Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-025-00033-3

Original Article

Sentiment Classification in Low-Resource, Code-Mixed Languages: Comparative Study of Llama 3 and Other Large Language Models

Md Samiur Rahman [1], Rashedur Mohammad Rahman [2]

1. *Computer Science and Engineering, North South University, Dhaka, BGD*
2. *Electrical and Computer Engineering, North South University, Dhaka, BGD*

Received: June 29, 2025 | Review began: August 12, 2025 | Review ended: September 21, 2025 | Published: February 24, 2026

Abstract

Large language models have seen limited evaluation on code-switched languages, despite the increasing use of such linguistic patterns on social media platforms. This study presents a systematic evaluation of Llama 3 for sentiment analysis of "Banglish", the informal blend of Bengali and English widely used online in Bangladesh. We examine five adaptation strategies: zero-shot prompting, few-shot learning, a two-step translation and analysis pipeline, fine-tuning, and model ensembling. Our experiments, conducted on 11,673 posts from the Bengali_Banglish_80K dataset, show that fine-tuning Llama 3 delivers the best performance, achieving an accuracy of 66.87 percent. This result outperforms its own zero-shot (43.80 percent) and few-shot (49.14 percent) baselines, as well as the zero-shot results of GPT-3.5 (55.90 percent), GPT-4 (65.15 percent), and Claude 3.5 (47.68 percent). The dual-phase pipeline achieved 48.07 percent accuracy, and the ensemble method reached 66.78 percent, offering slight improvements but falling short of the fine-tuned model. We release our best-performing model, samiur-r/BanglishSentiment-Llama3-8B, to support further research. These findings highlight the effectiveness of task-specific fine-tuning in low-resource, code-mixed settings and emphasize the need for more comprehensive code-switching datasets and pre-training strategies tailored to linguistically diverse communities.

Categories: Ensemble Learning, Natural Language Processing (NLP), AI/ML-based decision support systems

Keywords: sentiment analysis, banglish, llama 3, nlp, large language models (llms), machine translation, text dataset, data annotation, fine tuning, ensemble methods

Introduction

The rapid advancement of Natural Language Processing (NLP) has been driven by large language models (LLMs) such as GPT-3, GPT-4, and Llama 3, which demonstrate exceptional capability in understanding and generating human-like text across diverse languages and domains. Sentiment analysis remains a key NLP application, supporting tasks such as social media monitoring, customer feedback assessment, and public opinion analysis. However, sentiment classification becomes more complex in code-mixed languages such as Banglish, a fusion of Bengali and English widely used in Bangladesh and its diaspora. This hybrid language is increasingly prevalent in online platforms, instant messaging, and digital forums, reflecting authentic cultural expression and nuanced sentiment. Despite its importance, Banglish remains underrepresented in NLP corpora, with limited annotated datasets and few dedicated analytical frameworks.

Banglish presents multiple challenges for sentiment analysis, including lexical variability from non-standardized spelling, syntactic irregularities that deviate from both Bengali and English norms, and semantic ambiguities that require rich contextual understanding. Most existing sentiment analysis models, including state-of-the-art LLMs, are trained on

monolingual, resource-rich languages, limiting their adaptability to code-switched contexts. Transfer learning has improved performance for underrepresented languages, but its effectiveness diminishes when linguistic structures differ greatly from the training data. Real-time applications, such as monitoring dynamic social media streams, are further hindered by a lack of optimized, mixed-language processing pipelines.

The evolution of sentiment analysis techniques from lexicon-based and rule-based approaches to machine learning classifiers like Naïve Bayes and Support Vector Machine, and later deep learning architectures such as recurrent neural networks, long short-term memory, and transformers, has greatly improved contextual understanding and adaptability. Multimodal sentiment analysis, as explored by Das and Singh [1], integrates textual, audio, and visual data to overcome the limitations of purely text-based methods, while Hartmann et al. [2] highlight the role of sophisticated algorithms in optimizing precision across domains such as social media and customer review analysis. The advent of LLMs marked a transformative shift. Foundational models like Bidirectional Encoder Representations from Transformers and Embeddings from Language Models introduced contextual embeddings, and transformer-based architectures such as Generative Pre-trained Transformer (GPT) and its successors expanded capabilities through large-scale pretraining and task-specific fine-tuning [3]. These advances enhance contextual comprehension, improve flexibility in processing informal and dialectal language, and enable zero-shot and few-shot learning, which is particularly valuable for low-resource languages. Zhang et al. [4] demonstrate that LLMs outperform traditional NLP methods across diverse datasets, while Rusnachenko et al. [5] show that fine-tuning on domain-specific data improves targeted sentiment detection.

The growing importance of multilingual sentiment analysis stems from the increasing prevalence of code-switching in global communication. Challenges include linguistic diversity, resource disparity between languages, and the syntactic complexity introduced by mid-sentence language switching. Strategies to address these challenges include cross-lingual models that leverage shared representations and transfer learning from high-resource to low-resource languages. Buscemi and Proverbio [6] compared ChatGPT, Gemini, and LLaMA in multilingual contexts, revealing how architecture and training data diversity influence adaptability. Chowdhury et al. [7] evaluated LLMs for Bengali social media sentiment, including depressive content detection, demonstrating the benefit of fine-tuning for underrepresented languages. Veeramani et al. [8] proposed hybrid frameworks integrating transformers with LLMs, which improved performance on code-switched datasets.

