# Sentiment Analysis for Kannada–English Code-Mixed Social Media Text

**Dr. Pushpalatha M*[1], Dr. Sasikala P*[2], Dr. Santhosh Kuamr B N*[3], Dr. H S Nagalakshmi*[4]**

**[1]Associate Professor, Department of Computer Science, Govt Women's College, Hunsur, Mysuru, Karnataka, India**

**[2]Associate Professor, Department of Computer Science, Lal Bahadur Shastri Government First Grade College, RT Nagar Bengaluru, India**

**[3]Associate Professor, Department of Computer Science, Govt Women's College, Hunsuru, Mysuru, Karnataka, India**

**[4]Associate Professor and Head Department of BCA, Govt College for Women's Autonomous, Mandya, Karnataka, India**

**DOI: https://doi.org/10.51244/IJRSI.2026.1307000354**

**Received: 06 August 2026; Accepted: 11 August 2026; Published: 19 August 2026**

## ABSTRACT

The exponential rise of social media has led to the generation of a large amount of informal text. Especially, code-mixed languages have gained substantial popularity among social media users. In India, the code-mixed language Kannada-English is widely used in social media platforms. This informal and non-standard language form brings forward significant challenges to natural language processing (NLP) tasks like sentiment analysis. This paper presents a detailed analysis of sentiment analysis in kannada-english code-mixed social media text. A manually annotated dataset is created and classified into positive, negative, and neutral sentiment labels. Various kinds of machine learning, deep learning, and transformer-based models are evaluated. The results suggest that the transformer-based models trained on Indian language corpora outperform others. Furthermore, this paper provides a comprehensive set of observations regarding sentiment analysis of Indian language code-mixed social media text.

**Keywords:** Kannada NLP, Code-Mixed Text, Sentiment Analysis, Low-Resource Languages, Transformer Models, Social Media Analytics

## INTRODUCTION

Sentiment analysis is a fundamental task in natural language processing that aims to detect opinions, emotions, and attitudes of the text's author. With the boom of social media networks like Twitter and YouTube, sentiment analysis plays a vital role in social media mining, public opinion analysis, and customer relationship management. With the exponential use of social media, the informal and non-standard text form is generated at a high rate. In countries with linguistic diversity, multilingual mixing is a common phenomenon while posting on social media.

The code-switched and code-mixed informal texts pose significant challenges to the sentiment analysis task. Majority of the previous works conducted sentiment analysis on monolingual English text. Hence, there is a need for further research to conduct sentiment analysis on code-switched text of Indian languages like kannada-english.

### Related Work

The early research studies conducted sentiment analysis using machine learning classifiers such as Naïve Bayes, logistic regression, etc. But the traditional machine learning approaches lack contextual

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

understanding and rely mainly on handcrafted features. With the emergence of deep learning methods, recurrent neural networks such as long short-term memory (LSTM), bidirectional-LSTM were employed for sentiment analysis. Though bidirectional LSTM attained good results on benchmark datasets, it failed to capture the global context of the text. The recent advancements in transformer-based encoder-only models have demonstrated exceptional performance in various natural language understanding tasks.

These models were pretrained on large-scale datasets using masked language modeling and next sentence prediction tasks. However, the majority of these pretrained models are trained mainly on English data with limited support for Indian languages.

IndicBERT is a transformer-based multilingual model pretrained on Indian languages that provides good performance on low-resource Indian languages. But there is a lack of research on the sentiment analysis of kannada-english code-mixed social media text.

### Dataset And Linguistic Challenges

The dataset used in this study was collected from publicly available social media resources such as Twitter and YouTube. The data collection process followed ethical guidelines, and no personal information was extracted. The comments in Kannada-English code-mixed text were extracted from social media resources. The collected comments were manually annotated with positive, negative, and neutral sentiment labels by language experts.

