LLMs Against Digital Deviance: Scalable Hate Speech Detection in Low-Resource and Code-Mixed Social Media

Md Jahangir Alam[1][0009−0005−8731−7354], Ismail Hossain[1][0000−0001−8954−1150], Sai Puppala[2][0009−−0008−0334−5756], and Sajedul Talukder[1][0000−0001−8054−9770]

[1] University of Texas at El Paso, TX 79902 USA
{malam10, ihossain}@miners.utep.edu, stalukder.utep.edu
[2] Southern Illinois University Carbondale, IL 62901 USA
sai.puppala@siu.edu

Abstract. In this paper we present a comprehensive study on hate speech detection in Bengali (Bangla), a low-resource language with significant online presence. We explore the potential of large language models (LLMs) such as GPT-4, Qwen, and DeepSeek in identifying hate speech from social media content, including transliterated and code-mixed text. Using a consolidated dataset combining multiple public hate speech corpora, we evaluate LLM-based prompting and fine-tuning strategies alongside traditional deep learning and transformer models. Our findings show that fine-tuned LLMs like DeepSeek-67B and GPT-4 consistently outperform smaller models, achieving macro-F1 scores above 90%, with an ensemble of DeepSeek and XLM-R reaching 91.2%. These models also demonstrate stronger robustness to domain variation and better confidence calibration. This work highlights the value of LLMs in enhancing multilingual open-source threat intelligence and sets a new benchmark for hate speech detection in under-resourced language settings.

Keywords: hate speech detection · large language models · Bangla · low-resource languages · code-mixed text · open-source intelligence

1 Introduction

The proliferation of hate speech on online platforms poses serious risks to social cohesion, user well-being, and democratic discourse. Automatic hate speech detection systems have become essential tools to mitigate these risks. While progress in high-resource languages such as English has been significant, languages like Bengali (Bangla), spoken by over 230 million people, remain underserved due to limited annotated resources and model support [3,14].

Early efforts on Bangla hate speech detection were constrained by small datasets and traditional models. Hussain et al. [11] applied classical machine learning techniques like SVM on a few thousand Facebook comments and achieved around 75% accuracy. More recent work by Romim et al. [19] introduced a 30k-comment dataset (HS-Bangla) and reported an F1 score of 86.78% using transformer-based classifiers. Additional corpora such as HS-BAN [17], BD-SHS [18], and BanTH [10] have since emerged, addressing challenges such as domain diversity and transliterated “Banglish” content. Meanwhile, large language models (LLMs) such as GPT-3.5 and GPT-4 [5], Claude [2], DeepSeek [4], and Qwen [1], trained on massive multilingual corpora, offer a promising path forward through zero-shot and few-shot learning, even for low-resource languages. Faria et al. [8] showed that GPT-3.5, when prompted in zero-shot mode, outperformed fine-tuned BERT models on Bengali hate speech detection. In this paper, we systematically explore hate speech detection in Bangla using a broad family of models. These include classical deep learning models (CNN, LSTM, MLP), transformer-based encoders (mBERT, Bangla-BERT [3], XLM-R [6]), LLMs (GPT-4, GPT-3.5, Qwen, Claude, DeepSeek) via prompting and fine-tuning, and hybrid models (e.g., BERT+GPT2, mT5 [20], ensemble systems). We consolidate and experiment on all major publicly available Bengali hate speech corpora, encompassing over 120k samples across binary and multi-class settings. All models are evaluated under a unified framework for consistent comparison.

Our research aims to answer the following research questions:

- **(RQ1)** Do large language models achieve higher classification performance on Bangla hate speech detection compared to traditional and transformer-based models?
- **(RQ2)** How well do these models generalize across diverse datasets, including transliterated and code-mixed (Banglish) content?
- **(RQ3)** Are LLM-based approaches practical for deployment in real-world systems in terms of computational cost, latency, and confidence calibration?

