IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Comprehensive Review Of Machine Learning Approaches For Sentiment Analysis Of Code-Mixed Social Media Text

[1]Anita, [2]Ajit Kumar

[1]Research Scholar, [2] Associate Professor, Multani Mal Modi College, Patiala

[1]Department of Computer Science,

[1]Punjabi University, Patiala, India

Abstract: The rapid rise in social media use has led people to code-mix languages, mixing different languages in one sentence or conversation. This is common in multilingual communities and creates challenges for tasks such as sentiment analysis in natural language processing. Analyzing sentiment in code-mixed text is more challenging than in single-language text because of issues such as language identification, variations in writing, mixed grammar, and a lack of large annotated datasets. This study reviews existing research on analyzing sentiment in code-mixed social media text. It examines different methods, including traditional machine learning, lexicon-based methods, deep learning models, and transformer-based architectures such as BERT. It also examines commonly used datasets, evaluation metrics, and language processing techniques, such as language identification and part-of-speech tagging. The study compares different language pairs, such as Hindi–English, Tamil–English, and Punjabi–English, to show current research trends and limitations. The findings show that while progress has been made for some language pairs, research on Punjabi–English code-mixed sentiment analysis is still limited. Finally, this study points out key research gaps and suggests future research directions to improve sentiment analysis in code-mixed multilingual settings.

Keywords:- code mixed; sentiment analysis; machine learning; Part Of Speech (POS) Tagging; Natural Language Processing(NLP);

I. INTRODUCTION

The increasing use of social media platforms such as Twitter, Facebook, Instagram, and YouTube has increased the amount of user-generated text online. These platforms allow users to quickly share their thoughts and feelings, making them useful for sentiment analysis. Sentiment analysis, or opinion mining, is a key task in natural language processing. It identifies and categorizes opinions in text as positive, negative, or neutral (Sharma et al., 2025; Maurya & Jha, 2024).In multilingual communities, people often mix languages in a single sentence or conversation. This is known as code-mixing or code-switching (Sitaram et al., 2019). For example, in India and other multilingual areas, social media users often mix English with local languages, such as Hindi, Tamil, or Punjabi, in one post. This creates code-mixed social media text, which is challenging for sentiment analysis. Unlike single-language text, code-mixed data can have inconsistent grammar, transliterated words, spelling variations, and mixed language structures, making automatic processing difficult (Ahmad et al., 2022; Neetika et al., 2020).

In the past ten years, researchers have proposed different methods for handling sentiment analysis in code-mixed text. Early studies mainly used traditional machine learning algorithms, such as Naïve Bayes and support vector machines, along with handcrafted features, including n-grams and part-of-speech tags (Advani et al., 2020; Ranjan et al., 2019; Sharma et al., 2018). However, these methods often struggled to capture contextual and semantic relationships effectively.

Recently, deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have been widely used to improve performance. CNN-based methods have been applied to the sentiment analysis of code-mixed tweets (Sarkar et al., 2019), and LSTM-based models have shown promising results (Jamatia et al., 2020). Hybrid deep learning architectures further enhance performance by combining multiple techniques (Gupta et al., 2021). Additionally, transformer-based architectures like BERT have shown strong performance in multilingual sentiment analysis tasks (Choudhary et al., 2022; Sahoo et al., 2022; Verma et al., 2023).

Despite these advancements, sentiment analysis of code-mixed social media text remains an active research area with many challenges. Issues, such as language detection, transliteration normalization, lack of annotated datasets, and mixed grammatical structures, affect model effectiveness (Patwa et al., 2020; Solorio et al., 2020). In addition, while many studies have focused on common language pairs, such as Hindi–English and Tamil–English, there has been less research on Punjabi–English code-mixed sentiment analysis (Singh & Goyal, 2020; Tiwari et al., 2025). The widespread use of social media platforms has led to an increase in user-generated content that reflects people's opinions, emotions, and sentiments on various topics (Sampath & Supriya, 2024). Identifying emotions in social media posts has become a major research area, especially with the rise of code-mixed content (Perera & Caldera, 2024). Social media text is used in various applications, including speech recognition, machine learning, information retrieval, question answering, sentiment analysis, and named entity recognition (Giatsoglou et al., 2017). (Giatsoglou et al., 2017). However, Natural Language Processing (NLP) tools like Part of Speech (POS) taggers and parsers, which are trained on monolingual texts, often face difficulties with multilingual data. Assigning grammatical labels, such as verbs, adjectives, and nouns, becomes more challenging in code-mixed scenarios (Chiche & Yitagesu, 2022; Sunita et al., 2023). For instance, consider the following code-mixed social media text:

- “Ajj party ni miss karni, it's gonna be full fun!”
- “Exam di tension hor vi vadh gayi aa, can't focus at all!”

