How multilingual LLMs are! A case study of LLMs using multilingual Sentiment Analysis in Indian Language

Saurabh Kumar[1], Shifali Agrahari[1], Sanasam Ranbir Singh1[1], Arpit Saikia[2], Rhitwik Choudhury[3], and Anubhav Bhattacharya[4]

[1] Department of CSE, Indian Institute of Technology Guwahati, Assam, India
[2] Department of CSE, National Institute of Technology, Silchar, Assam, India
[3] Department of EE, National Institute of Technology, Silchar, Assam, India
[4] Department of CSE, Rajiv Gandhi Institute of Petroleum Technology, Jais, Uttar Pradesh, India
{saurabh1003,a.shifali,ranbir}@iitg.ac.in

Abstract. Large Language Models(LLMs) are computational models known for their ability to achieve general-purpose language understanding and generate text by training the model's billion parameters on massive amounts of text data. Multilingual LLMs are language models that support multiple languages and can understand and generate text in different languages. In the current scenario, there is a lack of Multilingual LLMs that have dedicated support for Indian languages. This paper has tried to assess the current performance of LLMs in Indian languages by hand-picking a few multilingual LLMs (like XLM-R, mBERT, MuRIL, Navarasa, IndicBERT v2, mT5) that support these Indian languages and fine-tuning them using the technique of prompt tuning and few-shot learning for multilingual Sentiment Analysis. Sentiment Analysis involves processing text to determine whether the emotional tone of the message is positive, negative, or neutral. It requires a model to understand and interpret the context, tone, and implicit meaning of text, which provides a robust test of the model's natural language understanding capabilities.

Keywords: LLMs · multilingual LLMs · sentiment analysis · prompt tuning · accuracy.

1 Introduction

The rise of multilingual model architectures has transformed natural language processing (NLP). This transformation has been spearheaded by the introduction of multilingual pre-trained language models such as mBERT [7], mT5 [19]. These models harness extensive unsupervised textual data from various languages, enabling zero-shot and few-shot cross-lingual transfer for a wide range of downstream NLP tasks. However, their performance can vary significantly depending on the specific task and language pair [10].

Alongside these developments, large language models (LLMs) like GPT-3 [5] have garnered significant attention. These LLMs are known for their ability to achieve general-purpose language understanding and generate text by training billions of parameters on massive amounts of textual data. Studies have highlighted the strong performance of LLMs in few-shot in-context learning, particularly for public English sentiment analysis tasks [20]. Despite being predominantly pre-trained on English corpora, some research has revealed fascinating multilingual capabilities in both public and proprietary LLMs [14].

However, the effectiveness of these multilingual LLMs in handling Indian languages, which vary significantly in syntax, script, and cultural context, remains underexplored. Our study aims to evaluate the performance of LLMs in Indian languages by using sentiment analysis as a case study. By examining these models in the context of India's linguistic diversity, we seek to provide insights into their strengths and limitations, identify areas for improvement, and suggest directions for future research. Ultimately, this will contribute to the development of more inclusive and accurate NLP technologies.

This study evaluates the performance of several multilingual LLMs, including XLM-R, mBERT, MuRIL, Navarasa, and IndicBERT v2, specifically for sentiment analysis in Indian languages. Our methodology involves fine-tuning these models using techniques like prompt tuning and few-shot learning. By highlighting the effectiveness of these models and identifying potential areas for improvement, we aim to advance the development of multilingual LLMs that are more inclusive of Indian languages, enhancing their utility in diverse linguistic contexts.

To evaluate the LLMs effectively, we have employed a combination of few-shot learning, fine-tuning, and prompt-based tuning. Specifically, we have tested these models with varying proportions of training data: 1%, 2%, and 10%. Few-shot learning enables the models to perform well with limited examples, while fine-tuning and prompt-based tuning allow us to adapt the models more closely to the specific nuances of Indian languages. This approach will help us assess how these models handle sentiment analysis with minimal training data and determine the most effective strategies for enhancing their performance.

2 Literature Review

Any LLMs analysis like sentiment analysis faces unique challenges when applied to Indian languages, primarily due to two major issues: the scarcity of low-resource datasets and the lack of support for Indian languages in most Large Language Models (LLMs). Because most of the LLMs trained on English language due to high resource language due to a deficit of annotated data in campare to indian low-resource language.

