International Journal for Modern Trends in Science and Technology
Volume 12, Issue 07, pages 01-10.
ISSN: 2455-3778 online
Available online at: http://www.ijmtst.com/vol12issue06.html
DOI: https://doi.org/10.5281/zenodo.20739608

A Comparative Study of Multilingual Transformer – Based Models for Classification of Transliterated Hindi Text

Pawan Tripathi | Ashish Kumar Pandey

Department of Computer Science and Engineering , IET Dr. R.M.L. Avadh University , Ayodhya, Uttar Pradesh, India.

To Cite this Article
Pawan Tripathi & Ashish Kumar Pandey (2026). A Comparative Study of Multilingual Transformer – Based Models for Classification of Transliterated Hindi Text. International Journal for Modern Trends in Science and Technology, 12(07), 01-10. https://doi.org/10.5281/zenodo.20739608

Article Info
Received: 06 June 2026; Revised: 29 June 2026; Accepted: 01 July 2026.

KEYWORDS

Hindi Transliteration, Multilingual Transformer Models, Dakshina Dataset, Text Classification, Natural Language Processing

ABSTRACT

Transformer-based Multilingual Models have made remarkable progress for multilingual Natural Language Processing (NLP) by enabling contextual representation learning and cross-lingual understanding. However, Indian languages continue to face challenges related to script diversity, transliteration ambiguity, code-mixed usage and the limited availability of low resources. Hindi, one of the most widely spoken Indian languages, is frequently represented in both its native Devanagari script and Romanized form, creating additional complexities for automated language processing systems. This study compares the performance of three multilingual transformer models- IndicBERT, mBERT, and XLM-R on Hindi language processing tasks using the Dakshina transliteration dataset. Dakshina is a dataset of aligned transliteration pairs of native Devanagari script and Hindi script in Roman script. Experimental results show that XLM-R achieves the maximum accuracy of 96.34% in classification tasks that require transliteration-aware contextual understanding, outperforming the other models mBERT and IndicBERT. The Comparative analysis indicates that XLM-R improves classification performance by 2.53% over mBERT and 1.22% over IndicBERT, demonstrating higher contextual understanding of transliterated Hindi text. The proposed framework enhances transliteration based multilingual NLP system for Indian regional languages, and serves as a basis for future low-resource Indic language research.

1. INTRODUCTION

Natural Language Processing (NLP) is a key area of Artificial Intelligence (AI) that enables machines to understand, process and generate human language. The emergence of transformer-based pre-trained language models has significantly advanced NLP by providing contextual representation learning and transfer learning capabilities for tasks such as text classification, sentiment

analysis, machine translation, and question answering [1], [2].

Recent multilingual transformer models further support cross-lingual knowledge transfer and multilingual language understanding through shared representations across languages [2], [4], [5]. Despite these advances, Indian languages present distinctive challenges due to linguistic diversity, multiple scripts, transliteration variations, and limited annotated resources [3], [6].

Hindi, one of the most widely used Indian languages, is commonly written in both Devanagari and Roman scripts, particularly in social media and digital communication. Romanized Hindi often exhibits multiple spellings for the same word because of phonetic and regional variations, creating ambiguity for NLP systems [3], [11].

Transliteration, the conversion of text from one script to another while preserving pronunciation, facilitates digital communication but introduces inconsistencies that affect language understanding and classification [3], [6], [11].

To address these challenges, multilingual transformer models such as mBERT, IndicBERT, and XLM-R have evaluated on the Dakshina Hindi transliteration dataset, using standardized preprocessing and performance metrics including accuracy, precision, recall, and F1-score to identify the most effective model for transliterated Hindi text classification.

Research Questions

The study is guided by the following research questions:

RQ1: What is the best multilingual transformer architecture for transliteration aware classification in Hindi?

RQ2: What is the impact of large-scale multilingual pre-training on contextual representation learning in transliterated Hindi text?

