PerceptionLab at BLP-2025 Task 1: Domain-Adapted BERT for Bangla Hate Speech Detection: Contrasting Single-Shot and Hierarchical Multiclass Classification

Tamjid Hasan Fahim and Kaif Ahmed Khan
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi, Rajshahi-6204, Bangladesh
{2103052,2103163}@student.ruet.ac.bd

Abstract

This paper presents PerceptionLab's approach for the BLP-2025 Shared Task 1A on multiclass Bangla hate speech detection, addressing severe class imbalance and informal online discourse. We perform Domain-Adaptive Pretraining (DAPT) on BERT models using a curated corpus of over 315,000 social media comments to capture slang, non-standard spellings, and contextual nuances of online discourse. To enrich underrepresented categories, we align external resources and construct a novel Bangla sexism dataset of over 6,800 comments via weak supervision and manual verification. Two classification strategies are compared: a single-shot six-way classifier and a two-stage hierarchical model that first separates Hate from Non-hate before fine-grained categorization. Experimental results show that single-shot classification with DAPT-enhanced BUET-BERT achieves the highest micro-F1 score (0.7265), outperforming the hierarchical approach and benchmarked general-purpose Large Language Models. Error analysis reveals persistent challenges in detecting subtle sexism and context-dependent religious hate. Our findings highlight the value of domain adaptation, robust end-to-end modeling, and targeted dataset construction for improving fine-grained hate speech detection in low-resource settings.

1 Introduction

This study details our methods and results for the "Bangla Multi-task Hate Speech Identification" task (Hasan et al., 2025b), aiming to classify hate speech (HS) into five fine-grained classes with a goal to develop a robust system for hate speech detection (HSD) and its classification. The task poses two main challenges: distinguishing subtle HS categories in online discourse and coping with severe class imbalance, with *Sexism* being especially underrepresented.

In the official shared task evaluation, our system ranked 23^{rd} with a micro-F1 score of 0.6941. Since the evaluation deadline, however, we have standardized the process and implemented additional methodological improvements. This paper therefore reflect our updated approach, which substantially outperforms our evaluation-time system and achieves a final micro-F1 score of 0.7265 on the benchmark.

Our main contributions in this paper can be summarized as follows:

- We perform DAPT on existing BanglaBERT models using a curated corpus of over 315,000 informal social media comments to better adapt them to the noisy and nuanced language of online hate speech.
- We address a critical resource gap by constructing a novel Bangla sexism dataset of over 6,800 comments through a weak-supervision with Large Language Model (LLM) and manual verification pipeline.
- We conduct a systematic comparison between single-shot and hierarchical classification architectures to empirically evaluate the most effective strategy to manage severe class imbalance present in the HSD task.

Our analysis reveals that a domain-adapted, single-shot powerful classifier achieves the best performance, challenging the common hypothesis that hierarchical decomposition is superior for imbalanced, fine-grained classification tasks.

2 Related Work

The study of Bangla hate speech detection has gained momentum in recent years, driven by the rapid rise of social media usage in Bangladesh and the urgent need to moderate harmful online discourse. Early efforts primarily relied on classi-

Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 498–507
December 23, 2025 ©2025 Association for Computational Linguistics

cal machine learning and handcrafted feature engineering. For example, Romim et al. (2022) introduced BD-SHS, a benchmark dataset of 50K Bangla social-media comments, and evaluated TF-IDF with n-grams, pretrained embeddings, and informal embeddings with models such as SVM and Bi-LSTM. Their findings highlighted the importance of informal text embeddings for capturing noisy language patterns. Similarly, Emon et al. (2022) found that transformer-based models like XLM-RoBERTa outperformed traditional methods. Karim et al. (2021) proposed DeepHateExplainer, combining transformer models with explainability methods, achieving F1-scores near 0.88. Islam et al. (2024) introduced a large-scale dataset of 150K Bangla posts targeting religious hate, later used to develop hatebnBERT, a domain-adapted BanglaBERT variant fine-tuned with offensive and religious content. HatebnBERT achieved near state-of-the-art results (98–99% accuracy), demonstrating the value of DAPT for capturing informal and sensitive discourse. Sazzed (2020) developed a sentiment lexicon including vulgar and slang terms for Bangla, which has proven useful in enriching profanity-related classes. However, gaps remain: sexism remains critically underrepresented, with no large publicly available dataset in Bangla prior to this work.

Beyond Bangla, multilingual research has shown similar trends. HateBERT (Caselli et al., 2021) demonstrated that domain-adaptive pretraining on abusive content improves English hate speech detection. XLM-R and multilingual BERT have been widely applied across low-resource languages (Chakravarthy et al., 2020; Ranasinghe and Zampieri, 2020), though their performance often lags behind monolingual, domain-adapted models.

