International Journal of Electrical and Computer Engineering (IJECE)
Vol. 16, No. 2, April 2026, pp. 973~980
ISSN: 2088-8708, DOI: 10.11591/ijece.v16i2.pp973-980

Cross-lingual semantic alignment and transfer learning using multilingual language models

Niranjan G C, Ramakanth Kumar P, Pavithra H, Minal Moharir
Department of Computer Science and Engineering, RV College of Engineering, Bengaluru, India

Article Info

Article history:

Received Aug 1, 2025
Revised Dec 21, 2025
Accepted Jan 16, 2026

Keywords:

Cosine similarity
Multilingual language models
Natural language inference
Natural language processing
Semantic alignment

ABSTRACT

Multilingual language models (MLMs) are widely used for cross-lingual tasks, yet their ability to achieve consistent semantic alignment and transfer to low-resource languages remains limited. This work examines cross-lingual semantic alignment and transfer learning through a comparative evaluation of MLMs at both the word and sentence levels. We analyze general-purpose models such as big science large open-science open-access multilingual language model (BLOOM) and task-specialized models including LaBSE and XLM-R across English, French, Hindi, and Kannada. Word-level experiments show that LaBSE achieves substantially higher cosine similarity scores of above 0.80 across languages. In sentence-level natural language inference, XLM-R outperforms other models, achieving an F1 score of 68.62% on Kannada and 74.81% on French. These results indicate that model specialization and training objectives play a crucial role in cross-lingual performance, particularly for low-resource languages, and should be carefully considered when deploying multilingual natural language processing (NLP) systems.

Corresponding Author:

Niranjan G C
Department of Computer Science and Engineering, RV College of Engineering
Mysore Road, RV Vidyaniketan Post, Bengaluru, Karnataka 560059, India
Email: niranjangc.scs23@rvce.edu.in

1. INTRODUCTION

Artificial intelligence (AI) has brought about significant changes in the way we interact with technology today. Among the various branches of AI, natural language processing (NLP) has stood out for its ability to help machines understand, interpret, and generate human language. With the rise of generative AI, the capabilities of language models have grown rapidly, enabling machines to perform tasks like translation, summarization, and even creative writing. One of the more recent advancements in this space is the development of multilingual language models (MLMs)-models that can process and understand multiple languages using a shared neural architecture.

In multilingual settings, these models are especially useful because they can handle input from different languages without requiring separate models for each one. This makes them ideal for use in global applications where language diversity is common. In the context of this project, we explore how MLMs can be applied to understand and process different languages effectively. The project aims to analyze whether these models truly capture the meaning of words across languages and how well they can transfer knowledge from a high resource language (English) to a low resource Indic language (Kannada).

MLMs such as BLOOM-1.7B and QWEN2 have been evaluated for their ability to align cross-lingual word embeddings, represent internal structures through probing tasks like sentence similarity and named entity recognition, and transfer knowledge across languages, with a focus on low-resource scenarios

[1]. Evaluation methodologies for models like mBERT, XLM-R, and InfoXLM have revealed that performance significantly drops in across-language tasks compared to within-language setups, especially in low-resource languages, exposing generalization weaknesses [2]. The translation performance of big science large open-science open-access multilingual language model (BLOOM) has been assessed in zero-shot and few-shot settings across high- and low-resource language pairs, showing notable improvements in few-shot scenarios when linguistic similarities exist between source and target languages [3]. A more efficient approach to cross-lingual pretraining has been introduced with XLM-E [4], which uses replaced token detection instead of masked language modeling and achieves faster, cost-effective training while maintaining high performance on benchmarks like XTREME [5].

