

Article

Not peer-reviewed version

Exploring Large Language Models for Multitask Learning in Bengali Text Classification

Md. Sajjad Hossain , [Kawsar Ahmed](#) , Suny Md Ashraf Khan , [Mohammed Moshikul Hoque](#) *

Posted Date: 3 April 2026

doi: 10.20944/preprints202604.0275.v1

Keywords: natural language processing; LLMs; low-resource languages; multitask text classification



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Exploring Large Language Models for Multitask Learning in Bengali Text Classification

Md. Sajjad Hossain ¹, Kawsar Ahmed ¹, Suny Md Ashraf Khan ²
and Mohammed Moshikul Hoque ^{1,*}

¹ NLP Lab, Chittagong University of Engineering and Technology, Bangladesh,

² Infratech Consultants Ltd, Bangladesh

* Correspondence: moshikul_240@cuet.ac.bd

Abstract

Text classification in low-resource languages has become increasingly important due to the rapid growth of user-generated digital content. While multitask learning has long been studied in NLP, the use of LLMs for multitask text classification in low-resource languages such as Bengali remains underexplored. Although LLMs are inherently multilingual and multitasking, their effectiveness in structured multitask classification settings for Bengali has not been systematically evaluated. In this work, we investigate how LLMs can be leveraged for multitask Bengali text classification across five domains: sentiment analysis, aggressive text detection, fake news detection, news categorization, and emotion analysis. We compare in-context learning strategies—including zero-shot, one-shot, and chain-of-thought prompting—with parameter-efficient fine-tuning approaches. Our findings show that CoT prompting does not consistently improve performance and often degrades performance, highlighting the instability of prompt-based adaptation in low-resource settings with limited pretraining exposure. Moreover, reasoning-optimized models such as DeepSeek-R1 exhibit substantial performance drops, indicating that enhanced reasoning capabilities alone cannot overcome the challenges posed by low-resource settings. Among the evaluated mLLMs, Gemma-3-4B demonstrates the most stable and balanced cross-task performance under both in-context learning and parameter-efficient fine-tuning, making it a strong backbone candidate for multitask Bengali text classification. These results provide empirical evidence on the limitations of prompting and the advantages of lightweight fine-tuning for low-resource multilingual NLP.

Keywords: natural language processing; LLMs; low-resource languages; multitask text classification

1. Introduction

With the rapid expansion of internet access and social media usage in Bangladesh and Bengali-speaking communities, large volumes of user-generated textual content are produced daily. This content often contains sentiment, misinformation, aggression, and emotionally expressive language, making automated text classification (TC) an essential tool for content moderation, information verification, and social media analysis. In recent years, several studies have addressed Bengali TC tasks individually, including sentiment analysis [1], fake news detection [2], emotion classification [3], and aggressive text detection [4]. These works have contributed valuable task-specific datasets and modeling approaches. However, most existing research focuses on single-task settings, where models are trained independently for each classification objective. In contrast, multitask learning (MTL)—which enables a unified model to learn shared representations across multiple related tasks—remains relatively underexplored in Bengali NLP. Although multi-task text classification has witnessed significant advancements in high-resource languages such as English and Chinese, progress in low-resource languages—including Bengali—remains limited [5]. The development of robust Bengali classification systems is hindered by the scarcity of standardized annotated corpora, limited domain-diverse

datasets, insufficient pre-trained embeddings, and comparatively underdeveloped NLP infrastructure. These challenges are further compounded in multitask settings, where a single model must generalize across heterogeneous domains with varying linguistic and semantic complexities.

Recent advances in Large Language Models (LLMs) have transformed NLP by enabling deep contextual understanding, cross-lingual transfer, and strong zero-shot and few-shot performance. LLMs have demonstrated impressive results across various classification tasks and languages [6,7]. Their multilingual pretraining suggests potential applicability to low-resource languages. Despite this promise, the effectiveness of LLMs for structured multitask text classification in Bengali remains largely unexplored. In particular, there is limited empirical evidence comparing different adaptation strategies—such as in-context learning and parameter-efficient fine-tuning—for Bengali classification tasks. Although modern LLMs are multilingual, their pretraining corpora are disproportionately dominated by high-resource languages. This imbalance often leads to suboptimal performance in low-resource contexts, where linguistic structures, morphology, cultural expressions, and script characteristics differ significantly from those of dominant languages. Understanding how different LLM architectures and adaptation techniques perform under these constraints is therefore essential for advancing multilingual NLP research.

In this work, we systematically investigate the use of LLMs for multitask Bengali text classification across multiple domains. By evaluating different adaptation paradigms and analyzing their task-specific behavior, we aim to provide empirical insights into the capabilities and limitations of LLMs in low-resource multilingual settings. The key findings of this work can be summarized as follows:

- The results demonstrate that Chain-of-Thought (CoT) prompting does not consistently enhance performance in Bengali text classification and, in several cases, leads to measurable degradation. This finding underscores the instability of prompt-based adaptation strategies in low-resource settings where pretraining exposure to the target language is limited.
- Reasoning-optimized large language models, such as DeepSeek-R1, exhibit significant performance deterioration across multiple tasks. This reveals a critical limitation of scaling reasoning capacity without sufficient linguistic grounding, particularly in low-resource language contexts.
- Among the evaluated models, Gemma-3-4B achieves the most stable and consistent performance across diverse tasks under both in-context learning and parameter-efficient fine-tuning. Its balanced cross-task generalization highlights its suitability as a backbone model for multitask Bengali text classification.

2. Related Work

There has been significant progress on TC tasks in high-resource languages (HRLs) using LLMs. In contrast, low-resource languages (LRLs) like Bengali have seen much less improvement.

2.1. Text Classification in HRLs

There are numerous ways of text classification using ML, DL, transformers, and LLMs. Most research on LLMs has focused on high-resource languages, such as English. Reference [8] proposed a few-shot text classification benchmark using GPT-3. They evaluated 11 tasks across 11 datasets in English. Reference [9] also performed 11 tasks, as before, and used the same dataset for each task. They used Pattern-Exploiting Training (PET), which performed nearly as well as non-expert humans in 7 out of 11 tasks. Their model outperformed that of Reference [8] on several tasks. However, they could not improve the accuracy of some tasks, which appeared to be very low. Reference [10] used context learning with GPT-3.5 and GPT-4 for classifying text of Banking77 datasets. This dataset comprises 77 classes and 13083 samples. Their approach outperformed fine-tuned, non-generative models even with fewer samples. They got an F1 score of 0.83. Reference [11] proposed a simplified approach for classifying text with LLMs (GPT-4, GPT-3.5, Gemini-pro, Llama-3-8b, etc) using zero and few-shot prompting. They demonstrated that zero-shot prompting is a more convenient approach for text classification than traditional methods (ML, DL, and Transformer models). Although several TC

methods in HRLs have demonstrated promising results, these models cannot be used for resource-constrained languages (e.g., Bengali) due to insufficient domain-specific training data, linguistic diversity, fine-tuned embeddings, and classification methods.

