Bahash-AI at BLP-2025 Task 1: Bangla Hate Speech Detection using Data Augmentation and Pre-trained Model

Sahinur Rahman Laskar
UPES, Dehradun
India
sahinurlaskar.nits@gmail.com

Bishwaraj Paul
Bahash-AI, Bahash Private Limited
Silchar, India
bishwaraj.paul@bahash.in

Abstract

In recent times, internet users are frequently exposed to Hate Speech on social media platforms that have long-lasting negative impacts on their mental wellbeing and also radicalizes the society into an environment of fear and distrust. Many methods have been developed to detect and stop propagation of Hate Speech. However, there is a limitation of annotated data available for Hate Speech in Bengali language. In this work, we have used a pretrained BanglaBERT model on an extended train dataset synthesized via data augmentation techniques. Our team Bahash-AI has achieved 20th, 20th and 17th position of the 3 subtasks out of total 37, 24 and 21 total number of teams who participated in the subtasks 1A, 1B and 1C respectively for Bangla Multi-task Hatespeech Identification Shared Task at BLP Workshop with F1 scores of 0.7028, 0.6954, 0.6969 respectively.

1 Introduction

Hate Speech is generally defined as public speech that conveys hate or incites harm towards an entity that is an individual or a group because of their race, religion, gender, sexual orientation and other characteristics (Cambridge Dictionary).

Hate speech has far-reaching consequences in society that causes the disintegration of social cohesion, incites harm, mental and physical and causes long-lasting psychological effects on individuals and suppresses the development of a nation altogether (Waldron, 2012). Hate speech also causes suppression of free speech of the affected entities, isolating them from public discourse out of fear. A fundamental part of a functioning society is skepticism and debate. People will fear to express new ideas out of fear of being targeted, and the exchange of ideas will become stagnant (Celuch et al., 2023). Political parties and their affiliated individuals and groups use hate speech to retract attention on critical issues and also to divide the population into voting groups to their advantage, hence the existence of hate speech becomes a crisis to the state of democracy (Putra and Damanik, 2021). While some claim that hate speech is protected under free speech and expression, hate speech in turn actually suppresses the expression and participation of free speech of the targeted individuals and communities. Hence, hate speech should be detected to protect societal interests and handled appropriately according to the situation (Waldron, 2012). Due to the advent of social media, it has become hard to detect hate speech manually. Rather, programmatic detection of hate speech has become a critical part of Natural Language Programming (NLP) techniques (Nascimento et al., 2023).

While significant progress has been made in English language due to the huge amount of data available online, Bengali language hate speech is lower as many individuals prefer to use English in social media due to unfamiliarity of using regional language keyboards or even using latinized version of Bengali. There is also the challenge of code-mixing, where English and Bengali are mixed (Das et al., 2022). The main theme of the shared task (Hasan et al., 2025b) is to detect and understand Hate Speech across a variety of subtasks, which is a reflection of real life scenarios where Hate Speech identification requires understanding not just its presence, but also its type, target, and severity.

2 Related Work

We briefly describe the recent related works related to Hate Speech detection techniques. Many works have used classical Machine Learning methods like Logistic Regression, Naive Bayes, Support Vector Machine, Decision Tree, Random Forests, etc. to detect Hate Speech after preprocessing the data with techniques like Lemmatization, Feature Extraction etc. (Davidson et al., 2017). The deep learning models like CNNs, RNNs and transform-

Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 402–407

ers are becoming more prominent nowadays (Malik et al., 2024), (Schmidt and Wiegand, 2017). BanglaBERT is a BERT model that was specifically pretrained on Bengali text to be able to perform tasks for Bengali language inputs (Bhattacharjee et al., 2021). There even exists a BanglaHateBERT specifically designed to detect Hate Speech in Bengali language (Jahan et al., 2022). LLMs are the latest type of models in recent years, which are the current state-of-the-art models. Due to resource unavailability, pretraining full precision LLMs like GPT-OSS, Deepseek R1, Qwen 3, Llama 3.1 etc., which have enormous requirements for training data and GPU resources, becomes unviable at the current stage, but we can use LoRa adapters to finetune a portion of the weights of the model to finetune on the training data albeit with low accuracy score (Albladi et al., 2025). They suggested that future works in LLM specifically on regional languages like Bengali language will possibly pave the way for hate speech detection on full precision LLM models. In our case, due to resource limitations, we have opted for BanglaBERT along with BanglaHateBERT for this task.

