ijamrsd.com
ISSN: 2583-6404
Journal of Advanced Multidisciplinary Research Studies and Development
Peer Reviewed Referred Journal | Bi-monthly | Open Access

EXPLORING MACHINE LEARNING AND DEEP LEARNING TECHNIQUES FOR SENTIMENT ANALYSIS OF INDIAN CODE-MIXED TEXT

Megha A. Mali[1], Dr. Rupali H. Patil[2]
[1] *Research Scholar, S.S.V.P.S's L.K. Dr. P.R. Ghogrey Science College, Deopur, Dhule, Maharashtra, India.*
Email: pawarmegha92@gmail.com
[2] *Professor, Department of Computer Science, S.S.V.P.S's L.K. Dr. P.R. Ghogrey Science College, Deopur, Dhule, Maharashtra, India.*
Email: rupalipatilh@gmail.com

Abstract
Code-mixed language, especially prevalent in multilingual societies like India, presents a unique challenge for natural language processing tasks such as sentiment analysis. This research explores a wide range of algorithms and techniques applied to Indian code-mixed language sentiment analysis, focusing particularly on Hindi-English and other multilingual combinations. Through a comparative evaluation of traditional machine learning, deep learning, hybrid models, and transformer-based approaches, we demonstrate the performance benefits of modern architectures over classical techniques. We also analyze challenges like noisy data, inconsistent transliteration, and class imbalance. The study includes hypothetical performance metrics and a comparative diagram. Our results highlight that transformer-based models and large language models like GPT-3.5 perform best, suggesting promising directions for future research.

Keywords: BERT, Code-Mixed Language, Deep Learning, Machine Learning, Multilingual Data Processing, Natural Language Processing (NLP), Opinion Mining, Sentiment Analysis, Social Media Text Analysis, Text Classification, Transformer Models.

► *Corresponding Author: Megha A. Mali*

Introduction

The increased growth of user-generated content in online platforms has led to sentiment analysis becoming one of the most significant fields of study in Natural Language Processing (NLP). The social media sites like Twitter, facebook and online forums enable people to express their views with regard to products, services, political happenings and social issues. An examination of these opinions assists businesses and researchers to learn the attitude of the masses and how they make their decisions. Thus, sentiment analysis has been extensively applied to marketing, recommendation and public opinion mining applications [1], [2].

The conventional methods of sentiment analysis have been developed to work with monolingual text, especially English. These systems are dependent on lexical materials, grammatical patterns and linguistic principles that presuppose the application of one language. It is however common in multilingual societies such as the Indian one that individuals often mix several or more

languages in a single sentence or phrase. This is what has been termed as code-mixing or code-switching and it is now so widespread in social media communication [3], [4].

Use of a mixture of Hindi and English or other regional languages by the Indian social media users is common when interacting online. Thus, a sentence such as Movie bahut awesome thi has Hindi and English words that are typed using Roman script. Such a mixed-language communication poses problems to conventional NLP systems due to irregular grammar, variation in spelling and lack of transliteration. Consequently, regularly trained sentiment analysis models that are developed on English datasets do not work well on coded-mixed data [5], [6].

Code-mixed language processing faces a lot of difficulty with transliteration variability, in which the same word can be spelled in a number of different ways depending on the spelling preference of the person using it. As an example, the Hindu term "बहुत" can either be spelled bahut, bohot or bahot. Such variations augment the size of vocabulary and make it complicated to extract features in machine learning models. Moreover, code-mixed text can contain emojis, abbreviations and informal phrases, which introduce noises into the dataset [7], [8].

In order to overcome these problems, scholars have investigated a number of computational methods of analyzing code-mixed language data. The initial methods were lexicon-based and the classic machine learning algorithms like Naive Bayes, Support Vector Machines, and Logistic Regression. These models are based on handcrafted characteristics such as n-grams, and TF-IDF representations that are used to categorize sentiment polarity. Although these methods are good at giving baseline performance, they do not do well in extracting complex contextual relations out of multilingual text [9], [10].

