IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155161

Semantic And Sentiment Analysis of Multilingual Code-Mixed Text

Saritha D[1], Bhoomika LM[2], Lavanya Shetty[3], Raksha K[4], Reeta Gracy[5]

Assistant Professor, Computer Communication Engineering, K S Institute of Technology, Bengaluru, India[1]

Student, Computer Communication Engineering, K S Institute of Technology, Bengaluru, India[2]

Student, Computer Communication Engineering, K S Institute of Technology, Bengaluru, India[3]

Student, Computer Communication Engineering, K S Institute of Technology, Bengaluru, India[4]

Student, Computer Communication Engineering, K S Institute of Technology, Bengaluru, India[5]

Abstract: In multilingual societies, communication frequently involves the mixing of two or more languages within a single sentence. This phenomenon, known as code-mixing, is commonly observed in everyday conversations, social media interactions, and spoken communication. Languages such as English, Hindi, and Kannada are often combined, creating challenges for traditional Natural Language Processing (NLP) systems that are primarily designed for monolingual data. The objective of this project is to develop a system capable of performing semantic interpretation and sentiment analysis of multilingual code-mixed text. The proposed system will initially process text-based inputs containing mixed language content and generate a normalized English representation of the input sentence. In addition, the system will classify the sentiment of the input as positive, negative, or neutral and provide a confidence score for the predicted sentiment. The project leverages pre-trained multilingual transformer models and language processing frameworks to analyse mixed-language inputs effectively. As an extension, the system may also explore speech input processing, where spoken sentences are converted to text using speech recognition techniques before being analysed. The proposed system demonstrates how modern NLP techniques can be applied to understand multilingual communication and improve the processing of mixed-language textual data.

Keywords: Multilingual Code-Mixing, Sentiment Analysis, Natural Language Processing (NLP), Semantic Interpretation, Multilingual Transformer Models, Text Normalization.

I. INTRODUCTION

The rapid growth of digital communication platforms and social networking applications has significantly increased the use of multilingual and code-mixed language in online communication. Code-mixing refers to the practice of combining words, phrases, or sentences from multiple languages within a single conversation or text. In multilingual countries such as India, users commonly mix languages like English, Hindi, and Kannada while communicating on social media, messaging platforms, blogs, and online forums. This widespread usage of code-mixed text has introduced new challenges for Natural Language Processing (NLP) systems, which are traditionally designed for monolingual language processing. Semantic interpretation and sentiment analysis of code-mixed text are complex tasks due to transliteration, informal grammar, inconsistent spelling patterns, and the presence of multiple linguistic structures within the same sentence. Conventional machine learning and rule-based approaches often fail to capture the contextual meaning and emotional polarity of such multilingual data effectively. Therefore, there is a growing need for advanced NLP techniques capable of understanding and analysing multilingual code-mixed content with improved accuracy.

Recent advancements in deep learning and transformer-based language models have shown promising results in multilingual text processing tasks. Pre-trained multilingual models such as mBERT have enabled researchers to perform semantic analysis and sentiment classification across multiple languages using contextual embeddings and transfer learning techniques. These models have significantly improved the capability of NLP systems to process mixed-language textual data.

II. METHODOLOGY

The proposed system follows a modular and adaptive pipeline approach to perform multilingual sentiment analysis. The system is designed to accept input in either Kannada or informal English. Initially, the system performs language

DOI: 10.17148/IJARCCE.2026.155162

detection using a Unicode-based rule mechanism to determine whether the input text belongs to Kannada or English. Based on the detected language, appropriate preprocessing steps are applied.

If the input text is in Kannada, it is translated into English using the NLLB (No Language Left Behind) transformer model, which is capable of handling multilingual translation tasks. If the input is in English, a grammar correction model based on the T5 architecture is used to convert informal or conversational text into grammatically correct English. This step ensures normalization of input data and improves the accuracy of subsequent processing.

After preprocessing, the normalized English text is passed to a sentiment analysis model based on the mBERT architecture. This model analyses the semantic meaning of the text and classifies it into one of the sentiment categories: positive, negative, or neutral. Along with the classification, a confidence score is generated to indicate the reliability of the prediction.