2.2. Text Classification in LRLs

In addition to HRLs, there has been extensive research on LRLs such as Telugu, Marathi, and Tamil. Reference [12] proposed MT-Text GCN, which combines graph autoencoder-based graph reconstruction and multi-task text classification in Telugu. They achieved F1 scores of 0.84 (Sentiment), 0.55 (Emotion), 0.83 (Hate Speech), and 0.66 (Sarcasm). Reference [13] also proposed a multi-task learning framework for sentiment analysis on various datasets, combining hard parameter sharing, a BERT-based shared encoder, and a task recognition mechanism. Their approach achieved an F1 score of 0.91. Reference [14] proposed a novel multi-task learning framework for detecting hateful memes. They integrated three pre-trained models: CLIP, BERT, and UNITER, as their work involves both text and images. On the MultiOFF and MAMI datasets, they achieved F1 scores of 0.79 and 0.84, respectively. A recent study [15] proposed a multimodal zero-shot multi-task learning framework that uses cross-attention to identify intent and emotion. Their model achieved an increase in F1 score of 0.06 for intent detection and 0.04 for emotion detection compared to other approaches.

2.2.1. Text Classification in Bengali

In contrast to the other languages, TC research leveraging LLMs in Bengali is currently at a rudimentary stage. A limited number of studies also utilized LLMs for TC tasks. For example, Reference [16] conducted multiple TC tasks such as question answering, paraphrasing, summarization, and text classification using LLMs (GPT-3.5, Claude-2, LLaMA-2). All the LLMs performed poorly with an F1 score of 0.49 (Claude-2), 0.48 (GPT-3.5), and 0.29 (LLaMA-2). Reference [17] explored LLMs with zero and few-shot techniques for Bengali sentiment analysis. The performance of GPT-4 was inferior in their study. The F1 score of GPT-4 in various settings was 0.60 (0-shot), 0.60 (3-shot), and 0.61 (5-shot). A recent work [18] explored various LLMs across four TC tasks, including text classification, sentiment analysis, summarization, and question answering. Their results demonstrated that the closed-source GPT-4o model, utilizing few-shot learning and chain-of-thought prompting, achieved the highest performance across multiple tasks. Few studies also focused on TC in various domains using ML and DL techniques, such as sports news classification [19], suspicious text classification [20], and authorship classification [21]. In recent days, transformer-based approaches have gained much attention in various Bengali TC research, including emotion classification [3,22], sentiment analysis [23], technical text classification [24], fake news classification [25], authorship attribution [26], and aggressive text classification [27]. Past research has shown that Bengali TC models struggle to comprehend new tasks that are not widely recognized.

Most past studies in Bengali have focused on a single task or domain, lacking a unified evaluation benchmark for multiple TC tasks or domains. To address this gap, this study explores five distinct Bengali text classification tasks using LLMs.

3. Datasets

The proposed method is evaluated on five Bengali tasks: Sentiment Analysis (SA), Aggressive Text Detection (ATD), Fake News Detection (FND), News Headline Categorization (NHC), and Textual Emotion Analysis (TEA). Each task utilizes its respective dataset from available sources, including SA [28], ATD [29], FND [30], NHC [31], and TEA [3]. For in-context learning, the evaluation dataset is used directly without modification. However, for PEFT [6], changes were made by manually adding task-specific instructions (prompts) to help the model identify the task being addressed. Additionally, to reduce output ambiguity, a class label name and a corresponding number were added to each output. Table 1 presents a summary of all datasets used in this study. This study encompasses 6,026 Bengali text samples across five distinct classification tasks for evaluation.

Table 1. Overview of Bengali text classification evaluation datasets. Here, LR refers to the range of minimum and maximum lengths of each sample, while T_s and T_w denote the total number of sentences and total number of words.

Task	Data	LR	T_s	T_w	Classes	Class Labels
SA [28]	434	1-562	456	10,909	2	Positive, Negative
ATD [29]	1,416	3-595	1897	31189	2	Aggressive, Non-Aggressive
FND [30]	2,551	390-3,761	54,508	6,67,351	2	Fake, Authentic
NHC [31]	1,000	2-9	1000	5932	6	Politics, Sports, National, Entertainment, International, IT
TEA [3]	625	4-97	832	14,508	6	Joy, Sadness, Fear, Disgust, Anger, Surprise
Total	6,026	-	58,693	729,889	-	-

For instruction tuning, we add task-specific samples to create a comprehensive training set. Each sample includes detailed task instructions, specific prediction guidance, and class names with corresponding labels for clarity and accuracy. Table 2 provides a summary of the instruction tuning dataset composition.

Table 2. Instruction tuning dataset statistics for tuning LLMs.

Task	Samples	T_s	T_w
SA	1,010	999	25,443
ATD	10,000	13,491	2,11,096
FND	2,551	1,26,004	1554914
NHC	15,000	15,000	88,512
TEA	4,000	6543	1,14,674
Total	32,561	162,037	1,994,639

The instruction tuning dataset is a significant expansion over the evaluation datasets, comprising 32,561 samples. To create an instruction dataset, we first inject a prompt before each sample in each task. The prompt provides an instruction to determine the accurate class of the task.

Table 3 provides an overview of the Alpaca-style instructions dataset for the tasks.

Table 3. Sample instructions, inputs, and outputs for each task.

Instruction	Input	Output
Find the sentiment of the following Bengali sentence.	[Bengali Text] (This movie was amazing.)	Positive
Detect whether the following Bengali sentence is aggressive or not.	[Bengali Text] (You're a useless donkey!)	Aggressive
Classify whether the following Bengali news is fake or real.	[Bengali Text] (The Prime Minister said that 5,000 taka will be given to each person tomorrow.)	Fake
Categorize the topic of the following Bengali news headline.	[Bengali Text] (The country was not liberated for their meetings and rallies: Hanif.)	Politics
Detect the emotion expressed in the following Bengali sentence.	[Bengali Text] (I'm feeling very sad today.)	Sad

The alpaca-style dataset guides the model to produce outputs in the desired format for Bengali classification tasks. In contrast, untuned LLMs often generate extraneous text. By training on this dataset, the model is constrained to provide concise, relevant responses tailored to the target classification task. Figure 1 illustrates the distribution of classes among all tasks in instruction tuning and evaluation datasets. It shows that the distributions are nearly identical across the two sets. Among all tasks, NHC has the most data (15,000) across six classes, whereas SA has the least (1,010) across two classes. In NHC, 35.9% of the samples belong to the International category, while only 1.9% of the samples belong to the IT category in the instruction tuning datasets.