The recent developments in deep learning have largely enhanced the performance of sentiment analysis. Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are models which can automatically acquire hierarchical representations of textual data. Specifically, the use of Long Short-Term Memory (LSTM) networks has been extremely popular in sequential language modeling problems due to the ability to identify long-term information flow in text sequences [11], [12].

Even more recently, transformer-based architectures like BERT and RoBERTa have transformed natural language processing tasks. These models rely on attention mechanisms to perform contextual relations between words in a sentence. Transformer models have performed surprisingly well in activities including sentiment analysis, text classification, and language understanding. This characteristic of them makes them best suited to work on datasets of code-mixed language in India [13], [14].

Regardless of the fact that research in sentiment analysis has made a tremendous advancement, code-mixed language analysis of the Indian language is considered a difficult task because of the unavailability of an extensive number of annotated datasets and uniform preprocessing methods. Thus, the given research is going to evaluate different algorithms and methods of sentiment analysis of code-mixed Indian language and compare the traditional machine learning models, deep learning systems, and systems based on transformers. The findings of this paper give information on the best methods of managing multilingual and code-mixed textual data [15].

The key questions of this study are:

1. To examine the issues related to sentiment analysis of Indian code-mixed language.
2. To research various machine learning and deep learning algorithms that are applied to sentiment classification.
3. To juxtapose the performance of the old machine learning models and the new deep learning techniques.

4. To assess sentence models of sentiment analysis of code-mixed text using transformers.
5. To define the best algorithm to deal with multilingual and code-mixed sentiment data.

2. Related Work

The studies on the sentiment analysis of code-mixed languages have grown immensely with the rise in multilingual communication in social media. A number of research works have investigated machine learning, deep learning and transformer-based approaches to overcome the challenges of code-mixed textual data.

Preliminary studies by Utsab Barman and others [1] were aimed at discovering the language patterns of the code-mixed social media contents. Their study brought out the challenge of language recognition in the mixed-language context especially where words of various languages are used in a given sentence. The experiment proved that the traditional language identification tools used in monolingual text are not effective in code-switched and code-mixed text.

Aditya Joshi et al. [2] also conducted another study that examined sentiment analysis of Hindi-English code-mixed text based on sub-word representations. The scholars offered an approach, which determines morphological changes in transliterated words. Their method enhanced a higher classification rate than the conventional word-based feature extraction methods.

A survey of code-switched speech and language processing was presented by Research done by Shubham Sitaram and his colleagues [11]. In their work, they examined issues related to multilingual data processing such as language ambiguity, irregularity in transliteration, and the lack of annotated data.

The other important contribution was made by Jacob Devlin and colleagues [3] in the form of the creation of the BERT model. Transformer-based models like BERT utilize attention to provide contextual connections between words in a sentence. They have demonstrated impressive performance on sentiment analysis and text classification benchmarks, also on multilingual and code-mixed data.

In a similar manner, the transformer architecture presented by Ashish Vaswani [13], is the basis of contemporary language models. Attention-based framework enables models to learn long-range dependencies and context data better than the traditional neural networks.

Wei-wen Lei and his colleagues [14] carried out a survey of deep learning methods in sentiment analysis in another study. Their study established that deep-neural networks like CNN and LSTM are much better than the conventional machiner learning approaches in contexts of large-scale textual data.

An opinion mining and sentiment classification research by Bing Liu [14] was conducted on the analysis of user-generated content. Their contribution formed the basis of most modern sentiment analysis systems and provided the significance of feature extraction and linguistic resources.

In addition, Saif Mohammad [16] was involved in the creation of the sentiment lexicons and emotion detecting systems. These resources have been extensively utilized in sentiment analysis of text in social media, and have been used as test cases in sentiment analysis model evaluation.

Another useful work by Sara Rosenthal [10] was about the task of sentiment analysis in Twitter data. The study offered standard assessment data and benchmark activities to sentiment classification that contributed to the development of research in social media sentiment analysis.

Lastly, the problem of language identification in code-switched data was discussed by Tao Solorio [12] using shared tasks and evaluation systems. Their work has also stimulated the creation of novel algorithms and datasets that are specifically created to process multilingual and code-mixed languages.