With the ongoing growth of social media platforms, users are increasingly engaging in informal and non-standard communication through code-mixed language. In multilingual countries such as India, it is typical to see a mix of English and local languages, such as Punjabi, in online interactions (Neetika et al., 2020). Code-mixing refers to the blending of two or more languages within a single sentence, shaped by context, culture, and the user's language skills (Ahmad et al., 2022). For example, in casual conversations, users often merge English and Punjabi using the Roman alphabet.

“Kal di participation was awesome yaar!”

This kind of content poses significant challenges for traditional sentiment analysis systems, which are typically designed for monolingual and grammatically correct texts (Pasupa & Ayutthaya, 2019). Consequently, this study offers a thorough review of existing research on sentiment analysis of code-mixed social media text. The aim of this review is to examine the methodologies, datasets, and evaluation techniques employed in previous studies, pinpoint key challenges, and identify research gaps that could inform future work in this area.

This study presents a comprehensive review of machine learning techniques for sentiment analysis of code-mixed text. Techniques and approaches of machine learning and deep learning for bilingual or multilingual text sentiment analysis are described along with their corresponding results in different scenarios and using different types of datasets. The key research highlights of this study are as follows:

- To present the results of existing literature on sentiment analysis of code-mixed social media texts.
- To provide a systematic review of studies performed on sentiment analysis of code-mixed social media on different Indian languages.
- To explore and report the current state of research in code-mixed using various machine learning and deep learning techniques.
- To present the results of various machine learning models in terms of their performance metrics used by the recent studies in English-Punjabi code-mixed text.

The remainder of this paper is organized as follows: Section II discusses related work, and Sections III and its subsections provide the machine learning and deep learning techniques and approaches used in the sentiment analysis of code-mixed social media text. Section IV presents the approaches used in English–Punjabi code-mixed social media text. Section V compares the study of non-POS tagged and POS tagged data and presents the results of the English and Punjabi work done in the study. Section VI discusses the observations and state-of-the-art methods. Section VII presents the research gap, and Section VIII presents the conclusions and future scope.

II. Related Work

Sentiment analysis is an important aspect of NLP. It focuses on finding and sorting opinions from the text. At first, research mainly used English data and traditional machine learning methods, such as support vector machines (SVMs) and Naïve Bayes classifiers. These methods relied on manually created features, such as n-grams and lexicon-based models. Reviews have shown a shift from these older methods to advanced deep learning techniques that are used in many fields (Alshamsi et al., 2020; Sharma et al., 2025).

The rise of social media has led to more mixed-language texts. Code mixing means the use of different languages in one sentence. This causes problems such as grammar issues, word confusion, and a lack of standard language resources. Some studies have examined these problems and reviewed the current methods for sentiment analysis in mixed-language settings (Perera & Caldera, 2024; Ahmad et al., 2022; Mahadzir et al., 2021). These studies emphasize the need for strong models to handle language diversity.

In India, much research has been conducted on Hindi-English and Dravidian mixed-language data (Mandalam & Sharma, 2021; Jhanwar & Das, 2018; Sharma et al., 2018). However, sentiment analysis for English–Punjabi mixed text has not been well studied. Current research mainly focuses on basic sentiment classification and language identification. It highlights major issues, such as a lack of labeled data, specific domain problems, and challenges with transliterated text (Singh & Goyal, 2020; Bansal et al., 2020; Yadav et al., 2020). Although there has been some progress in multilingual sentiment analysis, including Punjabi, this research is still new (Gill & Singh, 2024).