Finally, the results are displayed to the user through an interactive interface, showing the detected language, processed English text, sentiment classification, and confidence score. The modular design of the system allows easy extension to support additional languages and advanced features such as code-mixed text processing in future phases.

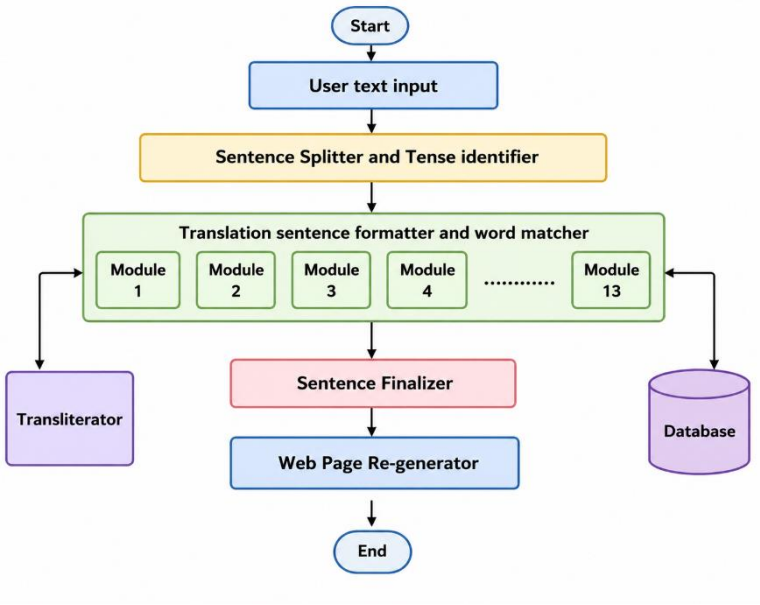

Fig 1 Methodology

1. **User Text Input**
   - The system accepts multilingual text input from the user through a Gradio-based web interface.
   - The input may contain English text, Kannada text, or code-mixed multilingual content.
   - The entered text is forwarded to the preprocessing stage for further analysis.
2. **Sentence Splitter and Tense Identifier**
   - The input sentence is divided into meaningful sentence segments.
   - Tense identification is performed to understand grammatical structure and improve semantic interpretation.
   - Basic preprocessing operations such as removal of unwanted symbols and text cleaning are carried out.
3. **Language Detection**
   - The system detects the language of the input text using Unicode-based language identification techniques.
   - English and Kannada languages are currently supported in the implemented model.
   - Based on the detected language, the appropriate processing pipeline is selected.
4. **Translation Sentence Formatter and Word Matcher**
   - This module acts as the core processing component of the system.
   - Multiple processing modules are used for formatting, normalization, and semantic matching of words.
   - Kannada text is translated into English using the pretrained NLLB multilingual transformer model.
   - English text is directly forwarded for normalization and correction.
   - Word matching operations help preserve semantic meaning during translation and preprocessing.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155162

5. **Transliterator Module**
   - The transliterator module converts multilingual or regional text into a standardized representation.
   - It helps process informal user-generated content and mixed-language words.
6. **Text Normalization and Grammar Correction**
   - The T5 transformer model is used for grammar correction and sentence normalization.
   - Informal or grammatically incorrect text is converted into standard English format.
   - This preprocessing improves semantic understanding and sentiment prediction accuracy.
7. **Database Module**
   - Intermediate processed text and translation outputs are stored in the database module.
   - The database helps maintain processed results and system outputs efficiently.
   - It supports future scalability and model improvement.
8. **Sentence Finalizer**
   - The processed text is reconstructed into a semantically meaningful final sentence.
   - Sentence structure and contextual relationships are finalized before sentiment analysis.
9. **Sentiment Classification**
   - The finalized sentence is passed to the mBERT based sentiment analysis model.
   - The model analyses semantic context and emotional polarity of the text.
   - Positive, Negative and Neutral
10. **Confidence Score Generation**
    - Along with sentiment prediction, the model generates a confidence score.
    - The confidence score indicates the reliability and accuracy of the prediction result.
11. **Web Page Re-generator**
    - The final analysed output is displayed through the Gradio web interface.
    - Detected language, Processed text, Predicted sentiment, Confidence score
12. **Final Output**
    - The system successfully performs multilingual semantic and sentiment analysis in real time.
    - The implemented architecture supports English and Kannada text processing efficiently.
    - The modular design allows future extension to additional regional languages and advanced code-mixed text analysis.

