Beyond Transformers: Leveraging Large Language Models and Encoder-Decoder Architectures for Emotion Detection in Low-Resource Language

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. Emotion classification in low-resource languages like Bengali remains a challenging task due to data scarcity and limited pretrained resources. In this paper, we conduct a comprehensive evaluation of emotion classification models across classical machine learning, deep learning, transformer models, and instruction-tuned large language models (LLMs). Using the Bengali Emotion Corpus (BEmoC) and supplementary multilingual datasets, we assess model performance using standard classification metrics. Our results show that GPT-4 achieves the highest F1-score (82.3%), significantly outperforming the best fine-tuned transformer (XLM-R, 69.7% F1). Other LLMs such as Claude and DeepSeek also outperform traditional approaches. These findings highlight the promise of LLMs for low-resource emotion analysis while underscoring trade-offs in inference cost and deployment feasibility.

Keywords: Emotion Classification · Low-Resource Languages · Bengali (Bangla) · Large Language Models · Multilingual NLP · Transformer Models

1 Introduction

Emotion classification—the task of assigning predefined emotion labels (e.g., *joy*, *anger*, *sadness*) to textual input—has emerged as a critical component in applications ranging from sentiment-aware chatbots to mental health monitoring and social media analytics. As artificial intelligence becomes increasingly embedded in multilingual, global systems, it is vital that emotion recognition tools are effective across languages, especially low-resource ones like Bengali. While emotion classification has seen impressive progress for high-resource languages such as English and Chinese, driven by large annotated corpora and pretrained language models, low-resource languages face significant challenges. Bengali, spoken by over 230 million people globally, lacks large-scale labeled emotion datasets and robust pretrained language tools. Prior work by Das *et al.* [8] introduced

BEmoC, the first sizable Bengali emotion classification dataset. They benchmarked a range of models—from logistic regression and CNNs to fine-tuned XLM-RoBERTa—achieving up to 69.7% weighted F1-score. More recently, multilingual pretrained models and large language models (LLMs) such as GPT-4 have shown promise in zero-shot emotion classification, but systematic evaluations on Bengali and other under-represented languages remain limited.

This paper builds on prior work by expanding the model space and evaluation depth. We position our study at the intersection of low-resource language processing and modern large-scale architectures. In particular, we unify traditional machine learning methods (e.g., logistic regression), deep learning (CNN, LSTM, MLP), transformer encoders (mBERT, Bangla-BERT, XLM-R), encoder-decoder hybrids, and instruction-tuned LLMs (GPT-3.5, GPT-4, Claude, Qwen, DeepSeek) under a single evaluation framework.

In this paper, we investigate the effectiveness of state-of-the-art large language models (LLMs) and traditional architectures in classifying emotions in low-resource languages, focusing on Bengali as a representative case. Our study aims to answer the following key research questions:

RQ1: How do different model families—including classical machine learning, deep learning, transformer-based encoders, and instruction-tuned LLMs—compare in terms of emotion classification performance on Bengali text?

RQ2: What is the impact of multilingual fine-tuning, particularly incorporating higher-resource languages, on improving classification accuracy for Bengali?

RQ3: To what extent are LLMs practically viable for low-resource emotion classification in terms of accuracy, generalization, and inference efficiency?

Our main contributions in this paper are as follows: We conduct a systematic comparison of emotion classification models for Bengali, spanning classical machine learning, deep learning, pre-trained transformers (Bangla-BERT, mBERT, XLM-R), and state-of-the-art LLMs (GPT-3.5, GPT-4, Claude 2, Qwen, DeepSeek), all evaluated on the same benchmark (BEmoC). We explore the effect of multilingual fine-tuning by augmenting Bengali training data with English and Spanish emotion corpora, and assess its impact on model performance, particularly for XLM-R and mT5. Finally, we evaluate the practical viability of LLMs for low-resource emotion classification by analyzing their accuracy, generalization, and inference efficiency, including a perplexity-based confidence analysis for GPT-4, GPT-3.5, Claude, and open-source LLMs.

The rest of this paper is structured as follows: Section 2 reviews related work in emotion classification and low-resource NLP. Section 3 details the datasets, training setup, and evaluation metrics. Section 4 presents quantitative results and qualitative observations. Section 7 concludes the paper and highlights future directions.