2.1 Multilingual pre-trained language models

In recent years, the rise of multilingual model architectures has been notable, particularly following the introduction of multilingual pre-trained language models

like mBERT [7], XLM-R [4], mT5 [19], and BLOOM [11]. These models harness the power of extensive unsupervised textual data spanning numerous languages, which facilitates zero-shot and few-shot cross-lingual transfer from one language to another across a variety of downstream NLP tasks, although the performance can vary [10].

More recently, large language models (LLMs) such as GPT-3 [5], Llama-2 [15], and Llama-3 [21] have garnered immense attention due to their exceptional performance in text generation. Research such as [20] has highlighted the strong capability of LLMs in few-shot in-context learning for public English sentiment analysis tasks. While most LLMs are pre-trained with corpora that predominantly feature English, some studies have uncovered intriguing multilingual capabilities in both public and proprietary LLMs [14]. Despite these advancements, to the best of our knowledge, the cross-lingual transfer capabilities of these LLMs have not been thoroughly examined for sentiment analysis tasks. It remains unclear how these LLMs compare to existing multilingual pre-trained models in the context of cross-lingual transfer.

2.2 Multilingual case study

[18] introduced the first code-switching corpus with sentiment labels, showing the robust- ness of a multilingual approach. [1] [16] have contributed to the sentiment analysis dataset for Hindi .

To confront these challenges, various studies have been conducted focusing on specific sentiment analysis tasks for Indian languages. Regarding sentiment analysis of "Hinglish" text, [13] [6] [12] proposed a dictionary-based approach using a dataset of movie reviews in this hybrid language. The method involved creating separate dictionaries for English and Hindi to accommodate word variations and case insensitivity. Feature extraction employed tf-idf with unigram, bigram, and trigram techniques, while sentiment classification utilized SVM, Naïve Bayes, Neural Network, and Logistic Regression algorithms to optimize feature selection and classifier performance for "Hinglish" text. In[17] introduced a sentiment analysis method using SVM with radial basis function for 2,866 Hindi-English code-mixed tweets annotated with six emotions, achieving an accuracy of 58.2%. In recent research [3], has emerged to address this issue by leveraging machine translation to augment data resources. [8] introduced indicSentiment, a dataset spanning 13 Indian languages including Assamese, Bengali, Bodo, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia, Punjabi, Tamil, Telugu, and Urdu. This dataset comprises 1,000 reviews per language, translated and annotated from English reviews collected from various e-commerce platforms. Recently,In [21] worked undertakes an empirical analysis to compare the cross-lingual transfer capability of public Small Multilingual Language Models (SMLM) like XLM-R, against English-centric LLMs such as Llama-3, in the context of sentiment analysis across English, Spanish, French and Chinese. Kumar et al. [9] published sentiment analysis datasets in 22 Indian languages.

Language	Review	Sentiment
Hindi	**लेकिन इसकी बैटरी कमज़ोर है ।**	**Negative**
	इसकी स्क्रीन अभूतपूर्व रंग और अच्छे दृश्य कोण देती है ।	**Positive**
English	But its battery is weak.	Negative
	Its screen gives unprecedented color and good visual angle.	Positive
Assamese	কিন্তু ইয়াৰ বেটাৰী দুৰ্বল।	Negative
	ইয়াৰ পর্দাই অসাধাৰণ ৰং আৰু ভাল দৰ্শন কোণ প্ৰদান কৰে।	Positive
Bengali	কিন্তু এর ব্যাটারি দুর্বল।	Negative
	এর স্ক্রিনটি অভূতপূর্ব রঙ এবং ভাল ভিজ্যুয়াল কোণ দেয়।	Positive
Tamil	ஆனால் அதன் பேட்டரி பலவீனமாக உள்ளது.	Negative
	இதன் திரையானது தனித்துவமான வண்ணங்களையும் நல்ல கோணங்களையும் வழங்குகிறது.	Positive
Telugu	కానీ దాని బ్యాటరీ బలహీనంగా ఉంది.	Negative
	దీని స్క్రీన్ అద్భుతమైన రంగులు మరియు మంచి వీక్షణ కోణాలను అందిస్తుంది.	Positive
Urdu	لیکن اس کی بیٹری کمزور ہے۔	Negative
	اس کی سکرین غیر معمولی رنگ اور دیکھنے کے اچھے زاویے فراہم کرتی ہے۔	Positive

Table 1. Samples from the dataset. Here, review in Hindi is considered as the source language and the rest as the target during translation.