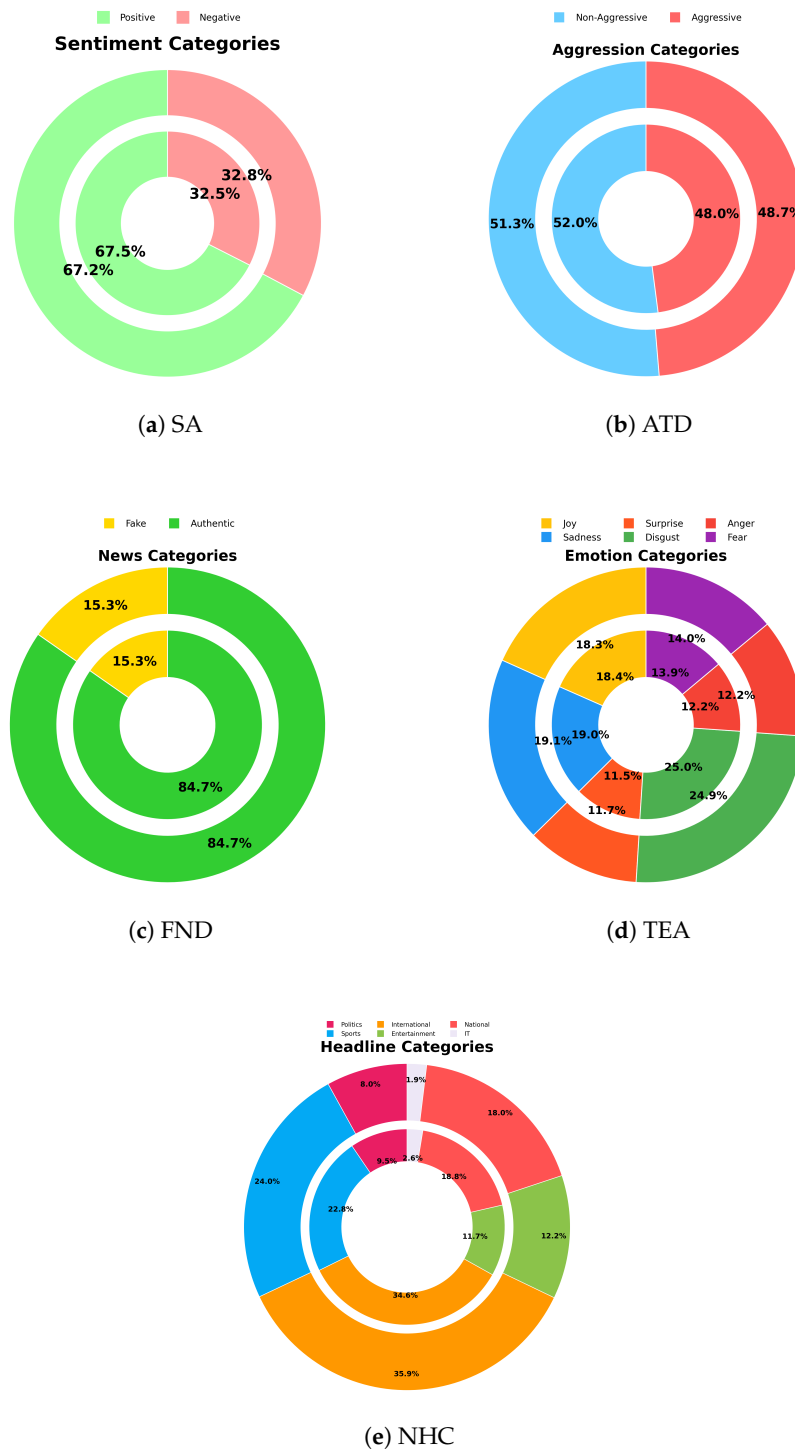


Figure 1. Class-wise distribution of data among all tasks in train and test set.

4. System Overview

This study explores both in-context learning and PEFT techniques to perform the multiple downstream tasks. Figure 2 provides an overview of the overall system.

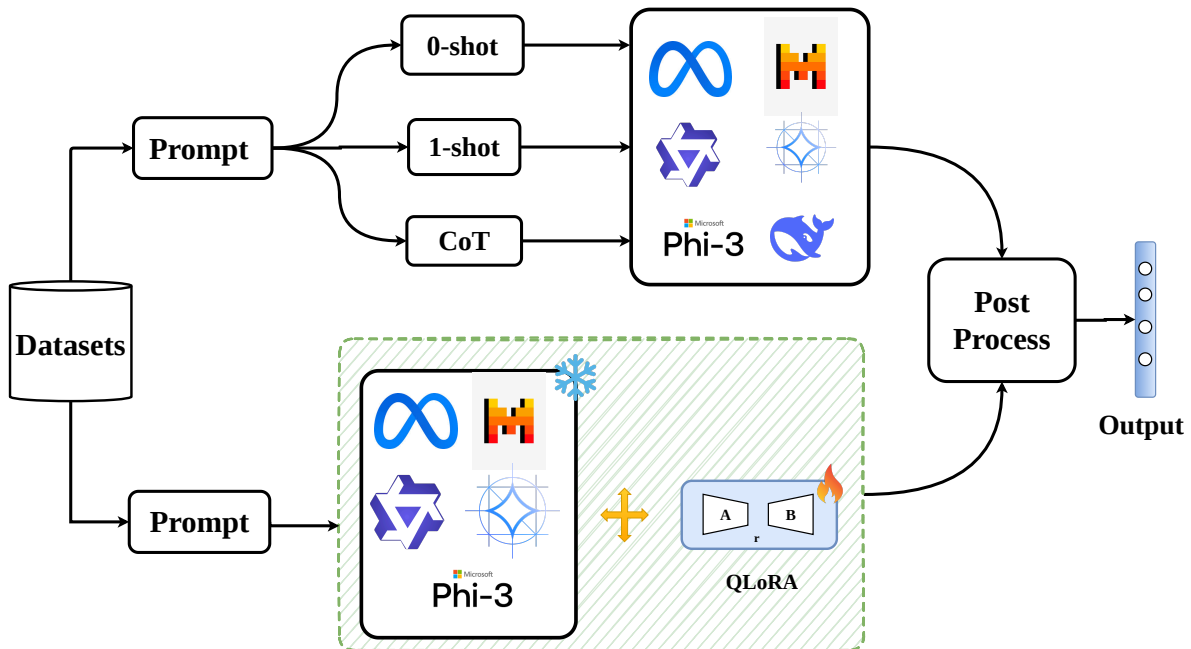


Figure 2. Schematic Diagram of LLM Evaluation for Bengali TC Tasks

4.1. Models

We have explored eight LLMs to evaluate five Bengali text classification tasks. Each model varies according to their internal settings. We have utilized Llama-3.2 3B [32], Mistral-7B [?], Qwen-2.5 [33], Phi-3 [34], Deepseek-7B [35], Gemma-3 [36] and DeepSeek-R1 [37]. For model implementation, we used 4-bit quantized models and QLoRA to tune instructions under limited GPU resources efficiently. Larger closed-source models like DeepSeek-R1 was accessed via their APIs. All other models, ranging from 3 billion to 7 billion parameters, were accessed and executed locally within the Kaggle environment.

4.2. In-Context Learning (ICL)

For ICL, we used different prompting techniques like zero-shot, one-shot, and CoT across five Bengali text classification tasks.

Appendix A illustrates the prompts used in this study.

- **Zero-shot Prompt:** The model is given only the task description without any examples.
- **One-shot Prompt:** The model is provided with a single example of the task along with its answer.
- **Chain-of-Thought (CoT) Prompt:** The model is first given the task context, and then guided to reason through the problem step by step, rather than producing a direct answer.

4.3. Parameter Efficient Fine-Tuning (PEFT)

In this approach, we have utilized different adapters for each task. We have trained five separate QLoRA [38] adapters, one for each of the five tasks, using the SA, ADT, FND, NHC, and TEA datasets. Formally, for a transformer layer weight matrix $W \in \mathbb{R}^{d \times k}$, the adapted weight W' is computed as:

$$W' = W + \Delta W, \quad \Delta W = AB^T \quad (1)$$

where $A \in \mathbb{R}^{d \times r}$ and $B \in \mathbb{R}^{k \times r}$ are trainable low-rank matrices with rank $r \ll \min(d, k)$. In our implementation, we set $r = 16$, which provides a balance between parameter efficiency and expressive capacity.

Following the LoRA formulation, the low-rank update is scaled as:

$$\Delta W = \frac{\alpha}{r} AB^{\top} \quad (2)$$

where α is a scaling factor. We set $\alpha = 32$ to ensure a stable update magnitude while preserving the benefits of low-rank adaptation. To improve memory efficiency, we adopt 4-bit quantization for the base model weights, following the QLoRA framework. This allows training large language models under limited hardware constraints by storing backbone weights in low-precision while maintaining adapter parameters in high-precision. We set the dropout rate to 0 based on empirical tuning, as preliminary experiments showed that additional dropout did not improve performance. During inference, the base model remains fixed, and the appropriate task-specific adapter is dynamically loaded.

