HateNet-BN at BLP-2025 Task 1: A Hierarchical Attention Approach for Bangla Hate Speech Detection

Mohaymen Ul Anam[1*], Akm Moshiur Rahman Mazumder[1*], Ashraful Islam[1], AKM Mahbubur Rahman[1], M. Ashraful Amin[1]

[1]*Center for Computational & Data Sciences, Independent University, Bangladesh (IUB)*

[*]Equal Contribution **Correspondence:** amazumder@iub.edu.bd

Abstract

The rise of social media in Bangladesh has increased abusive and hateful content, which is difficult to detect due to the informal nature of Bangla and limited resources. The BLP 2025 shared task addressed this challenge with subtask 1A (multi-label abuse categories) and subtask 1B (target identification). We propose a parameter-efficient model using a frozen BanglaBERT backbone with hierarchical attention to capture token level importance across hidden layers. Context vectors are aggregated for classification, combining syntactic and semantic features. On subtask 1A, our frozen model with hierarchical attention achieved a micro-F1 of 0.7178, surpassing the baseline of 0.71, while the unfrozen variant scored 0.7149. Our submissions ranked 15th (Subtask 1A) and 12th (Subtask 1B), showing that layer-wise attention with a frozen backbone can effectively detect abusive Bangla text. Our code can be found here https://github.com/MOSHIIUR/BLP_Subtask1A

1 Introduction

Hate speech detection has become an important research problem in Natural Language Processing (NLP) due to its social impact and the increasing spread of harmful content online. Many studies have focused on detecting hate speech, toxic language, and abusive text. While most of these studies have concentrated on high-resource languages like English (Lee et al., 2018; Albladi et al., 2025), low-resource languages such as Bangla have received far less attention (Romim et al., 2022; Das et al., 2022). The Bangla language presents unique challenges due to its complex vocabulary, lack of clear word separation, and regional language variations, making hate speech detection particularly difficult.

The primary objective of this paper is to detect hate speech in Bangla on social media using a layer-wise hierarchical transformer. The Second Bangla Language Processing Workshop (BLP), co-located with IJCNLP-AACL, organized a shared task (Hasan et al., 2025b) and provided a novel dataset for this purpose (Hasan et al., 2025a). The dataset is annotated for multiple subtasks, including subtask 1A, which identifies the type of hate with six classes comprising Abusive, Sexism, Religious Hate, Political Hate, Profane, and None, and subtask 1B, which identifies the target of hate with five classes consisting of Individual, Organization, Community, Society, and None.

To achieve this objective, we have experimented with two transformer-based models: XLM-RoBERTa, and BanglaBERT. Our three main contributions are:

- A parameter-efficient hierarchical attention model built on frozen BanglaBERT, reducing trainable parameters by nearly 89%.
- Improved results on subtask 1A (Micro-F1 = 0.7178), outperforming standard fine-tuning and other baselines.
- An ablation study showing that freezing the backbone improves both performance and efficiency.

2 Related Work

Early studies on Bangla hate speech detection mainly relied on classical machine learning models. For instance, (Alvi and Sharmin, 2019) collected 5,126 Bangla social media comments and achieved 70% accuracy using GRU-based models, outperforming traditional ML approaches. (Romim et al., 2022) created a larger dataset with over 50,000 offensive comments and demonstrated the limitations of standard LSTM and Bi-LSTM models on highly diverse and informal text.

Recently, transformer-based models have shown better performance for Bangla hate speech detection. (Alam et al., 2020) fine-tuned multilingual

Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 544–550

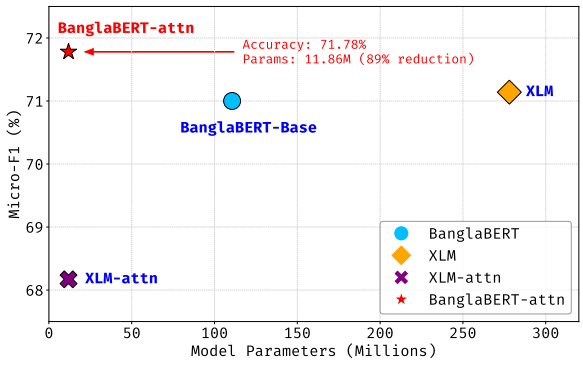


Figure 1: Micro-F1 vs. number of trainable model parameters for subtask 1A on test set. Our proposed BanglaBERT-attn model, using a frozen backbone, achieves the highest accuracy (Micro-F1) while reducing the trainable parameter count by 89% compared to the fully fine-tuned BanglaBERT.

transformers on Bangla text, achieving a 5–29% improvement over previous methods. (Das et al., 2022) developed datasets including both actual and Romanized Bangla posts, where XLM-RoBERTa achieved the best results. Similarly, (Keya et al., 2023) explored BERT-GRU models on Bangla social media comments, showing strong performance.