Our contributions are: (1) To the best of our knowledge, we provide the first comprehensive benchmark of Bengali hate speech detection using LLMs, extending prior transformer-based work [19,18] with models like GPT, Qwen, Claude, and DeepSeek. (2) We compile and leverage all public Bangla hate speech datasets, including code-mixed corpora, and summarize their characteristics. (3) We compare diverse model families: BERT-style transformers, classical DNNs, LLMs (prompted vs. fine-tuned), and hybrid or ensemble architectures. (4) We evaluate using standard metrics (accuracy, precision, recall, macro/micro-F1), plus LLM-specific measures like perplexity and Expected Calibration Error (ECE [9]). (5) We analyze trade-offs in model performance, cost, and reliability for deploying LLMs in Bangla hate speech detection. The remainder of the paper is structured as follows: Section 2 summarizes related work on hate speech detection and Bangla NLP. Section 3 outlines our model families, dataset integration, and training procedures. Section 4 discusses experimental setups and evaluation metrics. Section 5 reports results across models. Section 5.1 presents analysis and implications. Finally, Section 8 concludes the paper.

2 Related Work

2.1 Hate Speech Detection in Bangla

Bangla hate speech detection is a relatively recent research area. Early work, like Hussain et al. [11], used n-gram features and neural networks but was limited by small, informal datasets. Jahan et al. [12] addressed Banglish content using rule-based preprocessing and a CNN+LSTM, but data sparsity remained a challenge. Larger datasets appeared in 2020–2021. Karim et al. [14] released ∼4.5k labeled comments, while Romim et al. [19] compiled a 30k dataset (10k hate) achieving 79% accuracy and 85–87% F1 with FastText + Bi-LSTM. Later, Romim et al. [18] expanded to 40k+ samples (BD-SHS), reaching 86.78% macro-F1.

To handle transliteration, Jahan et al. [13] introduced BanglaHateBERT, outperforming multilingual models by 5–10% F1. Haider et al. [10] proposed *BanTH*, a 37.3k multi-label transliterated dataset, and showed GPT-4 prompting improved zero-shot accuracy. Overall, performance has improved with transformer models (87–90% macro-F1), but LLMs have not been comprehensively evaluated. Prior work also lacks analysis on calibration, explainability (e.g., HateXplain [16]), and robustness. Our study addresses these gaps using LLMs, hybrid models, and additional metrics like perplexity and ECE.

2.2 Large Language Models for Hate Speech Detection

Large Language Models (LLMs) have shown strong zero-shot performance on hate speech detection in high-resource languages like English, though bias and reliability remain concerns. For low-resource languages like Bangla, LLMs offer promise due to multilingual representation learning. Shibli et al. [8] found that GPT-3.5 and a Chinese LLM (Ernie/Gemini) performed well on Bengali hate speech in zero-shot without task-specific training. Haider et al. [10] showed translation-based prompting with GPT-4 improved Banglish hate detection.

Open-source LLMs such as *DeepSeek-Chat 67B* [4] and *Qwen-7B* [1] can be fine-tuned on Bengali, enabling adaptation to Bangla slurs or romanized text. However, their size demands significant compute and careful tuning to avoid overfitting. Hybrid models combining a BERT encoder with GPT-2 or T5 decoders have also been explored, allowing generative decoding with task-specific encoding. Our work is the first to benchmark both closed and open LLMs for Bangla hate speech, comparing prompting and fine-tuning. We also evaluate confidence calibration (ECE [9]) and perplexity to assess model trustworthiness, addressing reliability in LLM-driven moderation systems.

3 Methodology

3.1 Model Families

We evaluate a diverse set of models for Bangla hate speech detection, grouped into four major families:

Classical Deep Learning Models We begin with classical deep learning architectures that have historically served as baselines for text classification tasks. The first is a multilayer perceptron (MLP) trained on TF-IDF features extracted from character-level n-grams (1–3), allowing the model to capture morphological patterns in short social media texts. Next, we implement a convolutional neural network (CNN) based on the well-established architecture proposed by Kim [15], which applies filters of sizes 3, 4, and 5 to extract local features from token sequences and performs max-pooling to form a global sentence representation. Additionally, we train a bidirectional long short-term memory (BiLSTM) network that leverages 300-dimensional FastText Bengali word embeddings pretrained on Common Crawl. The BiLSTM captures sequential dependencies in the data, making it particularly useful for recognizing context-sensitive cues in hate speech. These classical models provide strong lexical and contextual baselines against which we compare more advanced transformer and LLM-based systems.