Traditional approaches to sentiment analysis in code-mixed text typically rely on techniques rooted in feature engineering and ensemble learning (Maurya & Jha, 2024; Ranjan et al., 2019). Although these approaches offer satisfactory baseline performance, their dependence on manual feature extraction restricts their capacity to capture the deep semantic and contextual nuances present in code-mixed data.

Deep learning techniques have greatly enhanced the current capabilities of sentiment analysis. Models, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention-based frameworks, have demonstrated superior performance by facilitating automatic feature extraction and capturing long-range dependencies in text (Ghosh et al., 2021; Gupta et al., 2021; Sarkar et al., 2019). These models are particularly adept at handling the intricacies of code-mixed language structures.

In recent times, transformer-based models have set new benchmarks for sentiment analysis tasks. Models like multilingual BERT and its variations have been thoroughly fine-tuned for sentiment classification in code-mixed languages, showcasing exceptional contextual comprehension across different languages (Sampath & Supriya, 2024; Kaur et al., 2024; Sahoo et al., 2022; Verma et al., 2023). Although these models are highly effective, they often necessitate large annotated datasets and significant computational power, which restricts their use in low-resource language environments.

In addition to neural methods, linguistic elements, such as POS tagging, have been incorporated to boost the effectiveness of sentiment classification. Research has shown that attention mechanisms informed by POS and contextual embeddings can enhance feature representation and make models more interpretable (Kanakkahewa et al., 2023; Tehseen et al., 2023; Sunita et al., 2026). However, achieving precise POS tagging in code-mixed text is difficult because of frequent language switching and the limited availability of annotated resources.

Moreover, the creation of benchmark datasets and shared tasks has been pivotal in advancing research in this field. Projects such as LinCE and SemEval have offered standardized datasets and evaluation protocols specifically for code-mixed sentiment analysis (Aguilar et al., 2020; Patwa et al., 2020). However, there is still a significant shortage of extensive, high-quality datasets for low-resource languages, particularly for Punjabi.

In short, while there have been notable advancements in sentiment analysis for code-mixed text through the use of deep learning and transformer-based models, several obstacles remain. These challenges include the scarcity of annotated datasets for Punjabi, insufficient management of intricate linguistic features, such as code-switching and sarcasm, and a lack of model interpretability. Overcoming these hurdles offers significant opportunities for future research on English–Punjabi code-mixed sentiment analysis.

III. Approaches / Techniques applied for code mixed social media text in sentiment analysis :

3.1. Machine learning based

There has been a consistent increase in research focused on processing monolingual Punjabi text, largely due to the demand for supporting social media, digital content, and multilingual interactions. A key area of interest has been POS tagging, which is essential for providing the syntactic framework needed for more advanced tasks. Lehal and Saini (2011) established the basis for Punjabi POS tagging systems by investigating discriminative sequence labelling techniques, such as conditional random fields (CRF); however, they did not report accuracy metrics.

The study of sentiment analysis in code-mixed social media text has attracted considerable interest, driven by the growing prevalence of multilingual communication on digital platforms. Initial research primarily aimed at understanding the linguistic features of code-mixed text and creating resources for computational analysis. For example, Das et al. (2015) conducted one of the first studies on mixed-script information retrieval using the Forum for Information Retrieval Conference 2015 (FIRE) dataset. Their research underscored several challenges related to code-mixed text, including inconsistent spelling, transliteration, and grammatical variations. Although the study did not directly address sentiment classification, it offered valuable insights that informed later research on code-mixed natural language processing tasks.

As the field evolved, traditional machine learning algorithms were extensively used for sentiment analysis of code-mixed text, especially for Hindi–English datasets. Sharma et al. (2018) examined classical machine learning models like SVM and Naïve Bayes for classifying sentiments in Hinglish tweets. Their experiments achieved an accuracy of approximately 72%, indicating that traditional machine learning techniques can effectively classify sentiments in code-mixed data when suitable features are employed. Similarly, Ranjan et al. (2019) utilized feature-based machine learning methods for sentiment classification and reported an accuracy of approximately 70%. These studies demonstrate the effectiveness of combining traditional machine learning algorithms with feature engineering techniques, such as TF-IDF, n-grams, and lexicon-based features.