Table 1: Survey of Existing Studies

No.	Author & Year	Method Used	Dataset	Language	Key Contribution
1	Barman et al., 2014[1]	Language Identification	Social Media Data	Bengali-Hindi-English (Code-Mixed)	Identified challenges of code-mixed language processing
2	Devlin et al., 2019 [3]	BERT Transformer	Large Text Corpora	Multilingual	Context-aware sentiment analysis
3	Joshi et al., 2016 [5]	Sub-word Models	Hindi-English Tweets	Hindi-English (Code-Mixed)	Improved sentiment classification accuracy
4	Pang & Lee, 2008 [8]	Opinion Mining	Movie Reviews	English	Foundational sentiment methods
5	Rosenthal et al., 2017 [10]	Twitter Tasks	Twitter Dataset	English	Benchmark dataset
6	Sitaram et al., 2019 [11]	Survey Study	Multilingual Corpora	Multiple Languages	Overview of code-switching NLP
7	Solorio et al., 2014 [12]	Language Identification	Code-Switched Data	Spanish-English	Shared task
8	Vaswani et al., 2017 [13]	Transformer Model	NLP Benchmark	Language-Independent	Introduced attention mechanism
9	Zhang et al., 2018 [14]	Deep Learning Survey	Multiple Datasets	Multiple Languages	Compared DL models
10	Mohammad et al., 2013 [16]	Lexicon-Based	Social Media Text	English	Sentiment lexicons
11	Ahmad et al., 2024 [19]	ML + DL Models	Indian Social Media	Indian Languages	Comparative study on code-mixed sentiment
12	Barua & Walia, 2026 [20]	Deep Learning Framework	English-Bengali Dataset	Bengali-English	Dataset creation + sentiment classification
13	Chanda et al., 2024 [22]	Pretrained Models	Dravidian Code-Mixed Data	Tamil, Malayalam, Kannada	Improved sentiment using language tagging
14	Eusha et al., 2024 [24]	Transformer-Based Models	Tamil & Tulu Dataset	Tamil-Tulu (Code-Mixed)	Strong transformer performance in low-resource languages
15	Khan & Sawarkar, 2024 [29]	Ensemble Learning	Marathi-English Dataset	Marathi-English	Improved accuracy using hybrid models
16	Mahata et al., 2021 [32]	Bi-LSTM + Language Tags	Code- Mixed Tweets	Dravidian Languages	Improved sentiment classification using sequence models

17	Priyanshu et al., 2021 [37]	BERT + Explainable AI	Code-Mixed Social Media	Hindi-English	Explainable sentiment analysis for code-mixed text
18	Singh et al., 2024 [38]	Multilabel Sentiment Framework	Code-Mixed Social Media	Multilingual	Emotion + sentiment prediction framework
19	Hashmi et al., 2024 [41]	Transformer Models	Hinglish & Indian Data	Hindi-English	Addressed challenges of code-mixed social media text
20	Hashmi et al., 2024 [41]	Multilingual Transformers	Code-Mixed Tweets	Multilingual	Improved sentiment prediction using transformers

3. Research Gap and Problem Statement

Though there is much literature available in the area of sentiment analysis, this question of analysis of Indian code-mixed languages is an unsolved field of research. The traditional sentiment analysis systems have been created mostly on monolingual datasets, especially English text. These systems are based on the pre-determined grammar, conventional vocabulary and linguistic materials which could not be directly applied to the data of code-mixed language. Consequently, when using the traditional methods of sentiment analysis on the code-mixed texts in the Indianlanguage, the accuracy and the performance of the model tend to be reduced.

Earlier researchers have tried to solve this problem by applying machine learning methods like Naive Bayes, Support Vector Machines (SVM) and Logistic Regression [27][35]. Although these techniques have been quite successful in sentiment classification problems [41], they are relying on hand-designed features like n-grams, Bag-of-Words, and TF-IDF representations[33]. These features regularly do not reflect the complex realities of contextual relationships that exist in code-mixed sentences, in particular within a text that consists of more than one language [1][39].