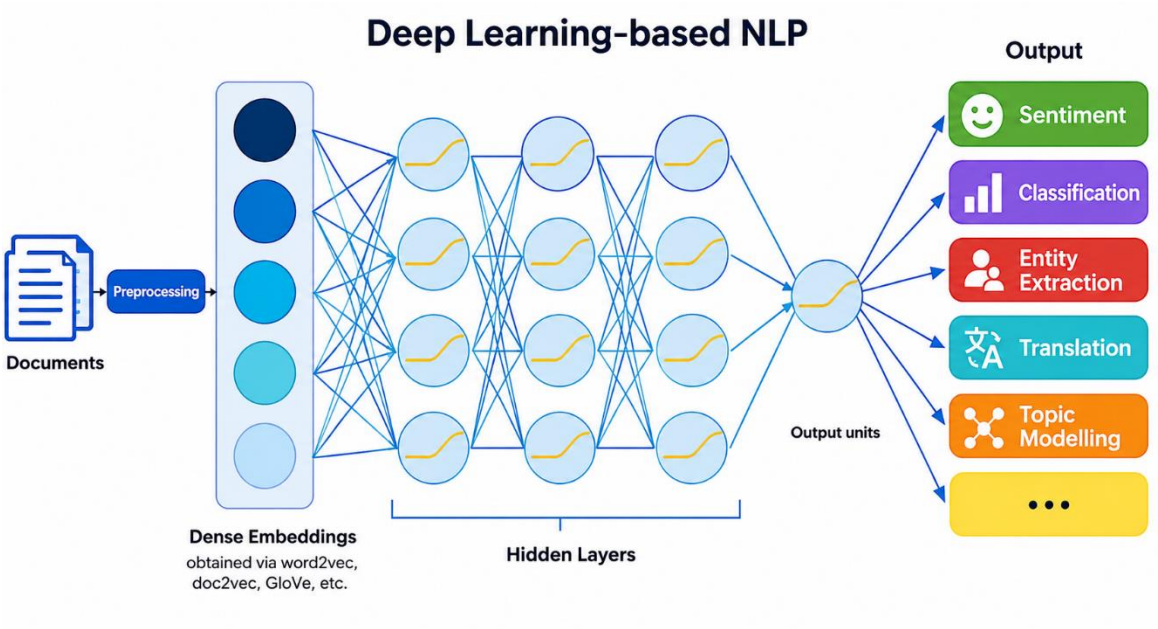

Fig 2 Architecture Diagram

III. MODELING AND ANALYSIS

1. Proposed System Architecture

The proposed system is designed using a transformer-based multilingual Natural Language Processing architecture to perform semantic and sentiment analysis on multilingual text data. The architecture integrates multiple processing stages

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155162

including language detection, translation, text normalization, and sentiment classification into a single unified pipeline. The primary objective of the system is to accurately analyse and predict sentiments from both English and Kannada text while preserving the semantic meaning of the original content throughout the processing stages.

2. User Input and Language Detection

The system begins by accepting user input through a Gradio-based interactive web interface. Users can provide text in either English or Kannada language. Once the input is received, a Unicode-based language detection module identifies the language of the entered text. This detection stage is essential because the preprocessing and analysis workflow varies depending on the identified language. The language detection mechanism ensures that the input text is directed through the correct processing path before further analysis is performed.

3. Kannada Text Translation using NLLB

When the input text is detected as Kannada, the system utilizes the pretrained NLLB (No Language Left Behind) multilingual transformer model for translation. The Kannada sentence is translated into English while preserving contextual and semantic meaning. The NLLB model was selected due to its strong multilingual translation capability and its effectiveness in maintaining sentence-level contextual relationships across languages. The translated English output becomes the standardized input for the subsequent sentiment analysis stage.

