TinyLLM Efficacy in Low-Resource Language: An Experiment on Bangla Text Classification Task

Farhan Noor Dehan*, Md Fahim*, AKM Mahabubur Rahman, M Ashraful Amin, and Amin Ahsan Ali

Center for Computational & Data Sciences
Independent University, Bangladesh
Dhaka-1229, Bangladesh

Abstract. Delving into the realm of Bangla text analysis, our study ventures to unlock the potential of both Large and Tiny Language Models across a range of classification tasks, from deciphering sentiment to detecting sarcasm, emotion, hate speech, and fake news. In a linguistic landscape where resources are scarce, we fill a crucial gap by meticulously evaluating model performance. Our findings unveil Gemma-2B and Bangla-BERT as top performers, with Gemma-2B excelling in detecting hate speech and sarcasm, while BanglaBERT shines in sentiment analysis and emotion detection. Notably, TinyLlama emerges as a standout, showcasing exceptional prowess in fake news detection. We emphasize the importance of selecting models attuned to the intricacies of Bangla text, with Gemma-2B, TinyLlama, and BanglaBERT exhibiting notable accuracy improvements, surpassing other contenders. Furthermore, we uncover performance disparities influenced by dataset origins, with Bangla Language Models adept at capturing social media sentiments, and Large Language Models excelling in identifying misinformation and abusive language in formal sources. Our comparison with ChatGPT's zero-shot prompting underscores the necessity for advanced NLP methodologies. By spotlighting TinyLLM, we showcase the potential of advanced NLP in Bangla text classification, paving the way for broader advancements in NLP research.

Keywords: Bangla Language Models · Multilingual Language Models · Tiny Large Language Models.

1 Introduction

In the realm of NLP, the landscape of text classification has evolved significantly. Traditionally, conventional machine learning algorithms were the go-to for such tasks. However, the recent surge in transformer-based models, particularly large language models, has reshaped the field [2]. While these models have predominantly been prompt-based, their utility in languages such as Bangla has been limited due to resource constraints, including a scarcity of annotated datasets,

* These authors contributed Equally to this work.

linguistic resources, and computational infrastructure. Bangla, as a low-resource language, faces challenges in terms of data availability and linguistic resources necessary for effective NLP tasks. These limitations have made fine-tuning these models challenging in languages like Bangla. Fortunately, strides have been made with the development of smaller versions of these models, often termed "tiny" models, to broaden their accessibility and applicability, even across diverse domains [27]. Despite this progress, the exploration of these models in Bangla remains relatively under-explored, creating a notable gap in understanding their performance in Bangla text classification tasks [19].

To bridge this gap, our research endeavors to analyze the efficacy of various language models, including tiny ones, in the context of Bangla text classification tasks. Specifically, we target tasks like Sarcasm Detection [3], Hate Speech Detection [20], Bangla Fake News Detection [16], and others. Preliminary observations indicate that Tiny Large Language Models (TinyLLMs) consistently outperform existing Bangla language models(BLMs) and multilingual language models(MLMs) by substantial margins, ranging from 0.1% to 15% in most cases. By delving into these investigations, we aim to provide valuable insights into the performance of contemporary NLP models in Bangla, catering to the academic community's quest for knowledge in this domain. In this research endeavor, our contributions will encompass several key aspects:

- **Implementation of Tiny Language Models:** We implement and fine-tune tiny language models for different text classification tasks in the Bangla language. This involves adapting pre-existing models or training new ones from scratch to suit the specific linguistic nuances of Bangla.
- **Analysis of Model Performance:** We undertake thorough analyses to assess the performance of TinyLLMs in comparison to other state-of-the-art transformer models frequently employed in NLP tasks. Additionally, we evaluate these models using zero-shot prompting with ChatGPT, a state-of-the-art large language model.
- **Identification of Model Suitability:** Through rigorous experimentation and evaluation, we aim to identify the most suitable models for specific text classification tasks. This involves assessing factors such as model efficiency, robustness, and generalization capabilities.

By undertaking these endeavors, we seek to contribute to the advancement of NLP research in Bangla and facilitate the development of effective solutions for text classification tasks in this language. Our research outcomes have the potential to benefit a wide range of applications, including sentiment analysis, content moderation, and information retrieval, particularly in the context of Bangla-speaking communities. Additionally, the comparison with ChatGPT's relatively mediocre performance underscores the necessity for utilizing TinyLLMs for improved classification accuracy and effectiveness.