3 Methodology

The objective of this work is to explore the multilingual transfer capability of pre-trained models within the context of a sentiment analysis task for Indian low resources language.

3.1 Datasets

The availability of datasets in Indian languages for sentiment analysis is limited. To address this scarcity, we have employed the machine translation method described by Kumar et al. [9] to create a sentiment analysis dataset in various Indian languages. We have used the IIT Patna product review dataset [2] in Hindi (HI) as the source dataset and translated it into several low-resource Indian languages, including Assamese (AS), Bengali (BN), Tamil (TA), Telugu (TE), and Urdu (UR). Additionally, for comparative purposes, we translated the dataset into English (EN). A few samples from the dataset are shown in Table 1. We have designated 80% of the dataset for each language for testing purposes. We have used 1%, 2%, and 10% of the dataset for each language for training the models, following the standard few-shot training approach.

3.2 LLM Models Details

We fine-tuned several multilingual LLMs on the IndiSentiment140 dataset:

– **XLM-R (Cross-lingual Language Model - RoBERTa)**: XLM-R is a transformer-based model pretrained on 100 languages, optimized for cross-lingual understanding. It leverages the RoBERTa framework, which is known

for its robust performance on monolingual tasks, and extends this capability to multiple languages. XLM-R is designed to handle a wide variety of languages with different linguistic structures, making it suitable for multilingual tasks such as sentiment analysis.

- **mBERT (Multilingual BERT)**: mBERT is the multilingual version of BERT, pre-trained on the top 104 languages with the largest Wikipedias. It shares the same architecture as BERT but is designed to handle multiple languages, providing a common framework for various language tasks. mBERT's training on diverse languages helps it to understand and generate text across different linguistic contexts, making it a versatile model for sentiment analysis.
- **MuRIL (Multilingual Representations for Indian Languages)**: MuRIL is specifically developed for Indian languages, addressing the unique linguistic properties and challenges of these languages. It is pre-trained on a large corpus of Indian languages, including code-mixed data. MuRIL aims to improve natural language understanding for Indian languages, making it particularly relevant for tasks such as sentiment analysis in the Indian context.
- **Navarasa**: Navarasa is a multilingual language model focused on Indian languages. It is fine-tuned to capture the nuances and variations in sentiment across different Indian languages. The model is named after the nine emotions (Navarasa) in Indian classical arts, highlighting its emphasis on understanding and classifying emotional content in text.
- **IndicBERT v2**: IndicBERT v2 is an updated version of IndicBERT, optimized for Indian languages. It incorporates improvements based on newer data and training techniques, providing better performance on tasks involving Indian languages. IndicBERT v2 aims to address the resource constraints and linguistic diversity of Indian languages, enhancing its applicability for sentiment analysis.

Each model was fine-tuned using prompt tuning and few-shot learning techniques to adapt to the sentiment analysis task in various Indian languages. These techniques involve using a small number of prompt parameters and a limited set of examples to guide the model's adaptation, enabling faster and more efficient training.

4 Results and Analysis

The evaluation of multilingual LLMs for sentiment analysis across various Indian languages, using different percentages of training data (1%, 2%, and 10%), provides several insightful observations. The models considered include XLM-R, mBERT, MuRIL, Navarasa, and Indic BERT v2, and the languages evaluated are English, Hindi, Bengali, Tamil, Telugu, Urdu, and Assamese. The performance in terms of accuracy for different languages considering different LLMs is tabulated in Table 2.