After generating an output, the LLM may include unnecessary text, which is removed during post-processing using regular expressions. Also, if the responses were found in the wrong format after post-processing, then they were treated as wrong answers.

4.4. Experimental Setup

This study was conducted using Python 3.10.12, PyTorch 2, and Unsloth to implement large language models (LLMs). Experiments were run on Kaggle using an NVIDIA Tesla P100 GPU, which requires 29 GB of RAM, 16 GB of VRAM, and 73.1 GB of storage. Unsloth was used for PEFT with QLoRA, while DeepSeek-70B APIs were employed. We evaluated performance using three prompting strategies and the metrics: precision (Pr), recall (Re), F1-score (F1), and accuracy (Acc).

5. Results and Analysis

Table 4 demonstrates the evaluation results of eight LLMs across five tasks.

How do the models behave in different ICL settings? Across zero-shot, one-shot, and chain-of-thought (CoT) prompting settings, no single model consistently outperforms others across all five tasks. This variability underscores the heterogeneous linguistic and semantic demands of multitask Bengali text classification, in which different tasks require distinct types of knowledge and reasoning.

Qwen demonstrates the strongest and most stable performance in Sentiment Analysis (SA), achieving F1 scores of 0.95–0.96 across all prompting strategies. The relative insensitivity to prompting variations suggests that sentiment classification primarily relies on coarse-grained polarity cues that transfer effectively across languages. This indicates that Qwen’s multilingual pretraining likely provides sufficient semantic grounding for polarity detection without requiring task-specific adaptation. In contrast, Gemma achieves superior performance on more structurally and semantically complex tasks, including Fake News Detection (FND), News Headline Categorization (NHC), and Text Emotion Analysis (TEA). These tasks demand contextual reasoning, pragmatic interpretation, and fine-grained label discrimination. Gemma’s comparatively strong instruction-following alignment appears better suited to handling multi-class classification scenarios that require nuanced semantic differentiation. Llama exhibits competitive zero-shot performance on Aggressive Text Detection (ATD). This observation further supports the notion that task complexity and cultural specificity significantly influence the effectiveness of cross-lingual transfer.

Collectively, these results indicate that model performance in Bengali multitask classification is highly task-dependent and closely tied to the interaction between pretraining distribution, instruction alignment, and the linguistic characteristics of each classification objective.

Does CoT help in these low-resource tasks? Chain-of-thought prompting does not consistently improve performance and frequently degrades it. Although CoT is theoretically beneficial for tasks requiring structured reasoning [39], its effectiveness depends on the availability of reliable internal knowledge representations. In Bengali, where pretraining exposure is comparatively limited, CoT often produces fluent but semantically unreliable reasoning chains. For example, models optimized for reasoning exhibit significant degradation under CoT on certain tasks. Instead of correcting errors, the additional reasoning steps amplify uncertainty. These findings suggest that CoT requires sufficiently

rich language-specific priors to be effective. In low-resource contexts, it may increase the risk of hallucinations by encouraging overconfident yet weakly grounded intermediate reasoning steps [40].

Table 4. Performance comparison of models across five tasks for zero-shot, one-shot, and chain-of-thought (CoT) prompting.

Task	Model	Zero-shot				One-shot				CoT			
		Pr	Re	F-1	Acc	Pr	Re	F-1	Acc	Pr	Re	F-1	Acc
SA	Llama	0.93	0.92	0.93	0.94	0.97	0.95	0.96	0.97	0.97	0.95	0.94	0.94
	Gemma	0.93	0.84	0.88	0.89	0.91	0.89	0.90	0.88	0.87	0.89	0.88	0.86
	Mistral	0.70	0.68	0.69	0.71	0.65	0.67	0.66	0.64	0.64	0.67	0.66	0.68
	Qwen	0.96	0.94	0.95	0.95	0.97	0.95	0.96	0.94	0.97	0.95	0.96	0.94
	Phi	0.71	0.73	0.72	0.70	0.64	0.67	0.66	0.68	0.65	0.67	0.66	0.64
	Dseek7B	0.81	0.79	0.80	0.83	0.84	0.82	0.83	0.85	0.84	0.82	0.83	0.85
	DseekR1	0.88	0.90	0.89	0.87	0.89	0.91	0.90	0.88	0.89	0.91	0.90	0.88
ATD	Llama	0.81	0.83	0.82	0.84	0.72	0.61	0.54	0.59	0.71	0.69	0.70	0.72
	Gemma	0.81	0.79	0.80	0.82	0.78	0.76	0.77	0.79	0.77	0.75	0.76	0.78
	Mistral	0.60	0.63	0.62	0.61	0.62	0.58	0.54	0.57	0.60	0.62	0.61	0.59
	Qwen	0.70	0.68	0.69	0.71	0.66	0.64	0.65	0.67	0.71	0.69	0.70	0.72
	Phi	0.50	0.52	0.51	0.49	0.51	0.49	0.50	0.52	0.50	0.48	0.49	0.51
	Dseek7B	0.58	0.60	0.59	0.57	0.58	0.56	0.57	0.59	0.45	0.43	0.44	0.46
	DseekR1	0.80	0.78	0.77	0.77	0.63	0.61	0.62	0.64	0.43	0.41	0.42	0.44
FND	Llama	0.62	0.70	0.63	0.73	0.61	0.69	0.48	0.51	0.77	0.75	0.76	0.78
	Gemma	0.87	0.89	0.88	0.86	0.85	0.83	0.84	0.86	0.79	0.77	0.78	0.80
	Mistral	0.60	0.54	0.49	0.58	0.50	0.48	0.49	0.51	0.49	0.47	0.48	0.50
	Qwen	0.70	0.68	0.69	0.71	0.64	0.62	0.63	0.65	0.74	0.72	0.73	0.75
	Phi	0.56	0.54	0.55	0.57	0.43	0.41	0.42	0.44	0.64	0.62	0.63	0.65
	Dseek7B	0.45	0.43	0.44	0.46	0.45	0.43	0.44	0.46	0.42	0.40	0.41	0.43
	DseekR1	0.92	0.93	0.87	0.83	0.67	0.65	0.66	0.68	0.69	0.67	0.68	0.70
TEA	Llama	0.61	0.42	0.32	0.33	0.63	0.58	0.52	0.53	0.48	0.47	0.45	0.46
	Gemma	0.57	0.56	0.55	0.64	0.66	0.63	0.57	0.59	0.56	0.54	0.49	0.50
	Mistral	0.51	0.31	0.25	0.29	0.55	0.39	0.29	0.36	0.34	0.30	0.24	0.30
	Qwen	0.51	0.45	0.42	0.44	0.46	0.44	0.37	0.39	0.51	0.51	0.47	0.48
	Phi	0.47	0.40	0.37	0.39	0.30	0.25	0.24	0.25	0.30	0.28	0.24	0.31
	Dseek7B	0.30	0.28	0.23	0.29	0.32	0.30	0.27	0.29	0.35	0.34	0.32	0.25
	DseekR1	0.48	0.47	0.45	0.46	0.44	0.42	0.41	0.47	0.38	0.37	0.36	0.40
NHC	Llama	0.57	0.55	0.56	0.58	0.34	0.32	0.33	0.35	0.53	0.51	0.52	0.54
	Gemma	0.71	0.76	0.71	0.71	0.69	0.67	0.68	0.70	0.72	0.70	0.71	0.73
	Mistral	0.14	0.12	0.13	0.15	0.37	0.35	0.36	0.38	0.29	0.27	0.28	0.30
	Qwen	0.23	0.21	0.22	0.24	0.43	0.41	0.42	0.44	0.43	0.41	0.42	0.44
	Phi	0.67	0.65	0.66	0.67	0.21	0.19	0.20	0.22	0.18	0.16	0.17	0.26
	Dseek7B	0.22	0.20	0.21	0.23	0.21	0.19	0.20	0.22	0.45	0.43	0.44	0.46
	DseekR1	0.65	0.63	0.64	0.66	0.67	0.65	0.66	0.68	0.68	0.66	0.67	0.69