Multilingual BERT has shown strong performance in high-resource languages for named entity recognition, but its effectiveness sharply declines in low-resource languages due to limited vocabulary coverage and insufficient training data, which are identified as critical factors affecting multilingual model performance [6]. Negative interference, where multilingual training degrades performance on individual languages, has been addressed through a meta-learning approach that introduces language-specific adapters to improve both monolingual accuracy and cross-lingual transfer [7]. A complementary strategy involves using language-specific subnetworks within multilingual models to control parameter sharing, which, when combined with meta-learning, yields significant improvements in few-shot transfer and low-resource dependency parsing [8]. In the domain of cross-lingual information retrieval, a novel method called OPTICAL uses optimal transport-based knowledge distillation to align token-level embeddings between languages, enabling high retrieval performance in low-resource settings without relying on cross-lingual relevance annotations [9].

The robustness of multilingual models like mBERT and XLM-R under adversarial perturbations has been evaluated for tasks such as named entity recognition, revealing that vocabulary overlap and linguistic proximity are key determinants of cross-lingual generalization in low-resource settings [10]. A comparative study involving multilingual and monolingual models on African languages Kinyarwanda and Kirundi showed that fine-tuned multilingual models like AfriBERT outperform others, demonstrating strong transfer capabilities between linguistically similar, low-resource languages [11]. A structured survey of cross-lingual word embedding techniques categorized methods by data alignment levels and type (parallel vs. comparable), offering a foundational understanding of multilingual semantic representation approaches [12]. Transfer learning strategies across various multilingual NLP tasks have been reviewed, highlighting that model architecture and pretraining data are more impactful than script similarity, with models like XLM-R out performing both mBERT and language-specific variants in zero-shot transfer scenarios [13].

ZeroShotTM demonstrated that multilingual contextualized embeddings can be effectively leveraged for zero-shot topic modeling, enabling cross-lingual topic inference without retraining or language-specific vocabularies [14]. In the domain of low-resource Indian languages, zero-shot translation using multilingual neural machine translation (NMT) models showed substantial improvement when incorporating training data from related languages, emphasizing the role of lexical proximity in enhancing translation performance [15]. To support NLP development for Indian languages, the IndicNLPSuite introduced large-scale corpora, pretrained models, and evaluation benchmarks, significantly advancing research in under-resourced Indic language processing [16].

To explore the aforementioned capabilities of MLMs, the project is divided into two core tasks: evaluating cross-lingual semantic similarity using word embeddings from models like BLOOM and LaBSE [17], and testing the zero-shot inference ability of MLMs through a natural language inference (NLI) task. By leveraging pretrained models, curated datasets, and visualizations, the study aims to assess the effectiveness and real-world applicability of MLMs in multilingual settings, particularly for low-resource languages.

Although MLMs are trained on large and diverse corpora, their ability to represent semantic meaning consistently across languages is not guaranteed, especially for low-resource languages. Prior studies have shown that multilingual alignment varies significantly depending on model architecture, training objectives, and language characteristics. However, many works either focus exclusively on high-resource languages or evaluate models using a single task. Indic languages such as Kannada remain underexplored in cross-lingual evaluation. This work addresses this gap by jointly analyzing word-level semantic similarity and sentence-level inference across multiple languages. By comparing general-purpose multilingual models with task-specific architectures, we aim to better understand the strengths and limitations of current multilingual models in low-resource Indic language settings.

The primary objective of this work is to evaluate how effectively MLMs capture cross-lingual semantic alignment and transfer this knowledge to downstream tasks. To achieve this, we present a unified evaluation framework that analyzes both word-level semantic similarity and sentence-level zero-shot transfer performance across English, French, Hindi, and Kannada. We compare general-purpose MLMs such as BLOOM with task-specialized models including LaBSE and XLM-R to study the impact of architectural

design and training objectives on cross-lingual performance. Our experiments provide empirical evidence that strong word-level alignment does not necessarily lead to improved sentence-level inference, particularly in low-resource languages such as Kannada.

2. METHOD

The system architecture of the proposed solution is designed to facilitate cross-lingual evaluation of MLMs through two core modules: word embedding similarity and NLI. It builds on the methodology proposed in [1], but with focus on Indic languages especially Kannada. Each module supports a specific downstream task and relies on pre-trained transformer-based models to measure multilingual alignment and semantic understanding. The interface acts as the user interaction layer, while backend logic handles translation, embedding generation, and inference using streamlined workflows.

