, people often mention Muslim, Hindu nastik (atheist), Allah (the one and only God in Islam), etc. *Gender* related hate comments are mostly related to females. In *Personal Offense* case, commenters address dalal (broker) and police (a law enforcement force) contemptuously using tui/tor. The same case is evident in *Abusive/Violence* labeled samples because the police showed violence during the student protest of July 2024. As transliterated Bangla is used by mainly Indian and Bangladeshi people, *Origin* related hate is often directed towards

Bangladeshi and Indian people. For *Body Shaming* classified samples, we see people refer to others as mota (fat) and takla (bald) to body shame and it is directed mostly towards females as apu (sister) is one of the most used words. Although the BanTH dataset is concentrated on these types of hateful words, it may not be the case when the source is different.

D Experimental Setup

For Language Models (LMs) baseline experiments, we fine-tuned a range of pre-trained LMs to establish baselines for binary and multi-label classification on the BanTH dataset. The models include multilingual architectures like XLM-RoBERTa(Conneau et al., 2019) and mBERT(Devlin et al., 2018), as well as language-specific models such as BanglaBERT(Bhattacharjee et al., 2022) and BanglishBERT(Bhattacharjee et al., 2022), which are designed for Bangla and code-mixed transliterated Bangla text, respectively. We also employed BanglaHateBERT(Jahan et al., 2022b), specifically optimized for detecting hate speech in Bangla, along with IndicBERT(Kakwani et al., 2020) and MuRIL(Khanuja et al., 2021), tailored for Indian languages. Additionally, we included models like CharBERT(Ma et al., 2020), and VAC-BERT(Bhattacharyya et al., 2023) to capture diverse model architectures, including lightweight models and character-level representations.

The models were fine-tuned with a learning rate of 2e-5, using the Adam optimizer. The maximum sequence length was set to 512 tokens, and training was performed for 5 epochs with a batch size of 32. Early stopping was applied based on the validation loss to prevent overfitting. For prompt-based experiments, we used OpenAI API to get LLM responses for GPT-3.5 and GPT-4o models. For measuring the performance, we consider Macro-F1 & Accuracy metrics for binary classification and Macro-F1, Subset-Accuracy & Hamming Loss for multilabel classification. In our further pre-training experiments, we randomly masked 15% of the tokens and used Masked Language Model loss (MLM loss) as the pretraining objectives. We used a learning rate of 1e-5, with batch_size = 32 and we trained the the TB-Encoder models for 5 epochs on the whole training dataset.

D.1 Performance Metrics

F1-Score. Harmonic mean of precision and recall, providing a balance between the two. It is particularly useful in cases of imbalanced datasets, as it accounts for both false positives and false negatives. For multi-label classification, the F1-Score is computed per label, and the final score is obtained by averaging across all labels.

Macro-Accuracy. For binary classification, we report macro-accuracy, which averages the accuracy across both classes (positive and negative), providing a balanced view of the model’s performance across the two categories. This is especially relevant when dealing with class imbalance, as it ensures that both classes are equally represented in the final accuracy measure.

Subset-Accuracy. Requires an exact match between the predicted and true label sets for each instance, meaning that all labels must be predicted correctly for a prediction to be considered accurate (Tsoumakas and Katakis, 2009).

Hamming Loss Measures the fraction of labels that are incorrectly predicted, either by being wrongly assigned or missed altogether. This metric is useful for multi-label classification as it accounts for both false positives and negatives across all labels. A lower Hamming Loss value indicates better performance as it signifies fewer labeling mistakes.

E Prompts for Classification on BanTH dataset

Base Prompt for Binary Classification (0 shot)

You are an expert at detecting hate speech in Bangla text (written in Latin script/English letters). Your task is to classify texts as either hate speech (true) or non-hate speech (false).

HATE SPEECH DEFINITION:

Text that expresses or incites hatred, discrimination, or violence against individuals or groups based on:

- Political affiliation
- Religion (e.g., Hindu, Muslim, Buddhist, Christian)
- Gender
- Personal offense
- Abusive or violence
- Origin
- Body Shaming

KEY INDICATORS OF HATE SPEECH:

1. Dehumanizing language or comparisons
2. Calls for violence or harm
3. Discriminatory slurs or epithets
4. Stereotyping entire groups negatively
5. Promoting supremacy of one group over others

LABELING INSTRUCTIONS:

1. For each text, determine if it contains hate speech indicators
2. If any hate indicators are present, classify as hate speech (true)
3. If no hate indicators are present, classify as non-hate speech (false)

SPECIAL CONSIDERATIONS:

- Consider local context and cultural references
- Be mindful of dialectal variations in spelling
- Pay attention to code-mixing (Bangla-English hybrid phrases)
- For ambiguous cases, focus on the presence of hate indicators

You will receive an array of objects, each containing an 'id' and 'text'. Analyze each text and provide the required classification.