RQ3: Can Indic-specific pre-training provide competitive performance compared with general multilingual transformer architectures?

Contributions of the Study

The major contributions of this research are summarized as follows:

1. A comprehensive comparative evaluation of IndicBERT, mBERT, and XLM-R using the Dakshina Hindi transliteration dataset.
2. Development of a preprocessing framework for transliterated Hindi text normalization and contextual representation learning.
3. Investigation of the impact of multilingual and Indic-specific pretraining strategies on transliteration-aware classification performance.

The remainder of this paper is organized as follows. Section 2 presents the related works and comparative analysis of existing multilingual transformer models. Section 3 discusses the research methodology, dataset statistics, data pre-processing, Section 4 describes the multilingual transformer models employed in the study. Section 5 introduces the experimental setup and evaluation metrics, Section 6 presents the experimental results and comparative performance analysis. Finally, Section 7 concludes the paper and outlines potential directions for future research.

II. Related Works

A. Evolution of Multilingual Transformer Models

Transformer-based models have revolutionized NLP by facilitating contextual representation learning and achieving cross-lingual transfer learning. With the introduction of BERT by Devlin et al. [1] a new paradigm in language modelling arose, with the introduction of bidirectional contextual embedding learned with masked language modelling. On top of this, multilingual BERT (mBERT) has been developed to represent over 100 languages with shared multilingual representations, enabling cross-lingual transfer learning and multilingual language understanding [1], [7].

With the goal of achieving the better quality multilingual representation, the subsequent studies centred on extending the training corpora size and diversity. XLM-R used the Common Crawl data to perform large-scale multilingual pre-training, and showed significant improvement on multiple multilingual NLP tasks [2]. Significant improvements in contextual understanding and transfer capabilities across languages, especially low resource languages [2, 10] were observed.

To overcome this general multilingual setting for Indian languages, researchers created a model called IndicBERT, designed for processing Indic languages. IndicBERT is trained on large-scale Indic language corpora and offers enhanced contextual representations for Indian languages with multiple script and few labelled data [3, 9].

These advancements have made multilingual transformers a strong instrument for language processing in low resource languages and multilingual NLP applications.

B. Hindi Transliteration and Dakshina Dataset

Hindi language processing is a challenging task as the language is often found in both Devanagari script and Romanization. Many of the above discussed social media platforms, messaging applications, and online communication systems use Romanized Hindi, which leads to transliteration ambiguities that hinder the automatic analysis for language understanding and text classification [6, 11].

Transliteration is the process of translating text between different writing systems, while maintaining phonetic information. But, there can be several romanizations of the same Hindi word and inconsistencies as well as complexity for the NLP systems.

To overcome this problem, Google Research introduced the multilingual benchmark of aligned transliteration pairs between native Indian scripts and Romanized text representations – 'Dakshina' dataset [12]. Dakshina has been made an important dataset to assess multilingual models on transliteration tasks due to the availability of standard transliteration pairs for multiple Indian languages.

Dakshina also provides a Hindi subset, which can be used to systematically explore the architectures of multilingual transformers in transliteration-aware language processing and contextual representation learning [12].

C. Recent Advances in Multilingual Language Models

Multilingual language models have been developed recently, which have significantly expanded the capacities of transformer models by large-scale pre-training, instruction tuning, and parameter-efficient adaptation methods. The use of modern multilingual models has been shown to work well in machine translation, question answering, sentiment analysis and multilingual conversational systems [14, 15].

Instruction-tuned multilingual models have demonstrated impressive zero-shot or few-shot learning ability, and the ability to transfer knowledge from one language to another without the need for significant task-specific training. These are ways to lessen the reliance on the large annotated data set and enhance the model's generalization on a wide range of linguistic contexts [14].