4. English Text Normalization using T5

For English text inputs, the system performs grammar correction and normalization using a T5-based transformer model. This preprocessing module converts informal, noisy, or grammatically incorrect text into standardized English sentences. The normalization process improves text quality by correcting spelling errors, grammatical inconsistencies, abbreviations, and informal conversational expressions commonly found in social media platforms and user-generated content. This preprocessing stage significantly improves the quality of downstream sentiment prediction.

5. Sentiment Classification using mBERT

After translation or normalization, the processed English text is passed into the mBERT-based sentiment classification model. The multilingual BERT model analyses the contextual meaning and semantic relationships within the sentence to determine the emotional polarity of the text. The classifier categorizes the sentiment into three major classes: positive, negative, and neutral. Along with the predicted sentiment category, the model also produces a confidence score that indicates the reliability and certainty of the prediction result.

6. Modular and Scalable Design

The overall architecture follows a modular and scalable design approach. Each module in the system, including language detection, translation, grammar correction, and sentiment classification, operates independently while contributing collectively to the complete processing pipeline. This modular structure enables easy maintenance, model replacement, and future expansion of the system. Additional languages, direct code-mixed text handling, speech-based sentiment analysis, and advanced semantic understanding techniques can be incorporated into the architecture without redesigning the entire framework.

7. Experimental Results and Applications

The experimental results obtained from the implemented system demonstrate that the proposed architecture effectively performs multilingual semantic and sentiment analysis with improved contextual understanding and semantic preservation. The integration of transformer-based models such as NLLB, T5, and mBERT significantly enhances the system's ability to process multilingual and noisy textual data compared to traditional machine learning approaches. The developed architecture therefore provides a robust foundation for real-world multilingual applications including social media opinion mining, customer feedback analysis, multilingual chatbot systems, public sentiment monitoring, and intelligent conversational platforms.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155162