Recent work has also focused on informal, transliterated, and multi-modal Bangla data. (Islam et al., 2021) collected controversial Bangla social media posts and achieved up to 88% accuracy using SVM. (Karim et al., 2022) explored multimodal hate speech detection with Conv-LSTM and XLM-RoBERTa achieving Micro-F1 scores up to 83%. (Haider et al., 2024) introduced a multi-label Bangla hate speech dataset using translation-based LLM prompting.

3 Methodology

3.1 Dataset Description & Preprocessing

We utilized the dataset offered by the BLP-2025 Shared Task 1 (Hasan et al., 2025a), which consists of manually annotated YouTube comments covering diverse topics such as politics, sports, international news, and major violent incidents. The dataset is structured in a multi-task setup with three subtasks: hate type, hate target, and hate severity. For subtask 1A (hate type), the data include six classes: *Abusive, Sexism, Religious Hate, Political Hate, Profane, or None*. For subtask 1B (hate target), the classes are: *Individual, Organization, Community, Society, or None*. Table 1 summarizes the class-wise distribution across training, development, and test splits for both subtasks.

	Class	Train	Dev	Test
Subtask 1A	Abusive	8212	564	2312
	Sexism	122	11	29
	Religious Hate	676	38	179
	Political Hate	4227	291	1220
	Profane	2331	157	709
	None	19954	1451	5751
	Total	**35422**	**2512**	**10200**
Subtask 1B	Individual	5646	364	1571
	Organization	3846	292	1152
	Community	2635	179	759
	Society	2205	141	625
	None	21190	1536	6093
	Total	**35522**	**2512**	**10200**

Table 1: Overview of the Data and Splitting Procedure for subtask 1A and subtask 1B

For preprocessing, we cleaned the text by removing all non-Bangla characters and punctuation, retaining only Bangla Unicode ranges. This decision was made to reduce input noise and focus the monolingual BanglaBERT model on the lexical features of the Bangla language.

3.2 Method

We propose a hierarchical attention model that leverages multi-level representations learned by pre-trained transformer-based language models. Our model, depicted in Figure 2, introduces a layer-wise attention mechanism to dynamically weigh the importance of hidden representations from each layer of the transformer backbone. The core of our methodology consists of three main stages: (1) a layer-wise attention module, (2) a feature aggregation and projection module, and (3) a final classification head. Hyperparameters for training can be found in Appendix A.1

Problem Formulation: Formally, given a social media comment $S = \{t_1, t_2, ..., t_n\}$ as a sequence of n tokens, the objective is to learn a function $f(S) \rightarrow \hat{y}$. For subtask 1A, $\hat{y}$ is a vector of $K = 6$ binary labels $Y = \{\text{Abusive, Sexism, } \ldots, \text{None}\}$. For subtask 1B (target identification), $\hat{y}$ represents one of $K = 5$ classes $Y = \{\text{Individual, } \ldots, \text{None}\}$. The model is trained to minimize a loss function (e.g., cross-entropy) between the predicted labels $\hat{y}$ and the ground-truth labels y.

Standard fine-tuning of BERT-based models typically addresses this task by feeding the final hidden-state representation of the *[CLS]* token into a classification head. Other common approaches involve pooling the hidden states of all tokens in the final