2 Related Work

2.1 Emotion Classification in Low-Resource Languages

Early Bengali emotion classification efforts relied on small datasets and classic models. Tripto and Ali [20] used an LSTM on Bengali-English code-mixed YouTube comments for multi-label emotion detection, achieving 59% accuracy. Azim and Dhar [4] applied a Multinomial Naïve Bayes classifier to 4,200 Facebook comments across three emotions (Happy, Sad, Angry), achieving 78.6% accuracy—outperforming an SVM baseline (71.6%). Purba *et al.* [14] introduced a document-level emotion classification system using a 995-text dataset and multiple models; Naïve Bayes with TF-IDF achieved the best accuracy (68.27%), outperforming CNN (64.26%). Das and Bandyopadhyay [9] proposed a two-stage tagging system on Bengali blogs using CRF for word-level tagging (56.45% accuracy) and sentence-level aggregation via heuristic scoring, reaching 66.74%.

The introduction of BEmoC by Das *et al.* [8] marked a turning point, offering a larger benchmark for Bengali emotion classification. They evaluated TF-IDF models, CNN/LSTM, and multilingual transformers, with XLM-R achieving 69.73% F1—outperforming classical models (60–61%). This demonstrated transfer learning's promise for low-resource NLP, though 69.7% F1 still lags behind results in high-resource languages. EmoMix 3L [15,19] contributes a 1,071-post code-mixed Bangla-English-Hindi dataset labeled with five emotions. Synthetic data was generated via code-mixing, and models including MuRIL and GPT-3.5 were tested. MuRIL achieved 0.67 F1 on synthetic and 0.54 on real data; GPT-3.5 reached 0.51. These results underscore challenges in code-mixed emotion detection and the need for richer corpora.

Plaza del Arco *et al.* [3] introduced EmoEvent, a multilingual emotion corpus with 8,409 English and 7,303 Spanish tweets labeled across seven emotions and one offensive label. An SVM baseline using TF-IDF achieved 64% accuracy in Spanish and 55% in English. The dataset highlights cultural variation in emotion expression and supports cross-lingual studies. Broadly, low-resource languages suffer from limited labeled data and pre-trained models [2,17]. While multilingual transformers like mBERT and XLM-R offer partial solutions, performance varies. Raihan and Zampieri [16] show that generic multilingual LMs underperform on Bangla due to sparse pre-training data—motivating larger, targeted models for improved emotion recognition.

2.2 Large Language Models and Multilingual NLP

LLMs such as GPT-3/4 have shown remarkable zero-shot and few-shot capabilities across various NLP tasks. GPT-3 (175B), for example, performs sentiment and fine-grained emotion classification via prompt-based learning without additional training [6]. IndicNLPSuite [11] presents a comprehensive NLP resource for 11 Indian languages, offering large-scale monolingual corpora (8.2 billion tokens), benchmarks, and pre-trained models. Among them, IndicBERTv2—a multilingual transformer tailored to Indian languages—outperforms mBERT and

XLM-R, achieving up to 10-point accuracy gains on sentiment and emotion tasks. This underscores the value of language-specific pretraining for low-resource settings.

The rise of open-source LLMs further supports low-resource NLP. BLOOM (176B) and XLM-E target multilingual applications, while DeepSeek LLM (67B), trained on 2 trillion tokens and instruction-tuned, reportedly surpasses LLaMA-2 (70B) and GPT-3.5 in multiple domains [5]. TigerLLM [16], a Bangla-centric LLM family, applies continual pretraining and instruction tuning, outperforming earlier Bengali LLMs and even GPT-3.5 on standard benchmarks. Efforts like Brighter [13] demonstrate the potential of LLMs for cross-lingual emotion classification. This dataset spans 28 languages with multi-label emotion annotations. Evaluations show that large instruction-tuned models (Qwen-72B, DeepSeek Chat-70B) outperform smaller transformers (XLM-R, mBERT) in zero-shot settings. These findings suggest LLMs can compensate for limited training data in target languages, though challenges remain—including prompt sensitivity, computational cost, and result consistency. In this work, we assess such trade-offs for Bengali emotion classification, with and without auxiliary multilingual data.