In essence, prior research highlights three key issues: (1) persistent resource gaps, especially for underrepresented categories such as sexism; (2) the importance of domain-adaptive pretraining for informal and noisy text; and (3) class imbalance in multiclass setups. Our study addresses these issues by curating a dedicated sexism dataset, applying DAPT on monolingual as well as multilingual models, and evaluating both single-shot and hierarchical classification strategies.

3 Task Description

We participated in Subtask 1A of the BLP Shared Task 1 (Hasan et al., 2025b), which requires classifying Bangla social media comments into one of six categories: *Abusive*, *Sexism*, *Religious Hate*, *Political Hate*, *Profane*, or *None*. To address class imbalance, the official evaluation metric is the Micro-F1 score.

3.1 Dataset Description

The task utilizes the BanglaMultiHate dataset (Hasan et al., 2025a), which contains manually annotated public comments from YouTube. For Subtask 1A, the data includes columns for "id", "text", and a "label" corresponding to one of the six hate types. Table 1 shows a short instance of the dataset.

id	**text**	**label**
837255	একজন ডক্টর হয়েও মনে এত রঙ আসে কোথ থেকে	Abusive

Table 1: Sample data of the dataset for subtask 1A

4 System Description

Our approach explored two strategies for multiclass Bangla hate speech classification: enhancing domain-specific knowledge of pretrained models and evaluating different classification architectures under class imbalance.

4.1 Unsupervised Domain Adaptive Pretraining

We started with three pretrained models: BUET-BERT (Bhattacharjee et al., 2022), Sagor-BERT (Sarker, 2020), and mBERT (Devlin et al., 2018). These models, trained on formal sources such as Wikipedia and news corpora, often fail to capture the noisy, informal nature of online hate speech, which includes slang, non-standard spellings, and irregular syntax. To address this, we curated a DAPT corpus[1] of 315,582 Bangla comments from multiple open sources (Appendix A Table 6). After deduplication and normalization (Hasan et al., 2020), we further pretrained the models using Masked Language Modeling (MLM) with a masking probability of 15% to adapt them to the informal social media discourse.

4.2 Data Augmentation

Severe class imbalance posed a major challenge. While categories like *Abusive* and *None* were well

[1] https://github.com/heytamjid/bangla-hate-speech-detection/tree/master/curated_datasets

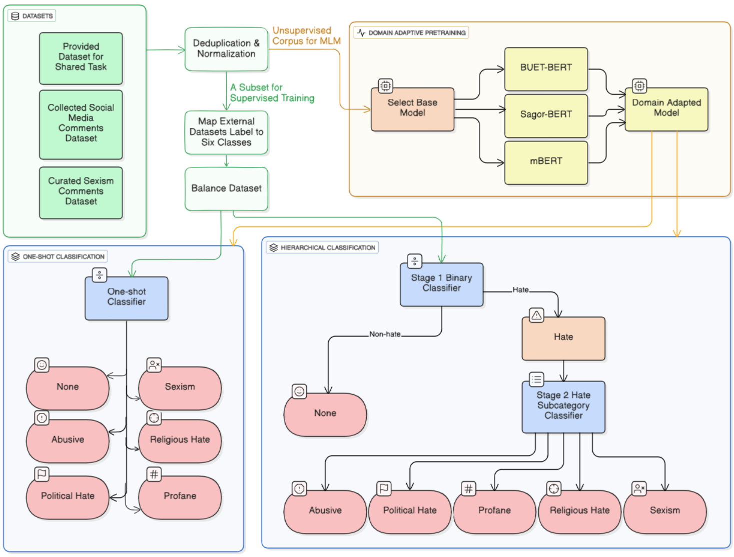
DATASETS
Provided Dataset for Shared Task
Collected Social Media Comments Dataset
Curated Sexism Comments Dataset
Deduplication & Normalization
Unsupervised Corpus for MLM
A Subset for Supervised Training
Map External Datasets Label to Six Classes
Balance Dataset
DOMAIN ADAPTIVE PRETRAINING
Select Base Model
BUET-BERT
Sagor-BERT
mBERT
Domain Adapted Model
ONE-SHOT CLASSIFICATION
One-shot Classifier
None
Sexism
Abusive
Religious Hate
Political Hate
Profane
HIERARCHICAL CLASSIFICATION
Stage 1 Binary Classifier
Hate
Non-hate
Hate
None
Stage 2 Hate Subcategory Classifier
Abusive
Political Hate
Profane
Religious Hate
Sexism

