Hate and Offensive Speech Detection in Bengali Language Using Machine Learning and Deep Learning Techniques

Sajjad Hossain Pappu
Department of
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
sajjad.hossain.pappu@g.bracu.ac.bd

Azizul Kabir Jayed
Department of
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
azizul.kabir.jayed@g.bracu.ac.bd

YASIR
Department of
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
yasir@g.bracu.ac.bd

***Abstract*—The rapid growth of social media has increased the spread of abusive and hateful content in low-resource languages such as Bengali, creating an urgent need for reliable automatic moderation tools. This work studies multi-class Bengali hate/offensive speech detection using a publicly available dataset of 3,418 Bengali comments labeled into five categories: *Gender abusive*, *Geopolitical*, *Personal*, *Political*, and *Religious*. We compare recurrent neural network (RNN) variants (SimpleRNN, LSTM, GRU and their bidirectional forms) initialized with Bengali GloVe embeddings, against transformer-based fine-tuning using BanglaBERT (Base and Large). Models are evaluated on a held-out stratified test split using accuracy and macro-averaged precision/recall/F1 to account for class imbalance. Results show that BanglaBERT Base achieves the best performance (80.85% accuracy, 0.776 macro-F1), while the best RNN baseline (GRU) reaches 67.54% accuracy and 0.601 macro-F1. We further analyze common error patterns using a confusion matrix and discuss why non-gated RNNs (SimpleRNN) underperform.**

***Index Terms*—Bengali NLP, hate speech detection, offensive language, RNN, BanglaBERT, macro-F1**

I. Introduction

Social media enables fast information sharing but also facilitates the spread of abusive language and hate speech. Manual moderation is costly and does not scale, motivating automated NLP-based detection systems.

Hate/offensive speech detection is well studied for high-resource languages (e.g., English) but remains challenging for Bengali due to limited labeled data, informal spelling variations, and complex morphology. In this project, we perform a comparative study of deep learning models for Bengali multi-class hate/offensive speech categorization.

Our key contributions are: (i) a consistent experimental pipeline (preprocessing, stratified splitting, and evaluation) for Bengali hate speech classification, (ii) a comparison of multiple RNN architectures vs. BanglaBERT fine-tuning, and (iii) an error analysis highlighting where models fail and why.

II. Related Work

Early Bengali hate and offensive speech studies primarily relied on classical machine learning pipelines and lightweight neural models due to limited labeled data. Ishmam and Sharmin collected comments from public Facebook pages and evaluated feature-based baselines as well as gated recurrent units (GRU) for Bengali hateful speech detection [1].

More recent work highlights the effectiveness of transformer representations and cross-lingual transfer for low-resource settings. Das *et al.* introduced an annotated dataset containing both Bengali and Romanized Bengali text and benchmarked multilingual transformers (e.g., XLM-R and MuRIL) for hate/offensive content detection [3].

Explainability has also emerged as an important direction for automated moderation. Karim *et al.* proposed DeepHateExplainer, combining transformer-based models with explanation methods such as sensitivity analysis and layer-wise relevance propagation to provide human-interpretable rationales for Bengali hate predictions [6]. Jobair *et al.* explored BERT-based Bengali hate speech detection alongside deep learning baselines [7], and Islam *et al.* studied a multi-step binary + multi-label setup for finer-grained offensive categorization using recurrent models [8]. A comprehensive survey by Maruf *et al.* summarizes Bengali hate speech datasets, modeling trends, and open challenges [9].

Beyond hate-only settings, toxicity and sentiment studies provide relevant modeling and preprocessing insights for Bengali social media text. Belal *et al.* proposed a two-stage pipeline for Bengali toxic comment classification with interpretability via LIME [11], and Haque *et al.* investigated multi-class Bengali social media comment classification and reported strong performance from hybrid CNN–LSTM architectures on a large labeled dataset [10].

Motivated by these trends, we compare word-embedding-based recurrent models (GloVe [5]) against transformer fine-tuning baselines, leveraging Bengali language resources such as BanglaBERT [4].