Figure 1 depicts the system architecture of the application. In the word embedding similarity module, the user begins by entering an English word through the interface. The system retrieves the corresponding translations in French, Hindi, and Kannada using a translation utility. The word and its translations are tokenized, and a preloaded MLM (LaBSE) generates vector embeddings for each word. These embeddings are compared using pairwise cosine similarity to measure semantic proximity across languages [18]. The similarity scores are accompanied by visualization outputs including heatmaps and t-SNE plots, providing insights into the degree of alignment between cross-lingual word representations.

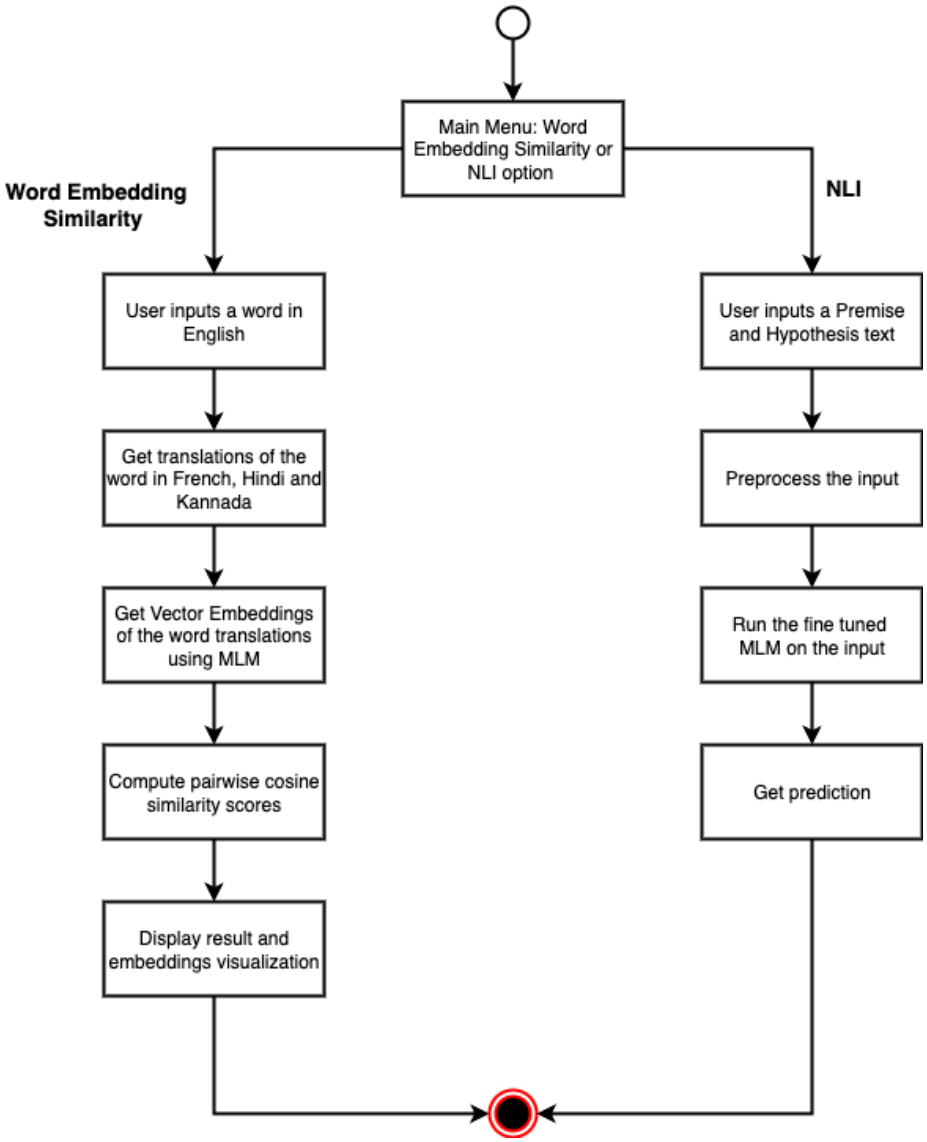


Figure 1. System architecture

In the NLI module, the user inputs a premise and a hypothesis, both of which are first preprocessed. The preprocessing involves cleaning the data, truncating or padding the input to a fixed length of 128 tokens, and converting the tokens into tensors along with attention masks. This formatted input is fed into a fine-tuned XLM-R model [19], which has been trained on the English XNLI dataset and evaluated on the French XNLI test dataset [20] and Kannada IndicXNLI test dataset [21]. The model processes the preprocessed input and classifies the relationship between the premise and hypothesis into one of three categories: entailment, contradiction, or neutral. The resulting prediction and confidence scores are displayed to the user in a readable format.

3. RESULTS AND DISCUSSION

This section presents the results obtained through a series of experiments conducted on various MLMs across the two modules: word embedding similarity and NLI. It provides a detailed account of the performance of each model, supported by relevant tables and visualizations. These evaluations offer insights into the strengths and limitations of each model in handling multilingual tasks and transferring knowledge across high- and low-resource languages.

3.1. Word embedding similarity results

For the word embedding similarity module, the models evaluated include BLOOM 560M, BLOOM 1.7B, and LaBSE. Their effectiveness is measured using the average pairwise cosine similarity score (1) computed over a dataset of 3000 English words and their translations in French, Hindi, and Kannada, which quantifies the semantic alignment between word embeddings across different languages.