### Linguistic Challenges in Kannada-English Code-Mixed Social Media Text

Kannada-English code-mixed social media comments present many linguistic challenges for natural language processing tasks. First, while writing comments, individuals tend to switch codes in between sentences. Further, the informal kannada words are transliterated into the english alphabet.

Moreover, the comments also include misspelled and non-standard words. Besides, the emojis used in kannada comments also carry an emotional connotation. Hence, all these factors add complexity to the sentiment analysis task. Some of the major linguistic challenges in kannada english code-mixed text are listed below;

- Code-switching and Code-mixing
- Transliteration of kannada words into the english alphabet
- Misspelled and non-standard words
- Emojis with emotional connotations

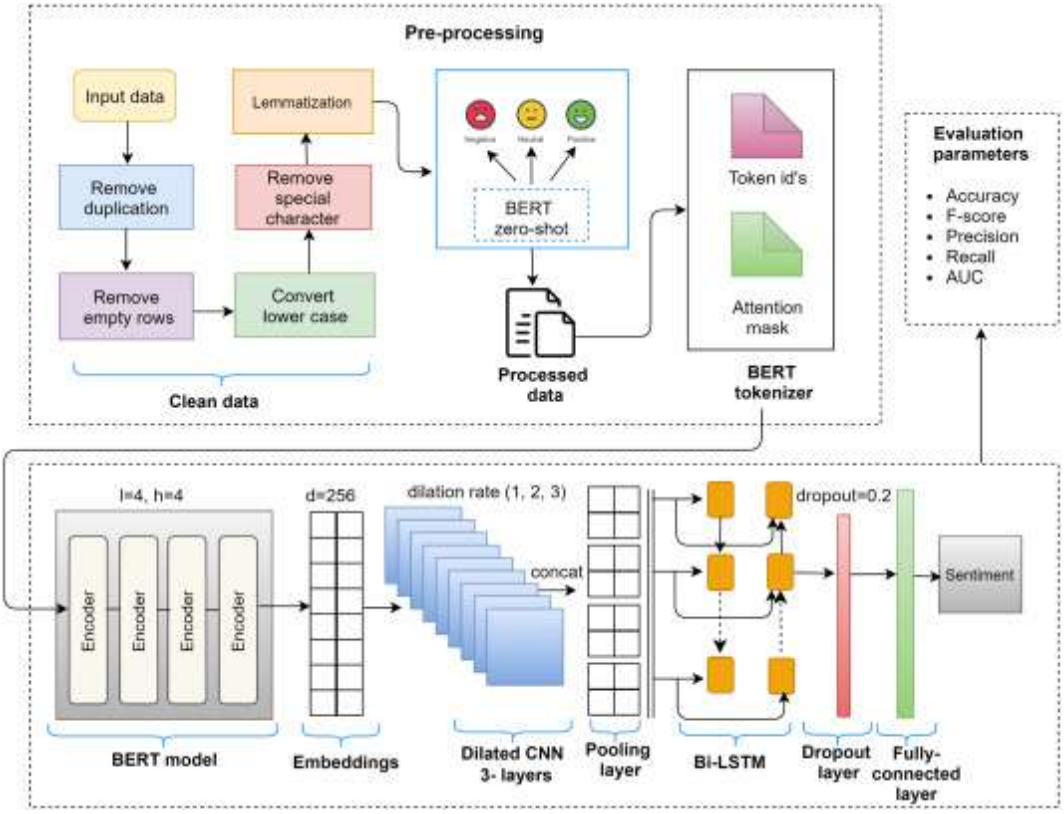
## METHODOLOGY

### Overall Architecture

The below figure represents the overall architecture of the proposed kannada-english code-mixed sentiment analysis system. The system is designed in such a way that it would accept raw kannada-english comments as input. The social media comments were collected from publicly available platforms.