III. Methodology

Fig. 1 summarizes the end-to-end pipeline used in this study.

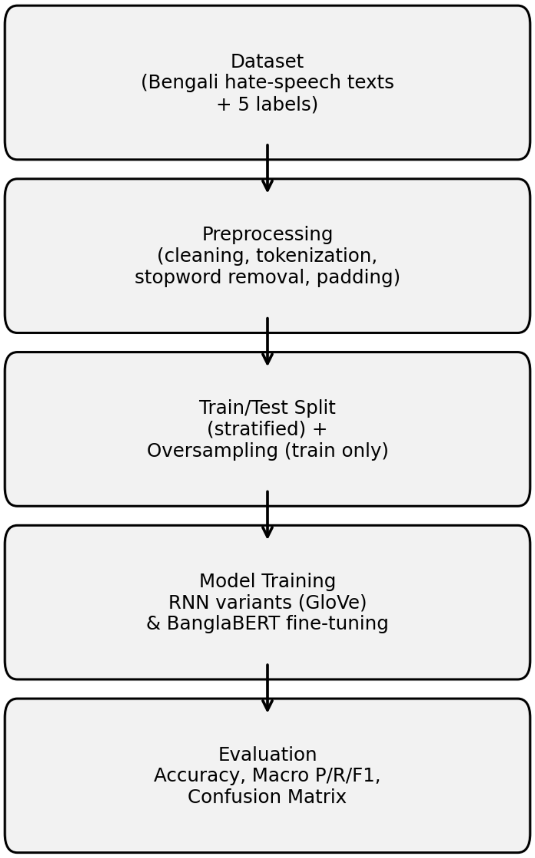


Fig. 1. Overall workflow of the proposed system.

A. Dataset and Preprocessing

We use a Bengali hate speech dataset containing 3,418 labeled comments across five categories: Gender abusive, Geopolitical, Personal, Political, and Religious [2]. We perform standard text cleaning (URL/noise removal, punctuation cleanup, whitespace normalization) and remove Bengali stopwords.

We split the dataset into 80% training and 20% testing using stratification to preserve label proportions. Because the labels are imbalanced (e.g., fewer Gender abusive samples), we apply random oversampling *only on the training split* to reduce bias toward majority classes.

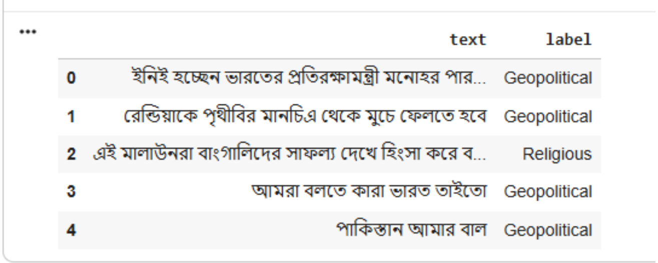
We also include a small qualitative snapshot of the dataset and an example notebook output for transparency.

B. Model Architectures and Training

We evaluate eight deep learning architectures:

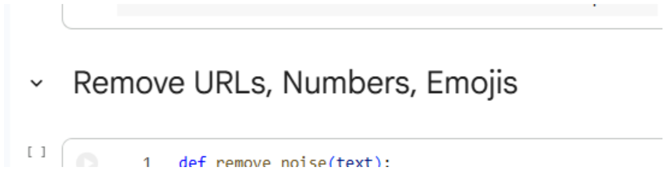
- **RNN baselines:** SimpleRNN, LSTM, GRU, and their bidirectional versions. These models use a word-level tokenizer with sequence length 50. The embedding layer is initialized with 300-dimensional Bengali GloVe vectors [5] and kept frozen to reduce overfitting on a small dataset.
- **Transformer models:** BanglaBERT Base and BanglaBERT Large [4], fine-tuned for multi-class classification. We tokenize with the official BanglaBERT tokenizer and use a maximum sequence length of 128.