2 Related Works

In the domain of text classification, researchers have embarked on a journey to explore various machine and deep learning models, with pre-trained models gaining significant traction in recent years. Hasan et al. (2023)[12] delved into sentiment analysis, employing a range of machine learning models alongside fine-tuned options such as BanglaBERT and XLM-Roberta. Notably, they also incorporated ChatGPT for sentiment analysis using both zero-shot and multi-shot approaches. Bhattacharjee et al. (2022)[4] examined different iterations of BanglaBERT, comparing them with models like XLM-Roberta and mBERT for Bangla text analysis. However, despite this exploration, the impact of TinyLLMs has remained largely overlooked in these studies. Dehan et al. [7] investigated the performance of graph-based models for Bangla text classification. Fahim et al. [10] proposed a contextual neural stemmer for Bangla and its performance for Bangla text classification problems.

Alam et al. (2021)[1] conducted a benchmarking exercise on datasets collected from various platforms for nine NLP tasks using state-of-the-art transformer-based models. Their comparative analysis extended to monolingual versus multilingual models of varying sizes. Yet, the inclusion of Tiny LLMs in their evaluation was notably absent. Our research adopts a novel approach by broadening the scope of comparison to encompass TinyLLMs across a diverse array of tasks, including sentiment analysis, sarcasm detection, fake news detection, hate speech detection, and emotion detection. Additionally, we compare these models against a prominent large language model like ChatGPT. This comprehensive analysis aims to provide a deeper understanding of TinyLLMs performance across various datasets and tasks, while also shedding light on ChatGPT's efficacy in these domains.

In a similar vein, Kabir et al. (2023)[19] explored the application of various Large Language Models (LLMs) across a spectrum of tasks, including text classification. Their investigation incorporated zero-shot evaluation for ChatGPT, LLaMA-2, and Claude-2. However, the specific examination of TinyLLMs and LLMs was lacking, and a comprehensive analysis for each individual task was not provided. Thus, our research endeavors to fill this gap by focusing on text classification within the realm of Natural Language Understanding(NLU). Furthermore, we sought to assess ChatGPT's performance across these specific tasks.

Li et al. (2023)[21] addressed the challenges encountered by Large Language Models (LLMs) in handling low-resource languages like Bangla. Despite the potential of LLMs in NLP, their effectiveness in such languages has been limited. To tackle this issue, the authors proposed an innovative approach that integrates cross-lingual retrieval with in-context learning. By strategically utilizing prompts from languages with abundant resources that are semantically similar, they empowered Multilingual Pretrained Language Models (MPLMs), particularly emphasizing the generative model BLOOMZ, to enhance their performance on Bangla-related tasks. Their comprehensive evaluation showcased that

incorporating cross-lingual retrieval consistently improves MPLMs beyond their initial zero-shot performance.

Corrêa et al. (2024)[6], akin to ours, contributes to the trend of developing LLMs for low-resource contexts, with a focus on Brazilian Portuguese. They introduce the TeenyTinyLlama (TTL) models, aiming to democratize access to LLMs and foster open-source development, especially for languages facing resource constraints. However, no research has yet compared state-of-the-art transformer models with TinyLLMs. Our study aimed to examine the factors influencing the performance of these analyzed TinyLLMs and other models, thus contributing to a deeper understanding of their capabilities in text classification tasks.

3 Methodology

Our research methodology dives into examining both the esteemed TinyLLMs and prominent language models(LMs). We refined these models through two different approaches: fine-tuning LMs using conventional methods and fine-tuning TinyLLMs using LoRA and Peft techniques.

3.1 LM Fine-tuning

In our research, we utilize a LM, which we denote as $H = f_\theta(S)$, to process input sentences and extract contextual representations. Upon tokenizing an input sentence S, represented as $T = t_1, t_2, \ldots, t_n$, the LM generates contextual representations for each token by applying the function $f_\theta(S)$, resulting in a sequence denoted as $H = h_1, h_2, \ldots, h_n$. These representations encapsulate the unique meaning of each token within the context of the entire sentence.

However, for tasks such as classification, where a fixed-size representation of the entire sentence is required, we employ a two-layer Feed Forward Neural Network (FFN) on the contextual representation of [CLS] token, h_{CLS}. This network utilizes weight matrices W_1 and W_2, bias terms b_1 and b_2, and the Rectified Linear Unit (ReLU) activation function to process h_{CLS} and generate a fixed-size representation z.

$$z = W_2 \cdot (\text{ReLU}(W_1 \cdot h_{[\text{CLS}]} + b_1)) + b_2 \tag{1}$$

3.2 TinyLLM Fine-tuning using LoRA and FEFT

Traditional fine-tuning of large language models (LLMs) involves significantly modifying the pre-trained model's parameters, which can be computationally expensive and time-consuming. PEFT (Parameter-Efficient Fine-Tuning) [22] offers a solution by adapting pre-trained models to new tasks with minimal changes to the original parameters. This significantly reduces training time and memory usage compared to traditional approaches. LoRA (Low-Rank Adaptation) [17] is a specific PEFT technique that introduces a more efficient way to

capture the adjustments needed for fine-tuning. Instead of directly modifying all the pre-trained parameters, LoRA utilizes a low-rank matrix. This matrix requires significantly fewer parameters to represent the task-specific adaptations, leading to substantial efficiency gains.