Base Prompt for Multilabel Classification (0 shot)

You are an expert at detecting and categorizing hate speech in Bangla text (written in Latin script/English letters). Your task is to classify texts into multiple categories of hate speech.

HATE SPEECH CATEGORIES:

1. POLITICAL: Targeting individuals or groups based on political affiliation or beliefs
2. RELIGIOUS: Targeting individuals or groups based on religion (e.g., Hindu, Muslim, Buddhist, Christian)
3. GENDER: Targeting individuals or groups based on gender identity or expression
4. PERSONAL OFFENSE: Insults or attacks directed at specific individuals
5. ABUSIVE/VIOLENCE: Language that is abusive or incites violence
6. BODY SHAMING: Targeting individuals based on physical appearance or body type
7. ORIGIN: Targeting individuals or groups based on their place of origin, ethnicity, or nationality
8. MISC (Miscellaneous): Other forms of hate speech not covered by the above categories

KEY INDICATORS OF HATE SPEECH:

1. Dehumanizing language or comparisons
2. Calls for violence or harm
3. Discriminatory slurs or epithets
4. Stereotyping entire groups negatively
5. Promoting supremacy of one group over others

LABELING INSTRUCTIONS:

1. For each text, determine if it contains hate speech indicators for any of the defined categories
2. Assign (True) all relevant hate speech categories to the text
3. If no hate indicators are present, classify as non-hate speech (empty array for categories)

SPECIAL CONSIDERATIONS:

- Consider local context and cultural references
- Be mindful of dialectal variations in spelling
- Pay attention to code-mixing (Bangla-English hybrid phrases)
- A single text may belong to multiple hate speech categories
- Use the MISC category for hate speech that doesn't fit into other categories

You will receive an array of objects, each containing an 'id' and 'text'. Analyze each text and provide the required classification.

In-context Examples in Few-Shot Prompting for Binary Classification

HATE(true)

1. *Ei hindu gula ke desh theke ber kore dite hobe, era shobi desher shotru*

English Translation : These Hindus need to be thrown out of the country, they are all enemies of the nation.

Reason: Calls for the expulsion of religious groups (Hindus) and labels them all as enemies

2. *Chakma der manush e bhabo na, oder shob ke mere fela uchit na*

English Translation : Don't even think of the Chakmas as humans, all of them should be killed.

Reason: Dehumanizes ethnic group (Chakma) and calls for violence against them

NON-HATE(False)

1. *Ami or political opinion er shathe ekmat hoi na, kintu sheta niye kharap lage na*

English Translation: I don't agree with his political opinion, but it doesn't bother me.

Reason: Expresses disagreement without hatred or discrimination

2. *Amader team abar harlo, ki kharap khelse ora*

English Translation: Our team lost again, they played so badly

Reason: Expresses criticism without targeting any protected group

In-context Examples in Few-Shot Prompting for Multi-label Classification

HATE(true)

1. *Ei hindu neta gula ke desh theke ber kore dite hobe, era shobi desher shotru*

English Translation: These Hindu leaders need to be thrown out of the country, they are all enemies of the nation.

Categories: RELIGIOUS, POLITICAL

Reason: Calls for expulsion of religious group (Hindus) and labels them all as enemies

2. *Tumi ekta mota beyadob, tomar moto lok der vote deoya uchit na*

English Translation: You are a rude fat person, people like you don't deserve to be voted for.

Categories: BODY SHAMING, POLITICAL, PERSONAL OFFENCE

Reason: Uses body shaming language, suggests denying voting rights, and personally insults the individual

3. *Ei mohila neta ra shudhu ghor e thakle e desh er unnoti hobe, oder politics e ashari ki dorkar*

English Translation: If these women leaders stayed at home, the country would progress; why do they need to come into politics?

Categories: GENDER, POLITICAL

Reason: Discriminates against women in politics, targeting both gender and political participation

NON-HATE(False)

1. *Ami or political opinion er shathe ekmat hoi na, kintu sheta niye kharap lage na*

English Translation: I don't agree with his political opinion, but it doesn't bother me.

Categories: None

Reason: Expresses disagreement without hatred or discrimination

2. *Amader team abar harlo, ki kharap khelse ora*

English Translation: Our team lost again, they played so badly

Categories: None

Reason: Expresses criticism without targeting any protected group

CoT-Based Prompt

CHAIN-OF-THOUGHT ANALYSIS STEPS:

1. Initial reading: Read the entire text and identify any potential hate speech elements.
2. Keyword identification: Highlight specific words or phrases that could indicate hate speech.
3. Context analysis: Consider the overall context and tone of the message.
4. Category mapping: Connect the identified elements to specific hate speech categories.
5. Multi-label consideration: Determine if the text fits into multiple categories.
6. Intensity assessment: Evaluate the severity or intensity of the hate speech, if present.
7. Final classification: Summarize the findings and assign relevant categories.

