iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.79891

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Sentiment Analysis of Code-Mixed Hinglish Text: A Machine Learning Approach for Social Media Reviews

Mannat Nandi, Khushal Patil, Yashika Patel, Aakansha Patil, Asst. Prof. Shital Cheke
Department of Computer Science & Engineering (Data Science) Vidyavardhini's College of Engineering & Technology, Vasai Road

Abstract: ***The rapid growth of social media has resulted in an enormous volume of user opinions expressed in textual form. In the Indian digital space, these opinions are often written in Hinglish, a code-mixed language that combines Hindi and English words using the Roman script. Conventional Natural Language Processing (NLP) techniques are typically designed for standard English text or pure Hindi text or any other specific language and therefore face difficulties when processing such mixed-language expressions. This research investigates the problem of sentiment identification in Hinglish comments commonly found on online platforms. The study explores the linguistic characteristics of code-mixed text, including inconsistent spelling, informal slang, and mixed grammatical structures. To address these challenges, the research examines machine learning–based approaches for classifying sentiments from Hinglish data collected from social media reviews. The proposed methodology includes data preprocessing, normalization of text, and feature extraction techniques to transform textual content into machine-interpretable representations. Various supervised learning models are analyzed to evaluate their ability to distinguish between positive, negative, and neutral sentiments. The findings highlight the importance of developing language-aware NLP techniques for multilingual environments and demonstrate how specialized models can improve sentiment analysis for Indian code-mixed communication.***

Keywords: Sentiment Analysis, Hinglish, Code-Mixed Language, Natural Language Processing, Machine Learning, Social Media Text, Text Classification.

I. INTRODUCTION

The way people express their opinions, share their personal experiences, and participate in public discourse has completely changed as a result of the exponential growth of social media platforms like Twitter, YouTube, Instagram, and numerous e-commerce review portals. Hinglish, a hybrid, code-mixed language that combines Hindi vocabulary with English syntax and is primarily written in Roman script, accounts for a sizable and expanding amount of communication in the Indian digital ecosystem. India has one of the biggest and fastest-growing social media user bases in the world as of 2024, making Hinglish a more prevalent and deeply ingrained form of online expression across all demographic groups.

Over the past ten years, sentiment analysis—the computational task of detecting and categorizing the emotional polarity embedded within text—has made significant progress for standard monolingual languages. For English, French, Mandarin, and a number of other well-resourced languages, reliable tools and pre-trained models are currently available. However, because of its intrinsic linguistic irregularities and dynamic vocabulary, traditional Natural Language Processing (NLP) frameworks are essentially ill-suited to handle code-mixed data. Inconsistent spelling, heavily transliterated words, informal slang unique to a given area, and fluid grammatical structures that simultaneously depart from Hindi and English norms are characteristics of Hinglish text.

The main goal of this study is motivated by the widening gap between the prevalence of Hinglish in online communication and the capacity of current NLP systems to process it. In order to create a trustworthy, language-aware sentiment analysis framework, this paper explores machine learning-based methods for sentiment classification of Hinglish social media reviews. The results aid in the development of more inclusive NLP systems that can address the linguistic needs of multilingual and multilingual communities in India and elsewhere..

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

II. PROBLEM DEFINITION

The code-mixed nature of Hinglish presents a distinct and complex set of issues that are still mostly unresolved within the NLP research community, despite notable and continuous advancements in the field of sentiment analysis for resource-rich languages like English and Mandarin. The majority of monolingual corpora with clear grammatical rules and uniform orthographic conventions are used to train existing sentiment analysis models.[1] When faced with mixed-language input, where vocabulary, grammar, and script conventions change fluidly within a single sentence or even within a single word, these models are unable to generalize effectively.
Accurately identifying and classifying sentiment as positive, negative, or neutral from Hinglish text taken from actual social media reviews and user-generated content is the main issue this study attempts to solve. This issue is made worse by a number of interrelated difficulties. First, a single Hindi word may appear in multiple phonetically correct but orthographically different Roman script variants across various users and platforms due to Hinglish's lack of standardized orthography.[2] Second, hashtags, emoticons, internet abbreviations, and informal slang are commonplace and add layers of semantic complexity that standard tokenizers are unable to effectively handle. Third, model training and benchmarking are severely constrained by the lack of large-scale, consistently annotated, high-quality Hinglish datasets.