Comparative studies provide critical benchmarks for model selection and advancement. Lossio-Ventura et al. [9] compared ChatGPT and fine-tuned open pre-trained transformers against conventional sentiment tools on COVID-19 survey data, finding that LLMs excel in capturing nuanced sentiment. Similarly, work by Ashraful Goni et al. [10] applied LLMs to machine translation and sentiment analysis in ethnic media, showing that fine-tuning with domain-specific data improved cultural and contextual accuracy. Despite these advancements, key gaps remain: a shortage of annotated Banglish datasets, limited evaluation of LLMs on code-switched text, and a lack of real-time solutions tailored for dynamic, mixed-language environments.

This study addresses these gaps by evaluating Llama 3 alongside other state-of-the-art LLMs for Banglish sentiment analysis, incorporating zero-shot and few-shot learning, a dual-phase pipeline combining real-time translation with sentiment classification, fine-tuning on annotated corpora, and ensemble-based inference. By integrating these techniques, we aim to establish a scalable, high-performance framework for mixed-language sentiment analysis, contributing both to practical applications and to the broader understanding of NLP in linguistically diverse contexts.

Materials And Methods

Dataset description and preprocessing

This study utilised the publicly available Bengali_Banglish_80K_Dataset (DOI: 10.17632/4dnrwbxt8n.2) [11], which contains 80,000 paired Bengali and Banglish sentences. From this corpus, 20,000 rows were randomly selected to serve as the initial dataset. The sequential workflow for preparing the Banglish sentiment dataset, from automated annotation to final manual verification, is illustrated in Figure *1*.

How to cite this article:
Rahman M, Rahman R (February 24, 2026) Sentiment Classification in Low-Resource, Code-Mixed Languages: Comparative Study of Llama 3 and Other Large Language Models. Cureus J Comput Sci 3 : es44389-025-00033-3. DOI https://doi.org/10.7759/s44389-025-00033-3

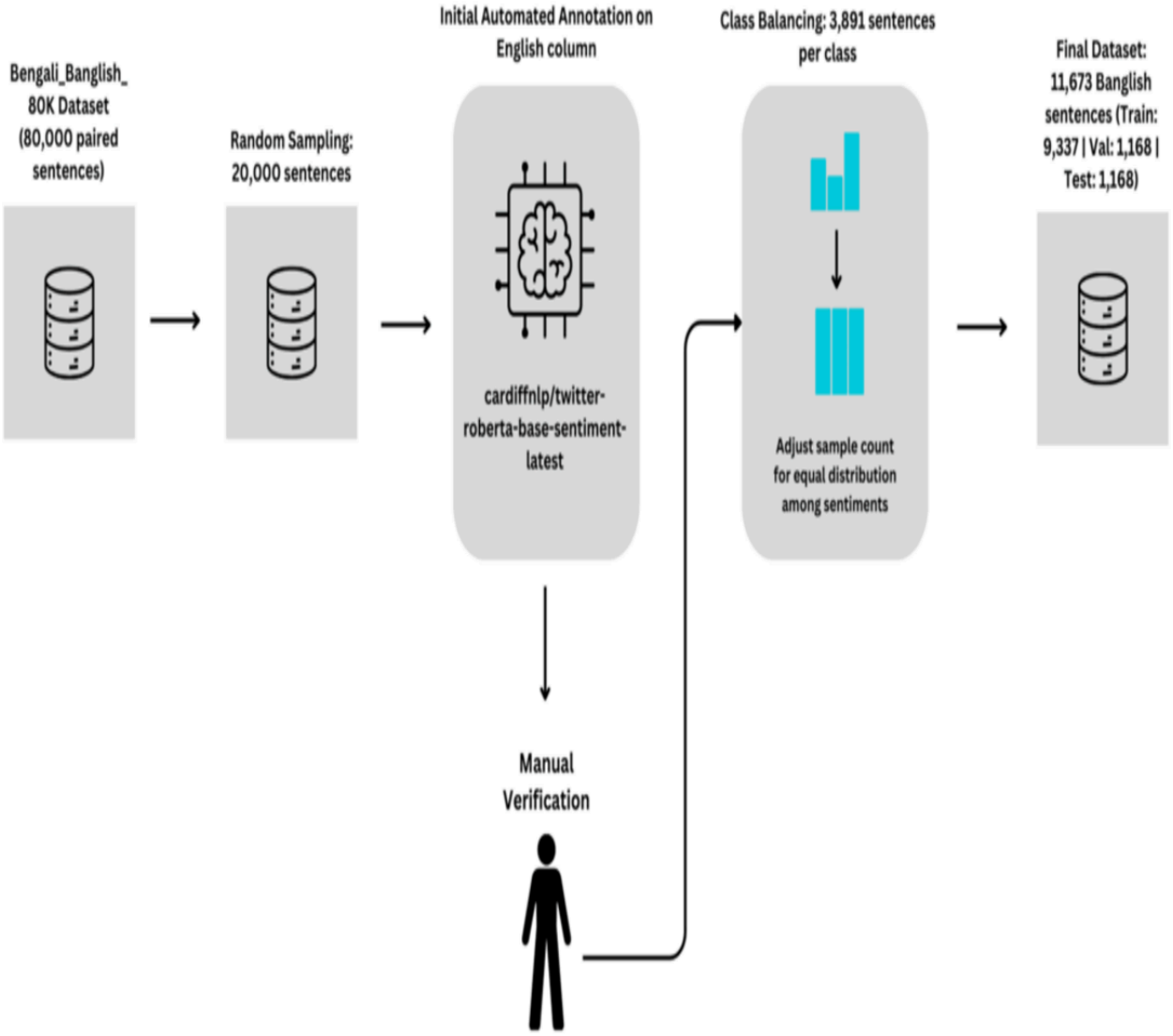

FIGURE 1: Sequential Steps in Preparing the Banglish Sentiment Dataset from Initial Annotation to Manual Verification

Database icons: Data storage and selection
Brain icon: Automated machine language processing (cardiffnlp model)
Human icon: Manual verification and quality control
Chart icons: Class balancing operations

The preparation process began with automated sentiment annotation of the English column using the cardiffnlp/twitter-roberta-base-sentiment-latest model. To achieve balanced class representation, the dataset was adjusted so that each sentiment category, positive, negative, and neutral, contained an equal number of entries, resulting in 11,673 sentences with 3,891 instances per class. The final class distribution is presented in Table *1* and visualised in Figure *2*.

