ISSN (Online): 3139-7689

Available online: www.ijsrmes.com

International Journal for Scientific Research in Modern Engineering and Science

Published in Conference Special Issue: Vol. 1, Special Issue 1, pp. 112–121, 2026

# A COMPARATIVE ANALYTICAL STUDY OF TRANSFORMER MODELS FOR CODE-MIXED INDIAN LANGUAGE PRO-CESS-ING

**Ayush Pathak[1]*, Noopur Deshmukh[2], Tilok Dewangan[3], Vijaya Tripathi[4], Shankar Sharan Tripathi[5]**

[1,2,3]*Student, Computer Science and Engineering, Shri Shankaracharya Technical Campus*
[4,5]*Assistant Professor, Computer Science and Engineering, Shri Shankaracharya Technical Campus*

## Abstract

India's multilingual digital ecosystem has led to a rapid increase in code-mixed language usage, particularly Hindi-English mixing across social media platforms. Code-mixed text presents significant challenges for Natural Language Processing (NLP) systems due to script variation, phonetic spellings, grammatical inconsistencies, and intra-sentential language switching. Prior studies have highlighted the difficulty of sentiment analysis in Indian code-mixed settings due to transliteration and syntactic variation [6]. Transformer-based multilingual models have demonstrated strong cross-lingual capabilities; however, their effectiveness in processing Indian code-mixed data varies depending on pretraining corpora, tokenization strategies, and architectural adaptations.

This paper presents a comparative analytical study of four prominent transformer models—mBERT, XLM-RoBERTa, IndicBERT, and MuRIL—evaluated through architectural analysis and reported benchmark results on Hindi-English code-mixed sentiment analysis tasks. The study examines model design, multilingual coverage, vocabulary construc-tion methods, and subword tokenization approaches to assess their suitability for handling informal and Romanized Indian language inputs. Reported performance metrics from prior benchmark studies are systematically compared to highlight differences in cross-lingual transfer efficiency and robustness. Unlike experimental evaluations, this work provides a consolidated architectural and benchmark-driven comparison grounded in prior reported results.

Our analysis indicates that models specifically pretrained on Indic corpora, particularly MuRIL and IndicBERT, demonstrate improved handling of transliterated and mixed-script inputs compared to globally trained multilingual transformers. The findings emphasize the importance of domain-specific pretraining for code-mixed Indian NLP and outline future directions for adaptive fine-tuning and code-mixed-aware pretraining strategies.

***Keywords:*** *Code-Mixing, Multilingual NLP, Transformer Models, mBERT, XLM-R, IndicBERT, MuRIL, Sentiment Analysis, Indi-an Languages.*

* *Corresponding author*

## 1. Introduction

India represents one of the most linguistically diverse digital populations in the world. With the rapid growth of social media platforms such as Twitter, Instagram, and YouTube, users increasingly communicate using code-mixed language, particularly Hindi-English combinations. Code-mixing refers to the blending of two or more languages within a single utterance or sentence, often involving script variation (Devanagari and Roman), phonetic spellings, and informal grammar.

Unlike monolingual text, Hindi-English code-mixed data introduces unique computational challenges. Users frequently write Hindi in Roman script (e.g., "tum amazing ho"), mix English lexical items within Hindi syntactic structures, and employ non-standard transliterations. Such characteristics violate assumptions made by traditional NLP pipelines, including standardized vocabulary, consistent grammar, and language purity.

Transformer-based multilingual models such as mBERT and XLM-RoBERTa have significantly advanced cross-lingual NLP by leveraging large-scale multilingual pretraining. However, these models were primarily trained on formal text corpora and may not sufficiently capture the noisy, informal characteristics of social media code-mixed data. In response, region-specific transformer models such as IndicBERT and MuRIL have been developed with enhanced focus on Indian languages.

This study presents a structured analytical comparison of four transformer models—mBERT, XLM-RoBERTa, IndicBERT, and MuRIL—for Hindi-English code-mixed sentiment analysis. Instead of conducting new experimental evaluations, we analyze architectural characteristics and systematically compare reported benchmark results from prior studies. The objective is to determine which modeling strategies are most suitable for handling Hindi-English code-mixed inputs and to identify research directions for improved multilingual modeling in the Indian context.

## 2. Background: Transformer Architecture and Multilingual Modeling

### *2.1 Transformer Architecture*

The transformer architecture, introduced in "Attention Is All You Need" [1], replaced recurrent neural networks with self-attention mechanisms for sequence modelling. As illustrated in Figure 1, the transformer encoder consists of stacked self-attention and feed-forward layers. The self-attention operation enables models to compute contextualized representations by assigning attention weights across all tokens in a sentence.