$$Cosine\ Similarity(\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \|\vec{B}\|} \quad (1)$$

where $\vec{A}$ and $\vec{B}$ are the word embedding vectors, $\vec{A} \cdot \vec{B}$ is the dot product, and $\| \vec{A} \| \ and \ \| \vec{B} \|$ are the magnitudes of the vectors. Table 1 presents the average pairwise cosine similarity scores obtained by the three MLMs-BLOOM 560M, BLOOM 1.7B, and LaBSE across different language pairs in the word embedding similarity module.

Table 1. Average pairwise cosine similarity scores

Language Pair	BLOOM 560M	BLOOM 1.7B	LaBSE
EN - FR	0.98	0.95	0.90
EN - HI	-0.003	-0.14	0.89
EN - KN	-0.16	-0.18	0.85
HI - KN	0.097	0.032	0.82

The comparatively higher similarity scores achieved by LaBSE across all language pairs indicate its superior ability to capture semantic alignment across languages and was thus chosen as the most suitable model for deployment in the Word Embedding Similarity module. Recent studies have emphasized that cross-lingual alignment is not a fixed property of multilingual models but can be influenced by architectural and training strategies. Techniques such as cross-lingual position encoding and bidirectional training have been shown to improve alignment by explicitly modeling word order differences across languages and strengthening bilingual representations [22], [23]. In contrast, the present work evaluates pretrained models without alignment specific fine tuning. The observed degradation in alignment for Indic languages, particularly in BLOOM embeddings, suggests that incorporating such alignment aware strategies could potentially enhance cross lingual robustness. Another important factor affecting performance in low resource languages is lexical frequency. Prior work has shown that multilingual training and knowledge distillation tend to favor high frequency words, leading to weaker representations for low frequency lexical items [24], likely contributing to the reduced performance observed for Kannada.

Figure 2 presents the embedding visualizations. Figures 2(a), 2(b), and 2(c) present the t-SNE visualization of the vector embeddings generated by the BLOOM 560M, BLOOM 1.7B, and the LaBSE model respectively, in a two-dimensional space. The significant overlap among the embeddings of translated words confirms that LaBSE effectively captures the semantic alignment of related words across different languages.