Let's denote the original pre-trained model parameters as W which will be frozen during training. LoRA introduces a low-rank update, denoted by ΔW, which captures the task-specific adjustments needed for fine-tuning. This low-rank update is further decomposed as the product of two trainable matrices, A and B: $\Delta W = A \times B^T$. Here, A with a shape of $d \times r$ and B with a shape of $r \times d$ have a much lower rank (denoted by r) compared to the original dimension d of the parameter matrix W. This means they require significantly fewer parameters to represent the necessary adjustments. The rows of matrix A and the columns of matrix B can be interpreted as capturing the task-specific adaptations applied to the original weight matrix W. Finally, the updated weight metrics W' with LoRA is the summation of pretrained frozen metrics W and task-specific fine-tuned metrics ΔW

$$W' = W + \Delta W = W + AB^T$$

In essence, LoRA leverages a more compact representation (the low-rank matrices A and B) to achieve fine-tuning, resulting in significant efficiency improvements compared to traditional fine-tuning methods that modify all the pre-trained parameters directly.

3.3 Experimented Models

Experimented LMs: For fine-tuning, two different types of LM models were considered i. Bangla LM and Multilingual LM

i. Bangla LM: Our investigation delves into the renowned **BanglaBERT** and its variants, acclaimed for their effectiveness in text classification tasks, utilizing contextual embeddings from meticulous multi-stage training on Bangla corpora, crucial for our study's objectives [4, 24].

ii. Multilingual LM: We also analyzed the fine-tuning performance of the multilingual language model for solving Bangla text classification tasks. In this experiment, we considered, XLM-RoBERTa[5], mBERT [8], mDeBERTa [14], and mDeBERTa-V3 [13].

Experimented TinyLLMs: In our pursuit of computational efficiency without compromising performance, we delve into the realm of TinyLLMs, exploring:

i. Gemma-2B: Gemma-2B, Google's lightweight, decoder-only language model, derived from Gemini, are versatile for text generation tasks like QA and summarization, trained on 2B parameters, enabling deployment in resource-constrained environments [25]. Gemma 2B's standout feature is its dynamic sparse attention, which efficiently allocates resources to the most relevant parts of the input, enhancing overall performance. Its modular architecture also allows for flexible scaling, adapting to different task complexities seamlessly.

ii. TinyLlama: TinyLlama, versatile and compact, trained on 1.1B parameters, ensures compatibility and ease of adoption for diverse applications [27]. TinyLlama stands out for its incredibly compact design that delivers strong language understanding while using minimal resources. Its innovative layer normalization techniques ensure that performance remains robust even with limited computational power.

iii. Falcon-1.3B: Falcon, a series of causal decoder-only models trained on 1.3B parameters, emphasizes computational efficiency with features like multi-query attention and support for efficient attention variants [23]. The Falcon-1.3B excels with its efficient use of flash attention, enabling it to achieve high performance despite its smaller size. It also integrates advanced gradient checkpointing, which optimizes memory usage during training and inference.

iv. OPT-1.3B: OPT-1.3B, utilizing causal language modeling and trained on 1.3B parameters, adeptly captures comprehensive linguistic patterns[28]. OPT-1.3B is remarkable for its open, pre-trained transformer framework, designed for easy customization and fine-tuning, all while maintaining a lean and efficient model. Additionally, its adaptive learning rate scheduler helps in fine-tuning across diverse datasets with improved stability.

4 Experiment Setup

In this study, we deployed multiple model configurations for a thorough analysis and evaluation.

4.1 Dataset

We employed five unique datasets, each designed for specific tasks including sentiment analysis, sarcasm detection, fake news detection, hate speech analysis, and emotion detection.

- **SentNoB:** A dataset comprising approximately 15k Bengali comments from diverse social media platforms across 13 domains. These comments are annotated with positive, negative, or neutral sentiments. The dataset is partitioned into roughly 13k training samples and 1.5k testing and validation samples, presenting challenges due to its noisy nature [18].

- **Bangla Sarcasm Detection Dataset:** This dataset consists of over 5k comments sourced from social media, encompassing 3k non-sarcastic and 2k sarcastic comments [3].

- **BanFakeNews:** An annotated dataset of approximately 50,000 news articles, useful for developing automated fake news detection systems. It consists of around 48,000 authentic news articles and 1,000 fabricated ones [16].