For RNN models, we train for 20 epochs using Adam (learning rate 0.001) with a 10% validation split from the (oversampled) training data. For BanglaBERT models, we fine-tune for 3 epochs using an AdamW-style optimizer schedule; we use smaller batch sizes to fit memory constraints.



	text	label
0	ইনিই হচ্ছেন ভারতের প্রতিরক্ষামন্ত্রী মনোহর পার...	Geopolitical
1	রেন্ডিয়াকে পৃথীবির মানচিএ থেকে মুচে ফেলতে হবে	Geopolitical
2	এই মালাউনরা বাংগালিদের সাফল্য দেখে হিংসা করে ব...	Religious
3	আমরা বলতে কারা ভারত তাইতো	Geopolitical
4	পাকিস্তান আমার বাল	Geopolitical

Fig. 2. Sample labeled Bengali comments from the dataset (exploratory view).



Remove URLs, Numbers, Emojis

```
def remove_noise(text):
```

Fig. 3. Example notebook output highlighting a key preprocessing step (removing URLs, numbers, and emojis).

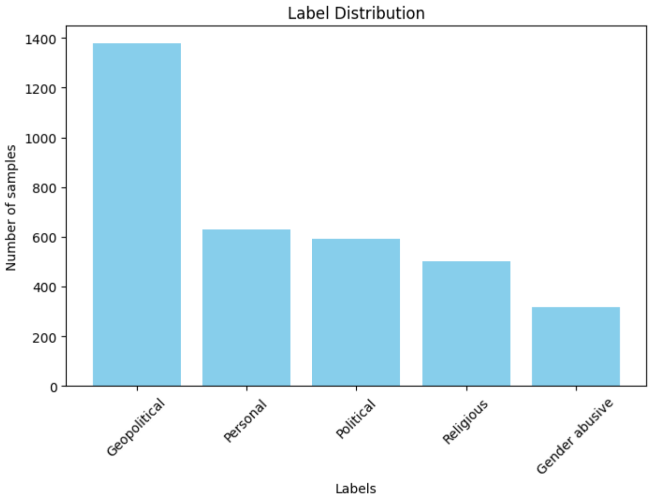


Fig. 4. Class distribution (support) in the held-out test set.

C. Evaluation Protocol

We report **accuracy** as an overall measure of correctness. Since the dataset is imbalanced, we also report **macro-averaged precision, recall, and F1-score**, which weigh all classes equally. In addition, we use confusion matrices for qualitative error analysis.

IV. Results

Table I summarizes the test-set performance. BanglaBERT Base achieves the best results, while GRU is the strongest RNN baseline. BanglaBERT Large slightly underperforms the Base model in this setting, suggesting that the larger capacity is harder to fine-tune effectively with limited labeled data.

TABLE I
Classification performance on the Bengali hate/offensive speech test set.

Model	Acc. (%)	Macro-P	Macro-R	Macro-F1
SimpleRNN	32.46	0.200	0.300	0.242
Bi-RNN	34.50	0.225	0.308	0.256
LSTM	51.32	0.514	0.510	0.504
BiLSTM	57.01	0.559	0.543	0.549
GRU	67.54	0.639	0.588	0.601
BiGRU	65.20	0.608	0.602	0.601
BanglaBERT Base	80.85	0.765	0.790	0.776
BanglaBERT Large	79.97	0.747	0.781	0.758

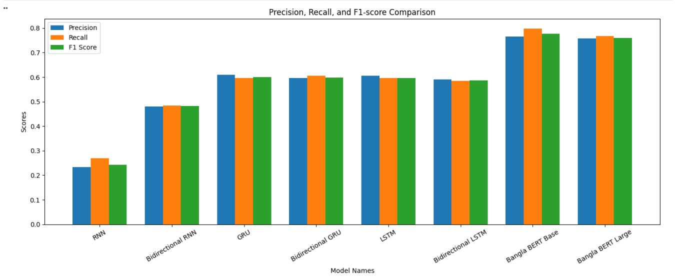


Fig. 5. Comparison of macro-averaged precision, recall, and F1-score across different models.

