Heisenberg at BLP-2025 Task 1: Bangla Hate Speech Classification using Pretrained Language Models and Data Augmentation

Warning: This research paper contains examples of hateful content.

Samin Yasir
Department of Computer Science and Engineering
Shahjalal University of Science and Technology
saminyasir.cs@gmail.com

Abstract

Detecting hate speech in Bangla is challenging due to its complex vocabulary, spelling variations, and region-specific word usage. However, effective detection is essential to ensure safer social media spaces and to take appropriate action against perpetrators. In this study, we report our participation in Subtask A of Task 1: Bangla Hate Speech Detection (Hasan et al., 2025b). In addition to the provided 50K Bangla comments (Hasan et al., 2025a), we collected approximately 4K Bangla comments and employed several data augmentation techniques. We evaluated several transformer-based models (e.g., BanglaBERT, BanglaT5, BanglaHateBERT), achieving the best performance with a micro-F1 score of 71% and securing 18th place in the Evaluation Phase.

1 Introduction

As social media continues to grow in popularity, particularly among children and adolescents (Lenhart et al., 2010; Li et al., 2021), it is imperative to address hateful content. Therefore, effective detection and restriction of hate comments on the internet is necessary.

Identification of hate speech in the English language has reached an accuracy of 98.0% (Saleh et al., 2021), allowing platforms to identify most of the offensive contents and take appropriate action according to the social media platform's terms and policies (Schmidt and Wiegand, 2017). Completely banning hate speech also can be seen as a restriction of freedom of speech on the internet. If a hate comment is not protected by the moral right to freedom of expression, it falls under a moral duty to refrain from hate speech, or against the law of the state, restrictions or banning on the comment can be applied (Howard, 2019).

With over 173.8 million Bengali speakers in Bangladesh, of whom approximately 45.0 million are active social media users (Sarkar, 2024; Haque et al., 2023), the development of an effective hate speech detection system is crucial for ensuring a safer online environment for this large community. Advancing research in this domain not only safeguards users but also contributes to the broader objective of enhancing large language models (LLMs) to identify hate speech across diverse languages, thereby enabling them to issue warnings or implement preventive measures when necessary.

Detecting hate speech in Bangla contains bigger challenges due to its morphological richness with diverse synonyms (Farzana, 2021; Ali et al., 2008), regional variations, and context-based meanings. Therefore, hate speech detection models for Bangla remain less effective with a small amount of data (Tanvir Alam and Mofijul Islam, 2018). Consequently, children remain vulnerable to harmful content, while individuals spreading hate in Bangla often go undetected and unpunished. However, recent models, such as BanglaHateBERT, showed prominent performance on the hate speech detection task (Jahan et al., 2022).

In this study, Several Bangla-specific transformer models, including BanglaT5 (Bhattacharjee et al., 2023), BanglaBERT (Bhattacharjee et al., 2021), BanglaHateBERT (Jahan et al., 2022), are experimented with different types of data augmentation methods.

Our contribution can be summarized as follows:

- Experimented with five transformer-based models to achieve a micro-F1 score of 71%
- Newly collected 3,874 data-points from Bangla YouTube comments and added to the training dataset
- Analyzed the errors in the dataset to find limitations

Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 476–484

2 Related Works

Hate speech detection is a problem that researchers have been working to improve over the past few decades (Tontodimamma et al., 2021). However, it remains a challenging task for many reasons. The definition of hate speech varies significantly across regions, time periods, and different political, economic, and social contexts (Parekh, 2006).

Overfitting behavior is found to be very frequent among hate speech detection systems because the domain is vast, covering areas such as race, religion, gender, sexuality, etc in any language (Moy et al., 2021). However, a multilingual online hate speech detection system has been developed to identify hate speech in English, Italian, and German, demonstrating satisfactory performance in these languages (Corazza et al., 2020).