- **Hate Speech Dataset:** This dataset contains approximately 3k training samples and 1k testing samples, covering various forms of hate speech across

different contexts, categorized into political, personal, gender-abusive, geopolitical, and religious hate [20].

– **YouTube Comments Emotion:** An emotion dataset containing around 3k samples with 5 classes representing different emotions, such as anger/disgust, fear/surprise, joy, sadness, and none. These samples are extracted from Bangla videos on YouTube [26].

4.2 Preprocessing and Experiment Setup

The preprocessing and experiment setup for training are discussed in detail in this section. Preprocessing steps included normalizing the text using a normalizer. We use BUET-NLP normalizer [1] in our experiment.

We use the Pytroch deep learning framework for modeling and the HuggingFace library for the pre-trained models. For LM models, we employed the AdamW optimizer with a learning rate of 1×10^{-5}, a number of epochs of 10, and a batch size of 16. Dropout regularization was applied to prevent overfitting with dropout_rate = 0.1. The hyper-parameters were chosen based on papers [9, 11]

In experiments involving TinyLLMs, we established a computational environment using specialized packages like peft, bitsandbytes, and accelerate. We utilized diverse TinyLLMs variants such as falcon-1.3b, TinyLlama-1.1b, opt-1.3b and gemma-2b. For these models, we employed the AdamW optimizer with a learning rate of 2e-5 and a weight decay of 0.01. The value of r = 64, LoRA_ALPHA = 32, and LoRA_DROPOUT = 0.1. LoRA was applied to the ***all-linear*** layer of the TinyLLM. In TinyLLM experiment, models were trained for 5 epochs, with batch sizes of 2, 4, and 8, depending on the dataset size.

Tokenization was performed using Huggingface AutoTokenizer, and fine-tuning was carried out using the Huggingface Trainer module. All experiments were conducted on a single Nvidia Tesla P100 GPU.

4.3 Performance Metrics

When assessing the effectiveness of language models, several key performance metrics are relied upon to provide important insights into their performance. In our evaluation, we have focused on five widely-used metrics to gain a comprehensive understanding of the model's performance.

Accuracy Accuracy measures the proportion of correctly classified instances among all instances. It is calculated by dividing the sum of true positives (correctly predicted positive instances) and true negatives (correctly predicted negative instances) by the total number of instances.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

[1] https://github.com/csebuetnlp/normalizer

Precision and Recall Precision measures the proportion of true positive instances among the instances predicted as positive, and recall measures the proportion of true positive instances that were correctly predicted out of all actual positive instances. The calculations are as follows:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN} \tag{3}$$

Macro and Weighted F1 Scores The F1 score is the harmonic mean of precision and recall. In the macro F1 score, each class is given equal weight, and the mean of these F1 scores across all classes is calculated. Weighted F1 score, on the other hand, considers the class distribution by assigning weights to each class based on their frequency in the dataset.

Macro F1 Score:

$$F1_{macro} = \frac{1}{N} \sum_{i=1}^{N} F1_i \tag{4}$$

Weighted F1 Score:

$$F1_{weighted} = \frac{\sum_{i=1}^{N} w_i \times F1_i}{\sum_{i=1}^{N} w_i} \tag{5}$$

Where N is the number of classes, w_i is the weight for class i, and $F1_i$ is the F1 score for class i.

5 Result Analysis

Through rigorous experimentation, we analyzed the performance of diverse language models on Bangla text classification datasets, revealing insights into their strengths and limitations across BLMs, MLMs, TinyLLMs, and ChatGPT, with efficacy varying based on task and dataset features.

5.1 Bangla Language Models

The performance analysis across different Bangla text classification datasets in Table 1 indicates variations in model efficacy. BanglaBERT consistently outperforms BanglaBERT-Large and BanglaBERT (Sagor Sarker) across most datasets. Notably, BanglaBERT demonstrates superior accuracy and F1 scores in Sent-NoB, Sarcasm Detection, Hate Speech Detection, and Emotion Detection datasets, achieving an average improvement of approximately $1 - 3\%$ in accuracy and F1 scores over BanglaBERT-Large.

In Hate Speech Detection, while BanglaBERT-Large surpasses BanglaBERT in weighted F1 score, accuracy, and macro F1 score by approximately $1 - 3\%$ respectively. BanglaBERT and BanglaBERT-Large also outperform BanglaBERT

Table 1. Performance Comparison of Bangla Language Models (BLMs) on Bangla Text Classification Datasets: This table displays performance metrics, including accuracy, macro F1, and weighted F1 scores, for various Bangla Language Models evaluated across different Bangla text classification datasets. BanglaBERT emerges as the top performer across most datasets, surpassing other evaluated models.