How to cite this article:
Rahman M, Rahman R (February 24, 2026) Sentiment Classification in Low-Resource, Code-Mixed Languages: Comparative Study of Llama 3 and Other Large Language Models. Cureus J Comput Sci 3 : es44389-025-00033-3. DOI https://doi.org/10.7759/s44389-025-00033-3

Sentiment	Count	Percentage
Positive	3,891	33.33%
Negative	3,891	33.33%
Neutral	3,891	33.33%
Total	11,673	33.33%

TABLE 1: Distribution of Sentiment Classes

How to cite this article:
Rahman M, Rahman R (February 24, 2026) Sentiment Classification in Low-Resource, Code-Mixed Languages: Comparative Study of Llama 3 and Other Large Language Models. Cureus J Comput Sci 3 : es44389-025-00033-3. DOI https://doi.org/10.7759/s44389-025-00033-3

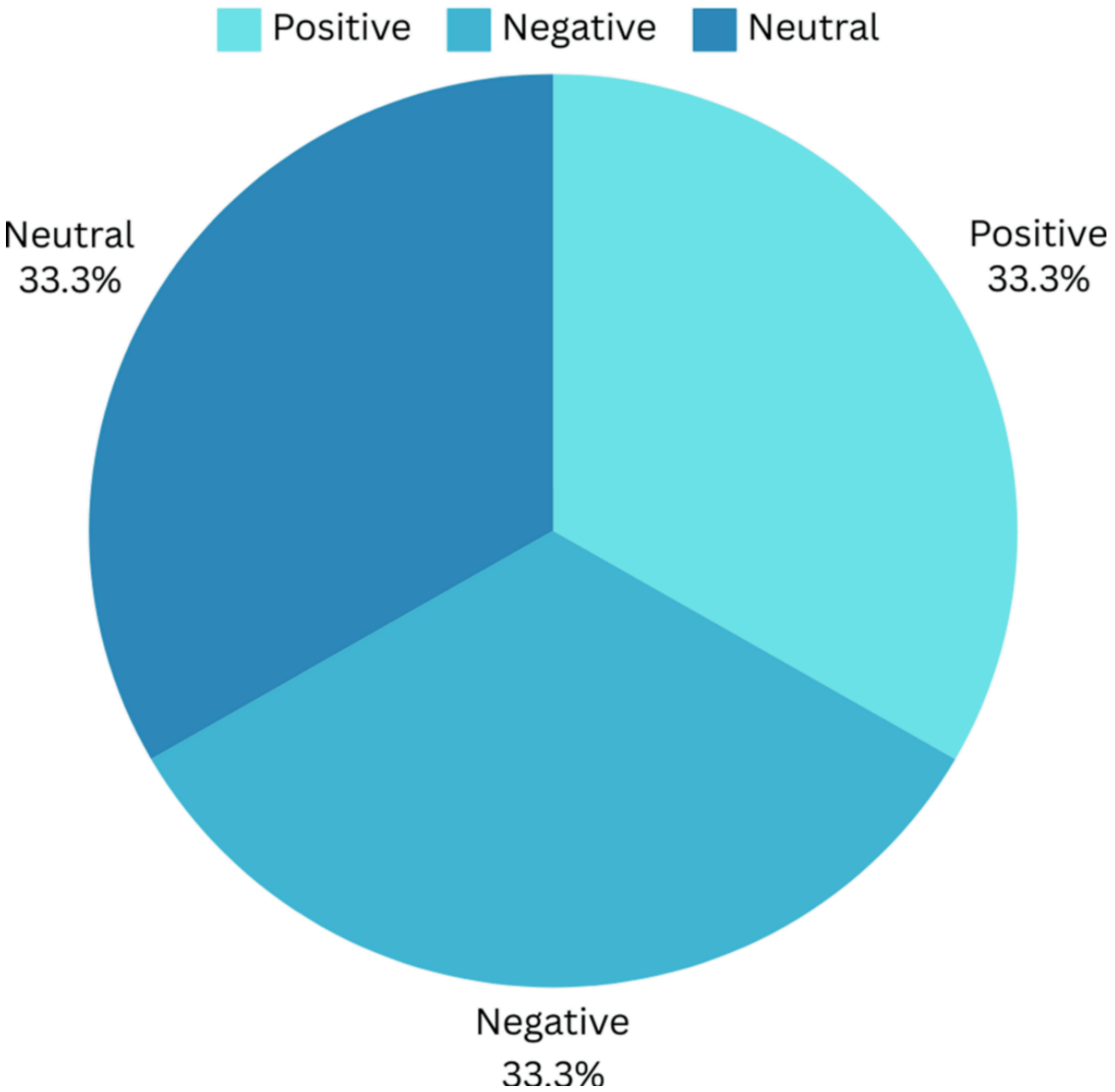

FIGURE 2: Distribution of Sentiment Classes in the Banglish Dataset (N = 11,673)

Following automated labelling, each Banglish sentence was manually reviewed to verify and, where necessary, correct its sentiment classification, ensuring that the dataset accurately reflected the linguistic and contextual nuances unique to Banglish. The dataset was then partitioned into training, validation, and test subsets in an 80:10:10 ratio while preserving the class balance across all splits. The split statistics are shown in Table *2*, with a visual breakdown in Figure *3*.