III. PROPOSED APPROACH

This study introduces a straightforward machine learning method for analyzing sentiment in Hinglish social media reviews.[2] The process consists of four main stages: collecting data, preprocessing it, extracting useful features, and training models while assessing their performance. Data is gathered from platforms like YouTube, Twitter, and product review websites. It is cleaned by removing URLs, special characters, and duplicate entries. The preprocessing stage also includes normalizing transliterated words, filtering mixed-language stopwords, and handling phonetic variations while carefully retaining important sentiment indicators such as negations and intensifiers. For feature extraction, a combination of TF-IDF, Word2Vec, and FastText embeddings captures both basic text patterns and deeper contextual meanings.[2] These features are fed into machine learning models including Logistic Regression, SVM, Random Forest, and Naïve Bayes.[3] Finally, the models are assessed using metrics like accuracy, precision, recall, and macro F1-score across three sentiment classes. This process helps identify the most effective method for Hinglish sentiment analysis.

IV. METHODOLOGY

This section outlines the end-to-end pipeline designed for sentiment classification of Hinglish social media reviews. The methodology is structured across five sequential stages: data collection, data preprocessing, feature extraction, model training, and evaluation.

Algorithm1: Hinglish Sentiment Classification
Input: Raw Hinglish text corpus D
Output: Sentiment label L ∈ {Positive, Negative, Neutral}
1. Preprocess D → remove noise, normalize, tokenize→T
2. Extract features from T:
   TF-IDF → term importance vectors
   Word2Vec→ semantic embeddings
FastText→subword/morphological vectors → F
3. Split F→ 80% train / 20% test (stratified)
4. Train classifiers on F:
   {SVM, Logistic Regression, Random Forest, Naïve Bayes}
5. For new input x:
   Apply steps 1–2 → extract features
   Predict L = argmax(class scores)
6. Evaluate: Accuracy, Precision, Recall, Macro F1
Select best model-feature combination

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

The complete sentiment classification pipeline for Hinglish social media text is described in Algorithm 1. The procedure starts with a multi-step preprocessing step that generates clean token sequences appropriate for feature extraction, eliminates noise, and normalizes transliterated spelling variations.

To capture deeper semantic and morphological information contained in code-mixed text as well as surface-level term importance, three complimentary feature representations are extracted: TF-IDF, Word2Vec, and FastText. A systematic cross-comparison of model-feature combinations is made possible by the separate training of four supervised classifiers on each feature set. Before sending the representation to the top-performing trained model, which generates a sentiment label from the three target classes, the pipeline performs the identical preprocessing and feature extraction procedures for any unknown input.In order to account for any class imbalance in the Hinglish dataset, model performance is quantified using macro F1-score as the key metric. The stratified 80/20 evaluation split guarantees that class distribution is maintained across training and testing.