3.2. Deep Learning-Based Approaches

With the development of neural networks, deep learning models have found increasing use in analyzing the sentiment of code-mixed text. Bohra et al. (2018) presented a dataset combining Hindi and English and utilized long short-term memory (LSTM) models for sentiment analysis, achieving approximately 78% accuracy. This study highlighted that deep learning models can more effectively capture contextual relationships in code-mixed text than traditional machine learning techniques. Further advancements have been made with other neural architectures. Sarkar et al. (2019) employed convolutional neural networks (CNNs) for sentiment analysis of Hindi-English code-mixed tweets, reaching an accuracy of approximately 79%. Similarly, Banerjee et al. (2019) investigated the use of FastText word embeddings for representing words in code-mixed text, achieving an accuracy of approximately 74%. Hybrid deep learning architectures have also been suggested to boost model performance. Lal et al. (2019) created a hybrid architecture that combines CNN and Bidirectional Long Short-Term Memory (BiLSTM) models to capture both local semantic features and long-range contextual dependencies in sentences. Their model achieved an accuracy of 83.54%, showing significant improvements over previous methods.

3.3. Benchmark Datasets and Shared Tasks

The year 2020 was a significant turning point in the field of code-mixed sentiment analysis, marked by the introduction of benchmark datasets and shared evaluation tasks. Jamatia et al. (2020) utilized LSTM-based models to analyze the sentiment of Hinglish tweets, achieving an accuracy of about 81%. During the same period, Aguilar et al. (2020) launched the LinCE benchmark, which offers standardized datasets for assessing linguistic code-switching tasks, and reported an F1 score of approximately 84%. Similarly, Khanuja et al. (2020) introduced the GLUECoS benchmark, a framework designed to evaluate code-switched natural language processing tasks, attaining an F1 score of approximately 82%.

In a significant study conducted by Singh and Lefever (2020), the researchers examined cross-lingual word embeddings for analyzing sentiment in Hinglish tweets, achieving an F1 score of approximately 70.5%.

3.4. Studies on Other Code-Mixed Language Pairs

Although most research has concentrated on Hindi–English datasets, there has been some investigation into other language combinations, such as English–Punjabi. Singh and Goyal (2020) were among the first to study sentiment analysis of English–Punjabi code-mixed social media text, employing a lexicon-based N-gram method and achieving an accuracy of about 83%. This study highlighted the potential of sentiment analysis for Punjabi–English code-mixed datasets. Similarly, Bansal et al. (2020) explored language

identification for English–Punjabi code-mixed social media text using machine learning classifiers, such as logistic regression, and achieved an accuracy of approximately 86.63%.

3.5. Hybrid Deep Learning and Attention-Based Models

Researchers have recently introduced hybrid architectures that integrate deep learning with attention mechanisms to enhance sentiment classification outcomes. Gupta et al. (2021) crafted a hybrid deep learning framework that merges several neural networks, achieving an accuracy close to 85%. Similarly, Ghosh et al. (2021) designed an attention-based neural network model that achieved an accuracy of approximately 84% when analyzing sentiment in code-mixed tweets. Additionally, Singh (2021) investigated ensemble machine learning techniques and documented an F1 score of approximately 69.07% for the sentiment classification of Hinglish tweets.

3.6. Transformer-Based and Ensemble Learning Models

Recent research has concentrated on ensemble learning methods and transformer-based models. Kumar et al. (2022) utilized ensemble learning techniques for the sentiment classification of code-mixed tweets, achieving an accuracy of about 86%. Choudhary et al. (2022) introduced a multilingual transformer-based model for analyzing sentiment in code-mixed social media text and reported an F1 score of approximately 85%. Sahoo et al. (2022) employed transformer-based architectures, reaching an accuracy of approximately 87%.

3.7. Recent Advances Using Multilingual Language Models

Recent studies have increasingly focused on developing context-aware and multilingual language models to enhance sentiment analysis outcomes. Reddy et al. (2023) introduced a context-aware deep learning model aimed at analyzing sentiment in code-mixed social media text, achieving an accuracy rate of about 88%. Similarly, Verma et al. (2023) utilized multilingual BERT (mBERT) models for classifying sentiment in Hinglish tweets, reporting an accuracy of approximately 89%. More recently, Kaur et al. (2024) developed a transformer-based model for sentiment analysis, which attained an accuracy of approximately 90%, highlighting the capability of modern transformer architectures in processing code-mixed language data effectively.