3 Dataset

The dataset (Hasan et al., 2025a) [1] used has been provided by the organizers of second Bangla Language Processing (BLP) Workshop for the shared task on Bangla Hate Speech Detection (Hasan et al., 2025b). The dataset consists of 3 subtasks. For all subtasks, each row has a Bengali text collected from YouTube comments. These are then labelled for each subtask appropriately. For subtask 1A, the label is the hate speech category of types *Abusive*, *Sexism*, *Religious Hate*, *Political Hate*, *Profane*, or *None*. For subtask 1B, the label is of the entity towards which the hate speech is directed and is of the following types: *Individuals*, *Organizations*, *Communities*, or *Society*. For subtask 1C, both of the labels from subtask 1A and 1C are included along with a new label categorizing the hate severity and is of the types: *Little to None*, *Mild*, or *Severe*.

The train, validation, devtest and final test consists of 35,522, 2,512, 2,512 and 10,200 instances respectively. For data augmentation, the Bengali train text sets were first translated to English. Next, the pegasus_paraphrase[2] model was used to paraphrase the sentences and then translated back to Bengali again. This added 28,220 instances to the train set.

4 System Description

The BanglaBERT (Bhattacharjee et al., 2022) model was used for the given Hate Speech Detection subtasks with specific finetuning for each of the subtasks. For subtask 1A and 1B, there only had a single output label, so straightforwardly Hugging Face trainer was used to predict the class. Whilst in subtask 1C, there were three outputs to predict, hence the code was modified by outputting the one hot encoded form of all the three labels in a single row, from where the required output was simply extracted. This is simple to implement as it can be simply extracted by slicing, considering the number of unique class labels for each output column. After this, the loss for each output column was calculated and added. By minimizing this combined loss function, the model learns to predict all the output columns correctly simultaneously. Data augmentation has been used to increase train set size, as mentioned in section 3. The model used batch size 16, dropout 0.1 while rest of parameters were set to default values. The models were trained on a single Nvidia L4 GPU with 24GB of VRAM for 10 epochs with early stopping. The training process took about one and half hours for each subtask. F1 score metric was then used to evaluate the models on the basis of the resulting outputs in every subtask. After BanglaBERT models were finetuned, BanglaHateBERT models were finetuned as well, but they had worse performance than BanglaBERT and therefore not submitted for the Shared Task.

5 Result and Analysis

The BLP 2025 shared task organizer (Hasan et al., 2025b) published the evaluation results of the three subtasks of Task 1. The shared task 1 includes three subtasks, namely, subtask 1A: Single label categorization of hate speech type, subtask 1B: Single label categorization of targeted entity of Hate Speech, subtask 1C: Multilabel categorization of hate speech type, severity and targeted entity. Herein, the participation of the research paper was present in all the 3 subtasks 1A, 1B and 1C with a team named Bahash-AI and achieved the 20th, 20th and 17th position of the 3 subtasks out of total 37, 24 and 21 total number of teams who participated

[1] https://github.com/AridHasan/blp25_task1

[2] https://hf.co/tuner007/pegasus_paraphrase

in the subtasks 1A, 1B and 1C respectively.

The evaluation metric used was the F1 score, which was calculated locally using the predictions and the corresponding gold label files provided by the organizers for each subtask. The computation was performed with a simple Python script available in the dataset's GitHub repository, as mentioned in Section 3. For the test set's gold labels which were not provided to us by the organizer before the competition ended, we had to upload the prediction files to Codabench website from where we got our F1 scores. We also compared with our models that were finetuned without data augmentation. The results of our system are marked in Table 1.