How to cite this article:
Rahman M, Rahman R (February 24, 2026) Sentiment Classification in Low-Resource, Code-Mixed Languages: Comparative Study of Llama 3 and Other Large Language Models. Cureus J Comput Sci 3 : es44389-025-00033-3. DOI https://doi.org/10.7759/s44389-025-00033-3

Set	Positive	Negative	Neutral	Total
Training	3,112	3,112	3,113	9,337
Validation	389	390	389	1,168
Test	390	389	389	1,168

TABLE 2: Dataset Split Statistics

How to cite this article:
Rahman M, Rahman R (February 24, 2026) Sentiment Classification in Low-Resource, Code-Mixed Languages: Comparative Study of Llama 3 and Other Large Language Models. Cureus J Comput Sci 3 : es44389-025-00033-3. DOI https://doi.org/10.7759/s44389-025-00033-3

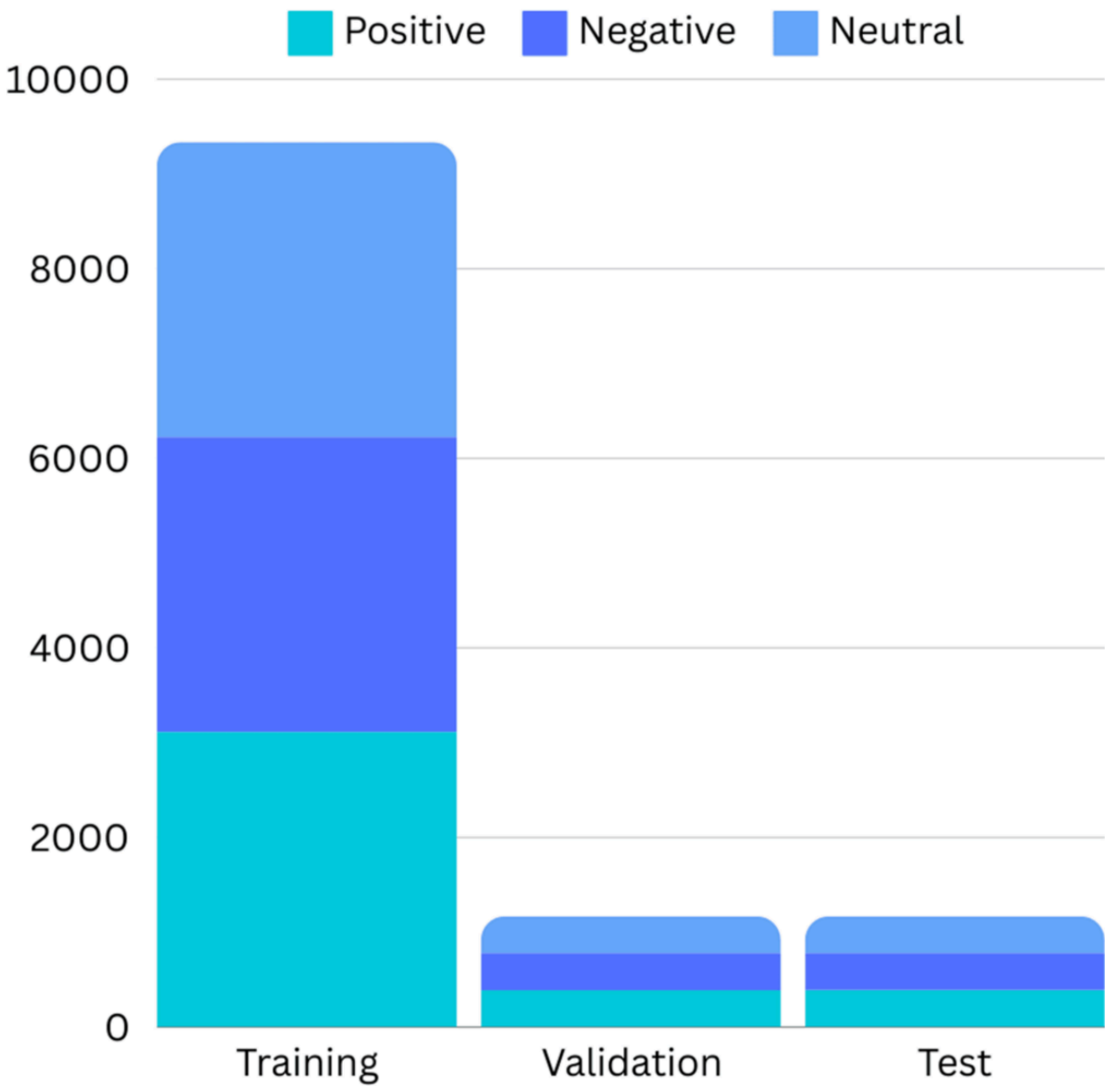

FIGURE 3: Distribution of Sentiment Classes Across Training, Validation, and Test Sets

To illustrate the diversity and structure of the corpus, Table *3* provides representative samples from each sentiment category, including the Bengali text, Banglish transliteration, English translation, and sentiment label.