At the same time, Parameter-Efficient Fine-Tuning (PEFT) methods like Low-Rank Adaptation (LoRA), Prompt Tuning, Prefix Tuning and Adapter-based learning have garnered significant interest. These methods allow for the adaptation of large multilingual transformer models to specific tasks, while also substantially lowering the computational demands.

This is especially useful in the case of low resource languages, where computational power and training data can be scarce [13, 15]. Also, the creation of sizeable multilingual corpora and Indic language benchmarks has made multilingual representation learning for Indic languages more robust.

All these developments have made it easier to deal with script diversity, transliteration variations, code-mixed text, and low resource linguistic environments, thus making multilingual NLP systems more useful in practice [3, 9].

D. Comparative Literature Analysis

Transformer based architectures like mBERT, IndicBERT and XLM-R have made significant progress in multilingual representation learning and various multilingual applications like multilingual classification, sentiment classification, machine translation and cross-lingual transfer learning tasks have been reported in the literature [1]–[5]. Likewise, the relevance of Indic-specific language models and multi-lingual corpora towards the issues of low-resource Indic languages has been pointed out recently [3], [6], [9]. Although these advances have been made, there are a few drawbacks.

Existing studies are mainly about multi-language representation learning and cross-lingual transfer tasks, and hardly any studies are available on

transliteration-aware Hindi classification. There is not much literature available to directly compare IndicBERT, mBERT and XLM-R on a common transliteration benchmark dataset. Moreover, the available literature does not have a uniform framework of pre-processing designed for transliteration ambiguity and variations in Romanised Hindi text [3, 6, 13].

Table1. shows the key findings and limitations of the previous studies and points out the gap in the research that is being filled by this research work. The gap identified motivates the need of comparing multilingual transformer architectures systematically using the Dakshina Hindi transliteration dataset using a common experimental framework.

Table1: Comparative Literature Analysis

RN	Model/ Dataset	Major Contribution	Limitation
[1]	BERT	Introduced contextual transformer embedding	Limited Indic language optimization
[2]	XLM-R	Large-scale multilingual pre-training	High computational complexity
[3]	IndicBERT	Indic language transformer architecture	Limited dataset size
[4]	Cross-lingual NLP	Cross-lingual transfer learning	Low-resource challenges
[5]	Multilingual Embedding	Improved multilingual alignment	Semantic inconsistencies
[6]	Indic NLP	Indic language benchmarking	Dataset imbalance
[7]	mBERT	Multilingual fine-tuning analysis	Lower Indic contextual depth
[8]	Transformer NLP	Enhanced multilingual representation	Large training cost
[9]	IndicCorp	Large Indic corpus generation	Resource intensive pre-processing
[10]	XLM-R	Improved multilingual representation	High GPU dependency
[11]	Hindi NLP	Transliteration-aware classification	Limited dataset diversity
[12]	Dakshina Dataset	Transliteration benchmark dataset	Limited downstream annotations
[13]	Indic Transformers	Transfer learning for Indic NLP	Generalization limitation
[14]	Multilingual LLMs	Advanced multilingual reasoning	Computational cost
[15]	LLaMA Multilingual	Efficient multilingual adaptation	Limited Indic fine-tuning

E. Research Gap

While the multilingual transformer models like mBERT, IndicBERT and XLM-R have achieved strong results on multilingual NLP tasks, there are some fundamental research questions related to the Hindi transliteration-aware language understanding.

On one hand, most of the previous works are limited to the focus areas such as machine translation, sentiment analysis, language modelling, and cross-lingual transfer learning, and on the other hand, comparatively not much attention has been paid to the transliteration-aware Hindi text classification.

Second, comparative assessment of IndicBERT, mBERT and XLM-R on a similar benchmark data with similar experimental setups is still limited.

Third, while the effects of Indic specific pre-training and large-scale multilingual pre-training on Romanized Hindi text is not yet fully understood. Moreover, a lot of previous works do not use standardized pre-processing frameworks which are tailored to transliteration ambiguity, spelling variation, and inconsistencies in Romanised Hindi.