How do the models perform after PEFT? Table 5 reports the performance of instruction-tuned models fine-tuned with QLoRA (PEFT), and the improvements over the best ICL baselines are consistent and often substantial. The most dramatic gains are observed on the tasks that proved hardest under ICL. For TEA, the best ICL F1 (Gemma, zero-shot) was 0.55; after PEFT fine-tuning, Gemma-3-4B achieves an F1 of 0.74 — a gain of 19 percentage points. Similarly, for NHC, the best ICL F1 score was 0.71 (Gemma, zero-shot), whereas PEFT raises it to 0.83. These results confirm that for culturally embedded tasks such as emotion analysis and news categorization in Bengali, task-specific adaptation to in-domain data is far more effective than any prompting strategy. For ATD, PEFT enables Mistral-7B to reach an F1 of 0.93, compared to a best ICL F1 of 0.82. On FND, Gemma-3-4B achieves an F1 of 0.97 — more than 9 points above its own best ICL performance. Across all five tasks, Gemma-3-4B consistently ranks among the top-performing PEFT models, achieving the highest F1 on SA (0.97), FND (0.97), TEA (0.74), and NHC (0.83). Mistral-7B leads only on ATD (F1: 0.93) but exhibits a catastrophic failure on FND after fine-tuning (F1: 0.06), suggesting a label prediction collapse that warrants further

investigation. The consistency of PEFT gains across models and tasks supports the conclusion that QLoRA fine-tuning efficiently surfaces latent multilingual capacity that prompting alone cannot elicit. By adapting the model’s lower-rank projection matrices to Bengali-specific lexical and morphological distributions, fine-tuning allows even relatively compact models (3B–7B parameters) to substantially close the gap with larger systems evaluated in zero-shot settings.

Table 5. Performance comparison of different instruction-tuned models across five tasks.

Task	Model	Pr	Re	F1	Acc
SA	Llama-3.2-4B	0.83	0.85	0.84	0.85
	Gemma-3-4B	0.96	0.97	0.97	0.97
	Mistral-7B-it	0.89	0.92	0.90	0.91
	Qwen-2.5-8B-it	0.94	0.94	0.94	0.94
	Phi-3	0.84	0.85	0.84	0.86
ATD	Llama-3.2-4B	0.89	0.89	0.89	0.89
	Gemma-3-4B	0.91	0.91	0.91	0.91
	Mistral-7B-it	0.93	0.93	0.93	0.93
	Qwen-2.5-8B-it	0.89	0.89	0.89	0.89
	Phi-3	0.91	0.91	0.91	0.91
FND	Llama-3.2-4B	0.89	0.89	0.89	0.89
	Gemma-3-4B	0.97	0.97	0.97	0.97
	Mistral-7B-it	0.67	0.16	0.06	0.16
	Qwen-2.5-8B-it	0.83	0.84	0.83	0.85
	Phi-3	0.93	0.73	0.81	0.73
TEA	Llama-3.2-4B	0.63	0.61	0.62	0.63
	Gemma-3-4B	0.74	0.74	0.74	0.74
	Mistral-7B-it	0.73	0.70	0.69	0.70
	Qwen-2.5-8B-it	0.67	0.67	0.67	0.67
	Phi-3	0.65	0.66	0.65	0.66
NHC	Llama-3.2-4B	0.76	0.77	0.76	0.77
	Gemma-3-4B	0.83	0.83	0.83	0.83
	Mistral-7B-it	0.81	0.81	0.81	0.81
	Qwen-2.5-8B-it	0.77	0.77	0.77	0.77
	Phi-3	0.27	0.40	0.30	0.40

Which task is most challenging for the models? Across both ICL and PEFT settings, a clear difficulty ordering emerges among the five tasks. SA is the least challenging, with top PEFT models achieving an F1 score above 0.97, whereas TEA remains the hardest, achieving a maximum F1 of 0.74 even after fine-tuning. This ordering reflects the varying degree to which task-relevant knowledge can be transferred from high-resource pretraining distributions. Sentiment polarity is a relatively universal semantic property; emotion, hate, and argumentation in Bengali, by contrast, are deeply tied to cultural register, code-switching practices, and community-specific linguistic conventions that multilingual LLMs are unlikely to have internalized from pretraining data alone.

Which model emerges as the most effective backbone for multitask Bengali text classification across tasks? Among the evaluated models, Gemma-3-4B emerges as the most suitable backbone for multitask Bengali text classification, demonstrating the most consistent cross-task performance under both in-context learning and parameter-efficient fine-tuning.

Why does DeepSeek-R1 perform poorly despite being a larger model? Despite being among the larger models evaluated, R1 underperforms relative to its scale on several tasks. In particular, its CoT performance is consistently inferior to its own zero-shot performance — most dramatically on ATD, where CoT reduces its F1 from 0.77 to 0.42. This behavior is a direct consequence of R1’s training objective, which optimizes for systematic chain-of-thought reasoning. When applied to Bengali classification tasks, this inductive bias is counterproductive: the model is compelled to reason

through a language and domain it has limited knowledge of, producing confident but incorrect reasoning chains. This finding highlights an important limitation of reasoning-optimized LLMs in low-resource languages.

5.1. Ablation Study

We conducted an ablation study on LoRA rank, LoRA α , and LoRA dropout for PEFT. Due to the high computational cost of training Gemma-3 4B Instruct, these experiments were limited to the ATD dataset only. Multiple combinations of LoRA rank, α , and dropout were evaluated, as summarized in Table 6.

Based on the ablation results, Set-8 from Table 6 was selected as the final configuration to develop the best-performing Gemma-3 4B model. Three key factors guided this decision. First, Set-8 achieved comparable or slightly better validation performance than higher-rank configurations (e.g., rank-32) and exhibited more stable generalization across validation splits. Second, it required significantly fewer computational resources than higher ranks (32, 64), including reduced GPU memory usage and faster training times, making it more practical for large-scale fine-tuning. Finally, increasing the LoRA rank beyond this setting did not yield consistent performance gains, indicating diminishing returns relative to the added computational cost. Therefore, Set-8 was chosen as the optimal trade-off between performance, generalization, and efficiency, and it was used for all subsequent experiments and analyses with all other models.

Table 6. Ablation Study on LoRA Hyperparameters (r , α , Dropout) in ATD

Set	LoRA r	LoRA α	Dropout	Pr	Acc	F1
1	32	64	0	0.92	0.91	0.91
2	8	8	0	0.91	0.90	0.90
3	8	32	0.03	0.89	0.88	0.89
4	16	128	0	0.91	0.91	0.91
5	8	8	0	0.87	0.84	0.84
6	16	64	0	0.91	0.91	0.91
7	32	32	0	0.91	0.91	0.91
8	16	32	0	0.91	0.91	0.91
9	16	16	0.01	0.91	0.90	0.90
10	64	64	0.05	0.91	0.87	0.87

5.2. Error Analysis

As shown in the previous section, the instruction-tuned Gemma-3 4B model exhibits the best overall performance for Bengali text classification.