XLM-R demonstrates a consistent improvement in accuracy as the training data increases from 1% to 10%. For example, its accuracy in English improves

LLM Model	Train Data (%)	Languages						
		EN	**HI**	**BN**	**TA**	**TE**	**UR**	**AS**
XLM-R	1%	0.553	0.591	0.538	0.546	0.599	0.594	0.751
	2%	0.620	0.665	0.563	0.636	0.632	0.621	0.751
	10%	0.653	0.722	0.567	0.689	0.678	0.719	0.751
mBERT	1%	0.602	0.511	0.586	0.553	0.491	0.509	0.751
	2%	0.690	0.613	0.625	0.604	0.615	0.528	0.751
	10%	0.824	0.676	0.720	0.664	0.698	0.703	0.742
MuRIL	1%	0.491	0.510	0.562	0.489	0.491	0.509	0.751
	2%	0.491	0.490	0.491	0.489	0.491	0.491	0.751
	10%	0.491	0.490	0.491	0.511	0.491	0.491	0.751
Navarasa	1%	0.713	0.518	0.364	0.575	0.539	0.704	0.579
	2%	0.314	0.519	0.581	0.466	0.554	0.623	0.455
	10%	0.703	0.582	0.729	0.714	0.702	0.705	0.731
Indic BERT v2	1%	0.747	0.665	0.650	0.283	**0.766**	0.251	0.554
	2%	**0.756**	0.242	0.236	**0.763**	0.659	0.662	0.262
	10%	**0.756**	**0.758**	**0.764**	0.762	**0.766**	**0.763**	**0.766**

Table 2. Accuracy of different LLMs for different Indian languages considering 1%, 2%, and 10% of data samples for training the model using Prompt tuning.

from 0.553 to 0.653 and in Hindi from 0.591 to 0.722. Other languages such as Bengali, Tamil, Telugu, and Urdu also see steady improvements, with Assamese maintaining a high and stable accuracy of 0.751 across all data percentages. Similarly, mBERT shows significant improvements with increased training data, particularly excelling in English where accuracy jumps from 0.602 to 0.824. While Hindi and Bengali see moderate improvements, Tamil and Telugu show steady accuracy increases. However, Assamese shows a slight decrease in accuracy at 10

In contrast, MuRIL's performance remains relatively unchanged across different training data percentages, with accuracy staying around 0.491 for most languages and only a slight increase for Tamil at 10%. Like other models, MuRIL maintains a stable high accuracy of 0.751 for Assamese. Navarasa exhibits variability in performance with mixed results across languages. For English, accuracy initially is high at 1% (0.713), drops at 2% (0.314), then recovers at 10% (0.703). Hindi and Bengali see significant improvements at 10%, with accuracy reaching 0.582 and 0.729 respectively. Tamil and Telugu also show strong improvements at 10%, while Assamese performs stably but not as high as some other models.

Indic BERT v2 demonstrates substantial performance improvements with increased training data. English and Hindi see significant jumps in accuracy, reaching 0.756 and 0.758 at 10%. Bengali and Telugu also show noticeable improvements, with accuracy reaching 0.764 and 0.766 respectively. Tamil and Urdu see high accuracy at 10%, indicating good adaptability, while Assamese shows consistent improvements, reaching 0.766 at 10%.

Overall, increasing the training data from 1% to 10% generally leads to better performance across all models, highlighting the importance of training data volume in fine-tuning LLMs for sentiment analysis. XLM-R and mBERT stand out as the most effective models, showing substantial improvements across most languages. Indic BERT v2 also performs well, especially with larger training data percentages. MuRIL's performance remains stable but unimpressive compared to other models, indicating potential limitations in adapting to sentiment analysis with limited training data variation. Some models like Navarasa show variability across different languages, suggesting that certain LLMs may be more suited to specific linguistic contexts. Almost all models maintain high accuracy for Assamese, potentially due to the dataset characteristics or the nature of the language data used.

5 Conclusion

This study underscores the effectiveness of multilingual LLMs in sentiment analysis across Indian languages. XLM-R and mBERT emerged as robust performers, particularly in high-resource languages like Hindi and Bengali. However, challenges remain in optimizing performance for low-resource languages. Future research should focus on enhancing model adaptability and performance in diverse linguistic contexts.Indic BERT v2 demonstrated superior performance compared to other language models, primarily due to its specialized training on Indian languages. This targeted training allowed Indic BERT v2 to effectively capture and leverage the nuances and complexities inherent to Indian linguistic structures and contexts.