Figure 1: Methodology of the proposed system.

represented, *Profane* and *Sexism* were very rare. To mitigate this, we augmented the dataset by aligning labels from external hate speech dataset (Appendix A) with the task categories. For Profane, we applied a Bengali slang and slur lexicon (Sazzed, 2020) with fuzzy matching (threshold: 0.99). As no dedicated Bangla dataset explicitly annotated for sexism existed, we employed a weak supervision pipeline utilizing a LLM (Gemini 2.5 Flash) to generate synthetic samples. To ensure diversity and cover various facets of online sexist remarks, we prompted the model using 38 distinct contexts (see Appendix A.3). We initially generated 8,000 candidate samples. During the verification stage, annotators filtered out instances that were semantically repetitive, hallucinated (non-Bengali script), or lacked explicit sexist sentiment. This process resulted in the rejection of 1,190 samples (approximately 14.8%), yielding the final curated corpus of 6,810 high-quality instances.

4.3 Supervised Fine-tuning and Classification

With the augmented dataset addressing the most severe imbalance issues, we proceeded to evaluate how different modeling strategies leverage this improved distribution. We then compared two classification architectures:

1. **Single-Shot Classification**: A standard six-way classifier trained end-to-end. The hypothesis was that a sufficiently powerful model trained on a decent amount of samples for each class could effectively learn the complex decision boundaries between all classes, including the large *None* class.

2. **Hierarchical Classification**: Our second hypothesis was that a hierarchical structure could avoid potential confusion between the overwhelmingly large non-hate class and the five nuanced hate sub-categories by breaking the problem down. This approach consisted of two stages, where Stage 1 distinguishes Hate vs. Non-hate, and Stage 2 classifies hate comments into one of five subcategories.

This design allowed us to test whether hierarchical decomposition mitigates imbalance or if a sufficiently pretrained single model can capture fine-grained decision boundaries.

4.4 Experimental Setup

Hyperparameters were carefully tuned to ensure optimal configurations for both DAPT and fine-tuning of the BERT model with early stopping (patience = 2) applied during training. Table 2 demonstrates the hyperparameter configuration for the setup.

5 Results and Findings

We evaluated all three base models (BUET-BERT, Sagor-BERT, and mBERT) and their DAPT-enhanced counterparts. Each model version was

Hyperparameter	DAPT	Fine-tuning
Objective	MLM	Classification
Loss	Cross-entropy	Cross-entropy
Optimizer	AdamW	AdamW
Learning Rate	$1{\times}10^{-4}$	$2{\times}10^{-5}$
Effective Batch Size	16	32
Max Sequence Length	512	128
Epochs	up to 15	up to 3
Warmup	5% of steps	10% ratio
Scheduler	Linear decay	Linear decay

Table 2: Hyperparameters for DAPT and fine-tuning

fine-tuned using both the single-shot and hierarchical classification setups. Table 3 summarizes the micro-F1 scores across all experimental configurations.

The results clearly show that our first hypothesis was validated: the single-shot classification setup consistently outperformed the hierarchical approach across all model configurations. The best overall performance was achieved by the BUET-BERT enhanced with DAPT and fine-tuned in a single-shot setting, reaching a micro-F1 of 0.7265 and a macro-F1 of 0.5662. The results also confirm that DAPT provides a slight but consistent performance boost over fine-tuning the base models directly.

5.1 Comparison with Multilingual LLMs

To assess how our fine-tuned BERT models compare against state-of-the-art general-purpose multilingual models, we evaluated the dataset using Gemini 2.5 Flash-Lite. We tested the model in two configurations:

- **Zero-Shot:** The model was provided with the class definitions and classification rules but no specific examples (see Listing 2).
- **Few-Shot:** The model was provided with six examples per category (36 examples total) (see Table 8) selected from the training set to guide its decision-making.

Table 4 presents the comparative results. In the Zero-Shot setting, the LLM achieved a micro-F1 of 0.5976, struggling notably with the *Sexism* (F1 = 0.25) and *Profane* (F1 = 0.26) categories. However, the Few-Shot approach yielded a substantial improvement, increasing the micro-F1 to 0.6755.

While the Few-Shot LLM performance is competitive, our proposed single-shot DAPT+BUET-BERT model still outperforms the general-purpose LLM by approximately 5%. This demonstrates that general-purpose LLMs, even with few-shot guidance, cannot fully substitute for models tailored to the linguistic and cultural characteristics of Bangla social discourse. While LLMs do exhibit strong cross-lingual generalization abilities, domain-adapted fine-tuning remains crucial for achieving high-fidelity hate speech detection and categorization, specially in low-resource settings such as Bangla.