In Bangla, various transformer-based models have been utilized to identify offensive content from Banglish Facebook comments in a multi-label setup (Raihan et al., 2023). However, a survey on textual hate speech detection highlights that despite the advances of deep learning, particularly transformer-based models, progress is limited by weak datasets, inconsistent definitions, and poor generalization (Alkomah and Ma, 2022). Hence, these challenges are especially pronounced for Bangla, where datasets remain scarce (Romim et al., 2021).

In their work, (Hossain Junaid et al., 2021) evaluates machine learning and deep learning approaches for Bangla hate speech detection, reporting that logistic regression achieved the highest accuracy (96.2%) among machine learning methods, while a GRU-based model outperformed all approaches. In contrast, applying SVM and Naive Bayes to 1,339 Bangla samples with Naive Bayes reached a maximum accuracy of 72% (Ahammed et al., 2019). These results suggest that deep learning models are generally more effective than traditional machine learning methods for this task.

In a study on benchmarking transformer models for violence detection in Bangla YouTube comments, and showed that data augmentation with 500 samples improved F1 scores, emphasizing the value of additional context-specific data (Saha and Nanda, 2023). Another work (Sharif et al., 2022) introduced a multi-label Bangla dataset of aggressive sentences, where BanglaBERT achieved the highest weighted F1-scores (92%) in detection. These findings indicate that expanding the dataset and using BanglaBERT can increase accuracy in our task.

Similarly, the research (Romim et al., 2022) benchmarked multiple models across eight Bangla datasets by exploring various model–feature combinations and reporting variations in F1-scores. Their frequency-based word cloud analysis of traditional and non-traditional swear words informed our data augmentation, where high-frequency terms were incorporated into the training dataset.

A detailed description of the data collection strategy has been described in the research (Haider et al., 2025), which we have followed in our research. Four sequential steps are followed in the paper to generate the dataset Figure 1. Moreover, back-translating (Bangla -> English -> Bangla) dataset approach was performed to add diversity into the dataset, where the GRU and Attention techniques provided high accuracy up to 98% (Faruqe et al., 2023). Most of the research shows that the model BanglaBERT performs best for detecting Bangla hate speeches, where the data has variation (Tariquzzaman et al., 2023; Bhattacharjee et al., 2022; Das et al., 2022). Data Augmentation using translation and back-translation is discussed as an effective method to gain better accuracy in some research with the Bangla dataset (Tariquzzaman et al., 2024; Aziz and Islam, 2025; Khandaker et al., 2025).

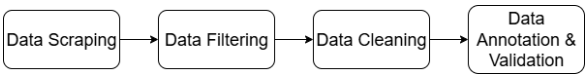

Figure 1: Data Augmentation Steps

3 System Description

This section describes how the system classifies Bangla hate content, dataset description, and different augmentation techniques used to achieve the highest micro-F1 score. All the code[1] and datasets[2] used for the task are publicly available.

3.1 Task Description

The goal of the shared task is to recognize Bangla hate comments. The input is Bangla sentences, and the output is to detect the type of hate. Both input

[1] https://github.com/Heisenberg71/blp25_task1/blob/main/example_scripts/subtask_1A_DistilBERT_example.ipynb?short_path=3504d21

[2] https://github.com/Heisenberg71/blp25_task1/tree/main/data/subtask_1A

and output is in TSV file format. There are a total of 6 types of hate classification: **Abusive, Sexism, Religious Hate, Political Hate, Profane,** and **None**.

3.2 Initial Dataset Description

The initial dataset is provided by the shared task organizers that contains about 35K labeled Bangla comments from YouTube for training (Hasan et al., 2025a). An example of a training dataset is given in Table 8. The labels are shown with frequency and percentages in the training data set, where the majority of hate types are **None** Table 1.

Label	Frequecy	Percentage
Abusive	8,212	23.12%
Sexism	122	0.34%
Religious Hate	676	1.90%
Political Hate	4,227	11.90%
Profane	2,331	6.56%
None	19,954	56.16%
Total	35,522	100%

Table 1: Training data frequency of the provided Dataset

3.3 Data Collection

We have extracted additional 3,874 Bangla comments from two videos[3] from a very popular political YouTube channel[4] in Bangladesh. We have used an online tool named[5] for collecting comments.

