IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, January 2026
Impact Factor: 7.67

Comparative Analysis of Word Embeddings on Transformer Model for Emotion Recognition in Indic Code-Mixed Hinglish

Pragati Mulik[1] and Dr. S. S. Sonawane[2]
M.Tech. Student, Department of Computer Engineering[1]
Professor, Department of Computer Engineering[2]
SCTR's Pune Institute of Computer Technology, Pune,India

Abstract: *Hinglish, a code-mixed blend of Hindi and English, has become increasingly common in digital communication across platforms such as WhatsApp, Instagram, and Twitter. Its informal grammatical structure, transliterated Hindi tokens in Roman script, and frequent language switching pose significant challenges for traditional NLP systems trained on monolingual corpora. This paper presents a comparative approach for Hinglish emotion recognition using four embedding-classifier combinations, namely Skip-gram + LLaMA, CBOW + LLaMA, BERT + LLaMA, and SBERT + LLaMA. A dataset of 16,000 Hinglish sentences annotated with seven emotion categories-joy, anger, sadness, fear, love, surprise, and neutral was used for experimentation. A specialized preprocessing pipeline was developed to address transliteration inconsistencies and spelling variations. The models were evaluated using accuracy and weighted F1-score. Among the methods tested, the CBOW + LLaMA model achieved the highest performance, followed closely by the BERT + LLaMA model. The study highlights the suitability of context-preserving embeddings for code-mixed Indic text and supports the development of practical emotion-aware NLP systems for multilingual Indian users.*

Keywords: Hinglish; Code-Mixed Text; Emotion Recognition; Multilingual NLP; Deep Learning; Transformer Models; LLaMA

I. INTRODUCTION

Hinglish, a hybrid of Hindi and English written in Roman script, has emerged as a common mode of communication on digital platforms such as WhatsApp, Instagram, and Twitter. In such environments, users frequently switch between languages, employ informal grammar, adopt phonetic spellings, and make use of transliterated Hindi words. These characteristics introduce significant linguistic irregularities, making emotion recognition more complex than standard monolingual NLP tasks.

Although recent NLP research has explored sentiment and emotion analysis for Indic languages, most approaches remain tailored to either English or Hindi, and therefore struggle to capture semantic cues present in code-mixed Hinglish text. The lack of standardized datasets, inconsistent transliteration patterns, and limited evaluation of embedding methods have further restricted progress in this area.

In response to this gap, the present work investigates a comparative framework for Hinglish emotion recognition using four embedding-classifier combinations: Skip-gram + LLaMA, CBOW + LLaMA, BERT + LLaMA, and SBERT + LLaMA. The objective is to determine which embedding strategy better preserves emotional semantics in code-mixed text and delivers improved performance for seven emotion categories. Experimental analysis demonstrates that contextstable embeddings yield higher accuracy than static word representations, highlighting the importance of embedding selection for code-mixed NLP applications.

DOI: 10.48175/IJARSCT-31012

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, January 2026
Impact Factor: 7.67

II. RELATEDWORK

Early work on Hinglish emotion analysis relied on classical machine learning approaches such as SVM, Naïve Bayes, and Random Forest using TF-IDF or bag-of-words features [11], [12]. These studies achieved moderate accuracy but were limited by small datasets and shallow lexical representations, which made it difficult to capture emotional nuances in transliterated and informal text.

With the introduction of deep learning, hybrid CNN-LSTM architectures were explored for Hinglish social media content and reported performance improvements, although challenges remained in handling slang, emoji-text blends, and transliteration inconsistencies [3]. Later, transformer-based models such as BERT and mBERT demonstrated stronger contextual understanding through subword tokenization and attention mechanisms [9], and multitask learning further improved performance for sentiment–emotion joint tasks, though minority emotion classes remained difficult due to class imbalance [7].

More recent studies have experimented with hybrid transformer pipelines trained on larger Hinglish datasets and achieved competitive results [2]. Survey work on Indic NLP continues to highlight persistent limitations such as lack of standardized datasets, inconsistent preprocessing pipelines for Romanized Hindi, and limited interpretability of model outputs [4]. Benchmark evaluations in SemEval-2024 and SIGTYP-2024 also show that multilingual language models still underperform on low-resource and code-mixed settings, reinforcing the need for Hinglish-specific modeling approaches [5], [6].