Fine-tuned Transformer Encoders Transformer-based models have become state-of-the-art in natural language understanding, and we fine-tune several such models specifically for Bangla hate speech detection. Our experiments include multilingual BERT (mBERT) [7], which is trained on Wikipedia and other corpora across more than 100 languages; Bangla-BERT [3], a monolingual variant trained solely on Bangla data; and XLM-RoBERTa (XLM-R) [6], which extends RoBERTa to cross-lingual learning with large multilingual corpora. We also include BanglaHateBERT [13], a fine-tuned BERT variant adapted specifically for Bangla hate speech classification. These models are fine-tuned with a standard classification head on the [CLS] token, and they serve as strong benchmarks for transformer-based sequence classification in both monolingual and multilingual contexts.

Large Language Models (LLMs) In addition to encoder-based transformers, we examine both prompted and fine-tuned large language models (LLMs). For the prompting setting, we use models such as GPT-3.5, GPT-4, Claude (v1.3), and Qwen-7B. These models are queried using zero-shot and few-shot prompts that define the task and ask the model to return a binary label ("Hate" or "Not Hate"), optionally with justification. In the fine-tuning setting, we adapt open-source generative LLMs including Qwen-7B, DeepSeek-7B, and DeepSeek-67B to the classification task by appending a label token to each input and training the model to generate that token. We also fine-tune GPT-3.5 using OpenAI's fine-tuning API. This generative classification approach aligns with the language modeling objective and allows LLMs to be used effectively for binary classification tasks, even in low-resource languages like Bangla.

Hybrid and Ensemble Models To explore whether architectural diversity leads to performance gains, we experiment with hybrid and ensemble approaches. We fine-tune the mT5-base model [20], a multilingual text-to-text transformer that performs sequence classification by generating class labels directly. Additionally, we build a hybrid model that combines a BERT encoder with a GPT-2 decoder; the BERT encoder produces contextual embeddings that are passed to the GPT-2 decoder, which is trained to output the classification label in tokenized form. Finally, we construct a simple ensemble by averaging the prediction probabilities of two top-performing models—XLM-R and DeepSeek-67B. This ensemble consistently improves macro-F1 score and expected calibration error, demonstrating that complementary model strengths can be harnessed through late fusion.

3.2 Preprocessing Strategy

To ensure consistency across all models and datasets, we apply a unified preprocessing strategy. All text is normalized by removing extraneous punctuation, URLs, and user mentions. For code-mixed or transliterated content, we retain the original script and avoid translation, preserving the linguistic and cultural nuances present in real-world Banglish expressions. Unlike English NLP tasks, we do not perform lowercasing since Bangla is a unicase language. In selected experiments, we also introduce a special token to explicitly flag offensive slurs, which helps the models focus attention on potentially hateful terms. This preprocessing pipeline is consistently applied across classical models, transformer-based encoders, and fine-tuned LLMs.

3.3 Prompting and Fine-tuning Strategy

For transformer-based models such as BERT variants and XLM-R, we follow a standard fine-tuning protocol by appending a linear classification head to the [CLS] token. These models are trained to predict one of two classes: "Hate" or "Not Hate." For LLMs used in prompting mode—such as GPT-3.5, GPT-4, Claude, and Qwen—we design task-specific prompts that clearly instruct the model to classify the input as hate speech or not. Prompts typically include the task description, possible labels, and optionally one or more examples in few-shot settings. For fine-tuned LLMs such as DeepSeek and Qwen, we formulate the classification task as next-token prediction. We append a classification token (e.g., "Hate" or "Not Hate") at the end of each input, and train the model to generate this token. This framing aligns the classification task with the generative nature of large language models and facilitates their effective use in binary prediction.

3.4 Training Details

All models are trained and evaluated using the same data splits to ensure a fair comparison across architectures. The training framework is implemented in PyTorch. For transformer-based encoders such as BERT and XLM-R, we fine-tune for 2–3 epochs using a batch size of 16 and a learning rate of $2e^{-5}$, tuned via a validation set. Fine-tuning of larger LLMs like DeepSeek and Qwen requires more aggressive learning rates (up to $1e^{-4}$) and includes learning rate warmup followed by cosine decay. The classical CNN and LSTM models are initialized with 300-dimensional FastText Bengali embeddings and use 128 filters or hidden units, respectively. The CNN model follows a multi-kernel design with filter sizes of 3, 4, and 5, while the MLP is trained on TF-IDF features derived from character n-grams. Dropout ranging from 0.3 to 0.5 is applied throughout to mitigate overfitting. All experiments are conducted on modern GPUs, and care is taken to maintain reproducibility and stability during training.