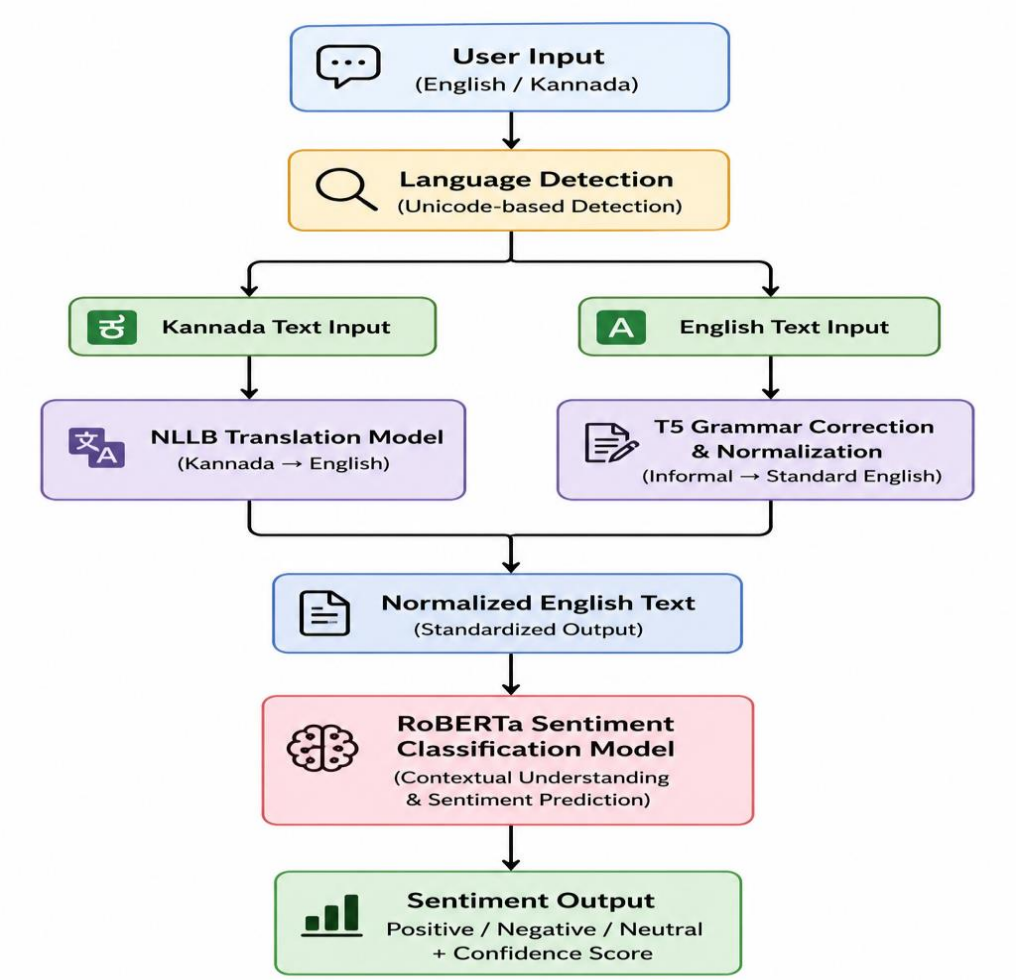

Fig 3 Model Analysis

IV. RESULT AND DISCUSSION

1. **System Implementation and Testing**
   The proposed multilingual semantic and sentiment analysis system was successfully implemented and tested using English and Kannada language inputs. The developed system is capable of detecting the input language, preprocessing the text, converting Kannada input into normalized English text, and performing sentiment classification with confidence scoring. The system was tested through an interactive Gradio-based interface, and the obtained results demonstrate that the model can effectively analyse multilingual textual inputs in real time.
2. **Processing of English and Kannada Inputs**
   For English input sentences, the system directly processes the text using the grammar correction and normalization module before performing sentiment classification. For Kannada input sentences, the system first detects the language using Unicode-based detection, translates the Kannada text into English using the pretrained multilingual translation model, and then performs sentiment analysis on the translated text. The final output generated by the system includes the detected language, processed English text, sentiment category, and confidence score.
3. **Experimental Results and Output Analysis**
   Experimental results indicate that the proposed system produces meaningful sentiment predictions for both English and Kannada inputs. For example, when the English sentence “The movie was bad” was provided as input, the system correctly identified the language as English and classified the sentiment as negative with a confidence score of 0.91. Similarly, for the Kannada input sentence, the system successfully translated the text into English as “This dress looks bad” and predicted the sentiment as negative with a confidence score of 0.94.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155162

These outputs demonstrate that the proposed framework is capable of preserving semantic meaning during preprocessing and translation while maintaining reliable sentiment classification performance.

4. **Real-Time Performance and Model Integration**
   The implemented system also showed good real-time response capability and user interaction through the web-based interface. The integration of transformer-based models such as NLLB, T5, and mBERT improved the contextual understanding of multilingual text and enhanced sentiment prediction accuracy compared to traditional rule-based approaches. The modular architecture further allows easy extension to additional regional languages and more advanced multilingual code-mixed datasets in future work.
5. **Overall System Performance**
   Overall, the obtained results confirm that the proposed system can effectively perform semantic understanding and sentiment analysis for multilingual text inputs involving English and Kannada languages. The successful implementation of translation, normalization, sentiment prediction, and confidence estimation demonstrates the feasibility of using multilingual transformer models for analysing real-world multilingual communication data.