III. DATASET DESCRIPTION

The dataset used in this study consists of 16,000 Hinglish (Hindi-English code-mixed) text samples, each annotated with one of seven emotion categories: joy, anger, sadness, fear, love, surprise, and neutral. The text samples are short, informal social media-style utterances that include code-switching, transliterated Hindi written in Roman script, and nonstandard spellings. Example expressions include *"mujhekisi se p hogaya and I am very happy"* (joy) and *"main thodabechainhoon because I became upset"* (anger).

The dataset spans a mixture of English lexical tokens and Romanized Hindi tokens (e.g., *udaasi, mujhe, hairan*), making it suitable for evaluating both static embedding models and contextual transformer-based approaches. The data was stored in CSV format with two fields: sentence (text input) and emotion (target label). For experimentation, the dataset was partitioned into 70% training, 15% validation, and 15% testing splits. The label distribution reflects naturally occurring usage frequencies in social media communication, rather than artificially balanced sampling.

IV. METHODOLOGY

A. Preprocessing

Hinglish text samples were normalized through a preprocessing pipeline designed to handle informal and code-mixed expressions. This pipeline removed URLs, hashtags, emojis, user mentions, repeated characters, and HTML artifacts, followed by lowercasing and punctuation standardization. Transliterated Hindi spellings were corrected to reduce variability, and tokenization procedures were adapted for both static word embedding models and transformer-based models. This ensured the generation of consistent and clean input suitable for Word2Vec and transformer architectures.

B. Embedding Approaches

Four embedding–classifier configurations were evaluated in this work: Skip-gram + LLaMA, CBOW + LLaMA, BERT + LLaMA, and SBERT + LLaMA. Skip-gram and CBOW were used to generate static word embeddings, which were aggregated at the sentence level through mean pooling. In contrast, BERT provided contextualized representations using the pooled **[CLS]** token, while SBERT generated fixed-length semantic sentence embeddings optimized for downstream classification tasks. All embeddings were subsequently processed by the LLaMA classifier to predict one of the seven emotion categories through a softmax layer.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, January 2026
Impact Factor: 7.67

1) Skip-gram + LLaMA

The Skip-gram model learns word embeddings by predicting surrounding context words given a target word. It is effective for infrequent and transliterated Hinglish tokens.

Working Example

For the example sentence, Skip-gram generates (target → context) pairs such as:

bahut → {mujhe, gussa}

gussa → {bahut, aa}

raha → {aa, hai}

Each word is mapped to a dense vector. Example embeddings:

bahut → [0.18, –0.21, 0.44, …]

gussa → [0.52, 0.11, –0.33, …]

A sentence embedding is computed using simple averaging:

$$\vec{S} = \frac{1}{N}\sum_{i=1}^{N} \vec{w}_i$$

This sentence vector is fed into the LLaMA classifier, which learns emotional patterns.

2) CBOW + LLaMA

Continuous Bag of Words (CBOW) predicts the center word from surrounding context words. It produces stable embeddings for frequently occurring Hinglish tokens.

Working Example

CBOW forms training examples of the form:

{mujhe, gussa} → bahut

{bahut, aa} → gussa

{aa, hai} → raha

CBOW computes a context vector:

$$\vec{C} = \frac{1}{k}\sum_{i=1}^{k} \vec{w}_i$$

and predicts:

$$P(w_t \mid context) = \text{softmax}(W \cdot \vec{C})$$

A final sentence vector is produced by averaging all word embeddings, and passed into LLaMA for classification. Strength: Performs well on frequent Hinglish patterns such as *bahut gussa, aa rahahai*.

3) BERT + LLaMA

BERT generates contextual embeddings using the self-attention mechanism. It captures meaning based on the entire sentence, which is crucial for code-mixed and noisy Hinglish text.