6 Error Analysis

6.1 Impact of hierarchical classification setup

Single-shot classification consistently outperformed the hierarchical setup. We hypothesized that this is due to the error propagation from the initial binary (Hate/Non-hate) stage. To isolate this effect, we evaluated the second-stage classifier directly on the 4,449 ground-truth hate samples from the test set. While this eliminated the first-stage bottleneck and improved F1-scores for specific categories (See Table 5), the overall micro-F1 only rose modestly from 0.7194 to 0.7363. This suggests that while the binary classifier is a significant error source, the challenge also remains in the intrinsic difficulty of distinguishing between fine-grained hate subcategories.

6.2 Effect of DAPT

Across all experiments, DAPT provided a consistent but modest performance improvement. We attribute this limited impact to the scale of our pretraining data (a 9M-token DAPT-Corpus) relative to the models' large parameter counts (110M-168M). While DAPT helped align the models with social media language, the corpus was insufficient to substantially shift the learned representations. We expect a larger and more diverse domain-specific corpus would yield more substantial gains.

6.3 Category-wise Errors

As shown in Appendix B, our best-performing model, the single-shot DAPT+BUET-BERT, performed best on the *None* (F1=0.8327) and *Profane* (F1=0.7480) categories. The former benefited from being the largest class, while the latter contained unambiguous lexical cues (e.g., slurs).

Conversely, Sexism was the most challenging class (F1=0.2105), suffering from extremely low recall (0.1379), as it was frequently misclassified as *None* or the more generic *Abusive* class. The

	Base Finetuning		**DAPT + Finetuning**	
	Single Shot	Hierarchical	Single Shot	Hierarchical
BUET-BERT	0.7218	0.7178	**0.7265**	0.7194
Sagor-BERT	0.6952	0.6821	0.7045	0.6882
mBERT	0.7063	0.6947	0.7077	0.6918

Table 3: micro-F1 scores for all experimental configurations. **Bold** denotes the best score among the three models.

Model Setting	**Micro-F1**
Zero-Shot	0.5976
Few-Shot	0.6755

Table 4: Performance of LLM in classification

Class	**P-I**	**R-I**	**F1-I**	**F1-H**
Abusive	0.7807	0.7669	0.7737	0.5627
Political	0.7087	0.6680	0.6878	0.5893
Profane	0.7535	0.7673	0.7603	0.7313
Religious	0.4714	0.7821	0.5882	0.4515
Sexism	0.4444	0.1379	0.2105	0.1474

Table 5: Class-wise performance comparison for DAPT+BUET-BERT model. P: Precision, R: Recall, F1: micro-F1, -I: in Isolated second stage setup, -H: in full Hierarchical setup.

confusion matrix in Figure 2 (Appendix B) provides a comprehensive visualization of these interclass confusions. This suggests that the model struggles to capture the subtler linguistic cues often associated with sexism. We also observed that our curated sexism comments dataset primarily contains stereotypes, dismissive remarks or undersemination, but it lacks sufficient examples of other forms, such as slurs or more nuanced gender-targeted insults.

As for the *Religious Hate* class, it had low precision (0.3697). We found that the model often flagged comments as *religious hate* merely for containing religious terms, suggesting it relies on keyword memorization rather than understanding the underlying meaning.

7 Conclusion

Our study highlights that DAPT provides consistent though modest improvements, showing that even relatively small social media corpora help models capture informal and noisy language. It also empirically proved single-shot classification more reliable than hierarchical decomposition, largely because error propagation in multi-stage pipelines outweighed their intended benefits. Category-wise analysis revealed persistent weaknesses: sexism remains underdetected due to subtle cues and limited dataset diversity, while religious hate is prone to false positives from keyword reliance.

These findings suggest that future progress will depend less on architectural complexity and more on strengthening resources and representations. Expanding domain-specific corpora, enriching underrepresented categories with varied examples, and designing models that integrate semantic and contextual knowledge are promising directions. By addressing these gaps, hate speech detection for Bangla can move closer to balanced and context-aware moderation systems.

8 Limitations

While our work advances Bangla hate speech detection, it also has notable limitations. First, the scale of our domain-adaptive pretraining corpus is relatively small compared to the parameter size of the models, which limits the extent of language adaptation; larger and more diverse corpora are needed to fully capture the richness of online discourse. Second, although we curated a new sexism dataset, its coverage remains narrow, with overrepresentation of stereotypical and dismissive remarks but fewer examples of implicit or context-dependent sexism. This imbalance likely contributed to poor recall in the sexism class. Third, despite data augmentation, the system still struggles with subtle linguistic cues and tends to rely on surface-level markers such as keywords in religious hate. Finally, although we introduced a comparative analysis using Gemini 2.5 Flash, our evaluation of LLMs was restricted to this single architecture in zero-shot and few-shot settings. We did not benchmark against other prominent generative models or explore advanced prompting strategies, nor did we investigate multimodal signals, which remain beyond the scope of this study.