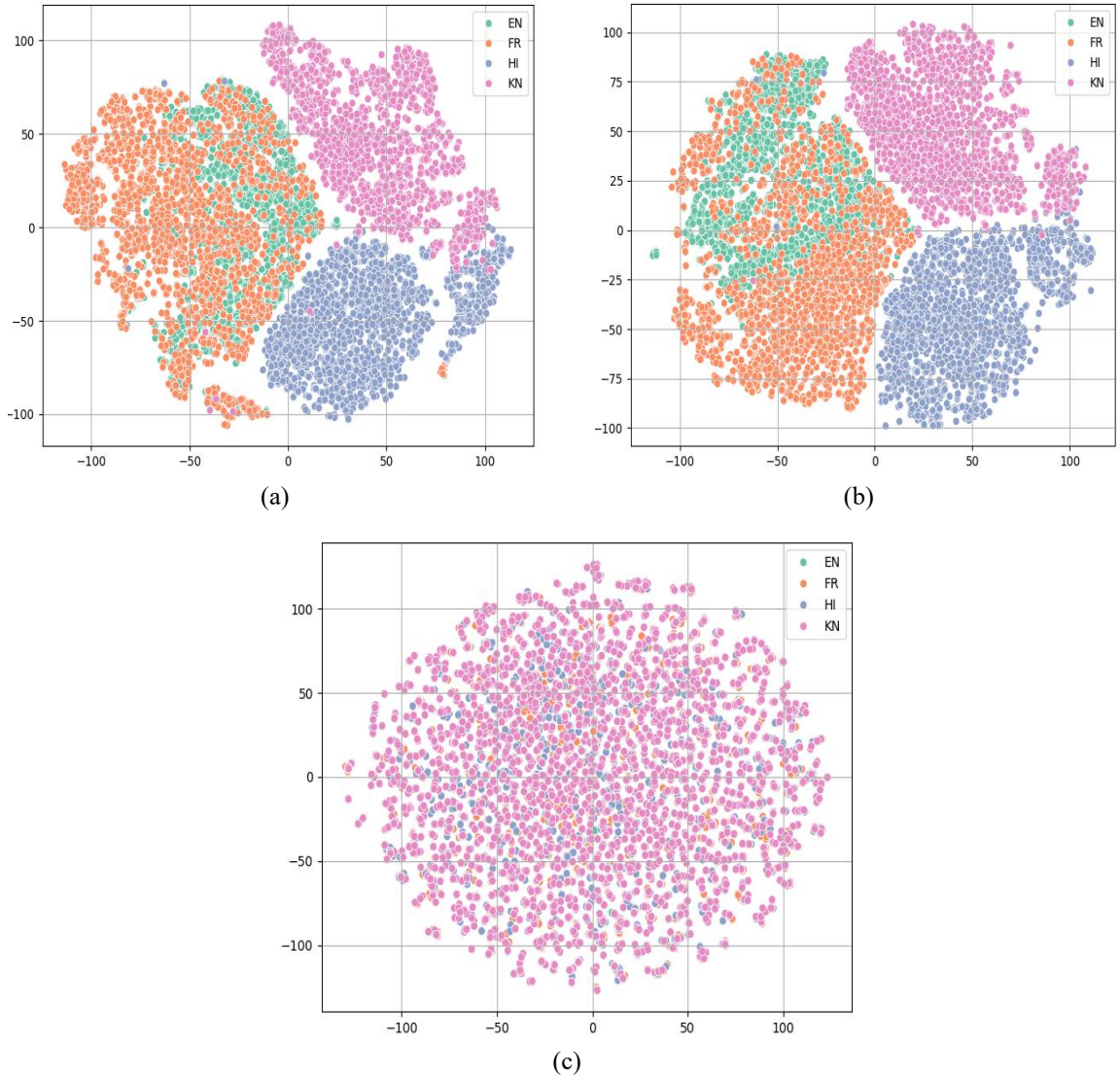


Figure 2. t-SNE visualization of embeddings obtained from the MLMs (a) BLOOM 560M, (b) BLOOM 1.7B, and (c) LaBSE