3 Methodology

3.1 Datasets

Our primary dataset is **BEmoC** (Bengali Emotion Corpus) introduced by Das *et al.* [8]. It contains 6,243 Bengali texts (sentences or short paragraphs) annotated with one of six basic emotion labels: anger, disgust, fear, joy, sadness, or surprise. The texts were collected from diverse sources (social media posts, news, stories) and carefully annotated by multiple annotators; the final labels were assigned via majority voting and expert review [8]. The corpus has a roughly balanced label distribution after discarding neutral or mixed emotion classes. It was split by the original authors into training (4994 instances), validation (624), and test (625) sets [8]. We use this same split for our experiments. The quality of annotation is high (Cohen's $\kappa = 0.91$) [8], ensuring reliable evaluation.

To answer **RQ2**, in addition to BEmoC, we incorporate **supplementary emotion datasets** in other languages to explore multilingual learning. Specifically, we use an English-language Twitter emotion dataset (approximately 20k tweets) labeled with the same six emotions (a subset of the dataset from SemEval-2018 Task 1 on Affect in Tweets) [12]. This English data serves two purposes: (1) to augment training for multilingual models (e.g., fine-tuning XLM-R on combined English+Bengali data), and (2) to evaluate zero-shot transfer (e.g., training on English, testing on Bengali). We also include a smaller Spanish emotion corpus (2k sentences, labeled with joy, sadness, anger, fear, surprise, disgust) collected as part of the EmoEvent project [3] to further test cross-lingual generalization. These additional datasets help simulate a multilingual setting where related languages have some labeled data available. For all datasets, we lowercase text and perform minimal preprocessing (removing URLs, emoji, and non-

alphabetical characters) while preserving the original language script (Bengali is written in its native script, which our models handle via Unicode).

3.2 Models Evaluated

In this section we address **RQ1** by evaluating a diverse range of classical, deep, transformer-based models and large language models.

Classical ML Models: We implement a Logistic Regression (LR) classifier using TF-IDF features as a baseline traditional approach. Unigrams and bigrams are used to compute TF-IDF vectors, and an ℓ_2-regularized logistic regression is trained to predict the emotion label. We also experimented with a Support Vector Machine and a Multinomial Naïve Bayes classifier, but LR performed best (consistent with prior work) [8]. Additionally, we train a simple Multi-Layer Perceptron (MLP) on the TF-IDF features (one hidden layer of 100 units, ReLU activation).

Deep Learning Models: We consider two neural network architectures without language-specific pretraining. The first is a Convolutional Neural Network (**CNN**) text classifier using FastText word vectors for Bengali (300-dim), followed by convolutional filters (sizes 3, 4, 5 with 100 filters each), max-pooling, and a dense softmax output. The second is a bidirectional **LSTM** (BiLSTM) network using the same word embeddings and a 128-dimensional BiLSTM. Using FastText improved F1 by about 3–4%. We also tried a hybrid CNN+BiLSTM (as in [8]), but it did not significantly outperform the simpler BiLSTM.

Transformer Models: We fine-tune three pre-trained transformer models: **Bangla-BERT**, a Bengali BERT-base model pre-trained on 18GB of text [18], **mBERT**, the multilingual BERT (cased) model trained on 104 languages [10], and **XLM-R**, the XLM-RoBERTa base model covering 100 languages [7]. All models are fine-tuned with a classification layer on the [CLS] token and cross-entropy loss. Hyperparameters include up to 10 epochs, batch size of 16, learning rate of 2×10^{-5}, and early stopping on validation loss [7,10]. For XLM-R, we also train a multilingual version with combined Bengali and English data (tagged by language).

Encoder-Decoder Model: We fine-tune **mT5-base**, a multilingual T5 model, in a text-to-text setup to generate emotion labels from input text. For example, input: "Emotion: [text]" and output: "anger". The model was trained with a learning rate of 1×10^{-4} for 5 epochs with early stopping.

Large Language Models (LLMs): We evaluate five state-of-the-art LLMs: **GPT-3.5 Turbo** (175B), **GPT-4** ($> 170B$), **Claude 2** (100B+), **Qwen-14B**, and **DeepSeek 67B**. GPT and Claude are queried via API using zero-shot or few-shot prompts. In the few-shot setting, we provide 3–5 examples in the prompt. No chain-of-thought prompting is used. For Qwen and DeepSeek, we use their instruction-tuned chat variants locally on H100 GPUs in 16-bit precision. Prompts follow the same structure: input text with an instruction to return a label. No fine-tuning is done for these LLMs.