Therefore, a systematic benchmark study along with a common experimental setting is required to examine multilingual transformer architectures for transliteration-aware classification of Hindi language tasks.

To overcome these shortcomings, this study aims to provide a comprehensive comparative analysis of IndicBERT, mBERT and XLM-R on Dakshina Hindi transliteration dataset and to explore the effectiveness of multilingual contextual representation learning for low-resource Indic languages.

III. Methodology

A. Dataset Collection

The goal of this work is to analyze the multilingual transformer architectures for Hindi transliteration-aware

classification, which needs a benchmark dataset with aligned transliteration pairs from the Hindi language.

Thus, this study leverages on the multilingual translation corpus publicly released by Google Research known as Dakshina Dataset [12].

The dataset consists of pairs of Hindi words written in their original Devanagari script and their transliterated Hindi in Roman script represented in table 2.

Table2: Data set representation

Devanagari Script	Romanized Hindi
नमस्ते	namaste
भारत	bharat
शिक्षा	shiksha
विद्यालय	vidyalaya
कंप्यूटर	computer

The aligned transliteration pairs help multilingual transformer models to learn semantic and phonetic relationships between the native script (Hindi) and the Romanized text (transliterations).

The Dakshina dataset is especially appropriate for this study since it offers: 1. a set of pairs for standardisation of transliteration for benchmarking purposes. 2. Support for multilingual, cross-lingual language processing tasks. 3. It enables reproducibility and ability to compare machine learning models fairly.

B. Dataset Splitting Strategy

To ensure unbiased model evaluation and prevent over fitting, the dataset is divided into three subsets:

- **Training Set** (80%): used for model learning and parameter optimization.
- **Validation Set** (10%): used for hyper-parameter tuning and model selection.
- **Testing Set:** (10 %): used for final performance evaluation.

The stratified splitting approach preserves the class distribution across all subsets and minimizes sampling bias. This strategy ensures that the multilingual transformer models are evaluated fairly and their generalization capability can be assessed accurately on unseen transliterated Hindi text.

C. Dataset Statistics

The dataset contains diverse transliteration patterns that reflect real-world spelling variations commonly observed in social media, online communication, and user-generated content. This diversity makes the dataset suitable for evaluating the contextual understanding capabilities of multilingual transformer architectures. Dataset Statistics represented in table 3.

Table 3: Dataset Statistics

Attribute	Value
Dataset Name	Dakshina (Hindi Subset)
Language	Hindi
Source	Google Research
Script Types	Devanagari and Romanized Hindi
Total Samples	9862
Number of Classes	4
Average Sequence Length	18 Tokens
Training Samples (80%)	7890
Validation Samples (10%)	986
Testing Samples (10%)	986
Task	Transliteration-Aware Classification

D. Significance of the Dataset

The Dakshina dataset plays a crucial role in evaluating multilingual transformer architectures because it reflects practical challenges encountered in Hindi language processing. The availability of aligned Devanagari and Romanized text allows researchers to investigate how effectively multilingual transformer models capture contextual and phonetic information across different script representations.

Furthermore, the dataset serves as an important benchmark for low-resource language research by enabling systematic comparison of multilingual models under identical experimental conditions. Consequently, it provides a reliable foundation for assessing the effectiveness of IndicBERT, mBERT, and XLM-R in transliteration-aware Hindi language understanding and classification tasks.

E. Data Pre-processing

Several pre-processing operations are applied before model training:

- Null value removal
- Duplicate elimination
- Unicode normalization
- Transliteration cleaning
- Lowercase conversion

- Tokenization
- Attention mask generation
- Sequence padding

Train–validation–test split these pre-processing techniques improve contextual representation capability and reduce training complexity.

F. Classification Task Formulation

To formulate the transliteration-aware classification task, transliteration pairs from the Dakshina dataset were categorized into four phonetic correspondence classes based on the similarity between Devanagari and Romanized representations represented in table 4.