Working Example

Input representation:

[CLS] mujhe bahut gussa aa rahahai but I am trying to stay calm [SEP] WordPiece tokenization handles Hinglish variations:

"raha" → ["ra", "##ha"]

"mujhe" → ["mu", "##jhe"]

BERT produces a contextual vector for every token, and the CLS token is used as the final sentence representation:

$$\vec{S}_{CLS} = \text{BERT}_{CLS}(sentence)$$

DOI: 10.48175/IJARSCT-31012

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, January 2026
Impact Factor: 7.67

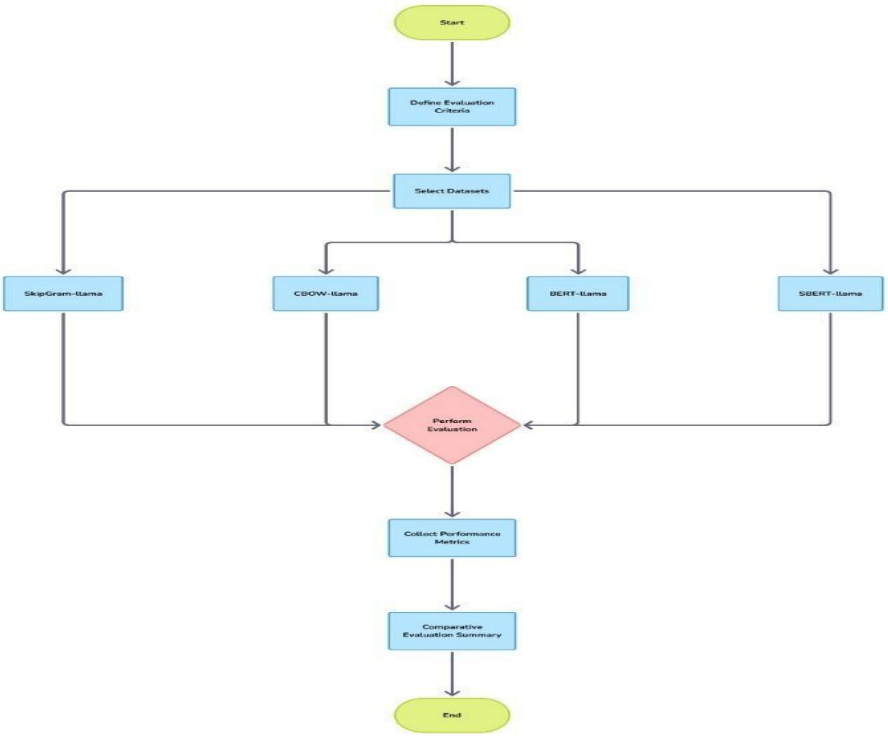
This CLS embedding is passed into LLaMA, which learns emotion-specific patterns such as:
"bahut gussa aa rahahai" → strong anger
"trying to stay calm" → secondary neutral context
Strength: Captures long-range dependencies and code-mixed semantics.

4) SBERT + LLaMA
SBERT creates a single fixed-length embedding for the entire sentence using a Siamese network architecture. It is optimized for semantic similarity and classification tasks.
Working Example
SBERT directly outputs one vector:

$$\vec{S}SBERT = f\text{SBERT}(sentence)$$

Example embedding produced:
[0.52, -0.09, 0.71, -0.33, …] (768 dimensions)
This embedding captures:
emotional polarity (anger),
mitigating expressions (stay calm),
code-mixed structure (Hindi + English).
The embedding is then fed into LLaMA for final classification.
Strength: Generates strong semantic representations and is robust to noisy Hinglish spellings.

Architecture Diagram



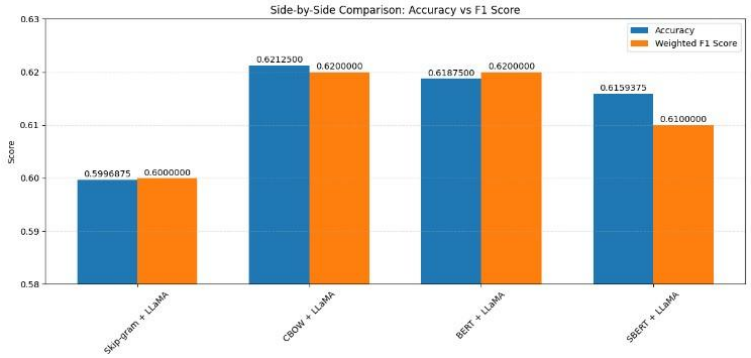
V. TRAINING AND EVALUATION

The dataset was partitioned into 70% training, 15% validation, and 15% testing splits. Model performance was assessed using commonly adopted evaluation metrics for multi-class classification, including accuracy, precision, recall, F1-

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, January 2026
Impact Factor: 7.67

score, and both macro and weighted F1. In addition, a confusion matrix was employed to analyze class-wise performance and identify prediction inconsistencies across emotion categories. These metrics provided a comprehensive assessment of the models under class imbalance and heterogeneous linguistic inputs present in Hinglish data.