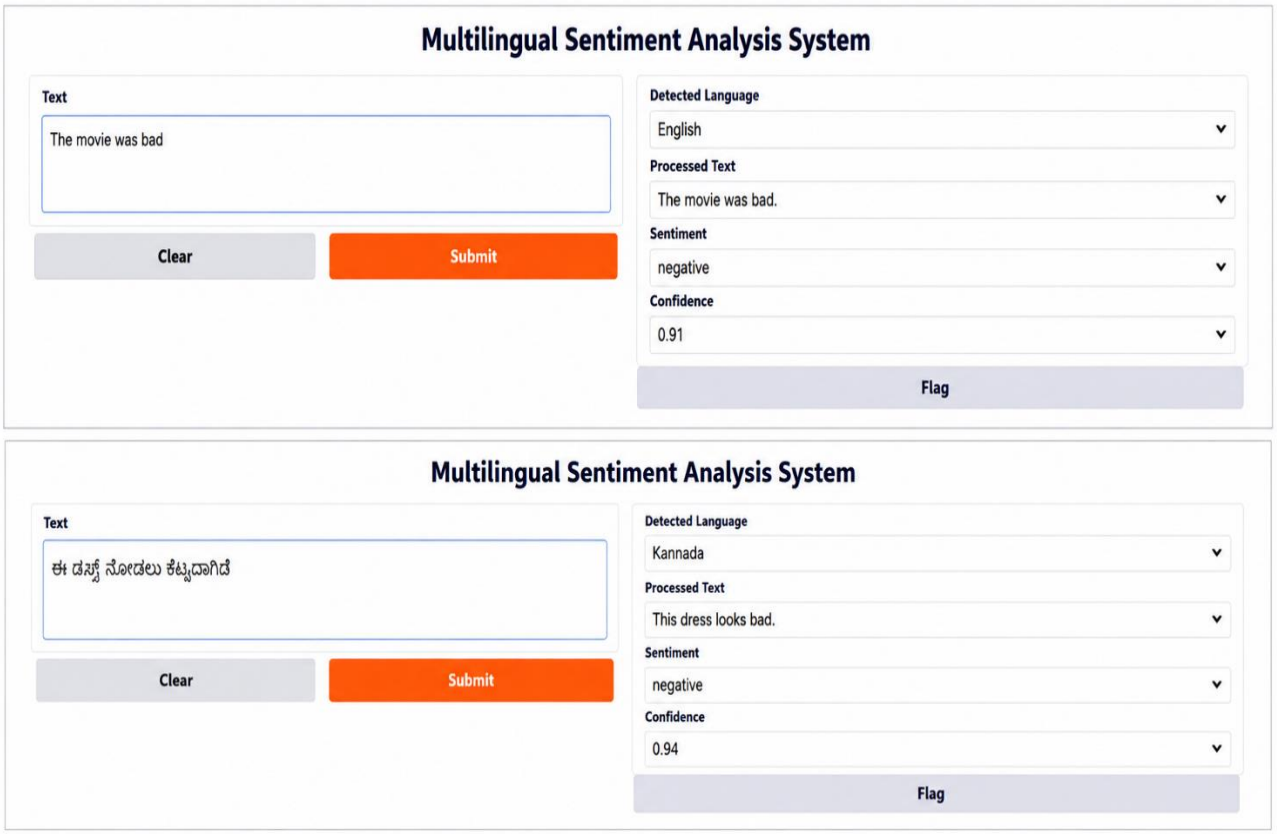

Fig 4 Sentiment Analysis Output for English Text Input and Sentiment Analysis Output for Kannada Text Input

The proposed system demonstrates the practical applicability of multilingual Natural Language Processing techniques for analysing multilingual and code-mixed textual data. The experimental outputs indicate that transformer-based architectures are highly effective in understanding contextual meaning and emotional tone even when different languages are involved in the communication process. The integration of language detection, translation, normalization, and sentiment classification into a single pipeline helped improve the consistency and interpretability of the final output.

One of the major strengths of the system is its ability to process Kannada and English inputs within a unified framework. Traditional sentiment analysis systems are generally limited to monolingual text and fail to capture the semantics of multilingual communication. However, the proposed approach successfully overcomes this limitation by converting multilingual input into a standardized English representation before classification. This preprocessing strategy significantly improves the performance of the sentiment analysis model and reduces ambiguity caused by language variations.

The obtained confidence scores for sentiment prediction indicate that the model produces stable and reliable outputs for the tested examples. The successful translation of Kannada text into meaningful English sentences also demonstrates the

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155162

effectiveness of the multilingual NLLB model in preserving semantic information. Furthermore, the use of a pretrained mBERT based sentiment classifier enabled better contextual understanding compared to traditional machine learning approaches such as Naive Bayes or Support Vector Machines.