3.8. Supporting Linguistic Research for Code-Mixed Text

In addition to sentiment classification, research has also explored other linguistic tasks associated with processing code-mixed text. For instance, Sunita et al. (2026) introduced an hidden Markov Model (HMM)-based method for part-of-speech tagging in English–Punjabi code-mixed social media content. Their research utilized annotated datasets sourced from platforms such as Facebook, YouTube, and WhatsApp, achieving a tagging accuracy of approximately 71.52%. Although this study focused on syntactic analysis rather than sentiment classification, it contributed to the development of linguistic resources for handling Punjabi–English code-mixed text.

In short, a diverse array of methods and techniques has been employed in code-mixed sentiment analysis, reflecting a clear shift from traditional machine learning approaches to sophisticated transformer-based models. Despite notable advancements, challenges such as limited data, linguistic diversity, and model interpretability continue to drive ongoing research in this field. Various strategies have been implemented for the sentiment analysis of code-mixed social media text, demonstrating clear improvements in performance over time. Initial studies using traditional machine learning techniques, such as support vector machines (SVMs) paired with TF-IDF features, achieved relatively lower accuracy, generally between 70% and 75% (Ranjan et al., 2019). The introduction of deep learning methods, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and hybrid CNN–RNN architectures, significantly enhanced performance, reaching accuracy levels of 80% to 85% (Sarkar et al., 2019; Gupta et al., 2021; Jamatia et al., 2020). Attention-based neural networks further improved results by concentrating on sentiment-relevant features, achieving approximately 82%–85% accuracy (Ghosh et al., 2021). More recently, transformer-based methods have set new benchmarks in code-mixed sentiment analysis. Models such as multilingual BERT (mBERT) and other transformer architectures have achieved accuracies between 85% and 90% when fine-tuned on code-mixed datasets (Verma et al., 2023; Choudhary et al., 2022; Sahoo et al., 2022). Similarly, transformer-based frameworks proposed by Sampath and Supriya (2024) and Kaur et al. (2024) reported accuracies of approximately 84%–88%, underscoring the effectiveness of contextual embeddings in managing multilingual and code-switched text. Additionally, models incorporating linguistic features, such as POS based attention and BiLSTM with contextual embeddings, have achieved competitive performances of approximately 83%–84% (Kanakkahewa et al.,

2023; Tehseen et al., 2023). Overall, the findings indicate consistent improvements in accuracy from traditional machine learning methods to deep learning and transformer-based models. However, it is crucial to recognize that reported accuracies vary based on the dataset, language pair, and preprocessing techniques employed, making direct comparisons across studies challenging.

IV Approaches/Techniques used in English–Punjabi Code-Mixed Social Media text

In contrast to other language combinations, such as Hindi–English, research on English–Punjabi code-mixed social media text has not been as extensively explored. Nevertheless, several studies have investigated natural language processing tasks, such as language identification, sentiment analysis, and part-of-speech tagging, for Punjabi–English code-mixed text. Singh and Goyal (2019) were among the pioneers in studying language identification for English–Punjabi code-mixed social media text. They gathered bilingual text from social media platforms where Punjabi words are often written in the Roman script and interspersed with English words. Machine learning classifiers were utilized to determine the language of individual tokens within code-mixed sentences. The study underscored significant challenges, including transliteration variations, spelling inconsistencies, and informal writing styles typical of social media communication.

In 2020, Singh and Goyal introduced a sentiment analysis system tailored for English–Punjabi code-mixed content on social media platforms, such as Facebook, Twitter, and YouTube. This dataset featured Punjabi words transcribed in the Roman script alongside English words. The authors employed a lexicon-based sentiment analysis method enhanced with n-gram features to categorize social media posts into positive, negative, and neutral sentiments. Their experimental findings indicated that the system achieved a classification accuracy of approximately 83%, highlighting the efficacy of lexicon-based methods in analyzing bilingual Punjabi–English social media text.