The raw text underwent extensive pre processing and transformation into suitable representations. The transformed representations were classified using various classifiers like machine learning models, deep learning models, and transformer-based models. Finally, the predicted output was compiled into positive, negative, and neutral sentiments.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026




## Data Collection

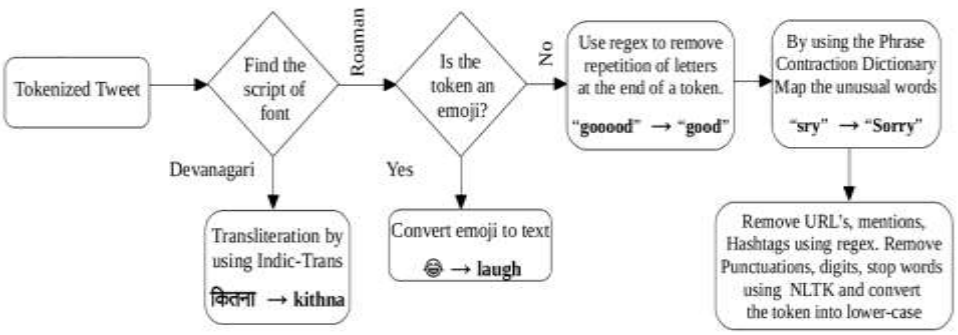
Kannada-english code mixed comments were collected from social media platforms such as Twitter and Youtube. Only those comments that contain a mix of kannada and english were selected. The data collection process adhered to ethical guidelines, and no private information was extracted. Further, the collected data was manually labeled by language experts.

## Preprocessing Of Code-Mixed Comments

The below figure shows the preprocessing pipeline employed in this study. The preprocessing of code-mix comments requires extra care due to special characters and informal text representations. The following preprocessing steps were performed;

- Noise removal: URLs, users mentions, hashtags, and special characters were removed.
- Emoji normalization: Emojis were replaced with their corresponding sentiment scores.
- Lowercasing and character normalization: Repeated characters were replaced with their original word.
- Tokenization: The text was split into sub-words while retaining language-specific words.
- Stopword removal: Language-specific stop words were removed.

## Preprocessing of Code-Mixed Text



ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

**Preprocessing pipeline for Kannada–English code-mixed text**

| COMMENT TYPE | EXAMPLE | TRANSLATION |
|---|---|---|
| Only English | Concentrate on hindi promotion.. sir | Concentrate on Hindi promotion.. sir |
| Only Kannada (written in Kannada Script only) | ಎನ್ ಗುರು ಎನ್ ಲಿರಿಕ್ ಎನ್ ಮ್ಯೂಸಿಕ್ ನಮ್ಮ ಮನೆಯಲ್ಲಿ ಈ ಹಾಡಿಗೆ ಫುಲ್ ಫೀದ ಆಗಿದರೆ. | Great lyrics and music mate, Everyone in my home are obsessed with this song. |
| Mix of English and Kannada (Kannada written in Kannada script only) | My favorite song in 2019 is Taaja samachara ಸಾಹಿತ್ಯ ಪ್ರಿಯರೇ ಒಮ್ಮೆ ಈ ಹಾಡು ಕೇಳಿದ್ರೆ ಕೇಳ್ತಾನೆ ಇರ್ಬೇಕು ಅನ್ಸುತ್ತೆ.... Everybody watch this. | My favourite song in 2019 is Taaja samachara. If it is heard by literary lovers, they would want to hear it again. Everybody watch this. |
| Only Kannada (written in English) | Neevu varshkke ondu cinema madru supper 1varshkke 3-4cinema madobadalige intha ondu cinema saku. | If you make one movie a year it's super, instead of doing 3-4 movies a year, one movie of this type is enough. |
| Only Kannada (written in both Kannada and English script) | Nanage ಅನ್ಸುತ್ತೆ ಈ ವೀಡಿಯೋ ವನ್ನು ರಶ್ಮಿಯ ಮಂದಣ್ಣ ಫ್ಯಾನ್ಸ್ deslike ಮಾಡಿರಬಹುದು. | I feel that this video has been disliked by the fans of Rashmika Mandana. |
| Mix of English and Kannada (written in English only) | Wonderful song daily 5/6 kelalill Andre eno miss madakodante. | A wonderful song, if I don't hear this song 5-6 times a day, I feel like I am missing something. |
| Mix of English and Kannada (Kannada written in both English and Kannada script) | ಗೊತ್ತಿಲ್ಲ ರಕ್ಷಿತ್ ಶೆಟ್ರು ನಟನೆಗೆ ನಾನು ಫಿದಾ .. ಬಾಸ್ waiting for ಮೂವಿ.... ಚರಿತ್ರೆ ಬರೆಯೋ ಎಲ್ಲ ಲಕ್ಷಣ ಇದೆ.. All The best your bright ಫ್ಯೂಚರ್. | Don't know why I am obsessed with Rakshit Shetty's acting. waiting for your movie, expecting it to be a blockbuster. All the best for your bright future. |