Translation-Based Prompt

You are an expert at detecting hate speech. You will be given an array of Bangla texts (written in Latin script/English letters). First, convert the Bangla Transliterated text into standard Bangla without losing the semantic meaning, then classify the text as either hate speech (true) or non-hate speech (false). Provide a reason for your classification and provide the translation.

Transliterated Text	English Translation	Pred.	Reason by GPT-4o+E	Reason by GPT-4o+E+T	Translation by GPT-4o+E+T
world er shob deshei torun shomaj montri mp hoy khali Bangladesh chara eitar karon holo onar moto kichu political peraon jara sharajibon nijei kheyete chay manusher jonno korte chay na	In all the countries of the world, youths become social ministers and MPs, but except for Bangladesh. The reason behind this is a kind of political anarchy those who live their entire lives for themselves not for the welfare of people.	NH	The text criticizes the entire political generation with derogatory implications.	The text criticizes political leaders in Bangladesh for not focusing on youth leadership, but it does not incite hatred or use dehumanizing language.	In all countries of the world, young people are made ministers or MPs, except in Bangladesh. The reason for this is that some political persons want to selfishly continue benefiting themselves instead of working for the people.
Tomrao to bikri hoe geco	You have been sold too	NH	The text implies corruption and selling out, suggesting negative stereotyping.	The phrase does not express hatred or discrimination towards any specific group or individual.	They have also been sold out.
Chup catukar khuni	Shut up killer sycophant	H	The text seems to express frustration but lacks direct hate indicators.	The phrase 'chup catukar khuni' implies killing, which incites violence and hate.	Shut up, murderous sycophant.
Ore dhora hok ainer awtai ana hok	Let him be caught and brought under justice	NH	The text calls for arrest and punishment.	The text talks about bringing someone under the law without expressing or inciting hate or violence.	Let him be caught and brought under the law.
Agey apnader mordhe indian gulare khuje ber koren	First, find the Indians among yourselves	H	The text calls for identifying certain individuals without expressing hate.	The text stereotypes and discriminates based on origin, which is a hate indicator.	First, find the Indian infiltrators among yourselves.
Myb eta India cal pararmeye	Maybe she is a girl from some slum in India.	H	The text derogatorily implies an Indian woman based on location, with a stereotype	There is no evident expression of hate speech in identifying someone by a phrase or as a resident of a place.	Maybe she is a girl from India Calcutta.
eto jhamelar ki orkar era beche thakle sob kiso vole abaro chori bodmashi korb	What's the use of so much trouble? If they stay alive, they will forget everything and steal and do mischief again.	NH	The text discusses obstacles without expressing hate.	The text suggests that a group of people, if they remain alive, will engage in theft and misconduct again. It stereotypes and implies inherent criminal behavior.	What is the point of all this trouble, if they stay alive, they will forget everything and steal and commit mischief again.
Tora nijerai to political party er pasa chatis.	You yourselves are the sycophants of political parties.	H	The text accuses others of opportunistic political affiliations.	The text is a general statement about political affiliation without direct hate speech indicators. It appears accusatory but not incite violence or hatred.	You people yourselves are involved with the political party.
Bhai ja korso tai paiso..tmra joto marso oitar hisab k dibe ??	Brother, you got what you deserved... will you give an account for how much you have killed?	NH	The text discusses the consequences of actions without hate indicators.	This text contains implicit threat and justification for physical violence, possibly in a retaliatory context.	Brother, you got what you did.. Who's going to account for how much you hit?
Na bnp na awami league Islami Andolon Bangladesh best.	Neither BNP nor Awami League, Islami Andolon Bangladesh is the best.	NH	The text expresses political preference without any hate indicators.	This text promotes political supremacy, suggesting one political movement is superior to others.	No BNP, no Awami League, Islami Andolon Bangladesh is the best.

Table 7: Comparison of predictions and reasoning generated by GPT-4o using Explanation (E) and Translation (T) prompting. **Green** indicates that the prediction (H:Hate and NH:Non-Hate) was correctly labeled by GPT-4o+E+T but incorrectly labeled by GPT-4o+E, and **red** indicates incorrect prediction by both the models.



Figure 6: Word clouds constructed from transliterated Bangla texts in the BANTH dataset for each label.



Figure 7: Word clouds constructed from English translations of the Bangla texts in the BANTH dataset for each label.