Task	Model	Dev	Test
1A	Fine-tuned	0.721	0.693
	Augmentation+ Fine-tuned	**0.724**	**0.703**
1B	Fine-tuned	0.715	0.689
	Augmentation+ Fine-tuned	**0.700**	**0.695**
1C	Fine-tuned	0.701	0.699
	Augmentation+ Fine-tuned	**0.695**	**0.697**

Table 1: BanglaBERT models comparison. The models and figures marked in bold are the latest submitted models for the task.

The n-gram baselines for the subtasks were provided by the organizers as 0.6075, 0.6279 and 0.6401 respectively for the dev-test set. These were thus clearly surpassed by our models, seen from the table 1. Minor improvement was observed on average in models finetuned on augmented data rather than models finetuned on original train dataset.

Now the original and augmented dataset's class label frequency for each subtask have been given in 4, 5 and 2. Since subtask 1C will also have the labels for 1A and 1B, only this class wise frequency is shown, while the rest are in appendix. The class wise accuracy (6) and confusion matrices (1, 2, 3, 4, 5) of test datasets are given in appendix A as well. From these, we can see that due to the None or Little to None categories being a majority proportion of the dataset have the most accuracy. While other categories which are in lower proportions have low accuracy. Sexism for instance in Task 1A and 1C have the lowest counts and hence have the lowest accuracy in both the models with 1A having 0 accuracy and 1C having 0.1034 accuracy. This unbalance ultimately causes the overall accuracy around the range of 0.70 (0.7028, 0.6954, 0.6969). Hence, these models can be said to be very poor in classifying the majority of labels caused by the unbalanced distribution. This could be potentially remedied by having balanced dataset distribution or by increasing data size. As we can see from 1, data augmentation using our method provided limited improvement in accuracy despite nearly doubling the train dataset. Hence, there is need to try other data augmentation procedures or look for further sources of Hate Speech data to include in our training.

Subtask	Label	Original	Augmented
Hate Type	None	19954	35877
	Abusive	8212	14701
	Political Hate	4227	7551
	Profane	2331	4175
	Religious Hate	676	1208
	Sexism	122	221
Hate Severity	Little to None	23489	42199
	Mild	6853	12262
	Severe	5180	9272
To Whom	None	21190	38085
	Individual	5646	10124
	Organization	3846	6909
	Community	2635	4719
	Society	2205	3896

Table 2: Class label frequency for Subtask 1C augmented training data.

Being the case that the augmentation technique used was highly susceptible to translation noise and semantic drift, further experiments were conducted on the original dataset along with oversampling and undersampling methods to get balanced output class frequency. The BanglaHateBERT model mentioned in Section 2 was utilized and finetuned, which is specifically made for Bangla language hate speech. This does solve the issue of highly unbalanced class dataset, but it doesn't yield performance improvement, and hence these models were not submitted to the Shared Task. The results are given in Table 3.

The reason for the BanglaHateBERT model being worse than BanglaBERT on the original dataset need to be investigated further, however it can

at least be understood that simple oversampling and undersampling strategies won't be helpful for BanglaBERT model and other synthetic data generation methods have to be explored in future works.

Task	Sampling Strategy	Dev	Test
1A	Original	0.654	0.659
	Oversampled	0.631	0.623
	Undersampled	0.433	0.436
1B	Original	0.664	0.649
	Oversampled	0.531	0.542
	Undersampled	0.612	0.597
1C	Original	0.622	0.607
	Oversampled	0.187	0.192
	Undersampled	0.315	0.311

Table 3: Comparison of different sampling strategies (Original, Oversampling, and Undersampling) for fine-tuning with BanglaHateBERT.

6 Conclusion

This paper showcases our research work in subtask 1, Bangla Hate Speech Detection at BLP Workshop. To tackle the problem of low number of train rows, we have used a data augmentation strategy alongside a specific language oriented pre-trained model, BanglaBERT that shows remarkable accuracy despite the lack of data and training resources. BanglaHateBERT along with oversampling and undersampling methods didn't help in performance improvement compared to original dataset and were worse than BanglaBERT based models. Future work can tackle the problem by introducing more sources of data along with code-mixed and latinized data, more methods of synthetic data generation and more sophisticated models which aligns more with real life situations, which would be specially useful in countries with diverse languages such as India.