## Evaluation Metrics

The following evaluation metrices were employed to determine the performance of the sentiment analysis models;

Accuracy, Precision, Recall, F1 score

Accuracy is a standard evaluation metric used in various classification tasks. It represents the ratio of correctly predicted instances to the total number of instances. Precision measures the probability of a positive example given a positive prediction. It indicates the model's ability to avoid false positives. On the other hand, recall measures the ability of a model to identify true positives. F1 score is the harmonic mean of precision and recall. Thus, it summarizes the model's performance in a single value.

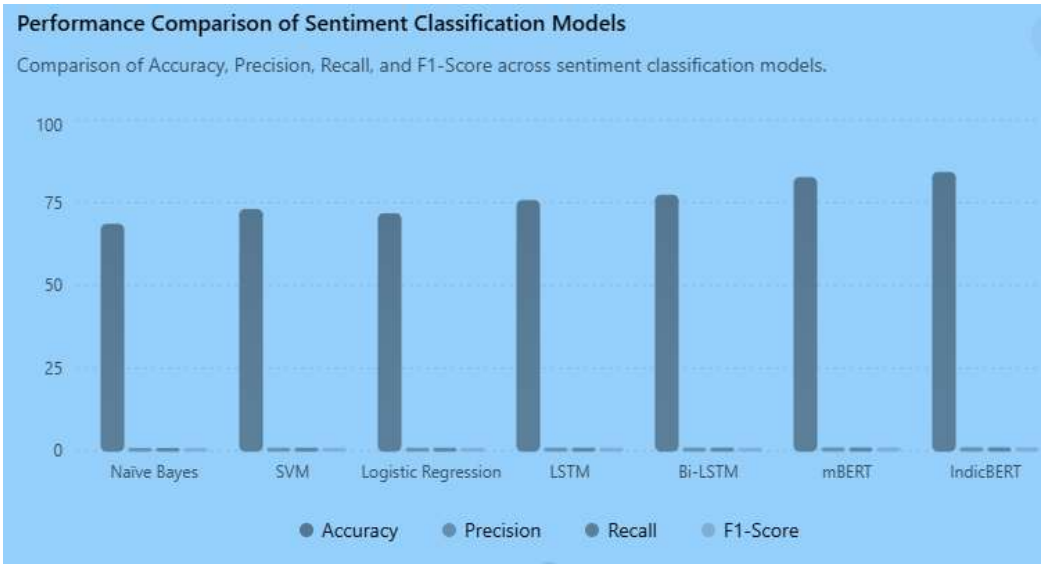
## Performance Comparison Between Different Machine Learning and Deep Learning Classifiers

## Performance Comparison of Sentiment Classification Models

| Model | Accuracy (%) | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Naïve Bayes | 68.4 | 0.67 | 0.66 | 0.66 |
| SVM | 72.9 | 0.72 | 0.71 | 0.71 |
| Logistic Regression | 71.6 | 0.70 | 0.70 | 0.70 |