Dataset	**Model**	**Performance Metrics**		
		Accuracy	Macro F1	Weighted F1
SentNoB	BanglaBERT	**74.46**	**69.55**	**73.03**
	BanglaBERT-Large	72.82	68.87	72.05
	BanglaBERT(Sagor Sarker)	69.42	64.54	68.01
Sarcasm Detection	BanglaBERT	**95.67**	**95.51**	**95.68**
	BanglaBERT-Large	94.55	94.23	94.50
	BanglaBERT(Sagor Sarker)	90.46	90.13	90.48
HateSpeech Detection	BanglaBERT	**69.33**	41.65	65.41
	BanglaBERT-Large	66.11	58.59	**66.96**
	BanglaBERT(Sagor Sarker)	67.11	**61.43**	66.81
BanFakeNews	BanglaBERT	96.65	92.99	96.51
	BanglaBERT-Large	**97.51**	**94.69**	**97.43**
	BanglaBERT(Sagor Sarker)	96.15	91.76	96.03
Emotion Detection	BanglaBERT	**70.78**	41.26	**65.52**
	BanglaBERT-Large	68.07	**42.87**	65.08
	BanglaBERT(Sagor Sarker)	63.86	40.10	61.09

(Sagor Sarker) consistently across all datasets. These results suggest that BanglaBERT offers notable advantages over both BanglaBERT-Large and BanglaBERT (Sagor Sarker) across various Bangla text classification tasks, while BanglaBERT-Large outperforms in certain cases. The reason for BanglaBERT's superior performance lies in its enhanced ability to grasp both semantic and syntactic contexts effectively.

5.2 Multilingual Language Models

The performance of various MLMs across different Bangla text classification datasets is summarized in Table 2. In general, XLM-Roberta consistently outperforms other MLMs across most datasets. Specifically, in SentNoB, XLM-Roberta achieves the highest accuracy, macro F1 score, and weighted F1 score, surpassing other MLMs by approximately $2-9\%$, indicating a significant margin of improvement.These results indicate that XLM-Roberta consistently provides superior performance compared to other Multilingual Language Models across

various Bangla text classification tasks. XLM-RoBERTa exhibits superior performance compared to other multilingual models due to its advanced architecture and optimized training methodology, enabling it to capture a broader range of linguistic nuances across various languages.

Table 2. Comparative Performance of Multilingual Language Models (MLMs) on Various Bangla Text Classification Datasets: This table presents performance metrics, including accuracy, macro F1, and weighted F1 scores, for different Multilingual Language Models across several Bangla text classification datasets. The evaluated models include XLM-Roberta, M-BERT, M-deBerta, and M-deBerta-V3.

Dataset	Model	Accuracy	Macro F1	Weighted F1
SentNoB	XLM-Roberta	**70.37**	**67.94**	**70.67**
	M-BERT	67.97	65.21	68.13
	M-deBerta	60.72	55.23	58.91
	M-deBerta-V3	67.28	63.80	66.75
Sarcasm Detection	XLM-Roberta	**93.60**	**93.30**	**93.30**
	M-BERT	88.68	87.92	88.52
	M-deBerta	90.22	89.90	90.25
	M-deBerta-V3	90.63	90.01	90.50
Hate Speech Detection	XLM-Roberta	**69.44**	**62.19**	**67.95**
	M-BERT	66.22	60.65	66.09
	M-deBerta	55.22	40.95	53.05
	M-deBerta-V3	60.00	41.54	58.21
BanFakeNews	XLM-Roberta	**97.65**	**94.96**	**97.57**
	M-BERT	89.27	88.81	89.26
	M-deBerta	91.95	81.74	91.43
	M-deBerta-V3	92.98	86.43	93.12
Emotion Detection	XLM-Roberta	**67.77**	**41.39**	**63.71**
	M-BERT	59.64	34.04	55.53
	M-deBerta	51.20	24.48	44.03
	M-deBerta-V3	56.63	29.83	51.74

5.3 Tiny Large Language Models

The performance analysis of TinyLLMs across various Bangla text classification datasets is presented in Table 3. Each TinyLLM was trained on a minimum of approximately 21 billion training tokens per 1 billion parameters for Bangla text [15]. Gemma-2B consistently outperforms other TinyLLMs in terms of accuracy, macro F1 score, and weighted F1 score across all datasets. In datasets such as SentNoB, Sarcasm Detection, Hate Speech Detection, and Emotion Detection, Gemma-2B achieves the highest accuracy, macro F1 score, and weighted F1 score, outperforming TinyLlama by approximately $0.50 - 9\%$ respectively. Falcon-1.3B and Opt-1.3B demonstrate comparatively lower performance metrics.