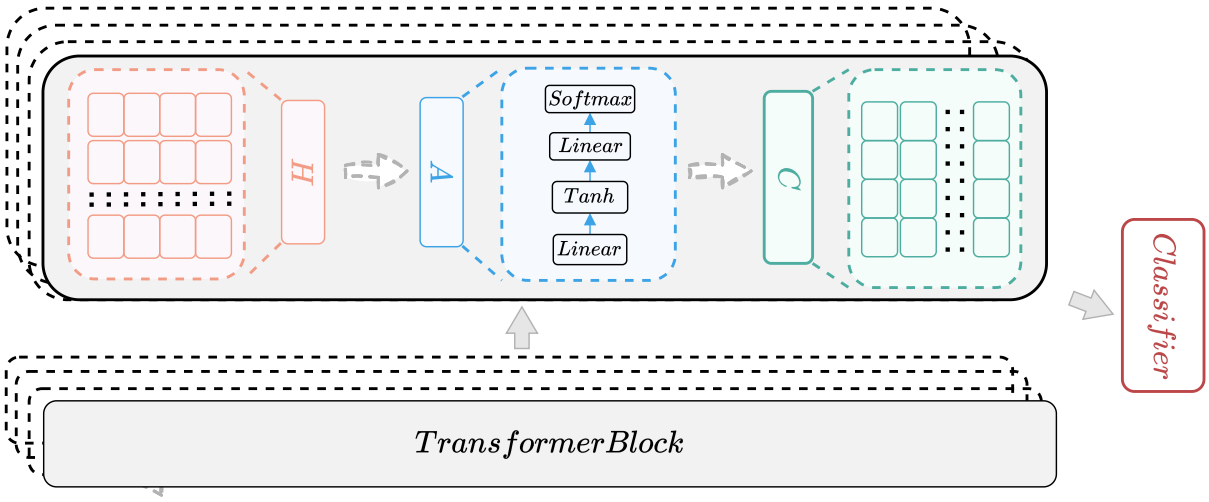


Figure 2: **Architecture of the proposed Hierarchical attention model**. For each of the L layers in the TransformerBlock, the corresponding hidden states, H are processed by a dedicated attention network, A to compute a layer-specific context vector, C. These context vectors are subsequently aggregated and projected into a single feature vector for the final Classifier.

layer. In contrast, our HateNet-BN model hypothesizes that valuable information for hate speech detection is distributed across all layers of the transformer. We posit that intermediate layers capture crucial syntactic and semantic cues that are diluted or lost in the final layer's representation. Our model, therefore, introduces a hierarchical attention mechanism to explicitly weigh and aggregate representations from each layer, creating a richer, multi-level feature vector for classification

Layer-wise Attention Mechanism: Instead of relying solely on the final layer's, our model attends to the representations from all intermediate layers. Let the set of hidden state matrices from the L layers of the transformer be $\{\mathbf{H}^{(1)}, \mathbf{H}^{(2)}, \ldots, \mathbf{H}^{(L)}\}$, where each $\mathbf{H}^{(l)} \in \mathbb{R}^{n \times d_h}$ contains the hidden states for a sequence of length n, and d_h is the hidden dimension.

We introduce a set of L distinct attention networks, $\{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_L\}$, one for each transformer layer. Each attention network, $\mathcal{A}_l$, is a feed-forward network that learns to assign an importance weight to each token's representation within its corresponding hidden state matrix $\mathbf{H}^{(l)}$.

For a given hidden state $\mathbf{h}_i^{(l)}$ (the i-th token at layer l), the attention network $\mathcal{A}_l$ computes a scalar weight $w_i^{(l)}$:

$$w_i^{(l)} = \mathcal{A}_l(\mathbf{h}_i^{(l)}) \tag{1}$$

These learned weights are then applied to their corresponding hidden states to produce a weighted representation for that layer. This allows the model to emphasize tokens that are more relevant to the classification task at that specific level of abstraction. A layer-specific context vector, $\mathbf{c}^{(l)} \in \mathbb{R}^{d_h}$, is then generated by computing the weighted sum of the token hidden states:

$$\mathbf{c}^{(l)} = \sum_{i=1}^{n} w_i^{(l)} \mathbf{h}_i^{(l)} \tag{2}$$

This process is repeated for all L layers, yielding a set of context vectors $\{\mathbf{c}^{(1)}, \mathbf{c}^{(2)}, \ldots, \mathbf{c}^{(L)}\}$, where each vector encapsulates the most salient information from its respective layer. This layer-wise approach is particularly advantageous for short social media posts. In such condensed texts, meaning is often implicit, and subtle cues (e.g., a single profane word, a sarcastic tone) are critical. By attending to all layers, our model can simultaneously leverage low-level features (like token-level profanity from early layers) and high-level semantic context (like sentence-level sarcasm from deeper layers), which might be overly smoothed in a final-layer-only representation.