Table 4: Transliteration Classification Categories

Class Label	Category	Description
0	Exact Transliteration	Romanized text closely follows standard transliteration conventions
1	Phonetic Variant	Minor spelling variations preserving pronunciation
2	Abbreviated Transliteration	Shortened Romanized forms commonly used in informal communication
3	Non-Standard Transliteration	Highly variant spellings differing from standard transliteration patterns

Examples of transliteration variations include:

Class Label	Category	Devanagari	Romanized Form
0	Exact Transliteration	नमस्ते	namaste
1	Phonetic Variant	भारत	bharat
2	Abbreviated Transliteration	धन्यवाद	thnxvad
3	Non-Standard Transliteration	विद्यालय	vidyalya

The resulting labelled dataset enables supervised fine-tuning of multilingual transformer architectures for transliteration-aware Hindi text classification.

IV. MODEL ARCHITECTURE

This research evaluates the performance of three transformer-based multilingual language models—IndicBERT, mBERT, and XLM-R—for Hindi transliteration-aware classification. Although these models differ in their pre-training strategies and multilingual corpora, they share a common transformer encoder architecture based on self-attention mechanisms. The objective is to investigate how different multilingual pre-training approaches influence contextual representation learning and classification performance for transliterated Hindi text.

The overall architecture of the proposed framework is illustrated in Figure 1. The input text is first pre-processed and tokenized according to the tokenizer associated with each transformer model. The tokenized sequences are then converted into contextual embeddings through the transformer encoder. The contextual representation corresponding to the special classification token ([CLS]) is extracted and passed to a fully connected classification layer followed by a Softmax activation function to generate the final class prediction.

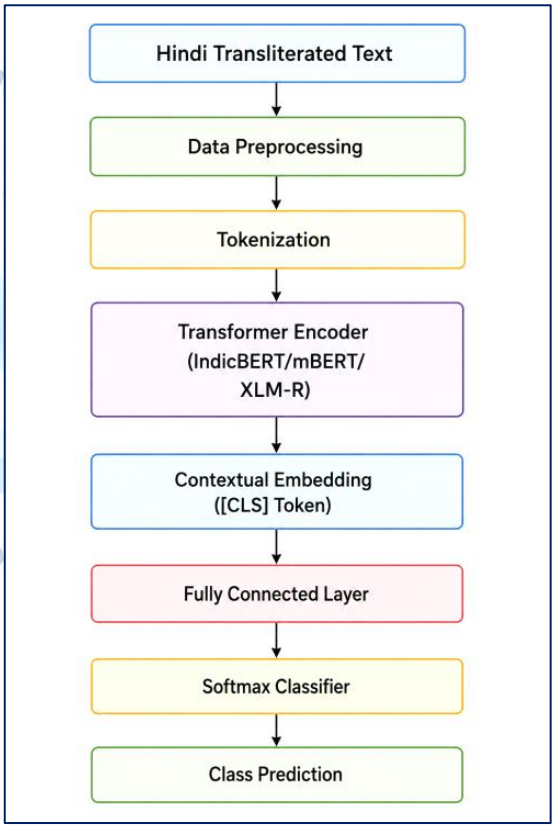

Fig.1 Proposed Multilingual Transformer-Based Classification Framework

The transformer encoder utilizes multi-head self-attention and feed-forward neural networks to capture semantic, syntactic, and contextual relationships among words. Through self-attention mechanisms, the model can learn long-range dependencies and contextual information from transliterated Hindi text, enabling more accurate classification decisions. A standardized pre-processing pipeline was designed

specifically for Romanized Hindi and transliterated text to reduce transliteration ambiguity.

A. Multilingual BERT (mBERT)

Multilingual BERT (mBERT) is an extension of the Bidirectional Encoder Representations from Transformers (BERT) model introduced by Devlin et al. [1]. It is pre-trained on Wikipedia corpora from more than 100 languages using Masked Language Modelling (MLM). The model learns shared multilingual representations that facilitate cross-lingual transfer learning and multilingual language understanding.