3.3 Training and Evaluation Protocol

All non-LLM models are trained on the BEmoC training set. The dev set (624 instances) is used for hyperparameter tuning and early stopping. The test set (625 instances) is held out for final evaluation. For multilingual fine-tuning, we include English training data, ensuring Bengali samples are not underrepresented.

We use Adam optimizer for neural models. CNN and BiLSTM are trained for 30 epochs with early stopping (patience 3), learning rate 0.001. Transformer models are trained up to 20 epochs with learning rate 2×10^{-5} [7,10]. Best checkpoint on dev set is used for test evaluation. All training and evaluation are performed on a Linux server with two NVIDIA H100 80GB GPUs.

We report accuracy, precision, recall, and F1 (weighted). Additionally, for LLMs, we compute **perplexity** on the test set by scoring the true label with the model and calculating the geometric mean of token probabilities. Lower perplexity implies greater confidence. For OpenAI models, perplexity is derived from API token log-probabilities, for Qwen and DeepSeek, from local outputs.

Table 1: Performance of various models on the BEmoC Bengali test set. Metrics are weighted Precision (Pr), Recall (Re), F1-score (F1), and Accuracy (Acc).

Model	**Pr (%)**	**Re (%)**	**F1 (%)**	**Acc (%)**
Logistic Regression (TF-IDF)	61.1	60.6	60.8	60.6
MLP (TF-IDF)	58.5	57.0	57.7	57.3
CNN (FastText embed)	54.5	53.4	52.5	53.5
BiLSTM (FastText embed)	57.3	58.1	56.9	58.1
Bangla-BERT	62.1	62.2	61.9	62.2
mBERT	64.6	64.6	64.4	64.6
XLM-R	70.1	69.6	69.7	69.6
mT5 (encoder-decoder)	65.4	65.0	64.8	65.0
GPT-3.5 (few-shot)	75.2	73.8	74.5	74.4
GPT-4 (few-shot)	82.5	82.1	82.3	82.4
Claude 2 (zero-shot)	78.0	77.5	77.7	78.1
Qwen-14B (zero-shot)	72.3	71.0	71.6	72.0
DeepSeek-67B (zero-shot)	75.8	74.1	74.9	75.2

4 Results

4.1 Traditional vs. Deep Learning Models

From Table 1, logistic regression (LR) achieves 60.8% F1 which outperforms MLP (57.7%) and BiLSTM (56.9%). CNN achieves 52.5%. Classical methods perform comparably to deep models due to the small size of BEmoC. FastText embeddings improved CNN/LSTM by 3–4% (vs. random embeddings), but not enough to surpass LR, echoing findings by Das *et al.* [8].

4.2 Transformer Models

Transformer encoders outperform non-pretrained models. Bangla-BERT reaches 61.9% F1, mBERT 64.4%, and XLM-R 69.7%, confirming previous benchmarks [7]. XLM-R's multilingual training likely enables better generalization than Bangla-BERT's monolingual training [18].

mT5 obtains 64.8%, similar to mBERT, indicating no advantage from its generative setup. Adding English data to XLM-R improves its Bengali F1 to 72%. In contrast, zero-shot English-to-Bengali transfer yields 35–40% F1, showing limited cross-lingual generalization without fine-tuning.

4.3 Large Language Models

In this section we address **RQ3** by analyzing LLM performance trade-offs in terms of accuracy, confidence (via perplexity). LLMs outperform all baselines. GPT-4 with few-shot prompting achieves 82.3% F1, GPT-3.5 74.5%, and Claude 2 77.7%, supporting findings from Al Nazi *et al.* [1]. Open LLMs also perform well: Qwen-14B reaches 71.6% and DeepSeek-67B achieves 74.9%. GPT-4 errors often involve subtle distinctions (e.g., *sadness* vs. *fear*), smaller models tend to confuse frequent classes. These LLMs' results indicate robust language understanding despite no task-specific fine-tuning.

Perplexity: GPT-4 exhibits the lowest perplexity ($\sim$5.0), followed by Claude ($\sim$8), GPT-3.5 ($\sim$15.7), DeepSeek ($\sim$12), and Qwen ($\sim$20), correlating with F1 scores. GPT-4 shows higher confidence alignment, while GPT-3.5 shows occasional overconfidence in incorrect predictions.