4 Experimental Setup

4.1 Datasets

Table 1 summarizes the datasets used in our experiments. We combine multiple Bangla hate speech datasets to ensure diverse coverage:

Table 1. Publicly available Bengali hate speech datasets used in this study.

Dataset	# Samples	Label Type	Reference
HS-BAN	50,314	Hate/Not	Romim et al.[17]
BD-SHS	30,000	Hate/Not	Romim et al.[19]
Bengali Hate v2.0	4,500	Multi-class†	Karim et al.[14]
BanTH	37,293	Multi-label‡	Haider et al.[10]

The primary dataset is **HS-BAN** [17], a binary-labeled corpus of ∼50k Bengali Facebook/YouTube comments (40.2% hate), covering domains like politics, religion, and sports. We converted it to binary labels: "Hate" for domain-labeled samples, "Not Hate" otherwise. **BD-SHS** [19,18] adds 30k comments (33% hate) and complements HS-BAN. **Bengali Hate v2.0** [14] has 4.5k comments labeled across five hate types and is used for fine-grained evaluation.

BanTH [10] contains 37.3k transliterated (Latin-script) multi-label YouTube comments. It tests model robustness to code-mixed Banglish, directly addressing **RQ2** on generalization. Preprocessing was minimal: emojis and excess punctuation removed for Bengali script; transliterated text kept intact. No translation was used in training, but LLM prompts included translation strategies (e.g., Banglish to Bangla/English) to improve understanding.

4.2 Training and Evaluation

We trained all models on a unified binary classification task by merging HS-BAN, BD-SHS, Bengali Hate v2.0, and BanTH. BanTH's multi-label annotations were binarized: any hate tag was labeled as 'Hate', others as 'Not Hate'. LLMs in prompting mode (GPT-3.5, GPT-4, Claude) were tested using zero- or few-shot prompts with fixed label formats (e.g., “Hate”, “Not Hate”) and were not fine-tuned. For evaluation, we report accuracy, precision, recall, macro- and micro-F1. Macro-F1 balances both classes; micro-F1 aligns with overall accuracy in binary settings. For generative LLMs, we compute **perplexity** to gauge word-level fit—lower values indicate better confidence. We also compute **Expected Calibration Error (ECE)** [9] using 10 bins. Lower ECE means better-calibrated outputs. Though temperature scaling was tested, reported ECE values use uncalibrated scores.

5 Results

In Table 2 we list the performance of various models on the combined dataset. We group the results by model category. In this section we address **RQ1**, evaluating whether large language models outperform classical deep learning and transformer baselines on the combined Bangla hate speech dataset.

Table 2. Performance of various models on combined dataset. We compute Accuracy (Acc), Macro-F1 (F1), Perplexity (Ppl), and Expected Calibration Error (ECE).

Model	**Acc (%)**	**F1 (%)**	**Ppl**	**ECE (%)**
Classical Models				
MLP (BoW)	81.5	77.0	–	12.8
CNN	87.4	83.7	–	11.5
BiLSTM (FastText)	89.5	86.0	–	10.9
Transformers (fine-tuned)				
mBERT	90.4	87.2	–	8.3
Bangla-BERT	91.1	88.1	–	7.9
XLM-RoBERTa	92.3	90.0	–	7.5
BanglaHateBERT	91.5	89.0	–	7.7
LLMs (Prompting)				
GPT-3.5 (zero-shot)	88.6	85.5	–	–
GPT-4 (zero-shot)	90.2	87.6	–	–
GPT-4 (few-shot)	91.0	88.6	–	–
Claude (zero-shot)	89.7	86.8	–	–
Qwen-7B (zero-shot)	84.3	80.6	–	–
LLMs (Fine-tuned)				
Qwen-7B	91.8	89.4	2.3	6.5
DeepSeek-7B	92.0	89.9	2.1	6.2
DeepSeek-67B	92.5	90.5	1.8	5.9
GPT-3.5 (API fine-tuned)	91.7	89.5	2.5	7.0
Hybrid & Ensemble				
mT5-base	90.0	87.2	3.5	8.0
BERT+GPT-2	90.8	88.4	3.8	8.5
XLM-R + DeepSeek (ensemble)	93.0	91.2	1.9	5.7