References

Aish Albladi, Minarul Islam, Amit Das, Maryam Bigonah, Zheng Zhang, Fatemeh Jamshidi, Mostafa Rahgouy, Nilanjana Raychawdhary, Daniela Marghitu, and Cheryl Seals. 2025. Hate speech detection using large language models: A comprehensive review. *IEEE Access*, 13:20871–20892.

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*.

Cambridge Dictionary. Hate speech. Accessed on 23 September 2025.

Magdalena Celuch, Reetta Oksa, Noora Ellonen, and Atte Oksanen. 2023. Self-censorship among online harassment targets: the role of support at work, harassment characteristics, and the target's public visibility. *Information, Communication & Society*, 27:1–20.

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.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

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.

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.

Jitendra Singh Malik, Hezhe Qiao, Guansong Pang, and Anton van den Hengel. 2024. Deep learning for hate speech detection: a comparative study. *International Journal of Data Science and Analytics*, pages 1–16.

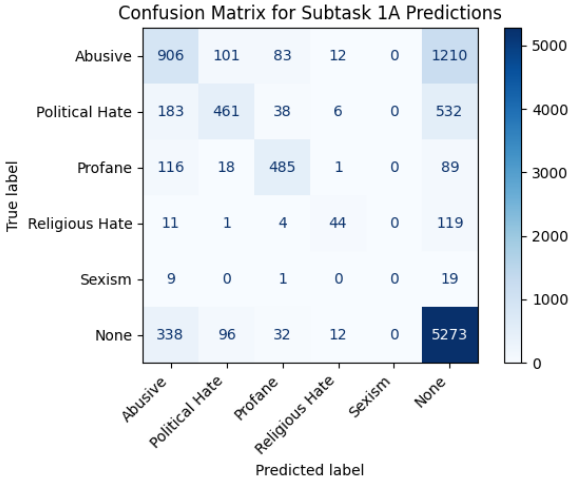

Figure 1: Subtask 1A

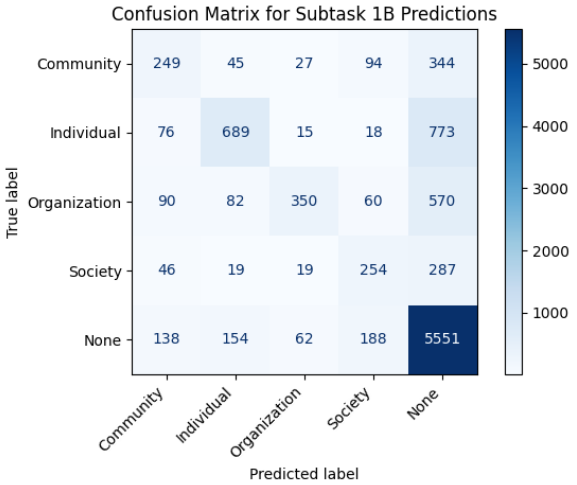

Figure 2: Subtask 1B

Francimaria Nascimento, George Cavalcanti, and Márjory Da Costa-Abreu. 2023. Exploring automatic hate speech detection on social media: A focus on content-based analysis. *SAGE Open*, 13.

Surya Putra and Sisila Damanik. 2021. Hate speech about politics in social media. *Talenta Conference Series: Local Wisdom, Social, and Arts (LWSA)*, 4.

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.

Jeremy Waldron. 2012. *Notes*, pages 235–278. Harvard University Press, Cambridge, MA and London, England.