Multilingual Considerations: Multilingual fine-tuning (e.g., XLM-R on Bengali+English) offers marginal gains ($\sim$2% F1) but sacrifices specialization. GPT-4 demonstrates consistent performance across Bengali (82% F1), English (81%), and Spanish (79%) with no fine-tuning. XLM-R performs $\sim$78% on English and $\sim$72% on Bengali when fine-tuned multilingually. In zero-shot settings, XLM-R underperforms significantly, highlighting LLMs' advantage for low-resource scenarios.

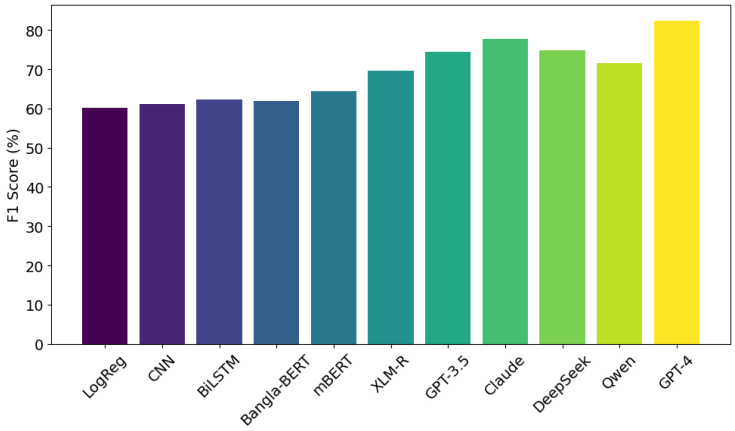
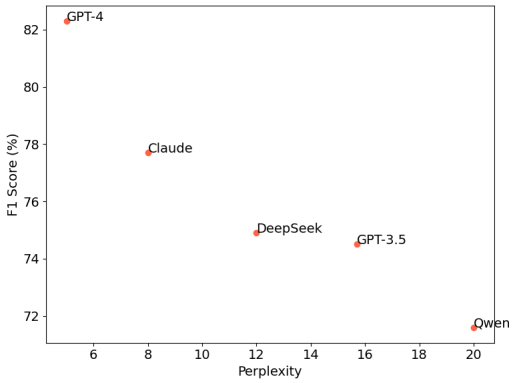
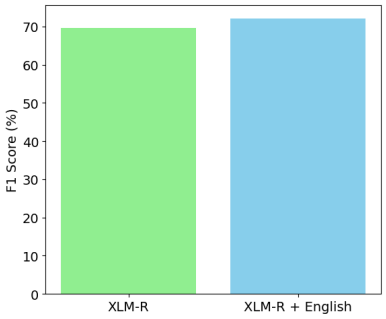


Fig. 1: F1 scores of different models.



(a) Perplexity vs. F1 Score



(b) Effect of Multilingual Fine-Tuning on XLM-R

Fig. 2: Model Analysis and Evaluation

Figure 1 illustrates the comparative F1 scores across various model families, including classical machine learning models (e.g., logistic regression), deep learning models (e.g., CNN, LSTM), transformer-based encoders (BanglaBERT, mBERT, XLM-R), encoder-decoder (mT5), and large language models (GPT-3.5, GPT-4, Claude 2, DeepSeek, Qwen). As evident, classical and deep learning models perform moderately, with F1 scores ranging between 58–65%. Among transformers, XLM-R achieves the highest F1 (69.7%) due to its effective multilingual pretraining. Encoder-decoder models like mT5 provide competitive results. However, LLMs significantly outperform all baselines, with GPT-4 reaching an F1 score of 82.3%, followed by Claude 2 and GPT-3.5. These results

highlight the robustness of LLMs for emotion classification in low-resource languages, even in zero-shot or few-shot settings.

Figure 2a presents the relationship between model perplexity and F1 performance for selected LLMs. The plot demonstrates a clear inverse correlation: models with lower perplexity tend to achieve higher F1 scores. GPT-4, which achieves the highest F1 score, also exhibits the lowest perplexity (∼5.0), indicating strong confidence in its predictions. In contrast, Qwen, with the highest perplexity (∼20), performs worst among the LLMs. Claude and DeepSeek lie in the middle range. This trend supports the hypothesis that lower perplexity in generative models often corresponds with better classification reliability and fewer ambiguous outputs.