5.2.1. Quantitative Error Analysis

Figure 3 presents the confusion matrices of the proposed Gemma-3 4B model for five tasks.

For the SA task, the model achieved 138 true negatives (TN) and 281 true positives (TP), indicating strong performance. For the ATD task, it recorded 693 TN and 694 TP; for the FND task, it reached 367 TN and 2118 TP. These results demonstrate excellent performance across all three binary classification tasks. However, performance declined on the EC and NHC tasks with six classes, which made them more complex. In these tasks, the Gemma-3 4B model exhibited a noticeable decline in true positives and true negatives compared with its binary classification performance, highlighting its limitations in making fine-grained multiclass distinctions.

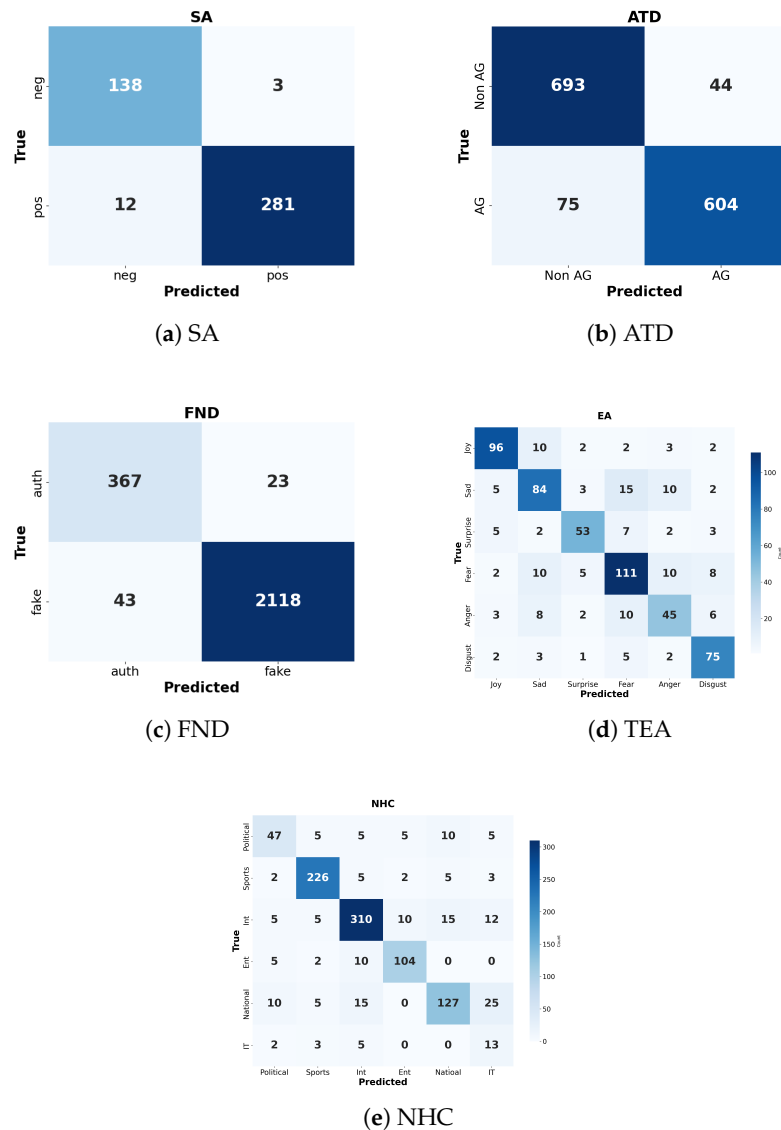


Figure 3. Confusion matrices of all five tasks for the proposed model.

5.2.2. Qualitative Error Analysis

Table 7 presents examples of both correctly and incorrectly classified predictions across the five tasks.

Table 7. Some correctly and incorrectly classified samples by Gemma-3-4B.

Text Samples	Actual	Predicted
[Bengali Text] (How does a publisher publish such a book.)	Negative	Negative
[Bengali Text] (Oh human! You create it yourself, and then worship it yourself.)	Aggressive	Aggressive
[Bengali Text] (Even after adopting Islamic ideology and circumcision, the distinguished poet, Hakimi physician, and new Muslim Farhad Mazhar did not escape criticism.)	Fake	Fake
[Bengali Text] (The country wasn't liberated for their rallies and protests: Hanif.)	Politics	National
[Bengali Text] (Riyad got angry at himself — what was the need to scare them? He could've just told the truth.)	Anger	Disgust

These predictions are generated by the Gemma-3 4B instruction-tuned version. In the first three samples, the model correctly identifies the class, demonstrating its effectiveness in clear, less-ambiguous cases. However, in the last two examples, the model fails to predict the correct label. The fourth sample, taken from the news headline categorization task, contains an inherently ambiguous headline that could reasonably belong to multiple categories. This ambiguity likely contributed to the model's misclassification. In the final example, from the emotion analysis task, the distinction between the predicted and actual class is subtle—even for humans. The text's emotional tone is nuanced, making it especially challenging for the model to interpret accurately. This highlights a common difficulty in fine-grained emotion classification: overlapping emotional cues can lead to confusion.

6. Conclusions

This paper presents a comprehensive evaluation of several LLMs, focusing on their instruction-tuned versions. Gemma-3 4B consistently outperformed other large language models, including Llama, Mistral, Qwen, and Phi, across most tasks, despite its relatively small size. Notably, the instruction-tuned Mistral-7B-it model achieved the best results in aggressive text detection. This study also compared prompting techniques, finding that instruction tuning generally improves performance, and that zero-shot prompting often yields better results than one-shot or chain-of-thought methods. Future work will focus on further extending evaluations to a broader range of domains and investigating more advanced instruction-tuning strategies. Additionally, the latest models, such as GPT-5 and Gemini-2.5 Flash, can be evaluated to assess their effectiveness in handling these tasks under in-context learning settings.

Limitations

Although the proposed method demonstrates satisfactory performance across various text classification tasks, several significant limitations remain unaddressed.

- The instruction-tuned Gemma-3 4B model achieves high accuracy on binary classification tasks; however, its performance declines on complex multiclass classification tasks.
- The model was adapted using Quantized Low-Rank Adaptation (QLoRA) rather than full fine-tuning, which may have potentially restricted task alignment and overall classification accuracy.
- Prompting strategies, including zero-shot, one-shot, and chain-of-thought approaches, yield inconsistent results across different tasks and model architectures.
- Several instruction-tuned models lack explicit training on Bengali-language data, which limits their ability to capture Bengali-specific linguistic features.
- This study relies on monolingual Bengali datasets and excludes Bangla-English code-mixed data, which is common in real-world usage.
- The research scope is restricted to classification tasks and does not investigate additional natural language processing (NLP) tasks, including text generation, summarization, or question answering.

Institutional Review Board Statement: This study used only publicly available, pre-existing datasets to evaluate five text classification tasks. No new human data were collected or annotated. All datasets were released for research under appropriate licenses, and all sources are correctly cited. Some tasks may contain sensitive or harmful language. To mitigate this risk, the datasets were used exclusively within their original research context, and dangerous content was not reproduced or promoted beyond what was necessary for model evaluation. All experiments adhered to responsible artificial intelligence research practices, emphasizing transparency, reproducibility, and fairness while minimizing potential misuse.