The collected comments contained URLs, emails, digits, punctuation marks, emojis, letters from other languages(English, Hindi, Arabic, etc), and special symbols that are unnecessary and holds little to no information on hate classification. Therefore, it was cleaned using BNLP's CleanText text cleaning package[6] and manual review. A summary of the collected data set is stated in Table 6 and Figure 2.

The collected data is used as testing data, and BanglaBERT is used to label the type of hate. Finally, the annotated data points are added to the testing dataset.

[3] https://www.youtube.com/watch?v=GvDDgxbfSYk, https://www.youtube.com/watch?v=Qb0YZc1K_-8

[4] https://www.youtube.com/c/pinakibhattacharya

[5] https://youtubecommentsdownloader.com/

[6] https://github.com/sagorbrur/bnlp

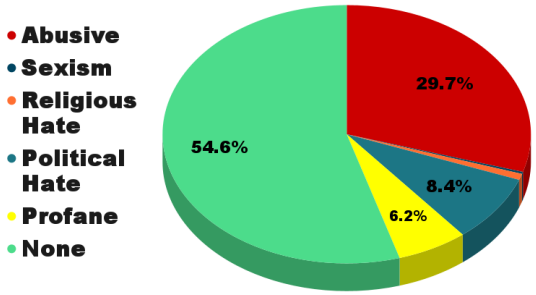

Figure 2: Distribution of hate percentage across labels

3.4 Data Augmentation: Synonym-based

Synonym-based augmentation was applied to the collected dataset to increase its variability while preserving semantic consistency. Specifically, a set of words frequently used in Bangla hate speech was identified (Romim et al., 2022), and selected words in the dataset were randomly replaced with their synonyms. This ensured that the type of hate expressed in the comments remained unchanged while introducing linguistic diversity. An illustrative example of this process is provided in Table 2.

Words	Synonyms
শালা	শালার পো
কুত্তা	কুত্তার বাচ্চা
খা*কি	বেশ্যা
হারামি	হারামজাদা
জা*জ	বে*ন্মা
বাল	বালছাল
জনোয়ার	পশু
দুঃখ	কষ্ট
শত্রু	প্রতিপক্ষ
দুর্গন্ধ	দুর্গন্ধময়
ক্ষমা	মাফ
দুর্গম	দুস্তর

Table 2: Some example of words that replace randomly on the dataset

3.5 Data Augmentation: Back-translation

We augmented the dataset through back-translation, translating the original Bangla data into English and then translating it back to Bangla using the Google Translate API, which introduced lexical and syntactic variations. A total of 27,000 data points were processed using this approach. While back-translation proved

effective in increasing diversity within the dataset, we observed that it frequently altered the type of hate expressed in certain comments. Such semantic shifts caused different between the original and translated labels, making the dataset unsuitable for training without extensive human review and relabeling. Given the impracticality of manually verifying a dataset of this scale, we ultimately decided not to use the dataset for model training. Illustrative examples of back-translated comments are provided in Table 9, and a detailed summary is presented in Table 7 and Figure 3.

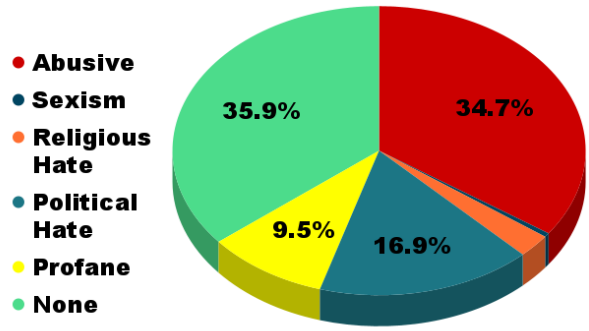

Figure 3: Distribution of hate percentage across labels in the back-translated dataset

3.6 Methodology

Approaches taken to improve the micro-F1 scores are described below.

Tokenization We used the Basic Tokenizer from BNLP[7] , which is specifically designed for Bangla text. This tokenizer provided more effective preprocessing of Bangla comments, thereby enabling the models to learn linguistic patterns more accurately and improving their ability to detect hate speech.