References

1. Akhtar, M.S., Ekbal, A., Bhattacharyya, P.: Aspect based sentiment analysis in hindi: resource creation and evaluation. In: Proceedings of the tenth international conference on language resources and evaluation (LREC'16). pp. 2703–2709 (2016)
2. Akhtar, M.S., Kumar, A., Ekbal, A., Bhattacharyya, P.: A hybrid deep learning architecture for sentiment analysis. In: Matsumoto, Y., Prasad, R. (eds.) Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 482–493. The COLING 2016 Organizing Committee, Osaka, Japan (Dec 2016), https://aclanthology.org/C16-1047
3. Araújo, M., Pereira, A., Benevenuto, F.: A comparative study of machine translation for multilingual sentence-level sentiment analysis. Information Sciences **512**, 1078–1102 (2020)
4. Barbieri, F., Espinosa Anke, L., Camacho-Collados, J.: XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France (Jun 2022), https://aclanthology.org/2022.lrec-1.27
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)

6. Cambria, E., Das, D., Bandyopadhyay, S., Feraco, A.: Affective computing and sentiment analysis. A practical guide to sentiment analysis pp. 1–10 (2017)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Khanuja, S., Bansal, D., Mehtani, S., Khosla, S., Dey, A., Gopalan, B., Margam, D.K., Aggarwal, P., Nagipogu, R.T., Dave, S., et al.: Muril: Multilingual representations for indian languages. arXiv preprint arXiv:2103.10730 (2021)
9. Kumar, S., Sanasam, R., Nandi, S.: IndiSentiment140: Sentiment analysis dataset for Indian languages with emphasis on low-resource languages using machine translation. In: Duh, K., Gomez, H., Bethard, S. (eds.) Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 7682–7691. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024), https://aclanthology.org/2024.naacl-long.425
10. Lauscher, A., Ravishankar, V., Vulić, I., Glavaš, G.: From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. arXiv preprint arXiv:2005.00633 (2020)
11. Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A.S., Yvon, F., Gallé, M., et al.: Bloom: A 176b-parameter open-access multilingual language model (2023)
12. Mangat, V., et al.: Dictionary based sentiment analysis of hinglish text. International Journal of Advanced Research in Computer Science **8**(5) (2017)
13. Pravalika, A., Oza, V., Meghana, N., Kamath, S.S.: Domain-specific sentiment analysis approaches for code-mixed social network data. In: 2017 8th international conference on computing, communication and networking technologies (ICCCNT). pp. 1–6. IEEE (2017)
14. Qin, L., Chen, Q., Zhou, Y., Chen, Z., Li, Y., Liao, L., Li, M., Che, W., Yu, P.S.: Multilingual large language model: A survey of resources, taxonomy and frontiers. arXiv preprint arXiv:2404.04925 (2024)
15. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
16. Uppal, S., Gupta, V., Swaminathan, A., Zhang, H., Mahata, D., Gosangi, R., Shah, R., Stent, A.: Two-step classification using recasted data for low resource settings. In: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing. pp. 706–719 (2020)
17. Vijay, D., Bohra, A., Singh, V., Akhtar, S.S., Shrivastava, M.: Corpus creation and emotion prediction for hindi-english code-mixed social media text. In: Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: student research workshop. pp. 128–135 (2018)
18. Vilares, D., Alonso, M.A., Gómez-Rodríguez, C.: Sentiment analysis on monolingual, multilingual and code-switching twitter corpora. In: WASSA 2015, the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. pp. 2–8. Association for Computational Linguistics (2015)
19. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mt5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934 (2020)

20. Zhu, W., Liu, H., Dong, Q., Xu, J., Huang, S., Kong, L., Chen, J., Li, L.: Multilingual machine translation with large language models: Empirical results and analysis. arXiv preprint arXiv:2304.04675 (2023)
21. Zhu, X., Gardiner, S., Roldán, T., Rossouw, D.: The model arena for cross-lingual sentiment analysis: A comparative study in the era of large language models. arXiv preprint arXiv:2406.19358 (2024)