A Appendix

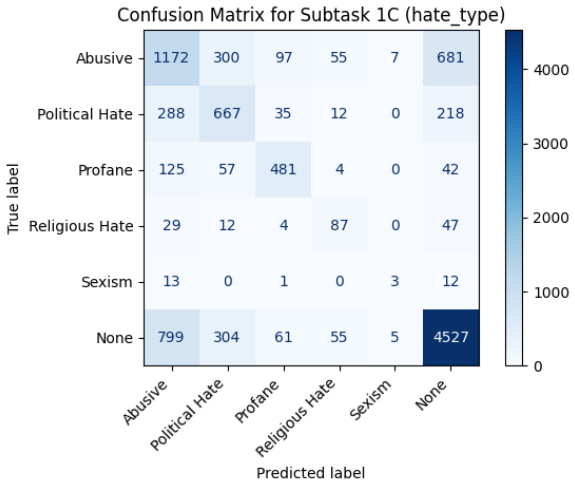

Figure 3: Subtask 1C Hate Type

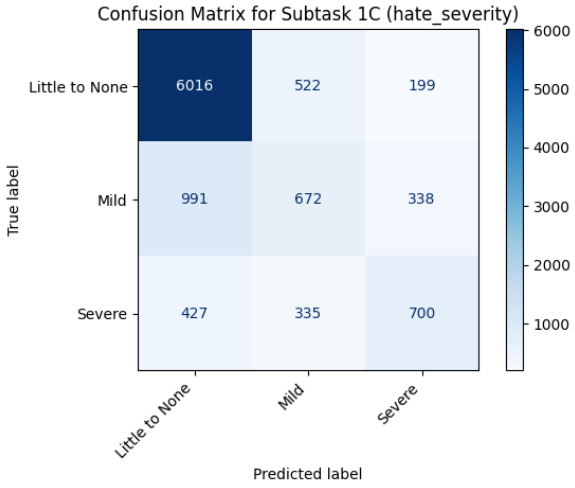

Figure 4: Subtask 1C Hate Severity

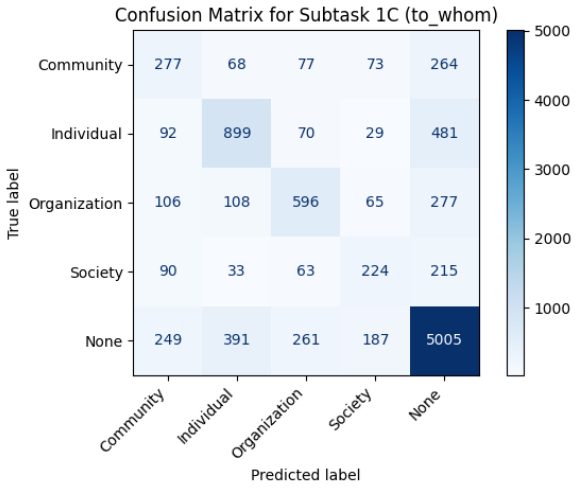

Figure 5: Subtask 1C To Whom

Label	Original	Augmented
None	19954	35877
Abusive	8212	14701
Political Hate	4227	7551
Profane	2331	4175
Religious Hate	676	1208
Sexism	122	221

Table 4: Subtask 1A: Frequency count of labels before and after augmentation.

Label	Original	Augmented
None	21190	38085
Individual	5646	10124
Organization	3846	6909
Community	2635	4719
Society	2205	3896

Table 5: Subtask 1B: Frequency count of labels before and after augmentation.

Subtask	Label	Accuracy
1A	Abusive	0.392
	Political Hate	0.378
	Profane	0.684
	Religious Hate	0.246
	Sexism	0.000
	None	0.917
1B	Community	0.328
	Individual	0.439
	Organization	0.304
	Society	0.406
	None	0.911
1C	**Hate Type**	
	Abusive	0.507
	Political Hate	0.547
	Profane	0.678
	Religious Hate	0.486
	Sexism	0.103
	None	0.787
	Hate Severity	
	Mild	0.336
	Severe	0.479
	Little to None	0.893
	To Whom	
	Community	0.365
	Individual	0.572
	Organization	0.517
	Society	0.358
	None	0.821

Table 6: Label-wise accuracy for different subtasks.