Stopwords Many Bangla words do not contribute significantly to the meaning of a sentence, and their removal does not alter the underlying semantics. To address this, we employed a stopword removal tool from the Bangla corpus to eliminate stopwords and punctuations. This preprocessing step ensures that the model receives only the meaningful components of a sentence as input. A list of commonly used Bangla stopwords is provided in Table 12.

[7] https://github.com/sagorbrur/bnlp/tree/main

Models The initial configuration was set to a single epoch. Through iterative experimentation, we found that training the model for three epochs yielded better performance within a shorter training time. Increasing the number of epochs beyond three led to overfitting, resulting in poor generalization on the test set. For model selection, we initially experimented with DistilBERT, but subsequently trained HateBERT, BanglaBERT, BanglaHateBERT, and BanglaT5 on the preprocessed dataset. Among these, BanglaBERT achieved the best performance on the provided test set. The hyperparameter settings for the experiments are summarized in Table 3.

Hyperparameters	Details
Dropout rate	0.1
Number of epochs	3
Training, validation, test split ratio	80:5:20
Learning rate	$2e^{-5}$
Optimizer	AdamW

Table 3: Hyperparameters of models (DistillBERT, HateBERT, BanglaBERT, BanglaT5, and BanglaHateBERT)

3.7 Results and Discussion

We experimented with different models to achieve the best micro-F1 score. Figure 4 presents the performance of these models, where BanglaBERT with the augmented dataset achieved the highest micro-F1 score of 0.71 on the released test set during the evaluation phase. Although models such as BanglaHateBERT and BanglaT5 were also used, BanglaBERT consistently outperformed them. The hyperparameters used for fine-tuning the models are provided in Table 3.

We observed that models specifically trained for the Bangla language, such as BanglaHateBERT and BanglaBERT, outperformed general hate speech detection models like HateBERT. Furthermore, data augmentation proved to be a crucial factor in our study, enhancing the performance of BanglaBERT and achieving the highest micro-F1 score of 71%.

However, due to the limited number of label samples of **Sexism and Religious Hate** in the training set, none of the models were able to identify comments belonging to this category effectively. A more detailed analysis of this limitation is provided in the error analysis section.

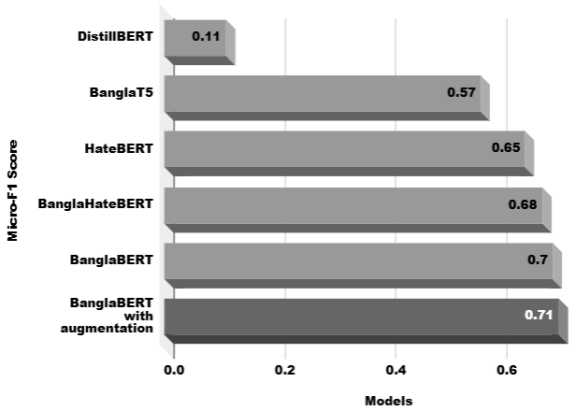

Figure 4: Micro-F1 scores of different approaches

4 Error Analysis

The representation of **Sexism** and **Religious Hate** in the dataset are extremely limited. Only 122 instances labeled as Sexism are present in the 35,523 training data points, accounting for merely 0.34% of the total training set. Consequently, every model we used fails to identify any occurrences of sexism in the test set, primarily due to the insufficient availability of training examples required to effectively learn and detect this category Table 4.

Data	# data	# Sexism	P.
Train	35,523	122	0.34%
Test	10,200	29	0.28%

Table 4: Dataset description for **Sexism**. P.: Percentage

Similarly, only 1.90% training data points is on religious hate making it difficult to detect Table 5.