Label	Bengali	Banglish	English	Sentiment
Sadness	চুরির চেষ্টা আবার	Curir ceshta abar	The theft attempt again	Negative
Joy	বাংলাদেশের সর্ব কালের সেরা প্লেয়ার	Bangladesher sorbo kaler shera player	Bangladesh's best player of all time	Positive
Sadness	সে বললো:- দেখুন আমি গরিব হতে পারি কিন্তু লোভী না	Se bollo:- dekhun ami gorib hote pari kintu lovi na	He said:- See I may be poor but not greedy	Neutral
Fear	রঞ্জনের স্ত্রী নিজের কংকালের ছবি দেখে আঁতকে উঠে বলেন আমি যেন সাক্ষাৎ মৃত্যুকে দেখতে পাচ্ছি কঙ্কের সামনে	Ronjner stri nijer kongkaler chobi dekhe atke uthe bolen ami jen sakkhat mritjuke dekhte pacchi cokher samne	Ranjan's wife is shocked to see the picture of her skeleton and says I am seeing death in front of my eyes	Negative

TABLE 3: Sample Sentences

Methodology

Following dataset preparation, the methodological pipeline was designed to address the specific challenges of sentiment analysis in Banglish through a comprehensive evaluation of multiple large language models with distinct architectural characteristics. The experimental framework incorporated four state-of-the-art large language models: Llama 3, GPT-3.5, GPT-4, and Claude 3.5. The approach was centred on the Llama 3 model with eight billion parameters, selected for its strong performance in preliminary evaluations, open-source accessibility enabling fine-tuning, and adaptability to linguistically diverse contexts. The key specifications of the model, including parameter size, training data composition, context length, and knowledge cutoff date, are summarised in Table *4*. Table *5* provides comprehensive architectural details for all evaluated models.

	Training Data	Parameter	Context Length	GQA	Token Count	Knowledge Cutoff
Llama 3	A new mix of publicly available online data	8B	8k	Yes	15T+	March 2023

TABLE 4: Overview of Llama 3 - 8B Model Specifications

GQA, Grouped Query Attention

How to cite this article:
Rahman M, Rahman R (February 24, 2026) Sentiment Classification in Low-Resource, Code-Mixed Languages: Comparative Study of Llama 3 and Other Large Language Models. Cureus J Comput Sci 3 : es44389-025-00033-3. DOI https://doi.org/10.7759/s44389-025-00033-3

Model	Architecture	Parameters	Context Length	Key Features	Implementation
Llama 3-8B	Transformer Decoder	8B	8,192 tokens	RMSNorm, SwiGLU, GQA, RoPE	Unsloth FastLanguageModel
GPT-3.5	Transformer	~175B	4,096 tokens	RLHF training, instruction following	OpenAI API v1
GPT-4	Transformer (multimodal)	~1.7T	32,768 tokens	Advanced reasoning, multimodal capabilities	OpenAI API v1
Claude 3.5	Constitutional AI	~200B	200,000 tokens	Safety-focused training, constitutional AI	Anthropic API v1

TABLE 5: Comparative Model Architectures and Implementation Details

GQA, Grouped Query Attention; RLHF, Reinforcement Learning from Human Feedback; RoPE, Rotary Positional Embeddings

Llama 3 employs several architectural innovations, including Grouped Query Attention for improved inference efficiency, RMSNorm for stable training dynamics, SwiGLU activation functions for enhanced performance, and Rotary Positional Embeddings for better positional understanding, particularly beneficial for code-mixed languages with variable syntactic structures. GPT-3.5 and GPT-4 utilize reinforcement learning from human feedback and advanced reasoning capabilities, while Claude 3.5 incorporates constitutional AI training for enhanced safety and cultural sensitivity. The Llama 3 implementation utilized the unsloth/llama-3-8b-bnb-4bit variant with 4-bit quantization for memory efficiency, configured with a maximum sequence length of 2,048 tokens and automatic dtype detection for optimal performance.

The first phase of experimentation involved zero-shot and few-shot evaluations to assess each model's ability to classify Banglish sentiment with minimal or no task-specific examples. Few-shot learning employed carefully selected representative examples from each sentiment class, formatted as input-output pairs within the prompt context using a structured template that instructed the model to analyze Banglish sentences and respond with single-word sentiment labels (positive, negative, neutral). Example selection was based on linguistic diversity, clarity of sentiment expression, and cultural relevance to ensure optimal in-context learning performance.

To further enhance performance, a dual-phase pipeline was implemented in which Banglish sentences were first translated into English before undergoing sentiment classification, leveraging the potential strength of models in English sentiment analysis while attempting to preserve the emotional context of the original Banglish input. This approach addresses the hypothesis that translation-mediated analysis might capture sentiment more effectively than direct code-mixed processing, though it introduces potential semantic drift. The workflow for this dual-phase approach is presented in Figure *4*.

How to cite this article:
Rahman M, Rahman R (February 24, 2026) Sentiment Classification in Low-Resource, Code-Mixed Languages: Comparative Study of Llama 3 and Other Large Language Models. Cureus J Comput Sci 3 : es44389-025-00033-3. DOI https://doi.org/10.7759/s44389-025-00033-3