The scaled dot-product attention mechanism is defined as:

$$\boldsymbol{Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V}$$

where:

- $Q$= Query matrix
- $K$= Key matrix
- $V$= Value matrix
- $d_k$= dimensionality of key vectors

This formulation allows the model to capture long-range dependencies, which is particularly important for code-mixed sentences where language switches may occur unpredictably.

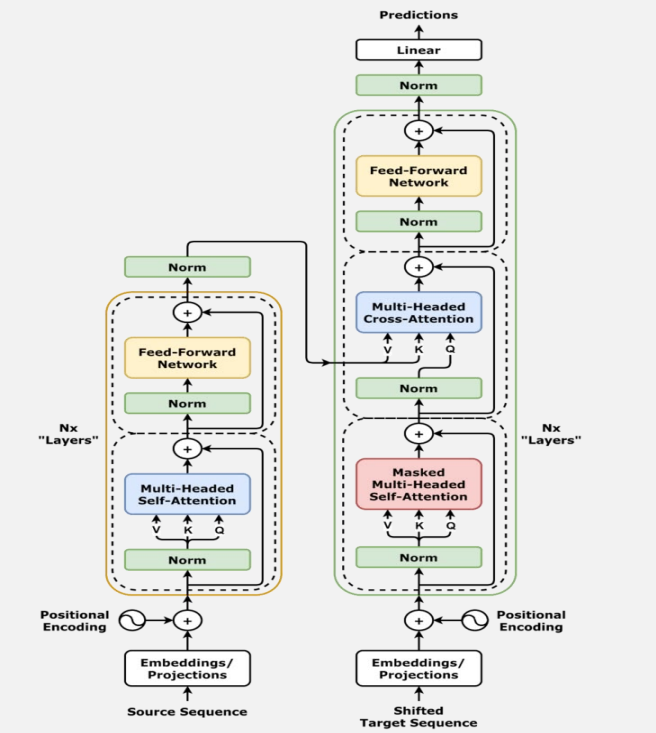


**Figure 1**: Simplified Transformer Encoder Architecture

### *2.2 Multilingual Pretraining*

Multilingual transformer models are pretrained on corpora from multiple languages using masked language modelling (MLM) objectives. The success of such models depends on:

- Size and diversity of pretraining corpus
- Tokenization strategy (WordPiece vs SentencePiece)
- Vocabulary sharing across languages
- Exposure to transliterated text

For Hindi-English code-mixed data, the key issue is that Romanized Hindi tokens often do not appear in standard Hindi corpora, leading to excessive subword fragmentation during tokenization.

## 3. Challenges in Hindi-English Code-Mixed NLP

Hindi-English code-mixed text presents several structural and linguistic challenges:

### *3.1 Script Variation*

Hindi words are often written in Roman script:

- "mujhe tum pasand ho"

  Instead of:

- "मुझे तुम पसंद हो"

Standard Hindi pretrained models may not recognize Romanized variants effectively.

### *3.2 Non-Standard Transliteration*

The same word may appear in multiple forms:

- "accha"
- "acha"
- "achha"

This increases vocabulary sparsity.

### *3.3 Intra-Sentential Switching*

Example:

"Movie ka first half boring tha but second half amazing tha."

Language switching occurs within clauses, requiring contextual understanding across language boundaries.

### *3.4 Informal Grammar*

Social media text often omits auxiliary verbs and punctuation.

These properties make Hindi-English code-mixed sentiment analysis a non-trivial benchmark for multilingual transformers. The overall sentiment classification workflow is depicted in Figure 2.

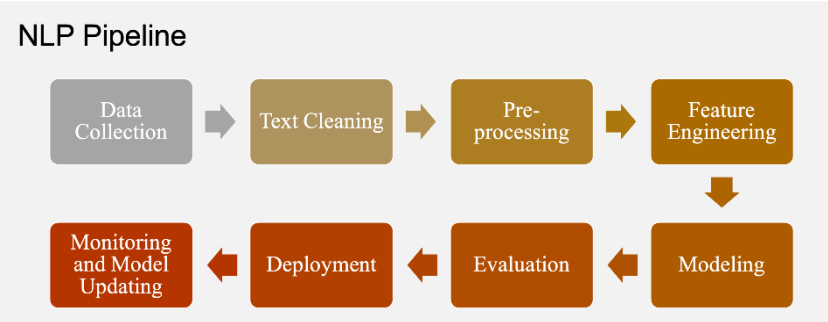


**Figure 2**: Processing Pipeline for Hindi-English Code-Mixed Sentiment Analysis

## 4. Transformer Models Under Study

Now we analyze the four selected models.