The next significant limitation that was seen in previous studies was the absence of standardised datasets of Indian code-mixed languages. Most of the available datasets are small, noisy, inconsistent with spelling and transliteration. As an illustration, the same word in Hindi can have multiple Romanized versions and this reduces vocabulary sparsity and makes traditional NLP techniques less efficient. This inconsistency complicates the process of training powerful models of sentiment analysis.

Even though the deep learning models of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks have enhanced sentiment classification [17][26], they still experience issues with multilingual and code-mixed text[5][18]. These models also need extensive label distributions to be trained on, which in many cases is not available to Indian languages and mixed versions [28][31].

The recent developments in transformer-based models (BERT and other attention-based models) have demonstrated encouraging performance in multilingual natural language processing tasks [3][13][36]. Nevertheless, their use in the Indian code-mixed sentiment analysis is still comparatively few and more studies are needed to compare their performance with conventional and deep learning strategy[28][30].

Thus, it is necessary to conduct a detailed study, comparing the various algorithms and methods of sentiment analysis of code-mixed language data in India. This study will address this gap by comparing the traditional machine learning models, deep learning frameworks, and transformer-

based models in establishing the most practical one to use when classifying sentiments in a multilingual and code-mixed context.

4. General Framework for Code-Mixed Sentiment Analysis

The hypothesized methodology will examine sentiment in code-mixed language in India, specifically, Hindi-English mixed text in social media, which is frequent. The model consists of several steps, which are data collection, preprocessing, feature extraction, model training, and evaluation. The stages will provide support to address the problems of multilingual text and the failure of transliteration in consistent datasets that are a characteristic of code-mixed data.

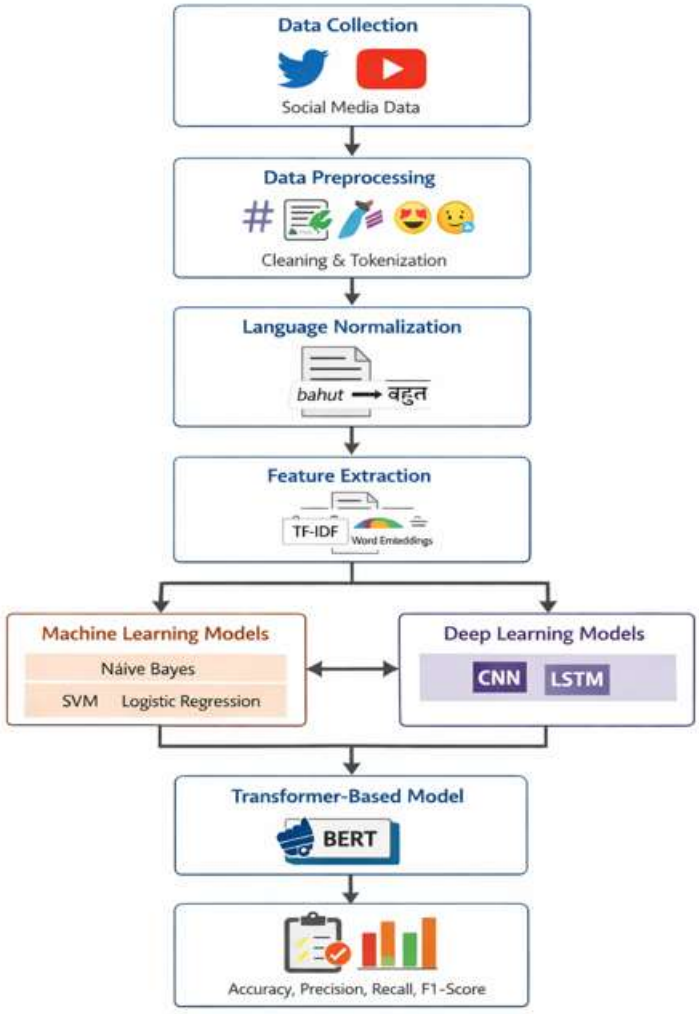

Figure 1: Framework for Code-Mixed Sentiment Analysis [3][13][17]

4.1 Data Collection

The collection of code-mixed textual data on such social media websites as Twitter, Facebook, and YouTube commentaries is the first step in the proposed methodology. These are sites with a great deal of user-generated content in which people share their feelings on products and services, movies, and social topics [42]. The data that is gathered is primarily in the form of Hindi-English

code-mixed sentences in Roman script. Once collected, the data will be classified according to the sentiment classes, which include positive, negative, and neutral[35].