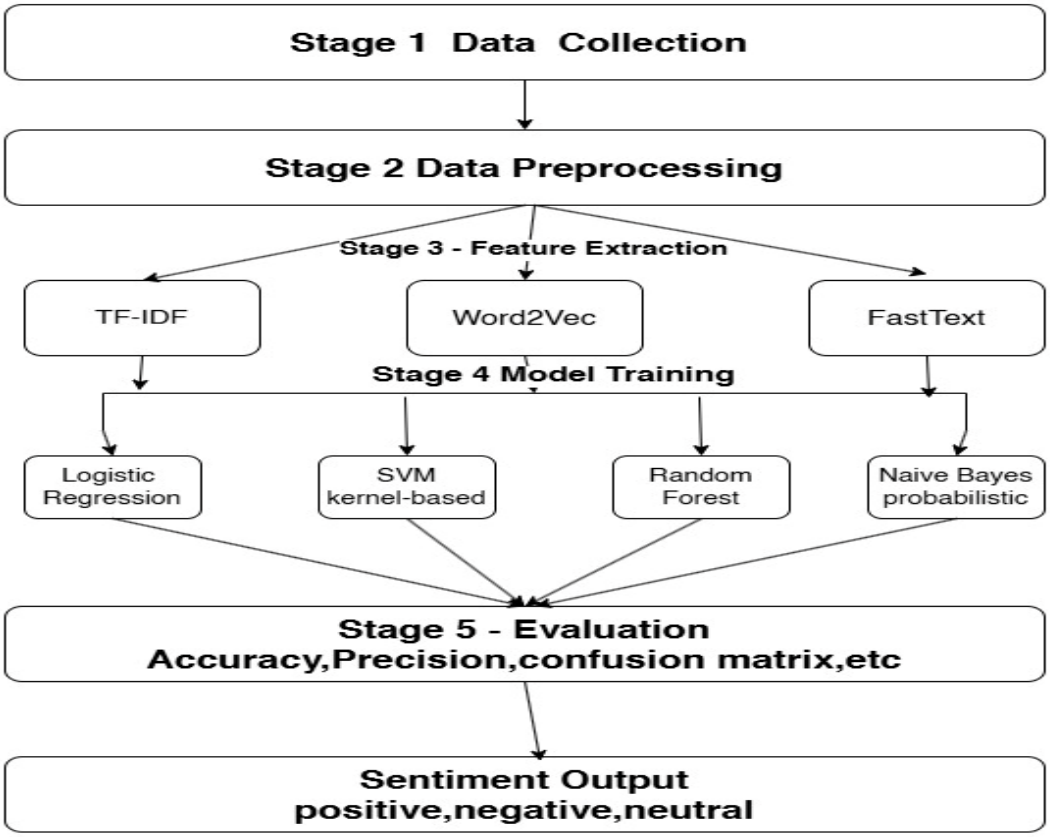


Figure 1. Proposed end-to-end sentiment analysis Architecture pipeline

Stage 1 : Data Collection: Hinglish user-generated content is collected from publicly accessible social media platforms including YouTube comment sections, Twitter/X posts, and Indian e-commerce review portals such as Flipkart. [2]The collected data encompasses a diverse range of topics, writing styles, and sentiment expressions. Each data instance is manually labelled as positive, negative, or neutral by annotators with native Hinglish proficiency to ensure high annotation quality.

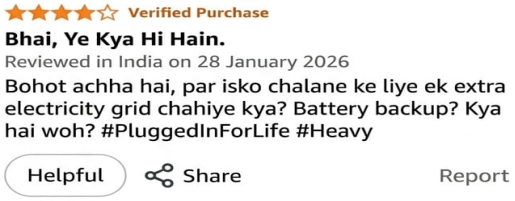


Figure 2. Sample Data 1 (Amazon review 1)

Figure 3. Sample Data 2 (Amazon review 2)

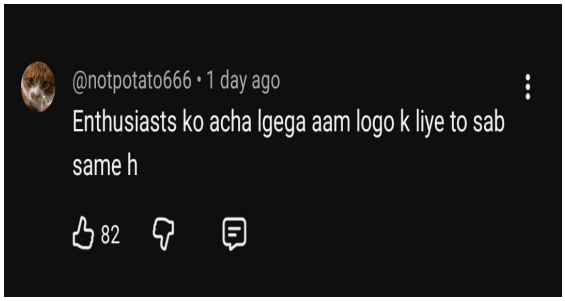


Figure 4. Sample Data 3(Youtube Comment 1)

Figure 5. Sample Data 4(Youtube Comment 2)

.Sample No.	Hinglish Text	Source	Sentiments	Key Signals
1	*"Bohot achhahai, par… extra electricity grid chahiyekya? #PluggedInForLife"*	Amazon Review	Neutral / Mixed	Sarcasm positive opener ironic hashtag
2	*"Beast hai… Sabko recommend karunga… dushmankeliye."*	Amazon Review	Negative	Irony rating mismatch ellipsis
3	*"Enthusiasts ko achalgegaaam logo k liye to sab same h"*	YouTube Comment	Neutral	factual tone abbreviated spelling no punctuation
4	*"twist pe twist de raha he's actually the victim tho"*	YouTube Comment	Positive / Excited	emoji signal intra-sentenceswitch slang

Table 1. Sample Hinglish Text Instances

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Stage 2 : Data Preprocessing:

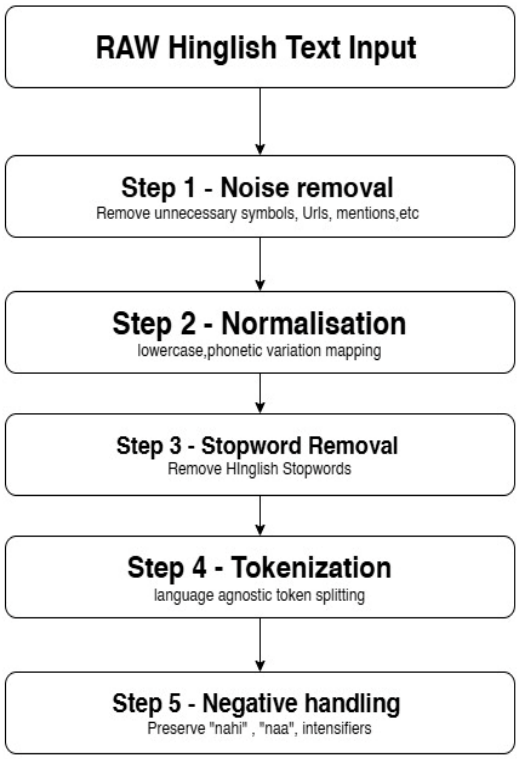


Figure 6. Preprocessing Flowchart

Raw Hinglish text undergoes multi-step preprocessing including removal of URLs, mentions, hashtags, and special characters; Unicode normalization; lowercasing; phonetic variant normalization (e.g., "kya", "kia", "kyaa"); code-mixed stopword removal; and tokenization using a language-agnostic tokenizer.[2][4]

Stage3 :Feature Extraction: Three complementary feature representations are applied: TF-IDF weighted bag-of-words (capturing term importance), Word2Vec embeddings (capturing semantic similarity), and FastTextsubword vectors (capturing morphological variations in transliterated Hindi words).

Stage 4 :Model Training: Four supervised classifiers are trained — Logistic Regression, Support Vector Machine (SVM), Random Forest, and Naïve Bayes — using each feature set independently to enable systematic comparison.

Stage 5 : Evaluation: Models are evaluated using a stratified 80/20 train-test split, with performance measured across accuracy, precision, recall, and macro F1-score for all three sentiment classes.[5]

V. EXPECTED OUTCOMES AND DISCUSSIONS

This section presents an analytical discussion of the anticipated results based on the proposed methodology and findings reported in closely related literature on Hinglish and code-mixed sentiment analysis. Since this work proposes a machine learning pipeline for future empirical evaluation, the outcomes described here are grounded in evidence from comparable studies rather than experimental data. [6]

Among the four classifiers proposed --Logistic Regression, SVM, Random Forest, and Naïve Bayes -- Support Vector Machine paired with TF-IDF feature representation is expected to yield the strongest overall performance. This expectation is well-supported by prior work; Singh (2021) and Ghosh et al. (2023) both report SVM as consistently competitive on code-mixed text classification tasks due to its effectiveness in high-dimensional sparse feature spaces. Logistic Regression is anticipated to perform comparably on well-balanced classes, while Random Forest is expected to offer moderate resistance to overfitting. [2] Naïve Bayes, despite its computational efficiency, is likely to underperform on the neutral sentiment class because its conditional independence assumption conflicts with the contextual dependencies inherent in Hinglish text. [3]

The neutral sentiment class is expected to be the most difficult to classify accurately across all models, a pattern consistently observed in existing Hinglish NLP literature. [5] Neutral expressions in code-mixed informal text are frequently ambiguous, context-dependent, and underrepresented in available labelled datasets, all of which contribute to lower recall for this class.