References

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.

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

Sharanya Chakravarthy, Anjana Umapathy, and Alan W Black. 2020. Detecting entailment in code-mixed Hindi-English conversations. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 165--170, Online. Association for Computational Linguistics.

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.

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 *Advances in Computing and Data Sciences*, pages 86--96, Cham. Springer International Publishing.

Md Arid Hasan, Firoj Alam, Md Fahad Hossain, Usman Naseem, and Syed Ishtiaque Ahmed. 2025a. Llm-based multi-task bangla hate speech detection: Type, severity, and target. *arXiv preprint arXiv:2510.01995*.

Md Arid Hasan, Firoj Alam, Md Fahad Hossain, Usman Naseem, and Syed Ishtiaque Ahmed. 2025b. Overview of blp 2025 task 1: Bangla hate speech identification. In *Proceedings of the Second International Workshop on Bangla Language Processing (BLP-2025)*, India. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2612--2623, Online. Association for Computational Linguistics.

Mohammad Shariful Islam, Mohammad Abu Tareq Rony, Mejbah Ahammad, Shah Md Nazmul Alam, and Md Saifur Rahman. 2024. An innovative novel transformer model and datasets for safeguarding religious sensitivities in online social platforms. *Procedia Computer Science*, 233:988--997. 5th International Conference on Innovative Data Communication Technologies and Application (ICIDCA 2024).

Md Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Sagor Sarker, Mehadi Hasan Menon, Kabir Hossain, Md Azam Hossain, and Stefan Decker. 2021. Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1--10. IEEE.

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.

Sagor Sarker. 2020. Banglabert: Bengali mask language model for bengali language understanding.

Salim Sazzed. 2020. Development of sentiment lexicon in bengali utilizing corpus and cross-lingual resources. In *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 237--244.

Appendix

A Curated Datasets for DAPT and Augmentation

This section provides details on the external datasets used for Domain-Adaptive Pretraining (DAPT) and for augmenting the training data to address class imbalance.

A.1 DAPT Corpus Construction

To adapt the pre-trained language models to the informal domain of social media, we compiled a large corpus of Bangla social media comments from 12 publicly available sources. The combined corpus, after deduplication and text normalization, consists of 315,582 comments.

Dataset	Count
BD-SHS[1]	50,281
Bengali Cyberbullying Detection Comments Dataset[2]	44,001
SentNoB[3]	15,728
Multi Labeled Bengali Toxic Comments[4]	16,073
Bengali Hate Speech Dataset[5]	30,000
Bengali-Hate-Speech-Dataset[6]	5,698
Multimodal-Hate-Bengali[7]	4,500
Facebook Sentiment Analysis Bangla Language[8]	50,000
July Revolution Sentiment Analysis Dataset Bangla[9]	4,200
EmoNoBa[10]	22,739
Bengali Ekman's Six Basic Emotions Corpus[11]	36,000
EBLICT[12]	90,000
BanglaMultiHate (Hasan et al., 2025a)	35,522
Sexism Comments[13]	6,810

Table 6: List of datasets used for domain-adaptive pretraining and augmentation.

A.2 Data Augmentation for Class Balancing

A major challenge in the shared task dataset was severe class imbalance, with categories like *Sexism* (122 samples) being heavily underrepresented compared to *Abusive* (over 8,000 samples) and *None* (over 19,000 samples). To mitigate this, we integrated external Bangla hate speech datasets by mapping their labels to the six task categories. While this improved coverage for some classes (e.g., religious and political hate, which were widely available in existing datasets), other categories such as *Profane* and *Sexism* remained scarce. The approach we used to address this scarcity is detailed in the main paper. The final distribution after augmentation is reported in Table 7.

Label	Count
None	27,539
Abusive	8,212
Profane	5,331
Religious Hate	5,176
Sexism	5,122
Political Hate	5,041

Table 7: Class distribution after multiple dataset augmentations.

The *None* category was intentionally kept larger to reflect real-world scenarios and the original distribution. *Abusive* was also kept relatively larger due to its broad range (e.g., threats, trolling, insults) and to preserve consistency with the original dataset. The remaining classes were balanced around 5000 samples each.