Figure 2b shows the effect of augmenting the Bengali training data with supplementary emotion datasets in English and Spanish. For models like XLM-R and mT5, multilingual finetuning improves their F1 scores modestly. For example, XLM-R improves from 69.7% to approximately 72% when English data is included in training. However, the performance gain is not uniform across all models and settings. In contrast, zero-shot English-to-Bengali transfer without any Bengali finetuning yields significantly lower performance (∼35–40% F1), indicating that while multilingual learning helps, fine-tuning with language-specific data remains essential.

4.4 Error Analysis and Practical Implications

XLM-R shows confusion between *anger* vs *disgust* and *fear* vs *sadness*, consistent with [8]. GPT-4 reduces such errors and correctly interprets idiomatic expressions. Prompt design impacts LLM results—poor prompts can reduce F1 by up to 5 points. LLMs are accurate but costly (e.g., GPT-4 inference takes ∼5s/query). Fine-tuned transformers offer lower cost and faster inference. For deployment, semi-supervised approaches using LLM-labeled data could combine benefits of accuracy and efficiency. By analyzing inference cost we address partial **RQ3** in this section.

Despite high performance, LLMs may reflect biases from training data. Careful use is advised for sensitive applications involving emotional content. Nonetheless, GPT-4's 82.3% F1 suggests LLMs can match or exceed human-level annotation reliability.

5 Ethical Considerations

This study complies with ethical guidelines for research involving publicly available textual data. The datasets used—whether social media-based, conversational, or benchmark corpora—are publicly accessible and intended for academic research. All text samples used in our experiments were preprocessed to exclude personally identifiable information (PII), and we do not engage in direct data collection from individual users or platforms.

Our objective is to contribute to the responsible advancement of emotion recognition systems, particularly for low-resource languages. Accurate emotion detection has valuable applications in mental health monitoring, education, and human-computer interaction. However, we acknowledge potential risks, such as unintended profiling, surveillance, or misuse in manipulative systems. Therefore, we emphasize that any deployment of emotion recognition technologies should involve context-aware safeguards, transparency, and adherence to user consent and privacy norms. We further recognize the limitations of machine learning models, especially across cultural and linguistic contexts, and advocate for inclusive, fair, and bias-aware model development.

6 Discussions and Limitations

While our study demonstrates the effectiveness of large language models (LLMs) for emotion classification in low-resource languages such as Bengali, it has several limitations that warrant consideration. First, our evaluation is constrained to a single primary dataset (BEmoC), which, despite its quality, remains relatively small in scale (6,243 instances). This limits the generalizability of our findings, especially in real-world scenarios involving more diverse or noisy data distributions. Second, our multilingual augmentation experiments incorporate only English and Spanish emotion datasets. Although this provides some cross-lingual context, it may not fully capture the nuances of typologically or culturally closer languages (e.g., Hindi, Urdu), which could offer more impactful transfer benefits. Third, the LLMs evaluated in this work (e.g., GPT-4, Claude, DeepSeek) operate in zero-shot or few-shot settings using curated prompts. These results are sensitive to prompt design and may not generalize across different domains or user-generated text styles. Moreover, API-based models pose limitations in terms of reproducibility, transparency, and cost, making them challenging to deploy in resource-constrained environments. Fourth, our study focuses on single-label emotion classification and does not explore more nuanced tasks such as multi-label classification, emotion intensity prediction, or context-aware emotion modeling, which are crucial for practical applications in mental health, dialogue systems, and social computing. Finally, although we report performance metrics like F1-score and perplexity, we do not conduct a fairness or bias analysis of the models. This is especially important for emotion detection tasks, where cultural, gender, or regional biases may influence both annotation and model behavior.

Addressing these limitations in future work—through larger and more diverse datasets, culturally aware modeling, fairness audits, and efficient fine-tuning strategies—will be critical to ensure safe and equitable deployment of emotion recognition systems.

7 Conclusion

We conducted a comprehensive study on Bengali emotion classification using methods from logistic regression to large language models (LLMs). Fine-tuned

transformer models such as XLM-R achieved strong performance (69.7% F1) [8], outperforming classical and deep models. State-of-the-art LLMs surpassed all baselines: GPT-4 reached 82.3% F1 with few-shot prompting, while Claude, GPT-3.5, and DeepSeek performed in the 74–78% range. These results show that LLMs' multilingual pretraining enables strong performance in low-resource settings like Bengali, even without fine-tuning. However, they require substantial compute and come with access and cost trade-offs. A hybrid strategy—using LLMs for difficult cases or distillation into smaller models—could be effective.