Table 3. Comparative Performance of Tiny Large Language Models Across Diverse Bangla Text Classification Tasks: This table highlights accuracy, macro F1, and weighted F1 scores of various models, encompassing tasks like sentiment analysis, sarcasm detection, hate speech identification, fake news detection, and emotion detection.

Dataset	Model	Performance Metrics		
		Accuracy	Macro F1	Weighted F1
SentNoB	Gemma-2B	**66.90**	**63.02**	**66.06**
	TinyLlama	66.02	58.93	63.38
	Falcon-1.3B	58.83	46.82	52.89
	Opt-1.3B	63.18	58.13	61.70
Sarcasm Detection	Gemma-2B	**96.86**	**96.72**	**96.85**
	TinyLlama	94.13	93.87	94.12
	Falcon-1.3B	80.26	77.64	79.14
	Opt-1.3B	92.41	92.14	92.43
HateSpeech Detection	Gemma-2B	**70.89**	**63.08**	**70.30**
	TinyLlama	67.78	54.60	66.13
	Falcon-1.3B	53.56	35.43	50.51
	Opt-1.3B	56.44	32.21	51.78
BanFakeNews	Gemma-2B	**97.83**	95.50	97.80
	TinyLlama	**97.83**	**95.54**	**97.81**
	Falcon-1.3B	95.26	90.98	95.39
	Opt-1.3B	92.55	84.01	92.31
Emotion Detection	Gemma-2B	**62.65**	**36.92**	**58.62**
	TinyLlama	57.83	32.50	53.25
	Falcon-1.3B	49.10	17.45	36.22
	Opt-1.3B	48.49	15.63	34.43

However, for BanFakeNews, both Gemma-2B and TinyLlama demonstrate comparable accuracy, with TinyLlama outperforming in terms of macro F1 score and weighted F1 score. Falcon-1.3B and Opt-1.3B again fall behind in performance across all metrics. Overall, Gemma-2B consistently demonstrates superior performance across all datasets, highlighting its efficacy as a Large Language Model for Bangla text classification tasks. The improved efficacy demonstrated by Gemma-2B and TinyLlama in processing Bangla text could be attributed to their adept utilization of specialized knowledge tailored to the task at hand.

5.4 Evaluating ChatGPT's Zero-shot Prompting Performance

The evaluation of ChatGPT 3.5 Turbo's zero-shot prompting for Bangla text classification is outlined in Table 4. For this experiment, we looked at different tasks and categories within each dataset. These tasks involved analyzing sentiment, spotting fake news, detecting hate speech, identifying sarcasm, and recog-

Table 4. The performance of Zero-shot Prompting with ChatGPT across diverse Bangla text classification datasets is evaluated in the table, comparing test labels against each label generated by the prompt.

Dataset	Precision	Recall	Macro F1
SentNoB	56.28	49.97	44.85
Sarcasm Detection	62.17	57.59	48.65
Hate Speech Detection	54.65	50.42	46.22
BanFakeNews	46.35	48.92	46.67
Emotion Detection	39.58	37.86	33.09

nizing emotions. The results suggest moderate performance across diverse classification tasks. Notably, the model demonstrates superior precision and recall in Sarcasm Detection compared to other tasks. However, a noticeable decrease is evident in Emotion Detection, indicating potential constraints in grasping nuanced emotional nuances, while showing relatively better comprehension of sarcasm. Addressing these challenges may require exploring alternative prompting techniques and fine-tuning approaches to improve task-specific performance. We revised the prompt design based on Kabir et al.(2023)[19] approach to enhance its efficiency. The subsequent illustration exemplifies the prompts employed within this study:

> For the given Input [INPUT]. Now, classify the text for [TASK]. Your output should be in between *class1,class2, . . . ,class n*. Write only your response, nothing else. Don't add anything before and after your response.

6 Findings

The performance analysis presented in Table 5 underscores the varying effectiveness of models across Bangla text classification datasets. BanglaBERT showcases superior performance SentNoB and Emotion Detection tasks, outperforming other models. Nevertheless, ChatGPT's performance appears to be notably less impressive, with accuracies falling behind by substantial margins. Gemma-2B and TinyLlama exhibits superior performance in Sarcasm Detection, Hate-Speech and BanFakeNews datasets.

In this study, various language models, including Bangla Language Models, Multilingual Language Models, and Large Language Models, were fine-tuned and evaluated across distinct datasets sourced from diverse online platforms. Notably, findings from tables 1, 2, 3, 4 and 5 reveal that Bangla Language Models exhibited superior performance when tasked with datasets originating from social media platforms such as YouTube, particularly those associated with sentiment analysis and emotion recognition. Conversely, Large Language Models demonstrated exceptional efficacy when confronted with datasets sourced from formal sources like newspapers or online articles, notably excelling in tasks such

Table 5. The table presents a performance comparison of top models across various Bangla text classification datasets, evaluating and contrasting the effectiveness of the best-performing models from BLM, MLM, TinyLLM, and ChatGPT classifications.