Bansal, Goyal, and Rani (2020) conducted a study to identify languages in English–Punjabi code-mixed text found on social media. They gathered data from Facebook and Twitter and utilized various machine learning classifiers, such as logistic regression, decision trees, and Naïve Bayes, incorporating N-gram features. Their findings revealed that logistic regression achieved the highest accuracy at 86.63%, demonstrating the effectiveness of machine learning models in detecting language tokens within code-mixed Punjabi–English text. Additionally, Bansal, Goyal, and Rani (2020) expanded their research to address the issue of word-level language identification in English–Punjabi code-mixed text. This aspect of the study aimed to categorize individual words as either Punjabi or English using machine learning algorithms and linguistic features. The research underscored the difficulties in managing transliterated Punjabi words written in the Roman script, which frequently appear with various spellings in social media text.

Sunita, Kumar, and Bansal (2026) introduced a hidden Markov model (HMM) method for POS tagging in English–Punjabi code-mixed social media content, alongside sentiment analysis and language identification. They compiled a dataset of approximately 900 annotated sentences sourced from Facebook, YouTube, and WhatsApp. The authors approached tagging as a sequence labeling task and employed the Viterbi algorithm for decoding. Their system achieved a POS tagging accuracy of 71.52%, establishing a foundational framework for the syntactic analysis of Punjabi–English code-mixed text.

The current body of research suggests that studies on English–Punjabi code-mixed social media text are still in their infancy. Most existing work has concentrated on fundamental tasks, such as language identification, sentiment analysis using lexicons, and part-of-speech (POS) tagging. It is only recently that more sophisticated methods, such as deep learning and transformer-based techniques, have started to be investigated. Consequently, further research is needed to create larger annotated datasets and employ modern machine learning methods to enhance sentiment analysis and other NLP tasks for English–Punjabi code-mixed social media text.

V Use of non-POS tagged and POS tagged Data in sentiment analysis code mixed social media text:

5.1. Non-POS Tagged data in sentiment analysis:

Conversely, most studies on sentiment analysis of code-mixed social media content, particularly for language combinations such as Hindi–English and Punjabi–English, do not employ POS-tagged datasets. Conventional machine learning methods, such as those used by Ranjan et al. (2019) for Hindi–English and Singh and Goyal (2020) for Punjabi–English, tend to yield lower accuracy rates, generally around 72.3% and 74.1%, respectively, due to their dependence on surface-level and manually crafted features. In contrast, deep learning models, including CNN and LSTM frameworks (e.g., Sarkar et al. (2019) and

Jamatia et al. (2020)) applied to Hindi–English code-mixed text, show marked improvement, achieving about 81.5% and 83.2% accuracy by effectively capturing sequential and contextual relationships in the data. More sophisticated pretrained transformer-based models, such as mBERT (Verma et al. 2023) for Hinglish (Hindi–English) and XLM-RoBERTa (Choudhary et al. 2022) for multilingual code-mixed data, further boost performance, attaining accuracies of approximately 89.4% and 91.2%, respectively, because of their robust contextual comprehension and cross-lingual representation abilities.

Nevertheless, even with these advancements, many of these methods overlook explicit linguistic elements, such as POS tagging, which could potentially boost the effectiveness of sentiment classification. In general, research suggests that while datasets with POS tagging offer crucial linguistic insights for feature extraction, most sentiment analysis studies in Indian languages predominantly use datasets without POS tagging, particularly as deep learning and transformer-based models become more prevalent.

5.2. POS tagged data in Sentiment analysis:

A thorough examination of the relevant literature indicates that only a few studies integrate POS tagged features in sentiment analysis, with a primary focus on monolingual or non-code-mixed datasets. For example, Kalarani and Selva Brunda (2019) employed POS features alongside SVM and ANN on English text, achieving an accuracy of 86.2%, which highlights the role of syntactic information in enhancing sentiment classification. Similarly, Srividya and Sowjanya (2019) used POS tagging with TF-IDF for aspect-based sentiment analysis in English, reporting an accuracy of 82.5%. In another instance, Pasupa and Ayutthaya (2019) worked with Thai language data, combining POS tagging with word embeddings to achieve an accuracy of approximately 85.6%. Kanakkahewa et al. (2023) expanded this concept to multilingual social media data by introducing a POS-based attention mechanism, resulting in an improved performance of approximately 86.8%. Collectively, these studies illustrate that POS tagged features enhance syntactic comprehension and sentiment classification accuracy; however, they are predominantly confined to monolingual or general multilingual contexts rather than code-mixed scenarios.