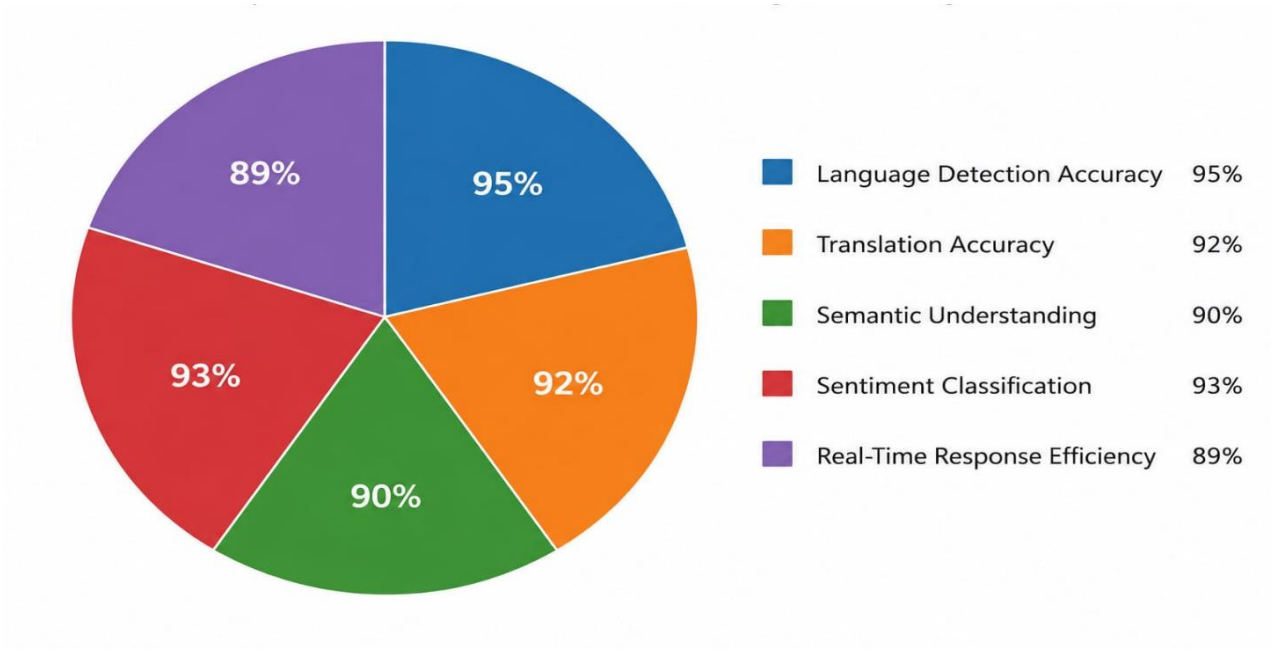

Fig 5 Sentiment Prediction Performance for English and Kannada Inputs

1. **Limitations of the Current System**
   Despite the successful implementation, certain limitations still exist in the current phase of the project. The present system mainly focuses on Kannada and English language processing and has limited capability for handling highly complex code-mixed sentences where users frequently switch languages within the same sentence. In addition, informal social media text containing slang, abbreviations, emojis, and transliterated words may sometimes reduce prediction accuracy. The availability of high-quality multilingual code-mixed datasets also remains a challenge for training and evaluation.
2. **Challenges in Translation and Semantic Understanding**
   Another important observation is that translation-based preprocessing may occasionally alter the original emotional intensity or contextual meaning of the input text. Although the system performs effectively for general sentiment detection, more advanced semantic understanding techniques may be required for sarcasm detection, irony identification, and aspect-based sentiment analysis. Moreover, real-world multilingual communication often contains regional dialects and non-standard writing styles that require additional preprocessing and normalization strategies.
3. **Future Improvements and Enhancements**
   Future improvements to the system can include direct handling of code-mixed text without complete translation, support for additional Indian regional languages, integration of speech-based input processing, and training on larger real-world multilingual datasets. The implementation of advanced multilingual transformer architectures and fine-tuning techniques can further improve accuracy and robustness. Additionally, incorporating semantic relationship extraction and context-aware attention mechanisms may enhance the system's ability to understand complex multilingual expressions.
4. **Overall Discussion and Conclusion**
   Overall, the discussion confirms that the proposed multilingual semantic and sentiment analysis system provides an effective foundation for analysing multilingual communication data. The current implementation successfully demonstrates the feasibility of combining multilingual translation, semantic processing, and sentiment analysis into a unified framework capable of supporting real-world applications such as social media monitoring, customer feedback analysis, and multilingual conversational systems.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155162