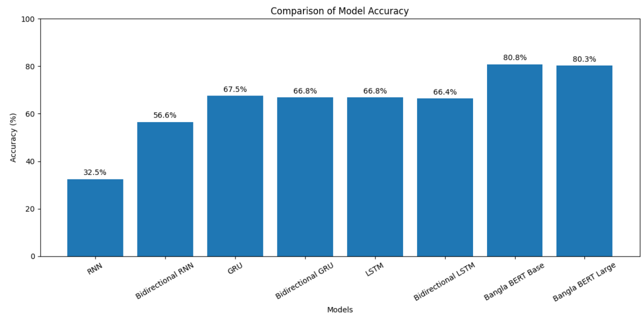


Fig. 6. Accuracy comparison of deep learning models on the Bengali hate and offensive speech dataset.

Fig. 5 illustrates the significant performance gap between SimpleRNN and more advanced architectures, where low precision and recall lead to poor macro F1-score. Fig. 6 further confirms that Transformer-based models substantially outperform RNN-based models in terms of overall classification accuracy.

V. Discussion

A. Why SimpleRNN is the Lowest Performing Model

SimpleRNN performs the worst because it uses an ungated recurrent cell that struggles with vanishing gradients and therefore has difficulty learning long-range dependencies. This limitation is amplified in noisy social media text where crucial cues can appear anywhere in the sequence. In addition, our RNN pipeline uses (i) a relatively short fixed sequence length (50 tokens), which can truncate context, and (ii) static, word-level embeddings that cannot adapt to misspellings and morphological variations in Bengali. Together, these factors lead to poor generalization; empirically, SimpleRNN shows very low macro-F1 and fails to learn some minority classes reliably.

B. Error Analysis

Fig. 7 shows the BanglaBERT Base confusion matrix. Most errors occur between semantically related categories (e.g., *Personal* vs. *Gender abusive*, and *Political* vs. *Geopolitical*), suggesting that fine-grained boundaries between hate types can be ambiguous without additional context.

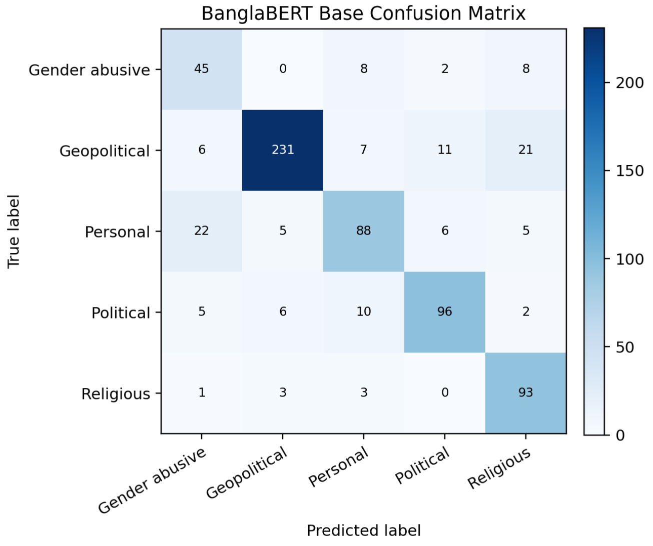


Fig. 7. Confusion matrix for BanglaBERT Base on the test set.

VI. Limitations and Future Work

A. Limitations

- **Dataset scope:** the dataset covers five hate/offensive categories and may not represent all domains, dialects, or Romanized/code-mixed Bengali.
- **Imbalance handling:** random oversampling duplicates training examples, which can increase overfitting and may not reflect real-world class priors.

- **Single split evaluation:** results are reported on one stratified train/test split; cross-validation and cross-domain testing were not performed.

B. Future Work

Future work can improve robustness and practicality by: (i) expanding data collection to include more platforms and Romanized/code-mixed Bengali, (ii) using cost-sensitive learning or data augmentation instead of simple oversampling, (iii) evaluating with cross-validation and out-of-domain test sets, and (iv) exploring lightweight/distilled transformers for deployment.