3.2. NLI results

For the NLI module, the zero-shot transfer learning performance of BLOOM 560M, BLOOM 1.7B, and XLM-R is compared using standard classification metrics-accuracy and F1 score on test datasets in French and Kannada. Table 2 summarizes the test accuracy scores and Table 3 summarizes the test F1 scores for the NLI task, obtained by evaluating the three aforementioned fine-tuned MLMs on French and Kannada test datasets. XLM-R achieves better test accuracy and F1 scores for both French and Kannada, indicating its superior performance in zero-shot transfer for the NLI task. Consequently, XLM-R was selected for deployment in the NLI module. The superior NLI performance of XLM-R can also be attributed to its architecture and pretraining strategy, which are better suited for sentence level reasoning. Cross-attention mechanisms play a critical role in enabling alignment between premise and hypothesis representations across languages. Prior analyses have identified limitations in cross-attention behavior that affect contextual alignment, particularly in non-autoregressive and multilingual settings [25]. Although this work does not perform attention-level analysis, the consistent improvement of XLM-R over BLOOM indicates that stronger and more context-aware cross-attention representations are beneficial for cross-lingual inference.

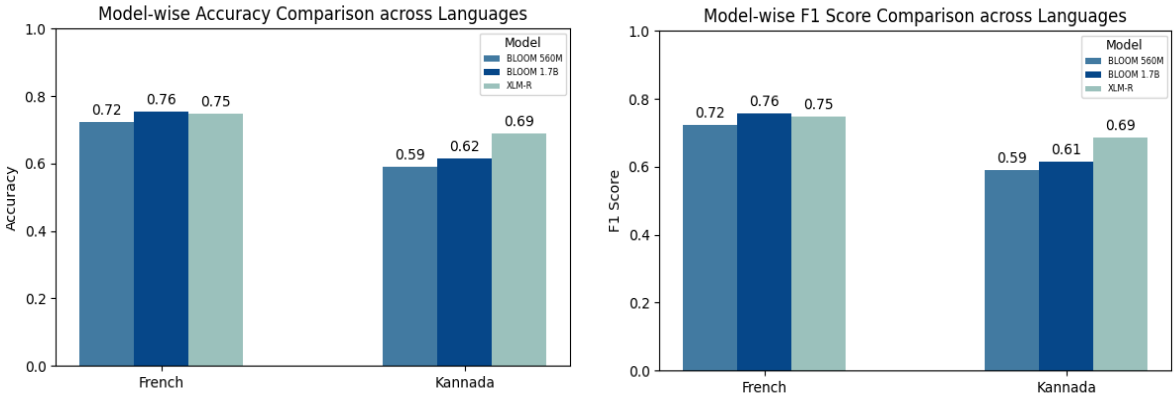
Figure 3 presents the model wise accuracy and F1 score comparisons. Figure 3(a) presents a bar chart of the test accuracy scores, and Figure 3(b) depicts a bar chart of the test F1 scores for the NLI task, obtained by evaluating the three fine-tuned MLMs-BLOOM 560M, BLOOM 1.7B, XLM-R on French and Kannada test datasets.

Table 2. Test accuracy scores

Language	BLOOM 560M	BLOOM 1.7B	XLM-R
French	72.27%	75.51%	74.83%
Kannada	59.08%	61.54%	68.8%

Table 3. Test F1 scores

Language	BLOOM 560M	BLOOM 1.7B	XLM-R
French	72.4%	75.59%	74.81%
Kannada	58.98%	61.46%	68.62%



(a) (b)