Feature Aggregation and Projection: The set of layer-wise context vectors must be aggregated into a single feature vector for the final classification. In this configuration, all layer-wise context vectors are concatenated to form a single, high-dimensional vector, preserving the distinct information from each layer.

$$\mathbf{c}_{\text{concat}} = [\mathbf{c}^{(1)}; \mathbf{c}^{(2)}; \ldots; \mathbf{c}^{(L)}] \tag{3}$$

where $\mathbf{c}_{\text{concat}} \in \mathbb{R}^{L \cdot d_h}$. This vector is subsequently

Subtask 1A									
Model	**Attn. Net.**	**Freeze**	**Abusive**	**Sexism**	**Religious Hate**	**Political Hate**	**Profane**	**None**	**Micro-F1**
BanglaBERT	×	×	0.4743	0	0.4225	0.5371	0.6641	0.8246	0.71
BanglaBERT	✓	✓	0.5227	0	0.4677	0.5794	0.6885	0.8283	**0.7178**
BanglaBERT	✓	×	0.4708	0	0.3643	0.5687	0.7202	0.8246	0.7149
XLM-RoBERTa	×	×	0.5172	0	0.4730	0.5743	0.7440	0.8179	0.7114
XLM-RoBERTa	✓	✓	0.4282	0	0.3547	0.5669	0.5938	0.8010	0.6817
XLM-RoBERTa	✓	×	0.5180	0.0606	0.4777	0.5859	0.7478	0.8212	**0.7143**

Subtask 1B								
Model	**Attn. Net.**	**Freeze**	**Individual**	**Organization**	**Community**	**Society**	**None**	**Micro-F1**
BanglaBERT	×	×	0.6269	0.5757	0.4375	0.4151	0.8299	**0.7187**
BanglaBERT	✓	✓	0.5973	0.5675	0.4068	0.3843	0.8271	0.7148
BanglaBERT	✓	×	0.6315	0.5755	0.4384	0.4173	0.8256	0.7142
XLM-RoBERTa	×	×	0.5976	0.5738	0.4062	0.4201	0.8231	0.7137
XLM-RoBERTa	✓	✓	0.5163	0.5527	0.3341	0.3179	0.8045	0.6857
XLM-RoBERTa	✓	×	0.6084	0.5915	0.4410	0.4204	0.8246	**0.7149**

Table 2: Comparison of Micro-F1 scores between baseline models and our proposed attention network (Attn. Net.) on both subtasks. Results are shown for the attention network with unfrozen(×) and frozen(✓) backbone.

passed through a linear projector layer to reduce its dimensionality back to the model's hidden size, d_h.

$$\mathbf{f} = \mathbf{W}_p \mathbf{c}_{\text{concat}} + \mathbf{b}_p \tag{4}$$

Here, $\mathbf{W}_p \in \mathbb{R}^{d_h \times (L \cdot d_h)}$ and $\mathbf{b}_p \in \mathbb{R}^{d_h}$ are the learnable parameters of the projector.

Classification and Training: The final feature vector $\mathbf{f}$ is fed into a linear classification head, which maps it to the logits for the K target classes.

$$\mathbf{z} = \mathbf{W}_c \mathbf{f} + \mathbf{b}_c \tag{5}$$

where $\mathbf{W}_c \in \mathbb{R}^{K \times d_h}$ and $\mathbf{b}_c \in \mathbb{R}^K$ are the weight and bias parameters of the classifier. The final class probabilities are then obtained by applying the softmax function to the logits:

$$P(y|S) = \text{softmax}(\mathbf{z}) \tag{6}$$

The entire model is trained end-to-end by minimizing the cross-entropy loss $\mathcal{L}$ between the predicted probabilities and the ground-truth labels. For a single training instance, the loss is given by:

$$\mathcal{L} = -\sum_{k=1}^{K} y_k \log(P(y_k|S)) \tag{7}$$

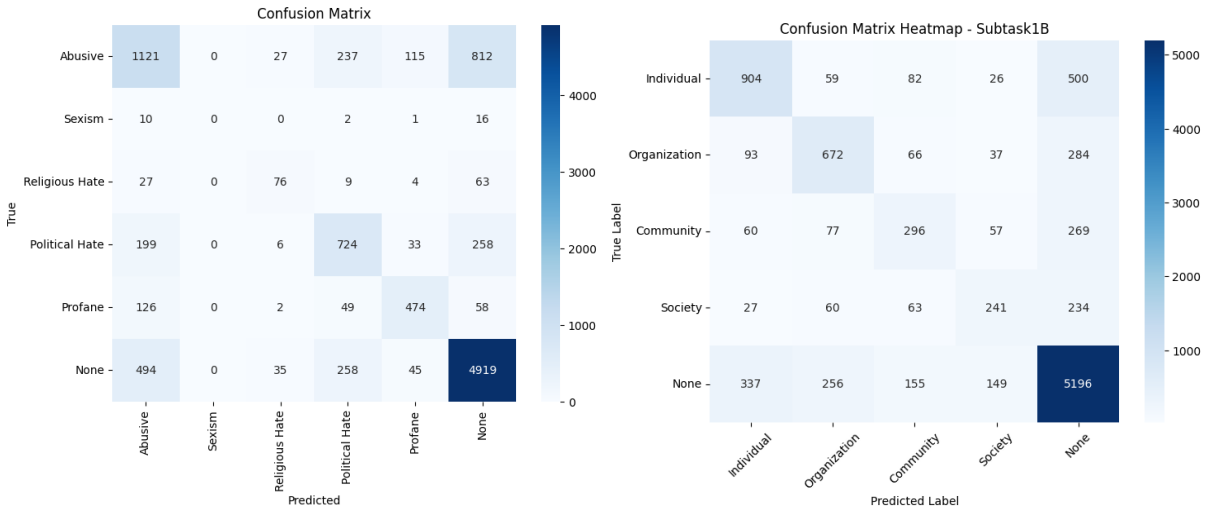
where y_k is a one-hot encoded vector representing the true class label. During training, we can optionally freeze the parameters of the underlying transformer backbone.

4 Results Analysis & Discussion

To evaluate the efficacy of our proposed layer-wise attention mechanism, we conducted a series of experiments on two distinct tasks: subtask 1A and subtask 1B. We benchmarked two transformer-based models, the monolingual BanglaBERT (Bhattacharjee et al., 2022) and the multilingual XLM-RoBERTa (Conneau et al., 2019). For each model, we evaluated three primary configurations: (1) A baseline model using standard fine-tuning without our attention network. (2) Our proposed model with the attention network applied to a fully fine-tuned (unfrozen) backbone. (3) Our proposed model with the attention network applied to a frozen backbone, where only trainable parameters are in our lightweight attention networks ($\mathcal{A}_l$) and the final classification layer (W_c, b_c) which resulting in 89% parameter reduction. For a 12-layer BERT-base model ($L = 12$, $d_h = 768$), this amounts to only 11.86M parameters. The performance of each configuration, measured by Micro-F1 score, is detailed in Table 2.

Performance on Subtask 1A: our proposed attention mechanism demonstrated a clear positive impact, particularly for the monolingual model. BanglaBERT equipped with the attention network and a frozen backbone achieved the highest score of all configurations, with a Micro-F1 of 0.7178.

This represents a notable improvement over its baseline counterpart (0.71). When the backbone was fully trained, the performance of BanglaBERT with our attention network remained strong at 0.7149, still outperforming the baseline. On the other hand, the multilingual XLM-RoBERTa model with standard fine-tuning baseline achieved a competitive score of 0.7114. In contrast to BanglaBERT, adding the attention network while keeping the model frozen led to a significant performance degradation (0.6817). However, with the fully finetuned version it does outperform the



(a) Subtask 1A

(b) Subtask 1B

Figure 3: Confusion matrix for BanglaBERT with hierarchical attention and frozen backbone.

baseline (0.7143).

Performance on Subtask 1B: The results for Subtask 1B reveal a more complex interaction between the models and our proposed method. For BanglaBERT, the standard fine-tuning baseline achieved the highest score of 0.7187. In this case, the introduction of the layer-wise attention network, in both frozen and unfrozen configurations, resulted in a minor decrease in performance. For XLM-RoBERTa, we observe a similar pattern to Subtask 1A. While the baseline result was competitive, adding the attention network to a frozen model again resulted in a comparatively low score of 0.6857. However, the synergy between our attention mechanism and a fully finetuned backbone was once again evident. This variant achieved the highest score among all XLM-RoBERTa configurations for this subtask, reaching a Micro-F1 of 0.7149.