The analysis underscores that although POS based methods achieve commendable accuracy in English, Thai, and multilingual contexts, they are seldom utilized in code-mixed sentiment analysis. Specifically, for English–Punjabi code-mixed social media content, current research primarily depends on machine learning, deep learning, and transformer-based techniques without the use of POS tagged features. This suggests that incorporating POS tagging into machine-learning models can potentially improve the performance and interpretability of sentiment classification in code-mixed environments.

Only a few published studies have utilized English–Punjabi POS tagged data specifically for sentiment analysis. Most studies either conducted sentiment analysis without incorporating POS tagging or performed POS tagging without focusing on sentiment analysis.

VI Observation and State of art

The review of 67 research articles highlights several key trends and gaps in the study of code-mixed social media content. Earlier studies, particularly between 2015 and 2019, mainly relied on traditional machine learning techniques, such as support vector machines, naïve Bayes, and logistic regression. These approaches depended heavily on manually engineered features. In recent years, however, there has been a clear shift toward deep learning and transformer-based models, including LSTM, CNN, and BERT, which are more effective at capturing context and semantic nuances in multilingual and code-mixed data. Despite these advancements, most research continues to focus on widely studied language pairs, such as Hindi–English and other high-resource combinations. In contrast, English–Punjabi code-mixed text remains largely underexplored, with only a limited number of studies addressing this area.

Another important observation from the review is the limited use of POS tagging in sentiment analysis. Only a few studies—such as those by Kalarani and Selva Brunda (2019), Srividya and Sowjanya (2019), Pasupa and Ayutthaya (2019), and Kanakkahewa et al. (2023)—have incorporated POS features into sentiment classification, reporting improvements in both accuracy and feature representation. In contrast, several studies have examined POS tagging in isolation without integrating it into sentiment analysis tasks. At the same time, most sentiment analysis research does not utilize POS information at all. This indicates that syntactic features remain largely underutilized, particularly in the context of code-mixed data.

Moreover, the literature consistently highlights several challenges in code-mixed sentiment analysis. These include ambiguity introduced by language mixing, the lack of large and standardized annotated datasets, the noisy and informal nature of social media text, and the difficulty of accurately performing POS tagging on mixed-language inputs. These challenges are even more pronounced for English–Punjabi data owing to the limited availability of linguistic resources and annotated corpora.

While advanced approaches such as transformer-based models, attention mechanisms, and hybrid frameworks have shown strong performance, they rarely incorporate linguistic features such as POS

tagging. As a result, an important gap remains in the development of sentiment analysis models that effectively combine deep learning techniques with syntactic information. Addressing this gap is particularly crucial for improving performance on English–Punjabi code-mixed social media texts.