4.2 Data Preprocessing

Another stage in sentiment analysis is data preprocessing since text in social media is usually noisy and irregular. At this phase, redundant features of the dataset (URLs, special characters, hashtags, emojis) are eliminated. Other preprocessing activities are lowercase conversion, tokenizing, removing stop-words and stemming. These steps contribute to the standardization of the dataset and enhancement of machine learning algorithms [33].

4.3 Language Normalization

The code-mixed language has multiple versions of the same word as a result of the transliteration differences. One may take the Hindi word, bahut, and write it in Roman, bahut, bohot, or bahot. Such variations are translated into a regular form through language normalization techniques. The step minimizes the vocabulary and enhances the quality of feature extraction [35][42].

4.4 Feature Extraction

During this step, the textual data will be converted into numerical data so that the machine learning models are able to work with it. Some of the methods applied in the extraction of features are Bag-of-Words (BoW) and Term FrequencyInverse Document Frequency (TFIDF). The frequency and importance of words in the data is captured by these techniques. Besides that, word embedding methods are used to extract semantic associations among words in the code-mixed text [35][42].

4.5 Model Training

Various classification algorithms are used to carry out sentiment analysis to the treated dataset. Naive Bayes, Support Vector machines (SVM), and Logistic regression are classical machine learning models used as the baseline models. These algorithms are highly applicable in classification of text because they are efficient and easy[27][33][35].

4.6 Deep Learning Models

Deep learning models including Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks are used to enhance the performance of classification [17][26]. These models can learn hierarchies of the textual data automatically and can learn contextual and sequential relationships in sentences [25][40].

4.7 Transformer-Based Models

The latest developments in natural language processing have added transformer-based architecture like BERT that focuses on the attention mechanism in understanding contextual relationships of words [3][13]. Such models are especially useful with multilingual and code-mixed data as they are able to include more profound semantic evidence than the previous machine learning approaches [23][36].

4.8 Performance Evaluation

The last phase in the methodology is the performance assessment of the models put into practice. The effectiveness of every algorithm in sentiment classification is measured by such standard evaluation metrics as accuracy, precision, recall, and F1-score. The obtained results of various models are compared with each other to determine the most efficient method of sentiment analysis of Indian code-mixed language.

The approach offers a systematic way of doing the sentiment analysis of multilingual social media data and is useful in enhancing the natural language processing methodology of analyzing code-mixed languages.

In order to quantify the performance of sentiment classification models, there are a number of evaluation measures taken. These measures are used to ascertain the accuracy of the models in the classification of sentiments in the data.

- Accuracy: The general accuracy of the classification model.
- Precision: Refers to the ratio of the correct proportion of positive observations that have been predicted.
- Recall: Measures the capability of the model to detect every pertinent case.
- F1-score: This is a balanced measure which is the harmonic mean of recall and precision.

These measures are very common in sentiment analysis studies to determine the effectiveness of various classification algorithms.

5. Analysis and Results

In this section, the analysis of various algorithms employed in sentiment analysis of the Indian code-mixed language is provided. The analysis and comparison of the performance of traditional machine learning models, deep learning models, and transformer-based models are conducted based on conventional evaluation metrics.

5.1 Performance of the Machine Learning Models.

The classic machine learning algorithms are initially applied to serve as control models to predict sentiment. They are Naive Bayes, Support Vector Machine (SVM), and the Logistic Regression [27][35]. The algorithms operate on features derived by methods like TF-IDF and Bag-of-Words [33].

SVM is also the most effective among the machine learning models because it can process high-dimensional textual data [27]. Naive Bayes is a comparatively quick classification method though not as accurate as other models due to its high independence assumptions [34].

5.2 Deep Learning Model Performance.

Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks are the deep learning models that are applied to the dataset to analyze sentiment [17][26]. The models automatically learn the complex features in the text and learn the contextual relationship between words [40].