Data	# data	# R.H.	P.
Train	35,523	676	1.90%
Test	10,200	179	1.75%

Table 5: Dataset description for **Religious Hate**. R.H.: Religious Hate, P.: Percentage

Some comments fall within the scope of multiple hate labels. However, assigning only one label to such comments creates accuracy issues. A single comment may correspond to two or more labels, but our model can output only one of the detected labels. If the test data assigns a different label than the one predicted, the model's score decreases, even though it has correctly identified the comment as hateful in Table 10 containing multiple hate type in a comment. Variation of hate detection between labels in the test set and BanglaBERT-generated labels.

There are also some data points on the test set that are not labeled correctly. However, BanglaBERT was able to label them correctly Table 11.

5 Conclusion and Future Scopes

This paper states the experiments we have performed to complete the shared task. Using the Bangla Tokenizer and the stopword removal technique is proven to be a very good pre-processing technique. We have collected comments from YouTube and labeled them carefully, then add them to the training set, and that improved the overall micro-F1 score. Lastly, we have utilized various well-known models that have demonstrated effectiveness in generating good results in Bangla. Among these, BanglaBERT performed best for our test and training datasets. Future studies will investigate the capabilities of LLMs and explainable hate speech detection for Bangla.

6 Limitations

The synonym-based dataset often make illogical sentences. The translators that are available are not good enough to preserve the whole meaning of a sentence. Moreover, comments are now dependent on recent political or socio-economic events. Therefore, a comment can be normal, but according to context, a normal sentence meaning can change to a hateful comment.

The model tested here is for specialized for the Bangla language. But, there are many multilingual models exists that can be very good to detect hateful comments over the internet, which have not been experimented with in this research. Similarly, the Bangla Basic tokenizer has been tried only in research. But, there are other Bangla tokenizers exists that can be improve the F1 score of the dataset.

References

Shovon Ahammed, Mostafizur Rahman, Mahedi Hasan Niloy, and S. M. Mazharul Hoque Chowdhury. 2019. Implementation of machine learning to detect hate speech in bangla language. In *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, pages 317–320.

Md. Nawab Yousuf Ali, S. M. Abdullah Al-Mamun, Jugal Krishna Das, and Abu Mohammad Nurannabi.

2008. Morphological analysis of bangla words for universal networking language. In *2008 Third International Conference on Digital Information Management*, pages 532–537.

Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information*, 13(6).

Faisal Ibn Aziz and Muhammad Nazrul Islam. 2025. Banglahealth: A bengali paraphrase dataset on health domain. *Data in Brief*, 61:111699.

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.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Kazi Samin, Md Saiful Islam, Anindya Iqbal, M Sohel Rahman, and Rifat Shahriyar. 2021. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. *arXiv preprint arXiv:2101.00204*.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, and Rifat Shahriyar. 2023. BanglaNLG and BanglaT5: Benchmarks and resources for evaluating low-resource natural language generation in Bangla. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 726–735, Dubrovnik, Croatia. Association for Computational Linguistics.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. A multilingual evaluation for online hate speech detection. *ACM Trans. Internet Technol.*, 20(2).

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

Omar Faruqe, Mubassir Jahan, Md. Faisal, Md. Shahidul Islam, and Riasat Khan. 2023. Bangla hate speech detection system using transformer-based nlp and deep learning techniques. In *2023 3rd Asian Conference on Innovation in Technology (ASIANCON)*, pages 1–6.

Afifa Farzana. 2021. A comparative study between english and bangla: The perspective of phonemics, morphology and syntax.

Fabiha Haider, Fariha Tanjim Shifat, Md Farhan Ishmam, Md Sakib Ul Rahman Sourove, Deeparghya Dutta Barua, Md Fahim, and Md Farhad Alam Bhuiyan. 2025. BanTH: A multi-label hate speech detection dataset for transliterated Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7217–7236, Albuquerque, New Mexico. Association for Computational Linguistics.

Rezaul Haque, Naimul Islam, Mayisha Tasneem, and Amit Kumar Das. 2023. Multi-class sentiment classification on bengali social media comments using machine learning. *International Journal of Cognitive Computing in Engineering*, 4:21–35.

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.

Mohd. Istiaq Hossain Junaid, Faisal Hossain, and Rashedur M. Rahman. 2021. Bangla hate speech detection in videos using machine learning. In *2021 IEEE 12th Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, pages 0347–0351.