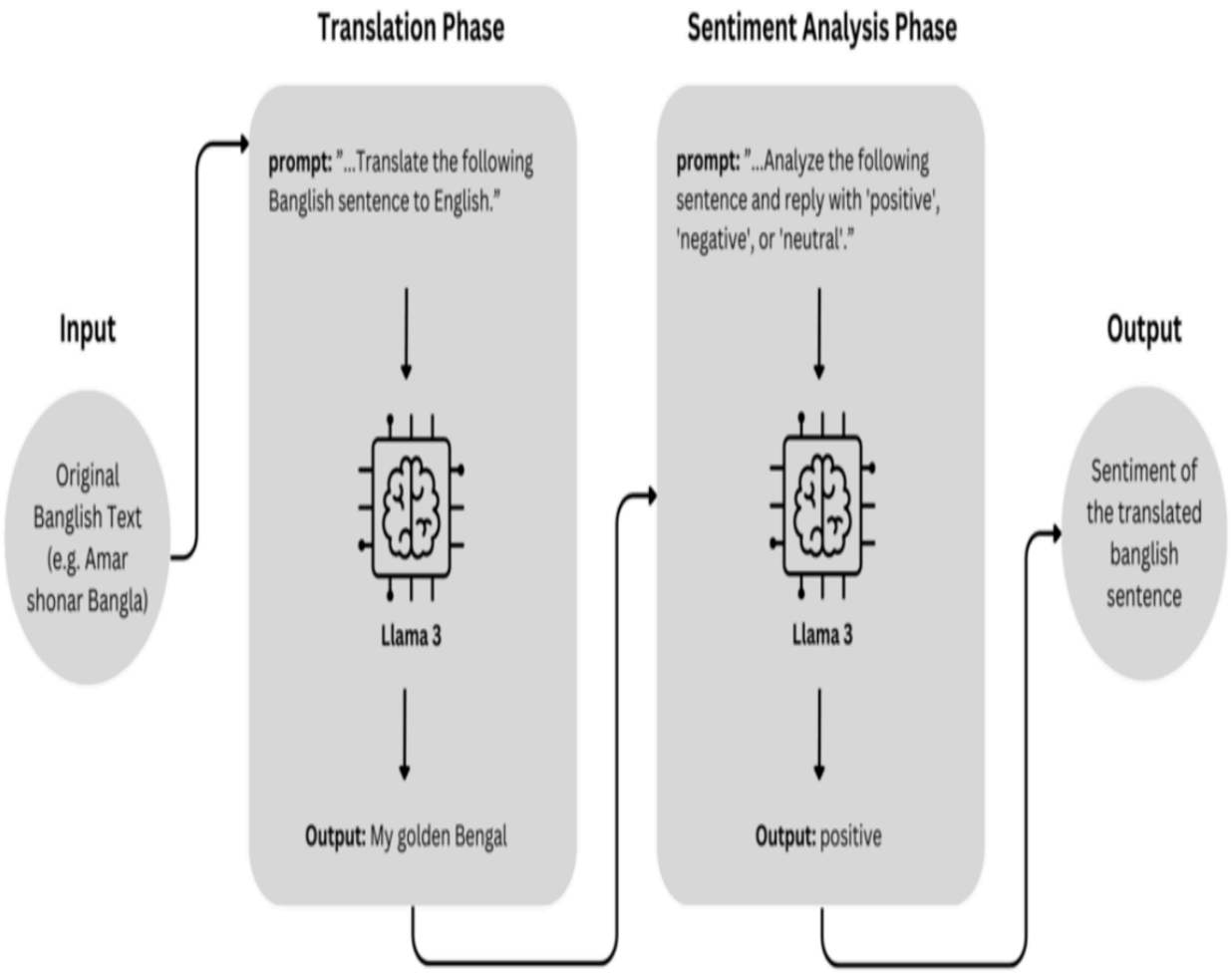

FIGURE 4: Workflow of Dual-Phase Approach

Targeted model adaptation was performed through fine-tuning the unsloth/llama-3-8b-bnb-4bit variant using Parameter-Efficient Fine-Tuning (LoRA) with rank and alpha set to 16, targeting key projection layers with zero dropout. Training was conducted on NVIDIA A100-80GB GPUs using PyTorch 2.0+, CUDA 12.1, and the Unsloth framework with mixed precision (FP16/BF16), AdamW 8-bit optimizer, batch size 4 with gradient accumulation (effective batch size 16), learning rate 5×10^{-5}, cosine scheduler, and 1,000 maximum steps with early stopping. The resulting fine-tuned model, published as samiur-r/BanglishSentiment-Llama3-8B on Hugging Face, is made publicly available for reproducibility.

To improve robustness, three Llama 3 models were fine-tuned independently with different seeds (3,407; 42, 2023) and learning rates (5×10^{-5}, 3×10^{-5}, 1×10^{-5}), then combined using a stacking ensemble strategy. Base model predictions were encoded numerically and fed into a Random Forest meta-classifier (100 estimators, random state 42) trained to

optimally combine individual predictions. This two-stage ensemble approach leveraged diverse learned representations to improve generalization performance.

Evaluation and error analysis

Model performance was assessed using standard multi-class classification metrics, including accuracy, precision, recall, and F1-score for each sentiment category as well as their weighted averages, with both the outputs of the individual fine-tuned models and the ensemble predictions evaluated for comparative purposes. To identify systematic patterns of error, an error analysis was conducted in which misclassified instances were isolated by comparing predicted labels with ground truth, and representative errors were manually reviewed to determine whether misclassifications arose from linguistic ambiguity, idiomatic expressions, or atypical code-switching patterns. This review provided insights into the limitations of the models and informed recommendations for refining preprocessing steps, augmenting dataset coverage, and adjusting model configurations in future work.

Results And Discussion

Experimental setup

The experimental evaluation was conducted on the curated Banglish sentiment analysis dataset described in the Methods section. Multiple modelling strategies were examined to determine their effectiveness in handling the linguistic complexity and cultural nuances of Banglish text. The models tested included zero-shot and few-shot configurations of Llama 3, GPT-3.5, GPT-4, and Claude 3.5, a dual-phase translation-sentiment approach, a fine-tuned Llama 3 model, and an ensemble of independently fine-tuned Llama 3 variants. All approaches were evaluated using accuracy, precision, recall, and F1-score, with the area under the receiver operating characteristic curve (AUC) reported for additional insight. Performance comparisons were carried out under consistent experimental conditions to ensure fairness across methods.