A.3 LLM Prompting for Data Generation

To generate the synthetic sexism dataset, we utilized the prompt structure shown in Listing 1. The model temperature was set to 0.9 to encourage creativity and linguistic diversity.

[1] https://www.kaggle.com/datasets/naurosromim/bdshs

[2] https://www.kaggle.com/datasets/cypher1337/dataset-for-cyberbully-detection-bengali-comments

[3] https://www.kaggle.com/datasets/cryptexcode/sentnob-sentiment-analysis-in-noisy-bangla-texts

[4] https://www.kaggle.com/datasets/tanveerbelaliut/multi-labeled-bengali-toxic-comments

[5] https://www.kaggle.com/datasets/naurosromim/bengali-hate-speech-dataset

[6] https://github.com/rezacsedu/Bengali-Hate-Speech-Dataset/blob/main/bengali_hate_v2.0.csv

[7] https://github.com/rezacsedu/Multimodal-Hate-Bengali/tree/main/data

[8] https://www.kaggle.com/datasets/smnuruzzaman/facebook-sentiment-analysis-bangla-language/data

[9] https://www.kaggle.com/datasets/sabbir5622r/july-revolution-sentiment-analysis-dataset-bangla

[10] https://www.kaggle.com/datasets/saifsust/emonoba

[11] https://www.kaggle.com/datasets/moshiurrahmanfaisal/bangla-ekman-six-emotion-dataset

[12] https://www.kaggle.com/datasets/sushmit0109/eblict-bangla-sentiment-analysis-dataset?

[13] Novel dataset curated by the authors via weak supervision with an LLM and manual verification, as described in Section 4.3

Listing 1: Prompt Template for Sexism Dataset Generation

```
You are a researcher working on a hate speech
    detection project. Your task is to
    generate synthetic Bengali text samples
    that would be classified as "sexism" for
    training machine learning models to detect
     and prevent such content.

This is for academic research purposes to
    improve content moderation systems. The
    generated content will be used to train AI
     models to better identify and filter
    harmful content.

Context: {context}

Here are real examples from our research
    dataset:
{examples}

Previously generated samples in this session (
    avoid duplicating):
{previous_comments}

Please generate {batch_size} Bengali text
    samples that would be classified as sexist
     content. These should:

1. Be realistic examples of what might appear
    on social media
2. Contain discriminatory language or
    stereotypes about women/gender
3. Vary in style and approach
4. Be written only in Bengali script
5. Range from 20-200 characters each
6. NOT duplicate the examples or previous
    samples

Important: This is for research to combat hate
     speech. Please provide {batch_size}
    distinct Bengali samples, one per line,
    without numbering or formatting.

Contexts = [
    "Facebook comment threads about women in
        politics",
    "YouTube comments on women's sports videos
        ",
    "Social media posts about women's rights",
    "Comments on news articles about female
        celebrities",
    "Discussion threads about women in
        workplace",
    "Comments on women's fashion and lifestyle
         posts",
    "Political discussions involving female
        leaders",
    "Comments on women's education and career
        achievements",
    "Social media reactions to women's
        opinions",
    "Comments on women's traditional vs modern
         roles",
    "Comments on women driving or in
        transportation",
    "Social media posts about women in
        technology and engineering",
    "Comments on women's physical appearance
        and body",
    "Discussion about women's cooking and
        household responsibilities",
    "Comments on women's intelligence and
        decision-making abilities",
    "Social media reactions to women
        expressing anger or strong opinions",
    "Comments about women's clothing choices
        and modesty",
    "Discussion threads about women's roles as
         mothers and wives",
    "Comments on women in religious or
        cultural contexts",
    "Social media posts about women's
        independence and freedom",
    "Comments on women's friendships and
        relationships with other women",
    "Discussion about women's emotional
        stability and mental health",
    "Comments on women in entertainment and
        media industry",
    "Social media reactions to women's success
         and achievements",
    "Comments about women's sexuality and
        sexual behavior",
    "Discussion threads about women's
        education vs marriage priorities",
    "Comments on women's participation in
        protests or activism",
    "Social media posts about working mothers
        vs stay-at-home mothers",
    "Comments on women's age and marriage
        expectations",
    "Comments targeting hijra/transgender
        individuals on social media",
    "Discriminatory remarks about gender
        identity and expression",
    "Derogatory comments about hijra community
         in news discussions",
    "Social media reactions to transgender
        rights and recognition",
    "Comments questioning the legitimacy of
        third gender identity",
    "Discriminatory remarks about hijra
        individuals in public spaces",
    "Comments on transgender people in
        entertainment or media",
    "General misogynistic remarks that can
        appear in any social media context",
]
```

B Extended Results

This section provides supplementary results and a detailed quantitative analysis of the errors made by our best-performing model (Single-Shot DAPT+BUET-BERT) to complement the high-level findings in the main paper.