Error Analysis: The confusion matrix for our best-performing model for subtask 1A(Figure 3a) reveals specific strengths and weaknesses. The model excels at identifying the *None* (4919 correct) and *Abusive* (1121 correct) classes. However, significant confusion exists between semantically similar categories. For instance, 126 *Profane* instances were misclassified as *Abusive*, and 199 *Political Hate* instances were also misclassified as *Abusive*. Furthermore, 812 *Abusive* instances were incorrectly labeled *None*, highlighting the difficulty in distinguishing low-severity abusive content from neutral text

Similarly, the confusion matrix for subtask 1B(Figure 3b) shows strong performance on the *None* (5196 correct) and *Individual* (904 correct) target classes. The primary sources of confusion are between *Organization* and *None* (284 misclassifications), and between *Community* and *None* (269 misclassifications), suggesting the model struggles to identify targets when the hateful content is not explicit.

5 Conclusion

In this paper, we introduced HateNet-BN, a parameter-efficient, hierarchical attention model designed to navigate the complexities of Bangla hate speech detection. Our experiments, conducted for the BLP-2025 Shared Task 1, demonstrated that a lightweight attention mechanism can effectively extract and weigh features from the different layers of a large, pre-trained transformer model. Our most significant finding was that the proposed model, when paired with a frozen BanglaBERT backbone, achieved a Micro-F1 score of 0.7178 on Subtask 1A. This result not only surpassed the standard, fully fine-tuned BanglaBERT baseline (0.7104 Micro-F1) but did so while reducing the number of trainable parameters by approximately 89%. This work successfully shows that a surprisingly small set of attention networks can effectively leverage the rich, multi-level representations of a large, frozen language model, offering a solution that is simultaneously high-performing and computationally efficient.

Limitations

Despite its effectiveness, this very efficiency raises critical questions about the scalability of our approach. While effective on the 12-layer architecture of BERT-base models, it is uncertain how this method would scale to significantly deeper models with 24, 48, or more layers. The linear increase in attention networks one for each layer could introduce its own optimization challenges and diminish the parameter-efficiency gains observed here. Future work should therefore investigate strategies for selective or shared attention mechanisms that can maintain this efficiency as model architectures continue to grow in scale.

References

Tanvirul Alam, Akib Khan, and Firoj Alam. 2020. Bangla text classification using transformers. *arXiv preprint arXiv:2011.04446*.

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*.

Md Ishmam Alvi and Sadia Sharmin. 2019. Hateful speech detection in public facebook pages for the bengali language. In *2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 555–560. IEEE.

Abhik Bhattacharjee, Tahmid Hasan, Kazi Mubasshir, Md. Saiful Islam, Wasi Ahmad Uddin, 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 North American Chapter of the Association for Computational Linguistics: NAACL 2022*.

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. 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.

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

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.

Tanvirul Islam, Nadim Ahmed, and Subhenur Latif. 2021. An evolutionary approach to comparative analysis of detecting bangla abusive text. *Bulletin of Electrical Engineering and Informatics*, 10(4):2163–2169.

Md Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Md Shajalal, and Bharathi Raja Chakravarthi. 2022. Multimodal hate speech detection from bengali memes and texts. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 293–308. Springer.

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

Ho-Suk Lee, Hong-Rae Lee, Jun-U Park, and Yo-Sub Han. 2018. An abusive text detection system based on enhanced abusive and non-abusive word lists. *Decision Support Systems*, 113:22–31.

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.

A Appendix

A.1 Hyperparameters

We employed two transformer-based models: XLM-RoBERTa base,BanglaBERT. Our approach involved fine-tuning these models on the preprocessed dataset. Each model was trained for three epochs, a duration sufficient for convergence

on the dataset and avoid model overfitting and underfitting. In order to enhance the model's performance, a batch size of 16 was utilized to accelerate the training procedure. The selection of a learning rate of 2e-5 was based on the principle that this rate facilitates more efficient learning of parameter estimates by the algorithm. Table 3 presents the hyperparameter used for this task.

Hyperparameter	**Value**
Learning rate	2e-5
Optimizer	Adam
Batch size	16
Number of epochs	3
Warmup ratio	0.1
Weight decay	0.01
LR scheduler	Cosine
Metric for best model	Micro-F1

Table 3: Hyperparameters used for fine-tuning transformer models