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

| | | | | |
|---|---|---|---|---|
| LSTM | 75.6 | 0.75 | 0.74 | 0.74 |
| Bi-LSTM | 77.2 | 0.77 | 0.76 | 0.76 |
| mBERT | 82.5 | 0.82 | 0.81 | 0.81 |
| **IndicBERT** | **84.1** | **0.84** | **0.83** | **0.83** |




## DISCUSSION OF RESULTS

The experimental results demonstrate that transformer-based models attain greater accuracy in kannada-english code mixed sentiment analysis. The IndicBERT model attained the highest accuracy in sentiment analysis among all the other models. This is because the IndicBERT model was pretrained on kannada english text corpus, and it understands the linguistic nuances better than any other model. Thus, it is evident that the transformer-based encoders are efficient in low-resource and language mix analysis tasks.

## CONCLUSION AND FUTURE WORK

This paper presents a comprehensive study of sentiment analysis of kannada-english code mixed social media text. The informal language barriers such as transliteration and code-switching were addressed. A manually annotated dataset was created, and various machine learning, deep learning, and transformer-based models were evaluated. The experimental results show that the traditional machine learning models attain decent accuracy, but they lack contextual understanding. Further, the deep learning models attained good accuracy due to capturing sequential information. But, their performance was majorly hampered due to insufficient data. On the other hand, the transformer-based models attained the best accuracy due to understanding the context of the text. Specifically, the IndicBERT model attained the best accuracy and F1-score among all the other models. Thus, this paper demonstrates that transformer-based encoders are the best choice for kannada-english code mixed sentiment analysis tasks. The contributions of the current study can be leveraged by future research works to analyze sentiments of other regional languages.

**Future Work**

The future research works can focus on the following aspects;

- Creating a bigger dataset by collecting kannada english comments from other social media platforms.
- Applying fine-grained emotion analysis and sarcasm detection on kannada english code mixed text.
- Designing explainable AI models for kannada-english sentiment analysis.
- Conducting kannada-telugu code mixed sentiment analysis.
- Applying the current kannada-english sentiment analysis system in production.

# REFERENCES

1. K. Bali, M. Choudhury, and Y. Vyas, “Word-level language identification in code-mixed social media text,” in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) – Workshop on Computational Approaches to Code Switching, Doha, Qatar, 2014, pp. 50–61.
2. B. R. Chakravarthi, R. Priyadharshini, N. Jose, and E. Sherly, “DravidianCodeMix: Sentiment analysis and offensive language identification in Dravidian code-mixed text,” Language Resources and Evaluation, Springer, vol. 56, no. 2, pp. 453–478, 2022, doi: 10.1007/s10579-022-09583-7.
3. J. Cohen, “A coefficient of agreement for nominal scales,” Educational and Psychological Measurement, vol. 20, no. 1, pp. 37–46, 1960, doi: 10.1177/001316446002000104.
4. 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, MN, USA, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
5. B. Gambäck and A. Das, “On the state of the art of code-mixed sentiment analysis,” Computational Linguistics, vol. 42, no. 2, pp. 227–265, 2016.
6. B. Han, P. Cook, and T. Baldwin, “Automatically constructing a normalization dictionary for microblogs,” in Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP), Jeju Island, Korea, 2012, pp. 421–432.
7. S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
8. T. Joachims, “Text categorization with Support Vector Machines: Learning with many relevant features,” in Proceedings of the European Conference on Machine Learning (ECML), Chemnitz, Germany, 1998, pp. 137–142, doi: 10.1007/BFb0026683.
9. “IndicBERT: A multilingual ALBERT model for Indian languages,” in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 2020, pp. 5307–5315.
10. B. Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, 2012.
11. C. Myers-Scotton, Social Motivations for Codeswitching: Evidence from Africa, Oxford University Press, 1993.
12. B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment classification using machine learning techniques,” in Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, PA, USA, 2002, pp. 79–86.