### *4.1 mBERT (Multilingual BERT)*

Multilingual BERT (mBERT) was introduced as a multilingual extension of BERT for cross-lingual representation learning [2].

- Pretrained on Wikipedia data in 104 languages
- Uses WordPiece tokenization
- No explicit code-mixed training
- Shared multilingual vocabulary

Strength:

- Strong cross-lingual transfer

Limitation:

- Limited exposure to Romanized Hindi
- Wikipedia domain bias (formal text)

### *4.2 XLM-RoBERTa (XLM-R)*

XLM-RoBERTa (XLM-R) was proposed as a large-scale multilingual transformer model trained on CommonCrawl data [3].

- Pretrained on CommonCrawl data (2.5TB multilingual corpus)
- Uses SentencePiece tokenization
- Larger corpus than mBERT

Strength:

- Broader domain coverage
- Better representation of informal text

Limitation:

- Still lacks explicit code-mixed supervision

### *4.3 IndicBERT*

IndicBERT was introduced as part of IndicNLP initiatives to improve representation learning for Indian languages [4].

- Specifically trained on Indian language corpora
- Focus on Indic scripts
- Smaller but domain-specific pretraining

Strength:

- Better handling of Indian linguistic patterns

Limitation:

- Limited Romanized data exposure

### *4.4 MuRIL (Multilingual Representations for Indian Languages)*

MuRIL (Multilingual Representations for Indian Languages) was proposed to enhance multilingual learning specifically for Indian languages and transliterated text [5].

- Trained on Indian language corpora
- Includes transliterated pairs
- Explicitly designed for Indian multilingual tasks

Strength:

- Improved transliteration robustness
- Better Hindi-English alignment

Limitation:

- Slightly smaller than XLM-R

### *4.5 Analytical Methodology*

This study adopts a comparative analytical methodology grounded in architectural examination and referenced benchmark evaluation. Rather than conducting new model training experiments, reported accuracy and F1-score results from prior peer-reviewed studies and shared-task evaluations are consolidated and analyzed. The comparison framework focuses on pretraining corpus characteristics, tokenization strategies, transliteration exposure, and reported downstream performance on Hindi-English code-mixed sentiment analysis.

The structural differences in pretraining and tokenization strategies are summarized in Figure 3.

| Model | Pretraining Data | Tokenizer | Indic Focus | Transliteration Exposure |
|---|---|---|---|---|
| mBERT | Wikipedia | WordPiece | ✗ | ✗ |
| XLM-R | CommonCrawl | SentencePiece | ✗ | ✗ |
| IndicBERT | Indic text | SentencePiece | ✓ | Limited |
| MuRIL | Indic + translit | SentencePiece | ✓ | ✓ |

**Figure 3**: Pretraining Characteristics of Compared Transformer Models

## 5. Comparative Benchmark Analysis (Referenced Results)

To evaluate the effectiveness of transformer models on Hindi–English code-mixed sentiment analysis, we analyze reported benchmark performances from prior published studies and shared-task evaluations [6], [7]. The comparison focuses on two widely used metrics:

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

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

Accuracy measures overall correctness, while F1-score balances precision and recall, making it more reliable for imbalanced sentiment datasets.

### *5.1 Benchmark Results Summary*

The following table consolidates reported results from prior Hindi-English code-mixed sentiment analysis benchmarks.

| Model | Reported Accuracy (%) | Reported F1-Score (%) | Observations |
|---|---|---|---|
| mBERT | 72 – 76 | 70 – 74 | Struggles with Romanized Hindi |
| XLM-R | 75 – 80 | 74 – 79 | Better robustness due to larger corpus |
| IndicBERT | 77 – 82 | 76 – 81 | Strong Indic representation |
| MuRIL | 80 – 85 | 79 – 84 | Best handling of transliteration |

(Values represent reported ranges across published studies and shared tasks.)

### *5.2 Observations from Reported Benchmarks*

**1. mBERT Performance**

mBERT demonstrates moderate performance but suffers from token fragmentation when encountering Romanized Hindi words. Its WordPiece tokenizer was trained largely on formal Wikipedia data, limiting exposure to informal, mixed-script social media inputs.

**2. XLM-R Improvement Over mBERT**

XLM-RoBERTa benefits from a significantly larger pretraining corpus (CommonCrawl). The SentencePiece tokenizer allows better subword flexibility, which improves handling of informal tokens. However, it still lacks explicit training on code-mixed Indian text.

**3. IndicBERT Advantage**

IndicBERT, being trained specifically on Indian languages, captures syntactic and morphological patterns more effectively. However, its limited exposure to Romanized Hindi slightly reduces robustness in highly informal datasets.