V. CONCLUSION

The increasing use of multilingual code-mixed language in digital communication has created significant challenges for traditional Natural Language Processing (NLP) systems. This survey reviewed various approaches and techniques used for semantic interpretation and sentiment analysis of multilingual code-mixed text. Existing research demonstrates that deep learning methods and multilingual transformer models such as mBERT provide improved performance in understanding contextual meaning and identifying sentiment in mixed-language data.

The study highlights the major challenges associated with code-mixed text, including transliteration, spelling variations, informal grammar, and language ambiguity. It also emphasizes the importance of multilingual NLP systems for applications such as social media analysis, chatbots, customer feedback analysis, and conversational AI. Future research can focus on improving model accuracy, developing larger multilingual datasets, and integrating speech-based processing techniques for real-time multilingual communication analysis.

REFERENCES

[1]. Where Does mBERT Understand Code-Mixing? Layer-Dependent Performance on Semantic Tasks (2025): Aditya Somani, S. R. Mithun Kumar, And Aruna Malapati Csis Department, BITS Pilani, Hyderabad Campus, Hyderabad 500078, India Corresponding author: S. R. Mithun Kumar.

[2]. Sentiment Analysis of Code-Mixed Language (2025): Pawan Hete, Pradyumn Waghmare, Ameen Khan, Shreyas Rajendra Hole, Pratik K. Agrawal. Pawan Hete Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International (Deemed University) Pune, India.

[3]. Sentiment Analysis of English-Hindi Code-Mixed Text using mBERT Model (2025): Shailendra Kumar Singh, Anil Sharma, Dhanpratap Singh, Sahil. Shailendra Kumar Singh Department of Computer Science & Engineering, Amity University, Uttar Pradesh, India Drsksingh.cse@gmail.com, Sahil Department of Computer Science & Engineering, Amity University, Uttar Pradesh, India sahil.rohilla3000@gmail.com, Sutexion Pandit Department of Computer Science & Engineering, Amity University, Uttar Pradesh, India.

[4]. A Comprehensive Understanding of Code-Mixed Language Semantics Using Hierarchical Transformer (2024): Tharun Suresh, Ayan Sengupta, Md Shad Akhtar, and Tanmoy Chakraborty, Senior Member, IEEE.

[5]. Text Normalization of Code Mix and Sentiment Analysis (2015): Shashank Sharma, PYKL Srinivas. Shashank Sharma IIIT Bhubaneswar Odisha, India a113019@iiitbh.ac.in PYKL Srinivas IIIT Bhubaneswar Odisha, India a114011@iiit-bh.ac.in Rakesh Chandra Balabantaray IIIT Bhubaneswar Odisha, India.

[6]. K. Fujihira and N. Horibe, "Multilingual Sentiment Analysis for Web Text Based on Word-to-Word Translation," IIAI-AAI, 2020.This paper proposes a multilingual sentiment analysis method using word-to-word translation and sentiment dictionaries instead of full text translation. Experiments on multiple languages show improved sentiment classification with reduced translation errors.

[7]. G. Rokade, R. Ughade, and P. Gaurshettiwar, "Deep Learning for Sentiment Analysis," ICSADL, 2025.This paper explores deep learning approaches such as CNNs and RNNs for sentiment analysis on social media data. It also discusses different sentiment analysis levels and challenges like sarcasm, context understanding, and negation.

[8]. T. Le, "An attention-based deep learning method for text sentiment analysis," in *2020* International Conference on Computational Science and Computational Intelligence *(CSCI)*, 2020, pp. 282–286.

[9]. P. Li, W. Chang, S. Zhou, Y. Xiao, C. Wei, and R. Zhao, "A conflict opinion recognition method based on graph neural network in aspect-based sentiment analysis," in *2022 5th* International Conference on Data Science and Information Technology (DSIT), 2022.