Data Availability Statement: The datasets and source code used in this study will be available at: <https://github.com/CUET-NLP-Lab/bengali-llm-multitask-classification>

Acknowledgments: This work was supported by the Directorate of Research & Extension (DRE), Chittagong University of Engineering & Technology (CUET), Chittagong, Bangladesh, under the Grant Number CUET/DRE/2023-2024/CSE/025.

Appendix A. Prompt Examples

In this task, we have explored three types of prompting. Tables A1, A2, and A3 illustrate all types of prompt templates used in this research. In Table A2, one-shot samples are provided. However, in this Table, we show only an example for one class. But while experimenting, we provided one example for each class.

Table A1. Prompt templates for zero-shot prompt across all the tasks

Task	Prompt
SA	Please classify the sentiment of this review. The answer should be either 1 (positive) or 0 (negative), based on the sentiment expressed. # Review: <review>
ATD	Please classify whether this Bengali sentence is Aggressive or non-aggressive. The answer should be either 1 (Aggressive) or 0 (Non-Aggressive), based on the sentence. # Sentence: <sentence>
FND	Please determine whether this Bengali news is genuine or not. The answer should be either 1 (Fake) or 0 (Authentic), based on the news. # News: <news>
NHC	Classify the following news headline into one of the predefined categories. Use the corresponding label number: politics: 0, sports: 1, international: 2, entertainment: 3, national: 4, IT: 5. # Headline: <headline> # Answer: [num]
TEA	Classify the emotion expressed in the following Bengali text into one of the predefined categories. Use the corresponding label number: 0: Joy, 1: Sadness, 2: Surprise, 3: Disgust, 4: Anger, 5: Fear. # Sentence: <sentence> # Answer: [label]

Table A2. Prompt templates for one-shot prompt across all the tasks

Task	Prompt
SA	Please classify the sentiment of this review. The answer should be either 1 (positive) or 0 (negative). Example: # Review: [Bengali Text] # Answer: 1 Now classify: # Review: <review>
ATD	Please classify whether this Bengali sentence is Aggressive or Non-Aggressive. The answer should be either 1 (Aggressive) or 0 (Non-Aggressive). Example: # Sentence: [Bengali Text] # Answer: 1 Now classify: # Sentence: <sentence>
FND	Please classify whether this Bengali news is Fake or Authentic. The answer should be either 1 (Fake) or 0 (Authentic). Example: # News: [Bengali Text] # Answer: 1 Now classify: # News: <news>
NHC	Classify the following news headline into one of the predefined categories. Use the corresponding label number: politics: 0, sports: 1, international: 2, entertainment: 3, national: 4, IT: 5. Example: # Headline: [Bengali Text] # Answer: 1 Now classify: # Headline: <headline>
TEA	Classify the emotion expressed in the following Bengali text into one of the predefined categories. Use the corresponding label number: 0: Joy, 1: Sadness, 2: Surprise, 3: Disgust, 4: Anger, 5: Fear. Example: # Sentence: [Bengali Text] # Answer: 0 Now classify: # Sentence: <sentence>

Table A3. Prompt templates for zero-shot CoT across all tasks with reasoning steps

Task	Prompt
SA	<p>Task: Classify the sentiment of the following Bangla review as 1 (positive) or 0 (negative). Steps: 1. Analyze Content: Break the review into phrases or sentences. 2. Identify Sentiment Indicators: Look for positive (e.g., “[Bengali Text]”) or negative (e.g., “[Bengali Text]”) words/phrases. 3. Evaluate Context: Consider how words are used, including sarcasm or mixed sentiment. 4. Determine Tone: Assess the overall tone based on key phrases. 5. Classify: Assign 1 for positive and 0 for negative. Now classify: # Review: <review></p>
ATD	<p>Task: Classify whether the following Bangla sentence is Aggressive (1) or Non-Aggressive (0). Steps: 1. Analyze Content: Break the sentence into meaningful parts. 2. Identify Aggression Indicators: Look for threatening or harmful expressions (e.g., “[Bengali Text]”). 3. Evaluate Intensity: Check the severity of the words used and their target. 4. Determine Tone: Assess whether the tone is hostile or neutral. 5. Classify: Assign 1 for Aggressive and 0 for Non-Aggressive. Now classify: # Sentence: <sentence></p>
FND	<p>Task: Classify whether the following Bangla news is Fake (1) or Authentic (0). Steps: 1. Analyze Content: Break the news statement into key claims. 2. Check Plausibility: Identify whether the claim sounds realistic or exaggerated. 3. Identify Unrealistic Elements: Look for impossible or illogical events (e.g., “[Bengali Text]”). 4. Evaluate Source-Like Tone: Assess if the sentence resembles factual reporting or satire. 5. Classify: Assign 1 for Fake and 0 for Authentic. Now classify: # News: <news></p>
NHC	<p>Task: Classify the following Bangla news headline into one of the predefined categories: politics: 0, sports: 1, international: 2, entertainment: 3, national: 4, IT: 5. Steps: 1. Analyze Content: Break the headline into key subjects and actions. 2. Identify Keywords: Detect topic-related words (e.g., “[Bengali Text]” → sports, “[Bengali Text]” → politics). 3. Match with Categories: Compare keywords with the predefined category list. 4. Resolve Ambiguity: If multiple categories fit, select the most dominant one. 5. Classify: Assign the corresponding label number. Now classify: # Headline: <headline></p>
TEA	<p>Task: Classify the emotion expressed in the following Bangla text into one of the predefined categories: 0: Joy, 1: Sadness, 2: Surprise, 3: Disgust, 4: Anger, 5: Fear. Steps: 1. Analyze Content: Break the sentence into emotion-bearing parts. 2. Identify Emotion Indicators: Look for explicit words or expressions (e.g., “[Bengali Text]” → Joy, “[Bengali Text]” → Anger). 3. Evaluate Context: Consider implied emotions or indirect expressions. 4. Determine Dominant Emotion: Choose the strongest emotion if multiple exist. 5. Classify: Assign the corresponding label number. Now classify: # Sentence: <sentence></p>

References

1. Bhowmick, A.; Jana, A. Sentiment analysis for Bengali using transformer based models. In Proceedings of the Proceedings of the 18th International Conference on Natural Language Processing (ICON), 2021, pp. 481–486.
2. Farhad, F.I.J.; Imran, S.; Santo, M.M.H.; Khan, M.; Sakib, A.; Rahman, M.S.; Islam, M.A.; Haque, R.; Rahman, S. Addressing Misinformation in Bengali Media: A Hybrid Deep Learning Solution. In Proceedings of the 2024 27th International Conference on Computer and Information Technology (ICCIT). IEEE, 2024, pp. 774–779.
3. Das, A.; Hoque, M.M.; Sharif, O.; Dewan, M.A.A.; Siddique, N. Temox: Classification of textual emotion using ensemble of transformers. *IEEE Access* **2023**, *11*, 109803–109818.
4. Rosni, T.R.; Hasan, M.; Mitra, T.; Ali, M.S.; Ferdous, M.H. Aggressive Bangla Text Detection Using Machine Learning and Deep Learning Algorithms. In Proceedings of the International Conference on Computation of Artificial Intelligence & Machine Learning. Springer, 2024, pp. 174–183.
5. Afroz, S.; Ahmed, K.; Hoque, M.M. Leveraging Multi-Task Learning for Detecting Aggression, Emotion, Violence, and Sentiment in Bengali Texts. In Proceedings of the 5th Muslims in ML Workshop co-located with NeurIPS 2025.
6. Ding, N.; Qin, Y.; Yang, G.; Wei, F.; Yang, Z.; Su, Y.; Hu, S.; Chen, Y.; Chan, C.M.; Chen, W.; et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature machine intelligence* **2023**, *5*, 220–235.
7. Wang, H.; Ren, C.; Yu, Z. Multimodal sentiment analysis based on multiple attention. *Engineering Applications of Artificial Intelligence* **2025**, *140*, 109731. <https://doi.org/https://doi.org/10.1016/j.engappai.2024.109731>.