On the test data, fine-tuned Transformer models outperform classical CNN/LSTM baselines. XLM-R (base) achieves 92.3% accuracy and 90.0% macro-F1, slightly better than Bangla-BERT and BanglaHateBERT, showing the strength of multilingual pre-training. Among classical models, the BiLSTM with FastText embeddings (prior state-of-art from [17]) reaches 86.0% F1, about 4 points below BERT-based models, highlighting the advantage of Transformer representations. Zero-shot LLMs show strong but slightly lower performance. GPT-4 with few-shot prompts reaches 88.6% macro-F1, comparable to fine-tuned BanglaHateBERT. Claude and GPT-3.5 also perform well (within 1–2% F1 of fine-tuned models). GPT-4 and Claude achieve higher recall than precision, suggesting some over-sensitivity to hate cues. Qwen-7B lags behind in zero-shot (80.6 F1), likely due to smaller size and less optimized prompting.

Fine-tuned LLMs improve substantially. Qwen-7B jumps to 89.4 F1, rivaling BERT models. DeepSeek-67B achieves the best performance: 90.5% macro-F1, slightly ahead of XLM-R. Fine-tuned GPT-3.5 (via OpenAI API) reaches 89.5 F1, close to Qwen and DeepSeek-7B. The ensemble of XLM-R and DeepSeek-67B further boosts performance to 91.2% F1. Fine-tuned LLMs show low perplexity (around 2.0), correlating with high accuracy. GPT-4 and Claude lack measurable perplexity due to API limitations. For calibration, DeepSeek-67B and the ensemble achieve the lowest ECE (5.7%), indicating better confidence alignment. Smaller models like CNN show higher ECE (11%), reflecting overconfidence. Fine-tuning improves calibration, but 5–6% ECE still implies some miscalibration, which can be addressed with temperature scaling if needed in deployment.

5.1 Error Analysis

To address **RQ3**, we analyze calibration (ECE), inference speed, and memory use. While DeepSeek-67B offers top accuracy, smaller models like XLM-R are more deployment-friendly. Error inspection showed that all models struggle with sarcasm and contextual insults. LLMs also better handled creatively spelled slang, which mBERT often missed. However, zero-shot LLMs sometimes overgeneralized, flagging aggressive non-hate comments as hate—a behavior reduced through fine-tuning. Training data size also plays a role: LLMs may benefit from their capacity to absorb the 35k examples. Still, performance gains beyond XLM-R were small—DeepSeek-67B and the ensemble only slightly improved over it, suggesting that 270M parameter models can be sufficient for binary classification. Practically, the 67B model required high GPU memory and was slow (2 samples/sec on H100), while XLM-R processed dozens per second. Thus, if 90% F1 suffices, smaller models are preferable. For higher accuracy or more diverse inputs, LLMs may be justified for offline or ensemble use. Also, LLMs showed better calibration, making their confidence scores more reliable for threshold-based flagging.

6 Ethical Considerations

This study adheres to ethical standards for computational research involving online data. All datasets used in this work—HS-BAN, BD-SHS, Bengali Hate v2.0, and BanTH—are publicly available and were created for research purposes with appropriate anonymization. We do not store or process any personally identifiable information (PII), nor do we collect user-generated content directly from individuals or platforms. Our experiments were conducted on preprocessed text data that lacks any direct identifiers.

The goal of this work is to advance the responsible use of machine learning and large language models (LLMs) for mitigating harm caused by hate speech. However, we acknowledge the dual-use nature of LLMs and the risks of misuse if such models are adapted for adversarial content generation. Therefore, we encourage future researchers and practitioners to consider both the benefits and potential consequences when deploying similar systems in real-world settings.

7 Discussion and Limitations

Our results demonstrate that fine-tuned LLMs such as DeepSeek-67B and GPT-4 achieve high accuracy and strong robustness in Bengali hate speech detection, outperforming traditional and transformer-based baselines. The ensemble approach also shows promise for further boosting performance through complementary strengths of different models.

Nonetheless, several limitations remain. First, the computational cost of fine-tuning and deploying large LLMs is substantial. While models like DeepSeek-67B yielded the highest macro-F1 scores, their inference speed and memory requirements make them impractical for real-time deployment on resource-constrained devices. Second, while our combined dataset enhances generalizability, domain-specific language—such as regional dialects, sarcasm, and creative transliteration—still poses challenges, especially for zero-shot LLMs.