Overview of experimental outcomes

The fine-tuned Llama 3 model achieved the highest accuracy (66.87%), narrowly surpassing the ensemble approach (66.78%). GPT-4 led among zero-shot methods (65.15%), while Llama 3 and Claude 3.5 performed notably lower at 43.80% and 47.68%, respectively. Few-shot learning provided moderate gains over zero-shot, and the dual-phase pipeline reached 48.07%. These findings indicate that task-specific fine-tuning yields clear improvements over baseline approaches. Complete results are shown in Table *6*.

How to cite this article:
Rahman M, Rahman R (February 24, 2026) Sentiment Classification in Low-Resource, Code-Mixed Languages: Comparative Study of Llama 3 and Other Large Language Models. Cureus J Comput Sci 3 : es44389-025-00033-3. DOI https://doi.org/10.7759/s44389-025-00033-3

Method Category	Method	Accuracy	Precision	Recall	F1-Score	AUC
Zero-Shot Methods	Llama 3	43.80%	46.68%	30.00%	48.00%	0.58
	GPT-3.5	55.90%	61.00%	56.00%	48.00%	0.62
	GPT-4.0	65.15%	68.00%	65.00%	61.00%	0.71
	Claude 3.5	47.68%	73.00%	48.00%	43.00%	0.60
Few-Shot Methods	Llama 3 Few-Shot	49.14%	52.01%	49.00%	48.00%	0.63
Translation-Based	Llama 3 Dual-Phase	48.07%	49.34%	48.00%	48.00%	0.61
Fine-Tuned Methods	Llama 3 Fine-Tuning	66.87%	67.00%	67.00%	67.00%	0.75
	Llama 3 Ensemble	66.78%	66.75%	66.78%	66.45%	0.76

TABLE 6: Key Performance Metrics Comparison for Different Sentiment Analysis Approaches

AUC, Area Under the Curve

The fine-tuned Llama 3 maintained balanced precision, recall, and F1, whereas some zero-shot models, like Claude 3.5, achieved high precision but low recall, suggesting under-identification of certain sentiment categories. AUC patterns were consistent with these results: zero-shot Llama 3 scored 0.58, few-shot reached 0.63, dual-phase 0.61, fine-tuning 0.75, and the ensemble 0.76.

Error analysis

Misclassification patterns (Table 7) showed that fine-tuned Llama 3 adapted best to Banglish, while GPT-4 was the strongest zero-shot performer. Other LLMs without adaptation, particularly zero-shot Llama 3 and Claude 3.5, struggled to generalise. Few-shot and dual-phase approaches improved slightly over zero-shot but underperformed fine-tuning, with the latter likely hampered by translation-induced loss of nuance. The ensemble offered negligible gains over a single fine-tuned model. Compared to prior multilingual and code-mixed sentiment analysis studies (55-65% accuracy without adaptation), the fine-tuned model's 66.87% score sets a competitive benchmark, offering balanced accuracy, recall, and F1 performance often missing in earlier work.

How to cite this article:
Rahman M, Rahman R (February 24, 2026) Sentiment Classification in Low-Resource, Code-Mixed Languages: Comparative Study of Llama 3 and Other Large Language Models. Cureus J Comput Sci 3 : es44389-025-00033-3. DOI https://doi.org/10.7759/s44389-025-00033-3

Method	Banglish Text	True Sentiment	Predicted Sentiment	Error Pattern
Zero-Shot	thik ache bondhu, thik ache,	Neutral	Positive	Neutral phrases with positive connotations misclassified as positive.
	sprshiyake val lagot. se je shorir bilay jan..	Positive	Negative	Positive sentiment misclassified due to ambiguous or mixed expressions.
Few-Shot	thik ache bondhu, thik ache,	Neutral	Positive	Similar to zero-shot, neutral phrases misclassified as positive.
	obilombe mukti hok..	Neutral	Positive	Neutral phrases with positive connotations misclassified as positive.
Dual-Phase	Allah amader upor rohomot borshon korun	Positive	Neutral	Translation errors or loss of nuance during the translation phase.
	ei rannaghor bondho hoye jabe	Negative	Positive	Negative sentiment misclassified as positive.
Fine-Tuning	niler mon kharap or bon ke paowa jayni mon kha...	Negative	Positive	Fine-tuned model struggles with subtle negative expressions.
	ei dhoroner vari desher jonjo omonggolojonok	Negative	Positive	Negative sentiment misclassified as positive due to cultural or linguistic nuance.
Ensemble	afridi bhai ek ranar dayitbo nile hobe na ei ...	Neutral	Positive	Neutral sentiment misclassified as positive due to positive connotations.
	papner potteyak cai	Neutral	Negative	Neutral sentiment misclassified as negative due to ambiguous phrasing.

TABLE 7: Error Analysis of Misclassifications Across Methods

Discussion

The experimental results demonstrate that fine-tuning remains the most effective strategy for Banglish sentiment analysis, with the fine-tuned Llama 3 model achieving the highest accuracy (66.87%) and balanced performance across metrics. While GPT-4.0 performed best among zero-shot methods, the substantial improvement from zero-shot to fine-tuned configurations underscores the necessity of task-specific adaptation for hybrid languages. The dual-phase translation approach underperformed, suggesting that translation may introduce semantic drift and obscure subtle sentiment cues. Ensemble methods offered only marginal gains over the fine-tuned single model, raising questions about their cost-benefit trade-off in this context. These findings collectively highlight the advantage of aligning model training with the linguistic and cultural characteristics of the target language, rather than relying solely on general-purpose architectures.