Figure 3. Bar chart for (a) test accuracy scores and (b) test F1 scores

4. CONCLUSION

This study shows that cross-lingual performance of MLMs is highly task-dependent and cannot be inferred solely from model size or multilingual coverage. Our experiments demonstrate that task-specialized models such as LaBSE and XLM-R consistently outperform general purpose multilingual models like BLOOM for word level semantic similarity and sentence level inference, particularly in low resource Indic languages such as Kannada. These findings highlight the importance of architectural design and training objectives when selecting multilingual models for practical applications. While this work focuses on evaluating pretrained models, future work could explore training strategies such as progressive multi-granularity learning, where models are trained from words to phrases and sentences, to further improve cross-lingual robustness and transfer performance in low resource settings. In addition, future enhancements may include investigating one-shot and few-shot transfer capabilities, capturing sentence-level semantic alignment more explicitly, and extending the analysis to other downstream NLP tasks.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Niranjan G C	✓	✓	✓	✓	✓	✓		✓	✓	✓				
Ramakanth Kumar P		✓		✓	✓					✓	✓	✓	✓	
Pavithra H				✓	✓					✓		✓		
Minal Moharir				✓	✓					✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The authors confirm that the data supporting the findings of this study are available within the article.

REFERENCES

[1] S. Kakarla, G. S. B. Venkata, and A. Gaddam, "How does a multilingual LM handle multiple languages?," *arXiv preprint arXiv:2502.04269*, 2025.

[2] S. Rajaee and C. Monz, "Analyzing the evaluation of cross-lingual knowledge transfer in multilingual language models," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 2895–2914, doi: 10.18653/v1/2024.eacl-long.177.

[3] R. Bawden and F. Yvon, "Investigating the translation performance of a large multilingual language model: the case of bloom," *arXiv preprint arXiv:2303.01911*, 2023.

[4] Z. Chi *et al.*, "XLM-E: cross-lingual language model pre-training via ELECTRA," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 6170–6182, doi: 10.18653/v1/2022.acl-long.427.

[5] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson, "Xtreme: a massively multilingual multi-task benchmark for evaluating cross-lingual generalisation," in *International conference on machine learning*, 2020, pp. 4411–4421, doi: 10.48550/arXiv.2003.11080.

[6] S. Wu and M. Dredze, "Are all languages created equal in multilingual BERT?," *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, no. July, pp. 120–130, 2020, doi: 10.18653/v1/2020.repl4nlp-1.16.

[7] Z. Wang, Z. C. Lipton, and Y. Tsvetkov, "On negative interference in multilingual models: findings and a meta-learning treatment," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4438–4450, doi: 10.18653/v1/2020.emnlp-main.359.

[8] R. Choenni, D. Garrette, and E. Shutova, "Cross-lingual transfer with language-specific subnetworks for low-resource dependency parsing," *Computational Linguistics*, vol. 49, no. 3, pp. 613–641, Sep. 2023, doi: 10.1162/coli_a_00482.

[9] Z. Huang, P. Yu, and J. Allan, "Improving cross-lingual information retrieval on low-resource languages via optimal transport distillation," in *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, Feb. 2023, pp. 1048–1056, doi: 10.1145/3539597.3570468.

[10] S. Manafi and N. Krishnaswamy, "Cross-lingual transfer robustness to lower-resource languages on adversarial datasets," *arXiv preprint arXiv:2403.20056*, 2024, doi: 10.48550/arXiv.2403.20056.

[11] H. Thangaraj, A. Chenat, J. S. Walia, and V. Marivate, "Cross-lingual transfer of multilingual models on low resource African Languages," *arXiv preprint arXiv:2409.10965*, 2024, doi: 10.48550/arXiv.2409.10965.

[12] S. Ruder, I. Vulić, and A. Søgaard, "A survey of cross-lingual word embedding models," *Journal of Artificial Intelligence Research*, vol. 65, pp. 569–631, Aug. 2019, doi: 10.1613/jair.1.11640.

[13] A. R. Jafari, B. Heidary, R. Farahbakhsh, M. Salehi, and M. Jalili, "Transfer learning for multi-lingual tasks--a survey," *arXiv preprint arXiv:2110.02052*, 2021.

[14] F. Bianchi, S. Terragni, D. Hovy, D. Nozza, and E. Fersini, "Cross-lingual contextualized topic models with zero-shot learning," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, pp. 1676–1683, doi: 10.18653/v1/2021.eacl-main.143.

[15] R. Huidrom and Y. Lepage, "Zero-shot translation among Indian languages," in *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, 2020, pp. 47–54, doi: 10.18653/v1/2020.loresmt-1.7.