Additionally, we evaluated only binary classification due to dataset constraints. Future work should explore fine-grained classification of hate types and include multilingual transfer across related South Asian languages. We also note that while LLMs improve calibration, a small but non-negligible Expected Calibration Error (ECE) remains, which could impact threshold-based moderation systems. Distillation into smaller, faster, yet effective Bangla-specific models is a promising direction for achieving practical deployment while retaining LLM-level performance.

8 Conclusion

We presented a comprehensive study on Bangla hate speech detection, evaluating models from classical baselines to state-of-the-art LLMs. Using all available datasets, we established new performance benchmarks, with fine-tuned LLMs like GPT-4 and DeepSeek exceeding 90% macro-F1. While these models outperformed traditional ones, the margin over fine-tuned BERT models was modest. LLMs also offered better calibration and handled nuances like sarcasm and transliteration more effectively. Despite their strength, high computational cost limits real-time use. Future work includes distilling LLMs into smaller Bangla models, improving interpretability, and extending to other low-resource languages and fine-grained hate classification.

9 Acknowledgments

This work was supported in part by the U.S. National Science Foundation (Award No. 2451946) and the U.S. Nuclear Regulatory Commission (Award No. 31310025M0012). ChatGPT was utilized to assist with language editing and clarity improvements in this work. No content was generated related to technical results, data, code, or analysis.

References

1. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
2. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al.: Constitutional ai: Harmlessness from ai feedback. arXiv preprint arXiv:2212.08073 (2022)
3. Bhattacharjee, A., Hasan, T., Ahmad, W.U., Samin, K., Islam, M.S., Iqbal, A., Rahman, M.S., Shahriyar, R.: Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. arXiv preprint arXiv:2101.00204 (2021)
4. Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z., et al.: Deepseek llm: Scaling open-source language models with longtermism. arXiv preprint arXiv:2401.02954 (2024)
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
6. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116 (2019)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
8. Ghosh, K., Senapati, A.: Hate speech detection in low-resourced indian languages: An analysis of transformer-based monolingual and multilingual models with cross-lingual experiments. Natural Language Processing **31**(2), 393–414 (2025)
9. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)

10. Haider, F., Shifat, F.T., Ishmam, M.F., Barua, D.D., Sourove, M.S.U.R., Fahim, M., Alam, M.F.: Banth: A multi-label hate speech detection dataset for transliterated bangla. arXiv preprint arXiv:2410.13281 (2024)
11. Hussain, M.G., Al Mahmud, T., Akthar, W.: An approach to detect abusive bangla text. In: 2018 International Conference on Innovation in Engineering and Technology (ICIET). pp. 1–5. IEEE (2018)
12. Jahan, M., Ahamed, I., Bishwas, M.R., Shatabda, S.: Abusive comments detection in bangla-english code-mixed and transliterated text. In: 2019 2nd international conference on innovation in engineering and technology (ICIET). pp. 1–6. IEEE (2019)
13. Jahan, M.S., Haque, M., Arhab, N., Oussalah, M.: 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. pp. 8–15 (2022)
14. Karim, M.R., Chakravarthi, B.R., McCrae, J.P., Cochez, M.: Classification benchmarks for under-resourced bengali language based on multichannel convolutional-lstm network. In: 2020 IEEE 7th international conference on Data Science and Advanced Analytics (DSAA). pp. 390–399. IEEE (2020)
15. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
16. Mathew, B., Saha, P., Yimam, S.M., Biemann, C., Goyal, P., Mukherjee, A.: Hatexplain: A benchmark dataset for explainable hate speech detection. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 14867–14875 (2021)
17. Romim, N., Ahmed, M., Islam, M.S., Sharma, A.S., Talukder, H., Amin, M.R.: Hs-ban: A benchmark dataset of social media comments for hate speech detection in bangla. arXiv preprint arXiv:2112.01902 (2021)
18. Romim, N., Ahmed, M., Islam, M.S., Sharma, A.S., Talukder, H., Amin, M.R.: Bd-shs: A benchmark dataset for learning to detect online bangla hate speech in different social contexts. arXiv preprint arXiv:2206.00372 (2022)
19. Romim, N., Ahmed, M., Talukder, H., Saiful Islam, M.: Hate speech detection in the bengali language: A dataset and its baseline evaluation. In: Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020. pp. 457–468. Springer (2021)
20. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mt5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934 (2020)