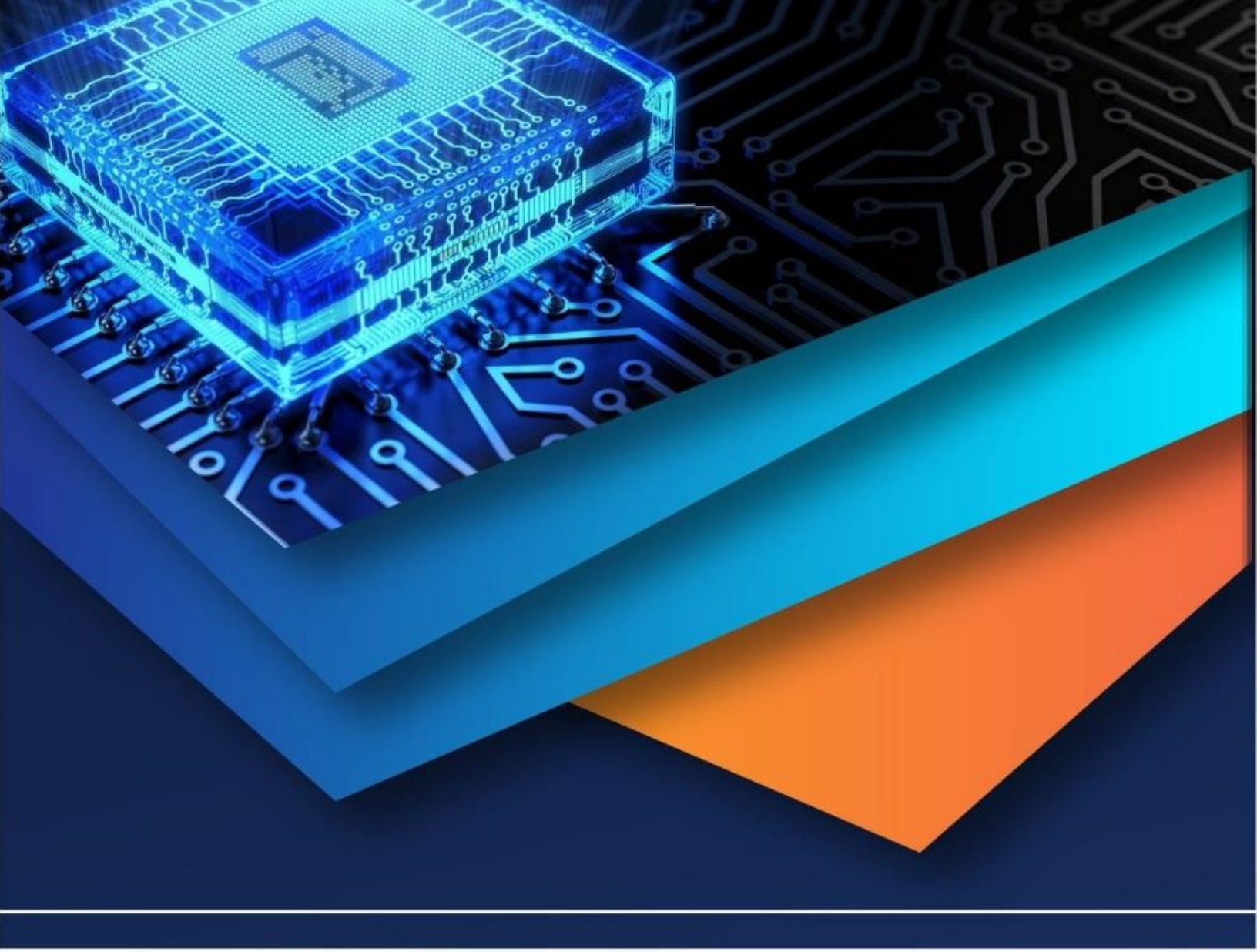
FastTextsubword embeddings are anticipated to outperform TF-IDF on short, slang-heavy reviews by capturing morphological variations in transliterated Hindi words — for example, treating "achha", "acha", and "accha" as semantically related. This advantage is expected to be especially prominent for user-generated content from platforms like YouTube and Twitter where spelling inconsistency is high. [7]

Overall, these anticipated findings highlight that preprocessing quality and feature representation choice are likely to be the primary determinants of model performance in Hinglish sentiment analysis. Empirical validation of this pipeline through actual model training and evaluation on a labelled Hinglish dataset constitutes the immediate next step of this research.

REFERENCES

[1] M. B. S. C. S. Muhammad Kashif Nazir, “Sentiment analysis for code-mixed low-resource languages: a systematic review of approaches, techniques, applications, challenges, and future directions,” Springer, 2026.
[2] G. Singh, “Sentiment Analysis of Code-Mixed Social Media Text (Hinglish),” p. 17, 2021.
[3] R. Baghel, “A Survey on Code-Mixed Sentiment Analysis Based on Hinglish Dataset,” in Lecture Notes in Networks and Systems ((LNNS,volume 664)).
[4] A. K. M. K. Pratibha, “Expanding Research Horizons for Hinglish Text by Tackling Challenges and Research Gaps,” Jisem Journal, 2025.
[5] A. P. A. E. P. B. Soumitra Ghosh, “Multitasking of sentiment detection and emotion recognition in code-mixed Hinglish data,” Science Direct, vol. 260, 2023.
[6] M. P. R. A. Adarsh Singh Jadon, “Hinglish Sentiment Analysis: Deep Learning Models for Nuanced Sentiment Classification in Multilingual Digital Communication,” IEEE explore, 2024.
[7] I. K. Brajesh Khare, “Optimized emotion classification in code-mixed Hinglish text using an mBERT based hybrid neural network with attention mechanisms,” International Journal of Information Technology, 2025.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)