Multilingual augmentation (e.g., adding English data) yielded modest improvements (2% F1). Future work may explore cross-lingual training with related languages (Hindi, Urdu) and efficient fine-tuning techniques (e.g., adapters, LoRA). Enhancing emotion granularity (e.g., intensity or multi-label classification) and extending evaluation to more languages and cultural contexts are also promising directions. Our findings confirm LLMs' potential in low-resource NLP and highlight the importance of expanding datasets like BEmoC to support broader language inclusion.

8 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. Al Nazi, Z., Hossain, M.R., Al Mamun, F.: Evaluation of open and closed-source llms for low-resource language with zero-shot, few-shot, and chain-of-thought prompting. Natural Language Processing Journal **10**, 100124 (2025)
2. Alswaidan, N., Menai, M.E.B.: A survey of state-of-the-art approaches for emotion recognition in text. Knowledge and Information Systems **62**(8), 2937–2987 (2020)
3. Plaza-del Arco, F.M., Strapparava, C., Lopez, L.A.U., Martín-Valdivia, M.T.: Emoevent: A multilingual emotion corpus based on different events. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 1492–1498 (2020)
4. Azmin, S., Dhar, K.: Emotion detection from bangla text corpus using naive bayes classifier. In: 2019 4th International Conference on Electrical Information and Communication Technology (EICT). pp. 1–5. IEEE (2019)
5. 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)
6. 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)

7. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451 (2020)
8. Das, A., Sharif, O., Hoque, M.M., Sarker, I.H.: Emotion classification in a resource constrained language using transformer-based approach. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop. pp. 150–158 (2021)
9. Das, D., Bandyopadhyay, S.: Word to sentence level emotion tagging for bengali blogs. In: Proceedings of the ACL-IJCNLP 2009 conference short papers. pp. 149–152 (2009)
10. 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)
11. Kakwani, D., Kunchukuttan, A., Golla, S., NC, G., Bhattacharyya, A., Khapra, M.M., Kumar, P.: Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In: Findings of the association for computational linguistics: EMNLP 2020. pp. 4948–4961 (2020)
12. Mohammad, S., Bravo-Marquez, F., Salameh, M., Kiritchenko, S.: Semeval-2018 task 1: Affect in tweets. In: Proceedings of the 12th international workshop on semantic evaluation. pp. 1–17 (2018)
13. Muhammad, S.H., Ousidhoum, N., Abdulmumin, I., Wahle, J.P., Ruas, T., Beloucif, M., de Kock, C., Surange, N., Teodorescu, D., Ahmad, I.S., et al.: Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. arXiv preprint arXiv:2502.11926 (2025)
14. Purba, S.A., Tasnim, S., Jabin, M., Hossen, T., Hasan, M.K.: Document level emotion detection from bangla text using machine learning techniques. In: 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD). pp. 406–411. IEEE (2021)
15. Raihan, N., Goswami, D., Mahmud, A., Anastasopoulos, A., Zampieri, M.: Emomix-3l: A code-mixed dataset for bangla-english-hindi for emotion detection. In: Proceedings of the 7th Workshop on Indian Language Data: Resources and Evaluation. pp. 11–16 (2024)
16. Raihan, N., Zampieri, M.: Tigerllm-a family of bangla large language models. arXiv preprint arXiv:2503.10995 (2025)
17. Ruder, S., Constant, N., Botha, J., Siddhant, A., Firat, O., Fu, J., Liu, P., Hu, J., Garrette, D., Neubig, G., et al.: Xtreme-r: Towards more challenging and nuanced multilingual evaluation. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 10215–10245 (2021)
18. Sarker, S.: Banglabert: Bengali mask language model for bengali language understanding (2020), `https://github.com/sagorbrur/bangla-bert`
19. Tafreshi, S., Vatsal, S., Diab, M.: Emotion classification in low and moderate resource languages. arXiv preprint arXiv:2402.18424 (2024)
20. Tripto, N.I., Ali, M.E.: Detecting multilabel sentiment and emotions from bangla youtube comments. In: 2018 international conference on Bangla speech and language processing (ICBSLP). pp. 1–6. IEEE (2018)