8. Alex, N.; Lifland, E.; Tunstall, L.; Thakur, A.; Maham, P.; Riedel, C.J.; Hine, E.; Ashurst, C.; Sedille, P.; Carlier, A.; et al. RAFT: A real-world few-shot text classification benchmark. *arXiv preprint arXiv:2109.14076* **2021**.
9. Schick, T.; Schütze, H. True few-shot learning with Prompts—A real-world perspective. *Transactions of the Association for Computational Linguistics* **2022**, *10*, 716–731.
10. Loukas, L.; Stogiannidis, I.; Malakasiotis, P.; Vassos, S. Breaking the bank with ChatGPT: few-shot text classification for finance. *arXiv preprint arXiv:2308.14634* **2023**.
11. Wang, Z.; Pang, Y.; Lin, Y. Smart Expert System: Large Language Models as Text Classifiers. *arXiv preprint arXiv:2405.10523* **2024**.
12. Marreddy, M.; Oota, S.R.; Vakada, L.S.; Chinni, V.C.; Mamidi, R. Multi-task text classification using graph convolutional networks for large-scale low resource language. In Proceedings of the 2022 international joint conference on neural networks (IJCNN). IEEE, 2022, pp. 1–8.
13. Zhang, J.; Yan, K.; Mo, Y. Multi-task learning for sentiment analysis with hard-sharing and task recognition mechanisms. *Information* **2021**, *12*, 207.
14. Kapil, P.; Ekbal, A. A transformer based multi task learning approach to multimodal hate speech detection. *Natural Language Processing Journal* **2025**, *11*, 100133.
15. Singh, G.V.; Firdaus, M.; Chauhan, D.S.; Ekbal, A.; Bhattacharyya, P. Zero-shot multitask intent and emotion prediction from multimodal data: A benchmark study. *Neurocomputing* **2024**, *569*, 127128.
16. Kabir, M.; Laskar, M.T.R.; Nayeem, M.T.; Bari, M.S.; Hoque, E. Benllmeval: A comprehensive evaluation into the potentials and pitfalls of large language models on bengali nlp. *arXiv preprint arXiv:2309.13173* **2023**.
17. Hasan, M.A.; Das, S.; Anjum, A.; Alam, F.; Anjum, A.; Sarker, A.; Noori, S.R.H. Zero-and few-shot prompting with llms: A comparative study with fine-tuned models for bangla sentiment analysis. *arXiv preprint arXiv:2308.10783* **2023**.
18. Nazi, Z.A.; Hossain, M.R.; Mamun, F.A. 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* **2025**, *10*, 100124. <https://doi.org/https://doi.org/10.1016/j.nlp.2024.100124>.
19. Barua, A.; Sharif, O.; Hoque, M.M. Multi-class sports news categorization using machine learning techniques: resource creation and evaluation. *Procedia Computer Science* **2021**, *193*, 112–121.
20. Sharif, O.; Hoque, M.M.; Kayes, A.S.M.; Nowrozy, R.; Sarker, I.H. Detecting Suspicious Texts Using Machine Learning Techniques. *Applied Sciences* **2020**, *10*. <https://doi.org/10.3390/app10186527>.
21. Hossain, M.R.; Hoque, M.M.; Dewan, M.A.A.; Siddique, N.; Islam, M.N.; Sarker, I.H. Authorship Classification in a Resource Constraint Language Using Convolutional Neural Networks. *IEEE Access* **2021**, *9*, 100319–100338. <https://doi.org/10.1109/ACCESS.2021.3095967>.
22. Hider, M.A.; Ahsan, S.; Hossain, J.; Hoque, M.M. Emotion Classification in Bengali-English Code-Mixed Data using Transformers. In Proceedings of the 2024 27th International Conference on Computer and Information Technology (ICCIT), 2024, pp. 3529–3535. <https://doi.org/10.1109/ICCIT64611.2024.11022096>.
23. Ahsan, S.; Tasnia, F.; Tabassum, N.; Das, A.; Hoque, M.M.; Siddique, N. Classifying Textual Sentiment Using Bidirectional Encoder Representations from Transformers. In Proceedings of the 2023 26th International Conference on Computer and Information Technology (ICCIT). IEEE, 2023, pp. 1–6.
24. Aodhora, S.R.; Hoque, M.M. TeTeC: Technical Text Classification in Bengali using Ensemble of Transformers. In Proceedings of the 2024 International Conference on Recent Progresses in Science, Engineering and Technology (ICRPSET), 2024, pp. 1–6. <https://doi.org/10.1109/ICRPSET64863.2024.10955931>.
25. Akther, A.; Alam, K.M.; Debnath, R. Automatic detection of manipulated Bangla news: A new knowledge-driven approach. *Natural Language Processing Journal* **2025**, *11*, 100155. <https://doi.org/https://doi.org/10.1016/j.nlp.2025.100155>.
26. Hossain, M.R.; Hoque, M.M.; Dewan, M.A.A.; Hoque, E.; Siddique, N. AuthorNet: Leveraging attention-based early fusion of transformers for low-resource authorship attribution. *Expert Systems with Applications* **2025**, *262*, 125643. <https://doi.org/https://doi.org/10.1016/j.eswa.2024.125643>.
27. Sharif, O.; Hoque, M.M. Tackling cyber-aggression: Identification and fine-grained categorization of aggressive texts on social media using weighted ensemble of transformers. *Neurocomputing* **2022**, *490*, 462–481. <https://doi.org/https://doi.org/10.1016/j.neucom.2021.12.022>.
28. Hossain, E.; Sharif, O.; Moshikul Hoque, M. Sentiment polarity detection on bengali book reviews using multinomial naive bayes. In *Progress in Advanced Computing and Intelligent Engineering: Proceedings of ICACIE 2020*; Springer, 2021; pp. 281–292.

29. Sharif, O.; Hossain, E.; Hoque, M.M. M-bad: A multilabel dataset for detecting aggressive texts and their targets. In Proceedings of the Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations, 2022, pp. 75–85.
30. Hossain, M.Z.; Rahman, M.A.; Islam, M.S.; Kar, S. Banfakenews: A dataset for detecting fake news in bangla. *arXiv preprint arXiv:2004.08789* 2020.
31. Hossain, E. Bangla News Headlines Categorization. GitHub repository, 2023. Accessed: Jun. 27, 2025.
32. Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* 2024.
33. Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* 2025.
34. Abdin, M.; Aneja, J.; Awadalla, H.; Awadallah, A.; Awan, A.A.; Bach, N.; Bahree, A.; Bakhtiari, A.; Bao, J.; Behl, H.; et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219* 2024.
35. 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.
36. Team, G.; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786* 2025.
37. Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* 2025.
38. Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems* 2023, 36, 10088–10115.
39. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D.; et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 2022, 35, 24824–24837.
40. Ahmed, K.; Osama, M.; Sharif, O.; Hossain, E.; Hoque, M.M. Bennumeval: A benchmark to assess llms' numerical reasoning capabilities in bengali. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 17782–17799.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.