For classification, the contextual representation associated with the [CLS] token is extracted from the final encoder layer and provided as input to a dense classification layer. Although mBERT demonstrates strong multilingual capabilities, its performance on Indic languages may be limited because of comparatively lower exposure to Indic-language-specific corpora during pre-training.

B. IndicBERT

IndicBERT is a multilingual transformer model specifically developed for Indian languages using the IndicCorp corpus [3]. The model is designed to address challenges associated with low-resource Indic language processing, including script diversity, limited annotated datasets, and linguistic variations.

Unlike general multilingual models, IndicBERT focuses on learning language representations from Indian language resources, thereby providing improved contextual understanding for Indic languages. Its relatively lightweight architecture also reduces computational requirements, making it suitable for resource-constrained environments.

C. XLM-R

XLM-R (Cross-lingual Language Model-RoBERTa) is a robust multilingual transformer architecture proposed by Conneau et al. [2]. It is pre-trained on a large-scale multilingual corpus comprising hundreds of gigabytes of text collected from multiple languages.

XLM-R employs a RoBERTa-based architecture and benefits from extensive multilingual pre-training, allowing it to learn richer semantic and contextual representations. The model has demonstrated state-of-the-art performance across numerous multilingual benchmarks and low-resource language tasks.

In this research, XLM-R is expected to provide superior performance due to its large-scale multilingual knowledge and enhanced contextual embedding capabilities, particularly when processing transliterated Hindi text.

D. Comparative Analysis of Selected Models

Table 5 presents a comparative summary of the multilingual transformer models evaluated in this study.

Table 5: Comparative Characteristics of Multilingual Transformer Models

Model	Archi-tecture	Pre-training Corpus	Strength	Limitation
Indic BERT	Transformer Encoder	Indic Corp	Efficient Indic language processing	Smaller multi-lingual coverage
mBERT	BERT based Transformer	Multi-lingual Wikipedia	General multi-lingual learning	Limited Indic specia-lization
XLM-R	RoBERTa-based Transformer	Common Crawl Multi-lingual Corpus	Strong contextual representation	High compu-tational cost

The comparative evaluation of these architectures enables systematic investigation of the effectiveness of Indic-specific pre-training versus large-scale multilingual pre-training for Hindi transliteration-aware classification.

V. Experimental Setup

A. Experiment Setup

The experimental implementation was carried out using the Google Colab Pro platform with NVIDIA Tesla T4 GPU support to facilitate efficient fine-tuning of multilingual transformer models. The proposed framework was developed using Python 3.10 and implemented with the PyTorch 2.2 deep learning framework. All experiments were conducted under identical training conditions to ensure fair comparison among the selected transformer architectures, namely IndicBERT, mBERT, and XLM-R.

B. Evaluation Metrics

The following evaluation metrics are used:

- **Accuracy** measures the proportion of correctly classified requirements among all predictions.

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

- **Precision (P)** indicates how many requirements predicted as a specific class actually belong to that class.

$$P = \frac{TP}{TP + FP}$$

- **Recall (R)** measures the model's ability to identify all relevant instances of a class.

$$R = \frac{TP}{TP + FN}$$

- **F1-Score** is the harmonic mean of precision and recall, providing balanced evaluation under class imbalance.

$$F1\ Score\ = 2 * \frac{P * R}{P + R}$$

Where: TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives.

VI. Results and Discussion

A. Results and comparative performance analysis

The comparative evaluation of transformer-based multilingual models represented in figure 2 demonstrates that XLM-R achieves the best overall performance among all evaluated models. XLM-R obtained the highest accuracy (96.34%), precision (96.18%), recall (95.97%), and F1-Score (96.07%), outperforming both mBERT and IndicBERT. Although XLM-R requires higher training time, its superior classification capability makes it the most effective model for multilingual text classification tasks.