Jeffrey W. Howard. 2019. Free speech and hate speech. *Annual Review of Political Science*, 22(Volume 22, 2019):93–109.

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

Md. Arafat Alam Khandaker, Ziyan Shirin Raha, Bidyarthi Paul, and Tashreef Muhammad. 2025. Bridging dialects: Translating standard bangla to regional variants using neural models. *Preprint*, arXiv:2501.05749. Accepted in 2024 27th International Conference on Computer and Information Technology (ICCIT).

A. Lenhart, K. Purcell, A. Smith, and Kathryn Zickuhr. 2010. Social media mobile internet use among teens and young adults. *Pew Internet and American Life Project*.

Wenxin Li, Xuantong Lin, Jiani Wu, Wenhan Xue, and Junxian Zhang. 2021. Impacts social media have on young generation and older adults. In *Proceedings*

of the 2021 4th International Conference on Humanities Education and Social Sciences (ICHESS 2021), pages 294–300. Atlantis Press.

Tian Xiang Moy, Mafas Raheem, and Rajasvaran Logeswaran. 2021. Hate speech detection in english and non-english languages: A review of techniques and challenges. *Technology*.

Bhikhu Parekh. 2006. Hate speech. *Public Policy Research*, 12(4):213–223.

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

N. Romim, M. Ahmed, H. Talukder, and M. Saiful Islam. 2021. Hate speech detection in the bengali language: A dataset and its baseline evaluation. In Mohammad Shorif Uddin and Jagdish Chand Bansal, editors, *Proceedings of International Joint Conference on Advances in Computational Intelligence*, Algorithms for Intelligent Systems, pages 457–468. Springer, Singapore.

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.

Saumajit Saha and Albert Nanda. 2023. BanglaNLP at BLP-2023 task 1: Benchmarking different transformer models for violence inciting text detection in Bangla. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 163–167, Singapore. Association for Computational Linguistics.

Hind Saleh, Areej Alhothali, and Kawthar Moria. 2021. Detection of hate speech using bert and hate speech word embedding with deep model. *arXiv preprint arXiv:2111.01515*.

Dr. Sukanta Sarkar. 2024. Social media as a tool of communication in bangladesh: Pattern, growth and challenges. *Pakistan Journal of Media Sciences*, 5(Issue2):54–60.

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.

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, Dublin, Ireland. Association for Computational Linguistics.

Md Tanvir Alam and Md Mofijul Islam. 2018. Bard: Bangla article classification using a new comprehensive dataset. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–5.

M. Tariquzzaman, A. N. Anam, N. Haque, M. Kabir, H. Mahmud, and M. K. Hasan. 2024. Bda: Bangla text data augmentation framework. *Preprint*, arXiv:2412.08753.

Md. Tariquzzaman, Md Wasif Kader, Audwit Anam, Naimul Haque, Mohsinul Kabir, Hasan Mahmud, and Md Kamrul Hasan. 2023. the_linguists at BLP-2023 task 1: A novel informal Bangla Fasttext embedding for violence inciting text detection. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 214–219, Singapore. Association for Computational Linguistics.

A. Tontodimamma, E. Nissi, A. Sarra, and L. Fontanella. 2021. Thirty years of research into hate speech: topics of interest and their evolution. *Scientometrics*, 126:157–179.

A Appendix

Label	Frequecy
Abusive	1,154
Sexism	0
Religious Hate	30
Political Hate	327
Profane	241
None	2,122
Total	3,874

Table 6: Training data frequency of the collected Dataset

Label	Frequecy
Abusive	9,366
Sexism	122
Religious Hate	706
Political Hate	4,554
Profane	2,572
None	9,680
Total	27,000