How to cite this article:
Rahman M, Rahman R (February 24, 2026) Sentiment Classification in Low-Resource, Code-Mixed Languages: Comparative Study of Llama 3 and Other Large Language Models. Cureus J Comput Sci 3 : es44389-025-00033-3. DOI https://doi.org/10.7759/s44389-025-00033-3

The study's scope is limited by its reliance on a balanced dataset drawn from a specific corpus, which may not capture the full diversity of Banglish usage across regions and demographics. Expanding the dataset and exploring alternative fine-tuning strategies or model architectures could further improve robustness. Nevertheless, this work provides evidence that targeted fine-tuning can substantially enhance performance in low-resource, code-mixed language settings. As hybrid languages become increasingly prevalent in digital communication, the ability to accurately process and interpret them will be essential for applications ranging from social media analytics to customer feedback systems.

Conclusions

This study establishes a performance benchmark for Banglish sentiment analysis, demonstrating that fine-tuning the Llama 3 model on a curated Banglish dataset substantially outperforms zero-shot and translation-based methods, achieving a 23.07% accuracy improvement over the best zero-shot configuration. The best-performing model, released publicly as samiur-r/BanglishSentiment-Llama3-8B, offers a reproducible resource for further research on hybrid language processing. The results underscore the importance of domain-specific adaptation for mixed-language contexts and highlight the limitations of relying solely on general-purpose models. While dataset scope and linguistic coverage remain constraints, expanding data diversity, exploring novel architectures, and optimising the balance between computational efficiency and predictive accuracy will further advance the field. These findings have practical relevance for applications such as social media analytics, customer feedback monitoring, and multilingual digital communication, while providing a foundation for extending tailored NLP strategies to other low-resource, code-mixed language settings.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Md Samiur Rahman, Rashedur Mohammad Rahman

Acquisition, analysis, or interpretation of data: Md Samiur Rahman, Rashedur Mohammad Rahman

Drafting of the manuscript: Md Samiur Rahman, Rashedur Mohammad Rahman

Critical review of the manuscript for important intellectual content: Md Samiur Rahman, Rashedur Mohammad Rahman

Supervision: Rashedur Mohammad Rahman

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

This work was supported by the Faculty Research Grant [CTRG-23-SEPS-09], North South University, Bashundhara, Dhaka 1229, Bangladesh. The datasets and code generated and/or analyzed during this study are included in this article or are available from the corresponding author upon reasonable request.

References

1. Das R, Singh TD: Multimodal sentiment analysis: A survey of methods, trends, and challenges. ACM Computing Surveys. 2023, 55:1-38. 10.1145/3586075
2. Hartmann J, Heitmann M, Siebert C, Schamp C: More than a feeling: Accuracy and application of sentiment analysis. International Journal of Research in Marketing. 2023, 40:75-87. 10.1016/j.ijresmar.2022.05.005
3. Fatouros G, Soldatos J, Kouroumali K, Makridis G, Kyriazis D: Transforming sentiment analysis in the financial domain with ChatGPT. Machine Learning with Applications. 2023, 14:100508. 10.1016/j.mlwa.2023.100508
4. Zhang W, Deng Y, Liu B, Pan S, Bing L: Sentiment analysis in the era of large language models: A reality check. Findings of the Association for Computational Linguistics: NAACL 2024. 2024, 3881-3906. 10.18653/v1/2024.findings-naacl.246
5. Rusnachenko N, Golubev A, Loukachevitch N: Large language models in targeted sentiment analysis for Russian. Lobachevskii Journal of Mathematics. 2024, 45:3148-3158. 10.1134/s1995080224603758
6. Buscemi A, Proverbio D: ChatGPT vs Gemini vs Llama on multilingual sentiment analysis. arXiv. 2024, 10.48550/arXiv.2402.01715
7. Chowdhury AK, Sujon MSR, Shafi MSS, Ahmmad T, Ahmed S, Hasib KM, Shah FM: Harnessing large language models over transformer models for detecting Bengali depressive social media text: A comprehensive study. arXiv. 2024, 10.48550/arXiv.2401.07310
8. Veeramani H, Thapa S, Naseem U: MLInitiative@WILDRE7: Hybrid approaches with large language models for enhanced sentiment analysis in code-switched and code-mixed texts. Proceedings of the 7th Workshop on Indian Language Data: Resources and Evaluation. 2024, 66-72.
9. Lossio-Ventura JA, Weger R, Lee AY, et al.: A comparison of ChatGPT and fine-tuned open pre-trained transformers (OPT) against widely used sentiment analysis tools: Sentiment analysis of COVID-19 survey data. JMIR Mental Health. 2024, 11:e50150. 10.2196/50150
10. Ashraful Goni MD, Mostafa F, Kee KF: Bangla AI: A framework for machine translation utilizing large language models for ethnic media. arXiv. 2024, 10.48550/arXiv.2402.14179
11. Faisal MR, Shifa AM, Rahman MH, Uddin MA, Rahman RM: Bengali & Banglish: A monolingual dataset for emotion detection in linguistically diverse contexts. Data in Brief. 2024, 55:110760. 10.1016/j.dib.2024.110760

How to cite this article:
Rahman M, Rahman R (February 24, 2026) Sentiment Classification in Low-Resource, Code-Mixed Languages: Comparative Study of Llama 3 and Other Large Language Models. Cureus J Comput Sci 3 : es44389-025-00033-3. DOI https://doi.org/10.7759/s44389-025-00033-3