VII. RESULTSANDANALYSIS



A comparative analysis of the four embedding–classifier configurations is shown in Fig. 2. Among the evaluated models, CBOW + LLaMA achieved the highest performance with an accuracy of 0.62125 and a weighted F1-score of 0.62000.

The BERT + LLaMA configuration performed competitively with an accuracy of 0.61875, followed by SBERT + LLaMA and Skip-gram + LLaMA, which yielded comparatively lower scores. These results indicate that context-stable embedding approaches such as CBOW and transformer-based embeddings provide better emotion recognition performance for Hinglish text than static word-level embeddings. The trends observed across accuracy and weighted F1scores further confirm the importance of preserving contextual and semantic cues in code-mixed environments.

VII. CONCLUSION

This study presented a comparative evaluation of four embedding–classifier configurations for Hinglish emotion recognition using Skip-gram + LLaMA, CBOW + LLaMA, BERT + LLaMA, and SBERT + LLaMA. The results indicate that embedding selection plays a significant role in modeling code-mixed Hinglish text, with CBOW + LLaMA achieving the highest accuracy among the tested approaches. The performance of CBOW suggests that stable word-level context, when combined with transformer-based classification, can be effective for emotion recognition in informal, transliterated Hinglish communication. These findings underscore the importance of accounting for linguistic properties of code-mixed text when designing NLP systems for multilingual social media environments.

REFERENCES

[1]. R. Mahajan, "Performance of ML and DL Models on Hindi–English Text," Procedia Computer Science, vol. 218, pp. 1652–1661, 2025.

[2]. S. Patankar and M. Kumar, "A CNN–Transformer Framework for Emotion Recognition in Code-Mixed English–Hindi Data," Journal of Big Data Analytics in Linguistics, vol. 12, no. 1, pp. 50–62, 2025.

[3]. V. N. Malavade, S. Gaikwad, S. Pawar, and S. Jadhav, "Analysis of Emotion Detection From Code-Mixed or Code-Switched Social Media Text Using Deep Learning," SSRN Electronic Journal, Nov. 2024.

[4]. H. Kaur and S. Gupta, "A Technical Review on Emotion Detection From Informal Text for Indian Regional Languages," Int. J. Emerg. Trends Eng. Res., vol. 8, no. 1, pp. 323–330, 2024.

[5]. H. Hidetsune, "An English-Based Approach to Emotion Recognition in Hindi–English Code-Mixed Conversations Using Machine Learning and Machine Translation," Proc. SemEval-2024 Task 10: Emotion Discovery & Reasoning Across Languages, 2024.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, January 2026
Impact Factor: 7.67

[6]. J. Ortega, J. Bjerva, and I. Augenstein, "Understanding Sociolinguistic Factors in Multilingual Embeddings: Insights From Code-Switching and Typology," Proc. SIGTYP-2024 Workshop, pp. 9–18, 2024.

[7]. S. Ghosh et al., "Multitasking of Sentiment Detection and Emotion Recognition in Code-Mixed Hinglish Data," Knowledge-Based Systems, vol. 260, p. 110182, 2023.

[8]. V. K. Jaiswal et al., "A Deep Neural Framework for Emotion Detection in Hindi Textual Data," Int. J. Interpreting Enigma Engineers, vol. 2, no. 2, pp. 36–47, 2025.

[9]. Wadhawan and A. Aggarwal, "Towards Emotion Recognition in Hindi–English Code-Mixed Data: A Transformer-Based Approach," Proc. WASSA-2021, pp. 192–198, 2021.

[10]. N. Garg and K. Sharma, "Multilingual Sentiment and Emotion Analysis Using MSIL-Based Architecture," Indian Journal of Science and Technology, vol. 13, no. 40, pp. 4216–4224, 2020.

[11]. T. T. Sasidhar, P. Premjith, and K. P. Soman, "Emotion Detection in Hinglish (Hindi + English) CodeMixed Social Media Text," Procedia Computer Science, vol. 171, pp. 1346–1352, 2020.

[12]. D. Vijay et al., "Corpus Creation and Emotion Prediction for Hindi–English Code-Mixed Social Media Text," Proc. NAACL Student Research Workshop, 2018.