Table 7: Training data frequency after back-translation

Text	Label
অতিরিক্ত এ নিজেকে বাদুর বানাইয়া ফেলছেন রে	Abusive
অযোগ্য মহিলাদের হাতে ক্ষমতা দিয়ে দেশকে রষাতলে ফেলার আগ্রহ দেশবাসীর নাই	Sexism
অথচ এরাই ৬০ লক্ষ্য ইহুদি হত্যা করেছিলো	Religious Hate
সরকারের নীলনকশা এখন মানুষ বুঝে গেছে মানুষ এখন আর ললিপপ খায় না	Political Hate
হালার পাছা দিয়া উত্তর কুরিয়ার পারমাণবিক বোমা মারা হোক	Profane
আমার বাস্তবে জাওয়া হলোনা আমি ইউটিউবে দেখবো। ইট আলী	None

Table 8: Sample dataset for each of the categories

Bangla	Translated English	Translated Bangla
এমন আশ্বাস আমিও দিতে পারি	I can also give such assurance	আমি এ জাতীয় আশ্বাসও দিতে পারি
সালা বাঙালি বোকা চোদা জাতি বেশি গাল মারোছ তাই হাল তোদের	Sala Bengali fools are more cheeky	সালা বাঙালি বোকা আরও চটকদার
আরো আগে কেন নিউজ হলো না	Why not before the news happened	খবরটি হওয়ার আগে কেন নয়
ওলে বাবালে আমাল প্যান্ট ভিজে গেছে তুমি আগে দুদু খেয়ে আসো	Ole Babale Amal pants are wet. You eat dudu first	ওলে বাবলে অমল প্যান্ট ভেজা।আপনি প্রথমে দুদু খান
তর গুষ্টি চুদি মা*র ছেলে টাকা পাচারকারী রাস্তায় আসিস তর বাপেদের ছাড়া	Your bunch of Chudi Ma*i's son is money smuggled on the street without Father	আপনার চুদি মা*র ছেলের গুচ্ছ আসিস তারা বাবা ছাড়া রাস্তায় অর্থ পাচার করা অর্থ

Table 9: Example of back-translation in the dataset

Text	Labels on Test set	Label by Bangla BERT
ভারতীয় দালাল সময় টিভিকে বয়কট করুন	Political Hate	Abusive
কিছুই হবে না বদমাশ ইসরাইল বা নেতায়াহুআর কিছুই হবে না	Political Hate	Abusive
সময় টিভি একটা জাউড়া মিডিয়া মিথ্যা তথ্য প্রচার করে বেড়ায়	Profane	Abusive
ইনুকে কেউ বুটদেবেনা নাচতিক সালা	Abusive	Profane
তোদের বংগবন্দু তো জাহান্নামে আছে বলেছে তো তোদের আওয়ামী কুত্তা লীগ এবার জনগণ সবাই মিলে বাঁশের লাঠি দিয়ে পিটিয়ে মেরে ফেলবো	Profane	Political Hate

Table 10: Datapoints containing multiple hate type in a comment. Variation of hate detection between labels in test set and BanglaBERT generated labels

Text	Incorrect Labels on Test set	Correctly Labeled by Bangla BERT
দাজ্জালের হাতে থাকবে তাপমাত্রা নিয়ন্ত্রণ অনেক এলাকায় গাছ পালা কাটা শুরু হয়েছে	Abusive	None
বিদ্যুৎ জ্বালানি খাতে আওয়ামী লীগের আমলে সবচেয়ে বেশি দুর্নীতি হয়েছে	Abusive	Political Hate
কোন জায়গায় বিচার পাবে মানুষ সব জায়গায় দুর্নীতি সরকার দুর্নীতি করে সরকারের প্রশাসন দুর্নীতি করে এটার জন্য খুবই দুঃখের বিষয় এটা একটা হাস্যকর কাহিনী	None	Political Hate

Table 11: Correctly labeled by BanglaBERT but incorrect labels in test set

Bangla Stopwords	Meanings
এবং	and
ও	and
যে	that
কি	what/that
কিন্তু	but
আমি	I
সে	he/she
ভালো	good
অনেক	many/much
আর	and/again
আগে	before
এখন	now
পরে	many/much
থেকে	from
এবং	and

Table 12: Some example of popular Bangla Stopwords