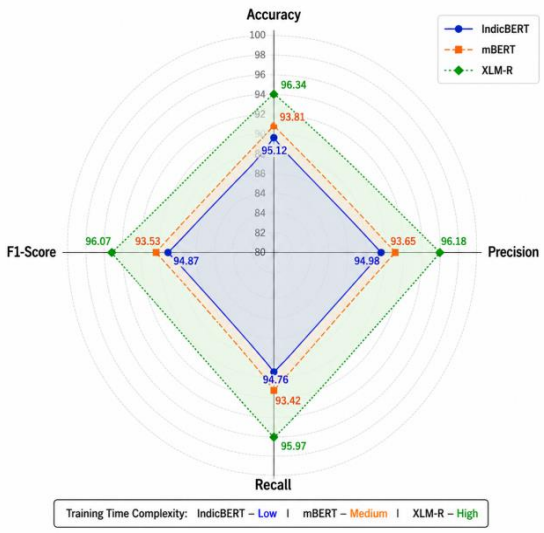

Fig.2 Evaluation of transformer-based multilingual models

XLM-R consistently achieves the highest values across all evaluation metrics, indicating strong generalization capability and robust classification performance. Compared with mBERT, XLM-R improves accuracy by 2.53%, precision by 2.53%, recall by 2.55%, and F1-score by 2.54%. Similarly, when compared with IndicBERT, the performance improvements are notable, with increases of 1.22% in accuracy, 1.20% in precision, 1.21% in recall, and 1.20% in F1-score.

The higher F1-score achieved by XLM-R demonstrates a better balance between precision and recall, making it more reliable for real-world multilingual NLP applications. Although IndicBERT requires low training time and mBERT requires medium training time, their predictive performance remains lower than that of XLM-R.

IndicBERT achieved an accuracy of 95.12%, precision of 94.98%, recall of 94.76%, and F1-score of 94.87%, demonstrating competitive performance while maintaining lower computational requirements. mBERT obtained an accuracy of 93.81%, precision of 93.65%, recall of 93.42%, and F1-score of 93.53%, showing moderate effectiveness in multilingual classification tasks.

Therefore, despite its higher computational complexity and training time requirements, XLM-R provides the most accurate and reliable results for multilingual classification systems. The experimental findings indicate that XLM-R is the most suitable model for achieving high-performance multilingual text classification, particularly in applications where predictive accuracy is prioritized over computational cost.

B. Answers to Research Questions

RQ1: Which multilingual transformer architecture achieves the best performance for hindi transliteration-aware classification?

The experimental results indicate that XLM-R obtained the highest accuracy (96.34%), precision (96.18%), recall (95.97%), and F1-Score (96.07%), outperforming both mBERT and indicBERT. Therefore, XLM-R is the most effective model for hindi transliteration-aware classification.

RQ2: How does large-scale multilingual pre-training influence contextual representation learning in transliterated Hindi text?

The higher performance of XLM-R demonstrates that large-scale multilingual pre-training significantly enhances contextual representation learning. Exposure to extensive multilingual corpora enables the model to better capture semantic and phonetic variations present in transliterated Hindi text.

RQ3: Can Indic-specific pre-training provide competitive performance compared with general multilingual transformer architectures?

IndicBERT achieved an accuracy of 95.12%, precision of 94.98%, recall of 94.76%, and F1-Score of 94.87%, demonstrating that Indic-specific pre-training is highly effective for Hindi language processing and multilingual text classification. Its performance is remarkably close to that of XLM-R, with a difference of only about 1.2% across all evaluation metrics. Although it did not surpass XLM-R. Nevertheless, indicBERT offers a computationally efficient alternative for resource-constrained environments.