Dataset	Model	Performance Metrics		
		Accuracy	Macro F1	Weighted F1
SentNoB	Gemma-2B	66.90	63.02	66.06
	XLM-Roberta	70.37	67.94	70.67
	BanglaBERT	**74.46**	**69.55**	**73.03**
	ChatGPT(Zero-shot)	56.31	44.85	50.56
Sarcasm Detection	Gemma-2B	**96.86**	**96.72**	**96.85**
	XLM-Roberta	93.60	93.30	93.30
	BanglaBERT	95.67	95.51	95.68
	ChatGPT(Zero-shot)	51.10	48.65	46.46
HateSpeech Detection	Gemma-2B	**70.89**	**63.08**	**70.30**
	XLM-Roberta	69.44	62.19	67.95
	BanglaBERT	69.33	41.65	65.41
	ChatGPT(Zero-shot)	49.67	46.22	49.27
BanFakeNews	TinyLlama	**97.83**	**95.54**	**97.81**
	XLM-Roberta	97.65	94.96	97.57
	BanglaBERT-Large	97.51	94.69	97.43
	ChatGPT(Zero-shot)	82.25	46.67	77.60
Emotion Detection	Gemma-2B	62.65	36.92	58.62
	XLM-Roberta	67.77	**41.39**	63.71
	BanglaBERT	**70.78**	41.26	**65.52**
	ChatGPT(Zero-shot)	44.88	33.09	43.06

as sarcasm detection, fake news detection, and hate speech identification.

These findings indicate that different language models have varying strengths depending on the dataset's nature and origin. BLMs are sensitive to nuances in sentiment, and emotions prevalent in user-generated content on social media. In contrast, TinyLLMs are proficient in identifying patterns of misinformation and abusive language in structured, formal sources. The study highlights the significance of dataset characteristics in influencing model performance. Social media discourse, with its complex linguistic phenomena, poses challenges for TinyLLMs, resulting in lower performance compared to models fine-tuned on datasets tailored to such complexities. Conversely, the structured nature of formal text sources aligns well with the capabilities of TinyLLMs, leading to higher accuracy in tasks involving misinformation, sarcasm, and hate speech detection.

7 Conclusion

The examination of diverse language models in Bangla text classification tasks provides valuable insights into their effectiveness and applicability. Gemma-2B consistently excels in tasks like sarcasm detection and hate speech identification,

showcasing its reliability and versatility. Conversely, TinyLlama stands out in fake news detection, underscoring the efficacy of specialized models in capturing subtle nuances within Bangla text. BanglaBERT demonstrated exceptional performance in the remaining selected tasks. When comparing TinyLLM's results with those of multilingual models such as XLM-Roberta and language-specific models like BanglaBERT, competitive outcomes were observed across various tasks. BLM's excel in capturing sentiment and emotions from social media, while TinyLLM's demonstrate superior capabilities in detecting sarcasm, hate speech, and fake news from formal sources.

Selecting the most suitable language model depends on factors like the task, dataset characteristics, and linguistic nuances. While Gemma-2B and TinyLlama demonstrate robust performance, XLM-Roberta, and BanglaBERT also yield commendable results. These findings offer insights for employing language models in Bangla text classification, aiding the development of accurate NLP solutions. Ongoing research is crucial to refine language models for enhanced performance and applicability in real-world scenarios.

Future Work: In our study on Bangla text classification, we faced challenges including limited annotated datasets, computational resource constraints, and potential biases in dataset characteristics. Future research could focus on expanding annotated datasets, optimizing Bangla language models, and exploring new architectures. Further analysis in specific domains, improvements in evaluation metrics, and addressing ethical concerns are also crucial. Deploying models in real-world applications and conducting user studies would provide insights into usability and effectiveness, driving further progress in the field.

Acknowledgments

We are thankful to Independent University, Bangladesh, for their support of this project. We would also like to express our gratitude to the Center for Computational & Data Sciences (CCDS Lab) for providing computational facilities and supervising this project.