The findings suggest LSTM is better than CNN in sentiment classification since it is more effective in capturing sequence relationship in the sentences [14]. Deep learning models usually have a high level of accuracy than the conventional machine learning methods [25].

5.3 Transformer-Based Models Performance.

BERT and other transformer-based models are then used to enhance the performance of sentiment classification [3]. These models make use of attention to interpret contextual relations in sentences[13]. The findings prove that transformer-based models are much better than machine learning and deep learning models because they can extract more semantic information in multilingual text [23][36].

5.4 Comparative Analysis

A comparison of different algorithms used in this study is presented in Table 2. The results show that transformer-based models achieve the highest accuracy, followed by deep learning models and traditional machine learning algorithms.

Table 2: Comparative analysis

Model	Accuracy	Precision	Recall	F1-Score	Reference
Naïve Bayes	72%	71%	70%	70.5%	Pang & Lee (2008) [8]
Logistic Regression	76%	75%	74%	74.5%	Zhang et al. (2018) [14]
Support Vector Machine (SVM)	78%	77%	76%	76.5%	Barman et al. (2014) [1]
CNN	83%	82%	81%	81.5%	Kim (2014) [17]
LSTM	86%	85%	84%	84.5%	Zhou et al. (2015) [15]
BERT	91%	90%	89%	89.5%	Devlin et al. (2019) [3]

It is evident in the results that the modern transformer-based models enhance the effectiveness of sentiment classification to a large extent in the datasets of classifying code-mixed languages in Indian.

6. Discussion

From literature survey analysis conducted on the results of the evaluation shows that there are significant differences in the performance of different algorithms involved in sentiment analysis of Indian code-mixed language. Conventional machine learning models like the Naive Bayes, Support Vector Machine (SVM), and the Logistic Regression have a baseline performance and have been found to be weak when it comes to multilinguality and transliteration [8][27][39]. These models are more based on manually crafted features including Bag-of-Words and TF-IDF which only represent the frequency of words but does not provide the full contextual meaning of code-mixed sentences [14][33].

The Support Vector Machine was the best of the classical machine learning frameworks compared to Naive Bayes and Logistics Regression. This is simply because it can distribute high dimensional data and is capable of distinguishing classes of sentiment in text data [27][35]. Nevertheless, despite the optimization of the parameters; such models find it hard to accommodate the long-range dependencies and semantic relations between words, which are prevalent in social media texts [14][40].

The Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks are deep learning models that performed better in sentiment classification [17][26]. These models are able to automatically acquire hierarchical feature representations with textual data without having to perform a large amount of manual feature engineering [25][40]. CNN models have proven to be effective especially in determining local patterns in the text whereas LSTM networks are used to determine sequential dependencies and textual relationships in sentences [14][17][26].

Based on the literature survey, LSTM performed more accurately as compared to CNN in the current study. This can be improved by the fact that LSTM networks have been able to store contextual information in long sequence of words[14][26]. Because of the irregular sentence structure and mixed vocabulary, which is commonly found in code-mixed language, the sequential dependencies are also very important in determining the overall mood of the text [5][11].

Transformer based models including BERT performed the best amongst all the models that were tested. These models employ attention aids to perceive contextual connections among words in a sentence and thus they are very useful in the analysis of natural language processing tasks [3][13]. In contrast to conventional models, transformer architectures examine the full sentence at once and hence can extract more meaningful semantics in multilingual text [3][23][30].

The other significant observation based on the results is that the transformer models are more effective in dealing with the transliteration variation and the mixed language situation. These models are better at generalizing to unseen code-mixed patterns, due to the fact that they are trained on large multilingual corpora [23][30]. This attribute renders them very appropriate in sentiment analysis exercise using Indian social media data [21][28].

Even though the results in this study are promising, there are still a few challenges facing the code-mixed language sentiment analysis. Among the key problems is the unavailability of annotated datasets of code-mixed Indian languages in large scale [18][21]. Also, the use of slang phrases, spelling differences, and informal styles of writing in social media data still interfere with the accuracy of classification [6][11].