VII. CONCLUSION AND FUTURE SCOPE

This study presented a comprehensive comparative evaluation of three multilingual transformer models- IndicBERT, mBERT, and XLM-R for Hindi transliteration-aware text classification using the Dakshina dataset. A standardized pre-processing framework incorporating text normalization, transliteration cleaning, tokenization, and sequence padding was employed to improve contextual representation learning.

Experimental results demonstrated that XLM-R achieved the highest overall performance, attaining an accuracy of 96.34% and significantly outperforming IndicBERT and mBERT across all evaluation metrics. The findings indicate that large-scale multilingual pre-training substantially enhances contextual understanding and classification performance in transliterated Hindi environments. Although IndicBERT achieved comparatively lower accuracy, it provided a computationally efficient alternative for resource-constrained Indic NLP applications.

The study contributes a benchmark comparison of multilingual transformer architectures for transliteration-aware Hindi classification and provides insights into the effectiveness of multilingual versus Indic-specific pre-training strategies. Future research can be focus on code-mixed hindi-english NLP systems and multilingual generative AI for Indic languages.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), Minneapolis, USA, 2019, pp. 4171–4186.

[2] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised Cross-Lingual Representation Learning at Scale," in Proceedings of ACL 2020, pp. 8440–8451.

[3] D. Kakwani, A. Kunchukuttan, S. Golla, G. N. C., A. Bhattacharyya, M. M. Khapra, and P. Kumar, "IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages," Findings of EMNLP 2020, pp. 4948–4961.

[4] 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, 2019.

[5] M. Artetxe and H. Schwenk, "Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond," Transactions of the Association for Computational Linguistics, vol. 7, pp. 597–610, 2019.

[6] B. Roark, L. Wolf-Sonkin, C. Kirov, S. J. Mielke, C. Johny, I. Demirsahin, and K. Hall, "Processing South Asian Languages Written in the Latin Script: The Dakshina Dataset," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 2413–2423.

[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv:1907.11692, 2019.

[8] K. Jain, A. Deshpande, K. Shridhar, F. Laumann, and A. Dash, "Indic-Transformers: An Analysis of Transformer Language Models for Indian Languages," arXiv:2011.02323, 2020.

[9] A. Kunchukuttan, D. Kakwani, S. Golla, G. N. C., A. Bhattacharyya, M. M. Khapra, and P. Kumar, "AI4Bharat-IndicNLP Corpus: Monolingual Corpora and Word Embeddings for Indic Languages," arXiv:2005.00085, 2020.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.

[11] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, et al., "Transformers: State-of-the-Art Natural Language Processing," in Proceedings of EMNLP: System Demonstrations, 2020, pp. 38–45.

[12] Y. Madhani, S. Parthan, P. Bedekar, G. N. C., R. Khapra, A. Kunchukuttan, P. Kumar, and M. M. Khapra, "Aksharantar: Open Indic-Language Transliteration Datasets and Models for the Next Billion Users," arXiv:2205.03018, 2022.

[13] A. Pfeiffer, I. Vulić, I. Gurevych, and S. Ruder, "MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer," in Proceedings of EMNLP 2020, pp. 7654–7673.

[14] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in International Conference on Learning Representations (ICLR), 2022.

[15] S. B. K. Vulić, G. Glavaš, and A. Korhonen, "Multilingual Language Models: A Survey," ACM Computing Surveys, vol. 57, no. 2, pp. 1–38, 2025.

[16] S. Khanuja, D. Bansal, S. Mehtani, et al., "MuRIL: Multilingual Representations for Indian Languages," arXiv:2103.10730, 2021.

[17] G. Singh, A. Kunchukuttan, P. Kumar, and M. M. Khapra, "IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages," Transactions of the Association for Computational Linguistics, 2024.

[18] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," Proceedings of ACL 2018, pp. 328–339.

[19] F. Chollet, Deep Learning with Python, 2nd ed. Shelter Island, NY, USA: Manning Publications, 2021.

[20] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.