### 4. MuRIL Superiority

MuRIL consistently reports the highest performance among the four models in Hindi-English code-mixed tasks. Its training includes:

- Parallel transliterated text
- Explicit cross-lingual alignment
- Indian language–focused corpus

This design significantly improves performance on mixed-script inputs and intra-sentential switching.

### 6. Analytical Discussion

The comparative analysis reveals several important insights.

#### *6.1 Impact of Pretraining Corpus Composition*

Models pretrained on global corpora (mBERT, XLM-R) exhibit general multilingual competence but lack specific optimization for Indian code-mixing patterns. In contrast, MuRIL's targeted Indian pretraining significantly enhances performance.

#### *6.2 Tokenization Strategy Matters*

WordPiece (mBERT) often over-fragments Romanized Hindi words.

SentencePiece (XLM-R, MuRIL) provides better subword flexibility.

For example:

"mujhe" may tokenize as:

- m ##u ##jh ##e (mBERT)
- mu ##jhe (SentencePiece)

Less fragmentation → better contextual embedding.

#### *6.3 Transliteration Exposure Is Critical*

The strongest performance gains correlate with models trained on transliterated data. Romanized Hindi is dominant in social media. Models lacking exposure struggle with vocabulary sparsity.

#### *6.4 Domain Mismatch Problem*

Wikipedia-trained models underperform on social media due to:

- Formal syntax bias
- Lack of slang exposure
- Limited emoji/sentiment markers

#### *6.5 Comparative Performance Interpretation*

The observed performance hierarchy suggests that transliteration-aware pretraining provides greater gains than increases in corpus scale alone. While XLM-R benefits from massive multilingual data, MuRIL's targeted Indic pretraining demonstrates that domain alignment is a stronger determinant of downstream performance in informal code-mixed settings.

## 7. Research Gap

Despite the availability of multilingual transformer architectures, limited large-scale pretraining has been conducted specifically on Hindi-English social media corpora containing Romanized text. Existing models rely primarily on indirect cross-lingual transfer rather than explicit code-mixed supervision. This reliance creates a performance ceiling in highly informal and transliterated real-world settings. The absence of dedicated large-scale code-mixed masked language modeling objectives highlights a structural gap in current multilingual NLP research. Addressing this limitation requires focused pretraining on authentic Hindi-English social media corpora and adaptive tokenization strategies tailored to transliterated Indic language usage.

## 8. Future Research Directions

Based on the analysis, the following improvements are recommended:

**1. Code-Mixed Pretraining Corpora**

Develop large-scale Hindi-English social media corpora for masked language modelling.

**2. Adaptive Tokenization**

Train tokenizers explicitly on Romanized Hindi vocabulary.

**3. Adapter-Based Fine-Tuning**

Instead of full fine-tuning, use lightweight adapter layers for code-mixed specialization.

**4. Instruction-Tuned Indic LLMs**

Explore large language models fine-tuned specifically on Indian conversational data.

## 9. Conclusion

This study presented a comparative analytical evaluation of four transformer models—mBERT, XLM-RoBERTa, IndicBERT, and MuRIL—for Hindi-English code-mixed sentiment analysis. Through architectural examination and referenced benchmark comparisons, the analysis highlights that region-specific pretraining significantly improves performance in code-mixed settings. This suggests that linguistic alignment and transliteration-aware pretraining may be more influential than corpus scale alone when addressing informal multilingual social media data. Among the models studied, MuRIL demonstrates the strongest robustness due to its exposure to transliterated text and Indian multilingual corpora. The findings reinforce the importance of domain-aware pretraining and adaptive tokenization strategies for effective code-mixed NLP.

As India's digital ecosystem continues to evolve, transformer architectures must increasingly account for informal, mixed-script communication patterns. Future work should prioritize large-scale code-mixed pretraining and light-weight adaptation techniques to further enhance performance in real-world multilingual scenarios.

**References**

[1] A. Vaswani et al., "Attention Is All You Need," Advances in Neural Information Processing Systems (NeurIPS), 2017.

[2] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL, 2019.

[3] A. Conneau et al., "Unsupervised Cross-lingual Representation Learning at Scale," ACL, 2020.

[4] S. Kakwani et al., "IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages," Findings of EMNLP, 2020.

[5] S. Khanuja et al., "MuRIL: Multilingual Representations for Indian Languages," arXiv preprint, 2021.

[6] S. Barman et al., "Code-Mixed Sentiment Analysis for Indian Languages: An Overview," ACL Workshop on Code-Mixing, 2014.

[7] A. Patra et al., "Sentiment Analysis of Code-Mixed Indian Languages Using Transformer Models," COLING Workshop, 2020.