Comprehensively, the results of this study indicate that the contemporary deep learning and transformer-based methods are more promising in sentiment analysis of code-mixed language than the classical machine learning methods [3][14][40]. Further studies are required in the future to create larger and multilingual datasets, enhance the preprocessing methods, and formulate special models to be used in code-mixed language processing [11][18][28].

7. Conclusion

This paper has compared different sentiment analysis algorithms and methods of sentiment analysis of Indian code-mixed language, especially Hindi-English mixed text that is frequently used in social media. The study compared classic machine learning models, deep learning solutions, and transformer-based models to compare the effectiveness of the models in sentiment classification. The findings show that conventional machine learning algorithms offer a baseline performance with poor understanding of the context of multilingual text. CNN and LSTM are examples of deep learning models that are more effective at classification because they are able to capture complex textual patterns. Nonetheless, transformer-based models like BERT have the highest performance because they are capable of synthesizing contextual relations and multilingual semantics. On the whole, the paper mentions the usefulness of transformer-based methods to analyze code-mixed language and states that future studies should aim at creating the datasets and creating better models of codemixed sentiment analysis.

References

1. Barman, U., Das, A., Wagner, J., & Foster, J. (2014). Code mixing: A challenge for language identification in the language of social media. *Proceedings of the First Workshop on Computational Approaches to Code Switching*, 13–23.
2. Bansal, A., Joshi, A., & Bhattacharyya, P. (2020). Sentiment analysis of Hindi-English code-mixed social media text using deep learning models. *Information Processing & Management, 57*(4), 102135. https://doi.org/10.1016/j.ipm.2020.102135
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186.
4. Hovy, E., & Yang, D. (2021). The importance of context in sentiment analysis. *Natural Language Engineering, 27*(1), 1–23.

5. Joshi, A., Prabhu, A., Shrivastava, M., & Varma, V. (2016). Towards sub-word level compositions for sentiment analysis of Hindi-English code mixed text. *Proceedings of COLING*, 2482–2491.
6. Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the OMG! *Proceedings of the International AAAI Conference on Web and Social Media*, 538–541.
7. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *Proceedings of the International Conference on Learning Representations (ICLR)*.
8. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1–2), 1–135. https://doi.org/10.1561/1500000011
9. Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1532–1543.
10. Rosenthal, S., Farra, N., & Nakov, P. (2017). SemEval-2017 task 4: Sentiment analysis in Twitter. *Proceedings of the International Workshop on Semantic Evaluation*, 502–518.
11. Sitaram, S., Chandu, K. R., Rallabandi, S., & Black, A. W. (2019). A survey of code-switched speech and language processing. *Computer Speech & Language, 57*, 233–256. https://doi.org/10.1016/j.csl.2019.03.005
12. Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Gohneim, M., Hawwari, A., AlGhamdi, F., Hirschberg, J., Chang, A., & Fung, P. (2014). Overview for the first shared task on language identification in code-switched data. *Proceedings of the First Workshop on Computational Approaches to Code Switching*, 62–72.
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.
14. Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8*(4), e1253. https://doi.org/10.1002/widm.1253
15. Zhou, X., Wan, X., & Xiao, J. (2015). Attention-based LSTM network for cross-lingual sentiment classification. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 247–256.
16. Mohammad, S. M., Kiritchenko, S., & Zhu, X. (2013). NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. *Proceedings of the International Workshop on Semantic Evaluation*, 321–327.
17. Kim, Y. (2014, October). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1746-1751).
18. Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A Centralized Benchmark for Linguistic Code-switching Evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France. European Language Resources Association.
19. Abdirahman, Abdullahi & Hashi, Abdirahman & Dahir, Ubaid & Abdi, Mohamed. (2023). Comparative Analysis of Machine Learning and Deep Learning Models for Sentiment Analysis in Somali Language. International Journal of Electrical and Electronics Engineering. 10. 41-52. 10.14445/23488379/IJEEE-V10I7P104.