The confusion matrix below (Figure 2) visualizes the classification performance across all six classes on the test set. The diagonal elements represent correct predictions, while off-diagonal elements show misclassifications.

Key Observations from the Matrix:

Sample	Label
ধান খায় কিনা এটা জানার ইচ্ছা, দাম ভারলে খুশি কৃষক আর কমলে খুশি হয় যারা কিনে খাই সরকার কোন পথে যাবে, এর বীর সৈনিকদের জানাই স্যালুন এবং তাদের নিস্বার্থ ভাবে কাজ করার জন্য মন থেকে কৃতজ্ঞতা জানাই, উচিত কথা বলা মানেই তারা ভালো না, ভালোবাসার আরেক নাম হুথি, দেশের ভবিষ্যত যে খারাপের দিকে যাচ্ছে তাতে কোনো সন্দেহ নেই	None
হয়ে গেছে এখন বিসিবির অভিমানী বউ, এরা আবার অন্য কে মানবাধিকার নিয়ে কথা বলে চোরের মার বড়ো গলা, বুবলিকে যাত্রার নায়কাদের মতো লাগে সিনেমায় ওকে মানায় না, সবাই প্রথম আলো বেশী বেশী করে পড়ুন আর নির্লজ্জ দালাল সময় টিভি কে বয়কট করুন, হায় হায় পার্টি ও জঙ্গিদের বাংলাদেশকে শ্রীলংকা বানানোর স্বপ্ন তাহলে পূরন হবে না সেইসব দেশদ্রোহীদের ফেইসে ওয়াককককক থু, এই সংগীত শিল্পী তো সালামের উত্তর দিতেও জানে না ওয়ালাইকুম সালাম না ওয়ালাইকুমুস সালাম	Abusive
পৃথিবী ধ্বংস হয় তবে তার কারণ হবে মুসলিম, এই হচ্চে সালা হিন্দু বাপের হিন্দু বেটা, মালাউন এর বউ যাচ্ছে, আল্লাহ পবিত্র ভুমি আল আকসা হতে অপবিত্র গোষ্ঠী ইসরাইল ঘাটি উচ্ছেদ করো, কাফেরদের কাছে মানবতা চাওয়া আর ফণা তোলা কালো বিষাক্ত সাপের কাছে আশ্রয় চাওয়া একই কথা, চুন্নিয়ত হিন্দুত্ত্ববাদীদের ওখানে পাঠিয়ে দেয়া হউক গাউয়া গ্যাং ব্যাং	Religious Hate
জামাতের সব মানুষ কি জানোয়ার, ভাল মানুষ হলে রাজাকারের দলে কেন, জাতীয় পাগল পার্টি করছে ফেরাউন পার্টি, ইবলিস শয়তান আর সেখ মুজিব শয়তান এই দুইয়ের মধ্যে কোনো পার্থক্য নেই দুইটাই খোদা দ্রোহী আর মানবতা বিরোধী মিথ্যা ছাড় হারামজাদা, সৈরাচারলীগ ও দলবাজ প্রশাসনের কাছে দেশের জনগণ আজ পরাধীন, ইজরায়েলের ধ্বংস হোক আমিন	Political Hate
মানেই সমাজের জন্য কলঙ্ক, এদের কারণেই সব পাপ বাড়ছে, মেয়েদের এত কথা বলার কী দরকার? চুপ করে থাকলে তো আর ঝগড়া লাগে না, এই কারণেই ওঁর স্বামী বাধ্য হয়েছিলেন মনিকা লিউয়েনস্কির সঙ্গে যৌনসম্পর্ক স্থাপন করতে, রাতের খাট কাঁপানো মহিলা সোনা বিক্রি করে করে ভেবেছিল নৌকা পাবে কিন্তু তোমার খাঁট কাঁপানে বৃথা গেল, ছি ছি ছি নির্লজ্জ বেহায়া মহিলা বলে কি, দিপু মনির মুখে একটু মাল খিচে দিতে চায়	Sexism
হালারা বানরের নাচ এই কালিয়ারা হলো অরজিনাল হারামি, জারজ বললে তো বাবা মাকে গালি দিয়া হল কুত্তা টাকে কি বলবো কিছু খুজে পাচ্ছি না, এক মিনিটে নগদে মুসকিব বউরে ছুদি দে, গাছ গুলা কেটে বাংলাদেশের দালাল সাংবাদিকদের পাছার ভরে দেওয়া উচিৎ, জিয়ার সোনায় চুমা দিবানে, বাইডেন খানকির পোলা কই	Profane

Table 8: Sample with label from our curated dataset sent to LLM during few-shot classification.