VII. Conclusion

This paper presented a comparative study of RNN variants and BanglaBERT models for Bengali multi-class hate/offensive speech detection. Experimental results show that BanglaBERT Base substantially outperforms RNN baselines, achieving 80.85% accuracy and 0.776 macro-F1. Our analysis indicates that SimpleRNN underperforms due to ungated recurrence, limited context retention, and sensitivity to noisy, morphologically rich Bengali text. The study highlights the benefit of pretrained transformers for low-resource language moderation tasks.

References

[1] A. M. Ishmam and S. Sharmin, “Hateful Speech Detection in Public Facebook Pages for the Bengali Language,” in *Proc. 18th IEEE Int. Conf. on Machine Learning and Applications (ICMLA)*, 2019, pp. 555–560, doi: 10.1109/ICMLA.2019.00104.

[2] M. R. Karim, B. R. Chakravarthi, J. P. McCrae, and M. Cochez, “Classification Benchmarks for Under-resourced Bengali Language based on Multichannel Convolutional-LSTM Network,” *arXiv:2004.07807*, 2020, doi: 10.48550/arXiv.2004.07807.

[3] M. Das, S. Banerjee, P. Saha, and A. Mukherjee, “Hate Speech and Offensive Language Detection in Bengali,” in *Proc. 2nd Conf. of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th Int. Joint Conf. on Natural Language Processing (AACL-IJCNLP)*, 2022, pp. 286–296, doi: 10.18653/v1/2022.aacl-main.23.

[4] A. Bhattacharjee, T. Hasan, W. U. Ahmad, K. Samin, M. S. Islam, A. Iqbal, M. S. Rahman, and R. Shahriyar, “BanglaBERT: Language Model Pretraining and Benchmarks for Low-Resource Language Understanding Evaluation in Bangla,” *arXiv:2101.00204*, 2022, doi: 10.48550/arXiv.2101.00204.

[5] J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global Vectors for Word Representation,” in *Proc. Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[6] M. R. Karim, S. K. Dey, T. Islam, S. Sarker, M. H. Menon, K. Hossain, B. R. Chakravarthi, M. A. Hossain, and S. Decker, “DeepHateExplainer: Explainable Hate Speech Detection in Under-resourced Bengali Language,” *arXiv:2012.14353*, 2020, doi: 10.48550/arXiv.2012.14353.

[7] M. Jobair, D. Das, N. B. Islam, and M. Dhar, “Bengali Hate Speech Detection with BERT and Deep Learning Models,” in *Lecture Notes in Networks and Systems*, vol. 867, 2024, doi: 10.1007/978-981-99-8937-9_56.

[8] M. Islam, M. S. Hossain, M. M. Islam, M. A. H. Murad, S. S. Niloy, M. Sanzidul Islam, and T. B. Islam, “Refining Bengali Hate Speech Detection: Multi-label Classification Using RNN and LSTM,” in *Proceedings of the 4th International Conference on Advances in Communication Technology and Computer Engineering (ICACTCE’24), Lecture Notes in Networks and Systems*, vol. 1313, 2025, pp. 242–253, doi: 10.1007/978-3-031-94623-3_21.

[9] A. A. Maruf, A. J. Abidin, M. M. Haque, Z. M. Jiyad, A. Golder, R. Alubady, and Z. Aung, “Hate speech detection in the Bengali language: a comprehensive survey,” *Journal of Big Data*, vol. 11, art. 97, 2024, doi: 10.1186/s40537-024-00956-z.

[10] R. Haque, N. Islam, M. Tasneem, and A. K. Das, “Multi-class sentiment classification on Bengali social media comments using machine learning,” *International Journal of Cognitive Computing in Engineering*, vol. 4, pp. 21–35, 2023, doi: 10.1016/j.ijcce.2023.01.001.

[11] T. A. Belal, G. M. Shahariar, and M. H. Kabir, “Interpretable Multi Labeled Bengali Toxic Comments Classification using Deep Learning,” *arXiv:2304.04087*, 2023, doi: 10.48550/arXiv.2304.04087.