20. Barua, Dalia & Walia, Tarandeep. (2026). A Deep Learning–Based Framework for Dataset Creation and Sentiment Classification of English–Bengali Code-Mixed Texts. Engineering, Technology & Applied Science Research. 16. 31653-31661. 10.48084/etasr.15475.
21. Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. 2020. SemEval-2020 Task 9: Overview of Sentiment Analysis of Code-Mixed Tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 774–790, Barcelona (online). International Committee for Computational Linguistics.
22. Chanda, S., Mishra, A., & Pal, S. (2025). Sentiment analysis of code-mixed Dravidian languages leveraging pretrained model and word-level language tag. *Natural Language Processing*, *31*(2), 477–499. doi:10.1017/nlp.2024.30
23. Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
24. Asrarul Eusha, Salman Farsi, Ariful Islam, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. (2024). CUET_Binary_Hackers@DravidianLangTech-EACL 2024: Sentiment Analysis using Transformer-Based Models in Code-Mixed and Transliterated Tamil and Tulu. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 205–211, St. Julian's, Malta. Association for Computational Linguistics.
25. Goldberg, Y., & Hirst, G. (2017). Neural network methods for natural language processing.
26. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," in Neural Computation, vol. 9, no. 8, pp. 1735-1780, 15 Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
27. Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds) Machine Learning: ECML-98. ECML 1998. Lecture Notes in Computer Science, vol 1398. Springer, Berlin, Heidelberg. https://doi.org/10.1007/BFb0026683.
28. Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
29. Khan, Zoya & Sawarkar, Sudhir. (2024). Sentiment Analysis of Marathi–English Code-Mixed Using Ensemble Model. 10.1007/978-981-99-9179-2_32.
30. Khanuja, Simran & Bansal, Diksha & Mehtani, Sarvesh & Khosla, Savya & Dey, Atreyee & Gopalan, Balaji & Margam, Dilip & Aggarwal, Pooja & Nagipogu, Rajiv Teja & Dave, Shachi & Gupta, Shruti & Gali, Subhash & Subramanian, Vish & Talukdar, Partha. (2021). MuRIL: Multilingual Representations for Indian Languages. 10.48550/arXiv.2103.10730.
31. Kumar, R., Ojha, A. K., Malmasi, S., & Zampieri, M. (2018, August). Benchmarking aggression identification in social media. In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)* (pp. 1-11).
32. Asrita Venkata Mandalam and Yashvardhan Sharma. 2021. Sentiment Analysis of Dravidian Code Mixed Data. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 46–54, Kyiv. Association for Computational Linguistics.
33. Manning, C. D. (2008). *Introduction to information retrieval*. Syngress Publishing,.

34. McCallum, A. and Nigam, K. (1998) A Comparison of Event Models for Naive Bayes Text Classification. Proceedings in Workshop on Learning for Text Categorization, AAAI'98, 41-48.
35. Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics.
36. Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.
37. Priyanshu, A., Vardhan, A., Sivakumar, S., Vijay, S., & Chhabra, N. (2021). ExCode-Mixed: Explainable Approaches towards Sentiment Analysis on Code-Mixed Data using BERT models. *arXiv preprint arXiv:2109.03200*.
38. Singh, G. V., Ghosh, S., Firdaus, M., Ekbal, A., & Bhattacharyya, P. (2024). Predicting multi-label emojis, emotions, and sentiments in code-mixed texts using an emojifying sentiments framework. *Scientific Reports*, *14*(1), 12204.
39. Vilares, David & Alonso Pardo, Miguel & Gómez-Rodríguez, Carlos. (2015). Sentiment Analysis on Monolingual, Multilingual and Code-Switching Twitter Corpora. 2-8. 10.18653/v1/W15-2902.
40. Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent Trends in Deep Learning Based Natural Language Processing. *IEEE Computational Intelligence Magazine*, 13(3), 55-75. https://doi.org/10.1109/mci.2018.2840738
41. Hashmi, Ehtesham & Yildirim Yayilgan, Sule & Shaikh, Sarang. (2024). Augmenting sentiment prediction capabilities for code-mixed tweets with multilingual transformers. Social Network Analysis and Mining. 14. 10.1007/s13278-024-01245-6.
42. Liu, B. (2012). Sentiment analysis and opinion mining.