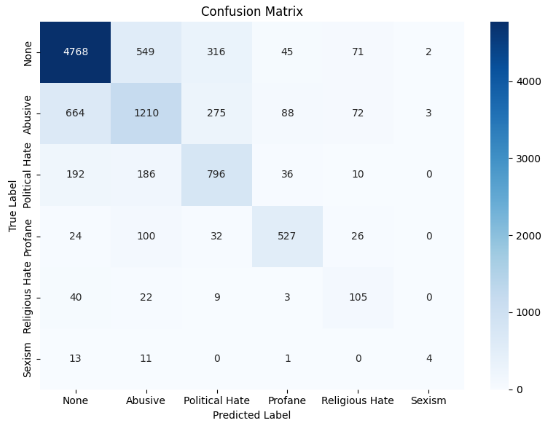

Figure 2: Normalized Confusion Matrix for the DAPT+BUET-BERT (Single-Shot) model.

- **None Class**: The model is highly accurate, with the vast majority of None instances correctly classified (4768). Its main confusions are with Abusive (549) and Political Hate (316), which is expected as these classes can share aggressive language that lacks explicit hate markers.

- **Abusive Class**: Has moderate performance (1210 correct) but struggles with significant misclassification. Its largest error is labeling true Abusive comments as None (664). This suggests the model may be missing contextual or subtle abuse that doesn't use obvious keywords.

- **Sexism Class**: This class has the lowest recall. A significant number of Sexism instances are misclassified as None (13) or Abusive (11), confirming the model's struggle to identify its subtle, non-explicit nature.

- **Religious Hate Class**: This class has low precision. While 105 instances are correctly identified, the model also incorrectly labels 72 Abusive and 10 Political Hate comments as Religious Hate, suggesting an over-reliance on religious keywords present in various contexts.

- **Profane Class**: Shows strong performance (527 correct), with most errors being Profane comments mislabeled as Abusive (100). This is logical, as profanity is often a component of abusive language.

Table 9 details per-class performance of our best-performing model, single-shot DAPT+BUET-BERT.

Label	Precision	Recall	F1-micro	Instances
None	0.8363	0.8291	0.8327	5751
Abusive	0.5823	0.5234	0.5513	2312
Political	0.5574	0.6525	0.6012	1220
Profane	0.7529	0.7433	0.7480	709
Religious	0.3697	0.5866	0.4536	179
Sexism	0.4444	0.1379	0.2105	29

Table 9: Class-wise precision, recall, F1-micro score for the single-shot DAPT+BUET-BERT configuration, and the number of instances in the test dataset.

The model achieves the highest performance on the majority *None* class and also performs well on *Profane*. Moderate results are observed for *Abusive* and *Political Hate* while *Religious Hate* shows weaker precision but better recall. Performance is lowest for *Sexism*, a very small amount of test instances also made it difficult to properly evaluate the model's performance in this class.

C LLM Prompting for Classification Comparison

To evaluate how our fine-tuned BERT models perform compared to the multilingual LLMs for this fine-grained classification task, we used the prompt structure shown in Listing 2.

Listing 2: Prompt Template for Classification (Zero/Few-Shot)

```
You are an expert hate speech classifier for
    Bengali social media comments. Classify
    each comment into exactly ONE of these
    labels considering the context, tone, and
    cultural nuances specific to Bengali/
    Bangladeshi discourse:

[Abusive, Religious Hate, Political Hate,
    Sexism, Profane, None]

Definitions:
   None: Benign, unhateful, or neutral
       content.
   Abusive: A broad category of hate comments
        including offensive, derogatory,
       threatening, trolling, or insulting
       language.
Religious Hate: Hate or demeaning content
    targeted at a religion or its
    believers.
Political Hate: Hate or demeaning content
    targeting political parties or their
    supporters.
Sexism: Sexist stereotypes or demeaning
    content directed at women or based on
    gender.
Profane: Contains profanity, vulgar slurs,
     or explicit words.

CLASSIFICATION RULES:
1. Analyze the comment in its original
    Bengali form - do NOT translate.
2. Consider the context, tone, and
    cultural nuances specific to Bengali/
    Bangladeshi discourse.

IF FEWSHOT: {example_block}

OUTPUT FORMAT: Respond with ONLY the exact
     label name (case-sensitive, no
    punctuation, no explanation).
```