[16] D. Kakwani *et al.*, "IndicNLPSuite: monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 4948–4961, doi: 10.18653/v1/2020.findings-emnlp.445.

[17] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic BERT sentence embedding," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 878–891, doi: 10.18653/v1/2022.acl-long.62.

[18] M. Farouk, "Measuring sentences similarity: a survey," *Indian Journal of Science and Technology*, vol. 12, no. 25, pp. 1–11, Jul. 2019, doi: 10.17485/ijst/2019/v12i25/143977.

[19] A. Conneau *et al.*, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8440–8451, doi: 10.18653/v1/2020.acl-main.747.

[20] A. Conneau *et al.*, "XNLI: evaluating cross-lingual sentence representations," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2475–2485, doi: 10.18653/v1/D18-1269.

[21] D. Aggarwal, V. Gupta, and A. Kunchukuttan, "IndicXNLI: evaluating multilingual inference for Indian languages," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 10994–11006, doi: 10.18653/v1/2022.emnlp-main.755.

[22] L. Ding, L. Wang, and D. Tao, "Self-attention with cross-lingual position representation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 1679–1685, doi: 10.18653/v1/2020.acl-main.153.

[23] L. Ding, D. Wu, and D. Tao, "Improving neural machine translation by bidirectional training," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 3278–3284, doi: 10.18653/v1/2021.emnlp-main.263.

[24] L. Ding, L. Wang, X. Liu, D. F. Wong, D. Tao, and Z. Tu, "Understanding and improving lexical choice in non-autoregressive translation," *arXiv preprint arXiv:2012.14583*, 2020.

[25] L. Ding, L. Wang, D. Wu, D. Tao, and Z. Tu, "Context-aware cross-attention for non-autoregressive translation," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 4396–4402, doi: 10.18653/v1/2020.coling-main.389.

BIOGRAPHIES OF AUTHORS

Niranjan G C is a post-graduate student at Visvesvaraya Technological University (VTU), pursuing M.Tech in computer science and engineering at Rashtreeya Vidyalaya College of Engineering (RVCE) Bangalore, India. With a strong inclination towards generative AI and large language models (LLM)s, his areas of interest include DL, NLP, and traditional machine learning (ML). He can be contacted at email: niranjangc.scs23@rvce.edu.in.

Ramakanth Kumar P is currently a professor and the dean of the CSE Cluster, Rashtreeya Vidyalaya College of Engineering (RVCE), Bangalore, India. He has taught courses on network programming and cybersecurity for Industry 4.0. He has published more than 100 research articles. He is a senior member at IEEE and has executed several funded research and consultancy projects sponsored by DRDO, ISRO, CAIR, LRDE, AICTE, GE India Pvt. Ltd., CABS, and HPE. His research interests include digital image processing, pattern recognition, and natural language processing. He can be contacted at email: ramakanthkp@rvce.edu.in.

Pavithra H is currently an associate professor in the Computer Science and Engineering Department at Rashtreeya Vidyalaya College of Engineering (RVCE), Bangalore, India. Her research interests are software defined networks, machine learning, deep learning, software engineering. She has executed projects sponsored by Samsung, Toyota. She can be contacted at email: pavithrah@rvce.edu.in.

Minal Moharir is currently a professor in the Department of Computer Science and Engineering at Rashtreeya Vidyalaya College of Engineering (RVCE), Bengaluru, with over 14 years of academic experience. She holds a Ph.D. in computer science and engineering and an M.Tech in computer network engineering. Her areas of expertise include computer networks, cybersecurity, and high-performance computing. She has published extensively, with over 70 papers in reputed international and national journals and conferences, and has guided numerous UG and PG research projects. She has led several funded research and consultancy projects in collaboration with organizations like NVIDIA, Citrix, Samsung, and DRDO labs, securing grants worth several lakhs. She can be contacted at email: minalmoharir@rvce.edu.in.