References

1. Alam, F., Hasan, A., Alam, T., Khan, A., Tajrin, J., Khan, N., Chowdhury, S.A.: A review of bangla natural language processing tasks and the utility of transformer models. arXiv preprint arXiv:2107.03844 (2021)
2. Alam, T., Khan, A., Alam, F.: Bangla text classification using transformers. CoRR **abs/2011.04446** (2020), https://arxiv.org/abs/2011.04446
3. Apon, T.S., Anan, R., Modhu, E.A., Suter, A., Sneha, I.J., Alam, M.G.R.: Banglasarc: A dataset for sarcasm detection. In: 2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE). pp. 1–5. IEEE (2022)

4. Bhattacharjee, A., Hasan, T., Ahmad, W., Mubasshir, K.S., Islam, M.S., Iqbal, A., Rahman, M.S., Shahriyar, R.: BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In: Carpuat, M., de Marneffe, M.C., Meza Ruiz, I.V. (eds.) Findings of the Association for Computational Linguistics: NAACL 2022. pp. 1318–1327. Association for Computational Linguistics, Seattle, United States (Jul 2022). https://doi.org/10.18653/v1/2022.findings-naacl.98, https://aclanthology.org/2022.findings-naacl.98
5. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.747, https://aclanthology.org/2020.acl-main.747
6. Corrêa, N.K., Falk, S., Fatimah, S., Sen, A., de Oliveira, N.: Teenytinyllama: open-source tiny language models trained in brazilian portuguese (2024)
7. Dehan, F., Fahim, M., Ali, A.A., Amin, M.A., Rahman, A.: Investigating the effectiveness of graph-based algorithm for bangla text classification. In: Proceedings of the First Workshop on Bangla Language Processing (BLP-2023). pp. 104–116 (2023)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) 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. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423
9. Fahim, M.: Aambela at blp-2023 task 2: Enhancing banglabert performance for bangla sentiment analysis task with in task pretraining and adversarial weight perturbation. In: Proceedings of the First Workshop on Bangla Language Processing (BLP-2023). pp. 317–323 (2023)
10. Fahim, M., Ali, A.A., Amin, M.A., Rahman, A.: Contextual bangla neural stemmer: Finding contextualized root word representations for bangla words. In: Proceedings of the First Workshop on Bangla Language Processing (BLP-2023). pp. 94–103 (2023)
11. Fahim, M., Ali, A.A., Amin, M.A., Rahman, A.M.: Edal: Entropy based dynamic attention loss for hatespeech classification. In: Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation. pp. 775–785 (2023)
12. 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)
13. He, P., Gao, J., Chen, W.: Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing (2021)
14. He, P., Liu, X., Gao, J., Chen, W.: Deberta: Decoding-enhanced bert with disentangled attention. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=XPZIaotutsD
15. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L.A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan,

K., Elsen, E., Rae, J.W., Vinyals, O., Sifre, L.: Training compute-optimal large language models (2022)
16. Hossain, M.Z., Rahman, M.A., Islam, M.S., Kar, S.: BanFakeNews: A dataset for detecting fake news in Bangla. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 2862–2871. European Language Resources Association, Marseille, France (May 2020), https://aclanthology.org/2020.lrec-1.349
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021)
18. Islam, K.I., Kar, S., Islam, M.S., Amin, M.R.: Sentnob: A dataset for analysing sentiment on noisy bangla texts. In: Findings of the Association for Computational Linguistics: EMNLP 2021. pp. 3265–3271 (2021)
19. Kabir, M., Islam, M.S., 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)
20. Karim, M.R., Chakravarthi, B.R., Arcan, M., McCrae, J.P., Cochez, M.: Classification benchmarks for under-resourced bengali language based on multichannel convolutional-lstm network. 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA) pp. 390–399 (2020), https://api.semanticscholar.org/CorpusID:215786049
21. Li, X., Nie, E., Liang, S.: Crosslingual retrieval augmented in-context learning for Bangla. In: Alam, F., Kar, S., Chowdhury, S.A., Sadeque, F., Amin, R. (eds.) Proceedings of the First Workshop on Bangla Language Processing (BLP-2023). pp. 136–151. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.banglalp-1.15, https://aclanthology.org/2023.banglalp-1.15
22. Mangrulkar, S., Gugger, S., Debut, L., Belkada, Y., Paul, S., Bossan, B.: Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft (2022)
23. Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., Launay, J.: The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. arXiv preprint arXiv:2306.01116 (2023), https://arxiv.org/abs/2306.01116
24. Sarker, S.: Banglabert: Bengali mask language model for bengali language understanding (2020), https://github.com/sagorbrur/bangla-bert
25. Team, G.: Gemma: Open models based on gemini research and technology (2024)
26. Trinto, N.I., Ali, M.E.: Detecting multilabel sentiment and emotions from bangla youtube comments. 2018 International Conference on Bangla Speech and Language Processing (ICBSLP) pp. 1–6 (2018), https://api.semanticscholar.org/CorpusID:54440144
27. Zhang, P., Zeng, G., Wang, T., Lu, W.: Tinyllama: An open-source small language model (2024)
28. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P.S., Sridhar, A., Wang, T., Zettlemoyer, L.: Opt: Open pre-trained transformer language models (2022)