Table 6.1: No. of the Reviewed Papers and the Source of Reviewed Papers

Source	No. of Papers	POS Tagged Papers	POS Used in Sentiment Analysis	English–Punjabi POS Tagged Papers	Authors (Year)
IEEE	10	2	1	0	Tiwari, A.; Sehgal, J.; Singh, M.; Mishra, A. (2025) Tripathi, N.; Singh, M.; Yadav, N. (2024) Nandakumar, R.; Pallavi, M. S. (2022) Yadav, K.; Lamba, A.; Gupta, D.; Gupta, A.; Karmakar, P.; Saini, S. (2020) Mandalam, A. V.; Sharma, Y. (2021)
Springer	8	2	2	0	Sharma, N. A.; Ali, A. S.; Kabir, M. A. (2025) Kalarani, P.; Selva Brunda, S. (2019) Chiche, A.; Yitagesu, B. (2022) Gupta, R.; Sharma, P.; Singh, K. (2021)
Elsevier (Procedia, Expert Systems, etc.)	9	2	2	0	Sampath, K. K.; Supriya, M. (2024) Maurya, C. G.; Jha, S. K. (2024) Pasupa, K.; Ayutthaya, T. S. N. (2019) Giatsoglou, M.; Vozalis, M. G.; Diamantaras, K.; Vakali, A.; Sarigiannidis, G.; Chatzisavvas, K. C. (2017)
ACM / ACL / Conference Proceedings	12	0	0	0	Maity, K.; Jha, P.; Saha, S.; Bhattacharyya, P. (2022) Aguilar, G.; Kar, S.; Solorio, T.; Diab, M. (2020) Patwa, P.; Sharma, S.; Pykl, S.; Guptha, V.; Kumari, G.; Akhtar, M. S.; Ekbal, A.; Das, A.; Chakraborty, T. (2020) Chakravarthi, B. R.; Muralidaran, V.; Priyadharshini, R.; McCrae, J. P. (2020)
ResearchGate / ResearchSquare / Preprints	4	2	1	0	Kanakkahewa, K. H. S. L.; Mohotti, W. A.; Subhashini, L. D. C. S. (2023) Sunita, S.; Kumar, A.; Neetika, N. (2023) Sunita, S.; Kumar, A.; Bansal, N. (2026) Jhanwar, M. G.; Das, A. (2018)
Other Journals	15	2	1	0	Srividya, K.; Sowjanya, A. M. (2019) Ahmad, G. I.; Singla, J.; Ali, A.; Reshi, A. A.; Salameh, A. A. (2022) Mahadzir, N. H.; Omar, M. F.; Nawi, M. N. M.; Salameh, A. A.; Hussin, K. C. (2021)
Punjabi / Regional Studies	9	3	0	2	Dinesh Kumar, M.; Josan, G. S. (2016) Lehal, G. S.; Saini, T. S. (2011) Lehal, G. S.; Gupta, V. (2012) Lehal, G. S.; Singh, S. (2012) Tehseen, S.; Singh, M.; Kaur, H. (2023)

VII RESEARCH GAPS

The existing literature reveals that several studies have been conducted on sentiment analysis in code-mixed text for other languages. After studying various research papers, we found the following:

- No studies have been conducted using POS tagged data for sentiment analysis in English–Punjabi code-mixed text from social media.
- Analysis tools perform well on POS tagged text and improve accuracy compared to untagged text.
- This is the first attempt (for the time being) to develop a sentiment analysis tool for English–Punjabi code-mixed social media text using POS tagged data.

VIII CONCLUSION AND FUTURE SCOPE

Despite significant progress in sentiment analysis for code-mixed social media content, several important challenges remain unresolved. Most existing research has focused on high-resource language pairs, such as Hindi–English, whereas low-resource combinations, such as Punjabi–English, have received comparatively little attention. This imbalance has led to a shortage of annotated datasets, which limits the development and proper evaluation of robust sentiment analysis models for these languages.

This review provides a detailed analysis of sentiment analysis approaches applied to code-mixed social media data, with a methods, deep learning architectures, and transformer-based models. The findings indicate that current studies largely rely on both pretrained and non-pretrained deep learning models, often without incorporating linguistic features such as POS tagging.

Although POS tagging has been shown to improve sentiment classification performance in monolingual and some multilingual contexts (e.g., Kalarani & Selva Brunda, 2019; Kanakkahewa et al., 2023), its application to English–Punjabi code-mixed data remains largely unexplored. Most existing work prioritizes deep learning and transformer-based approaches while overlooking syntactic features, revealing a notable gap in the use of POS-tagged information to enhance performance in code-mixed settings.

Overall, this review highlights that, despite advancements in multilingual and code-mixed sentiment analysis, there is still a lack of linguistically informed and robust models, particularly for low-resource language pairs such as English–Punjabi.

REFERENCES

[1] Tiwari, A., Sehgal, J., Singh, M., & Mishra, A. (2025, March). Sentiment Analysis in English–Punjabi Mixed Social Media Posts. In *2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)* (Vol. 3, pp. 1-6). IEEE.

[2] Sharma, N. A., Ali, A. S., & Kabir, M. A. (2025). Review of sentiment analysis: tasks, applications, and deep learning techniques. *International journal of data science and analytics*, *19*(3), 351-388.

[3] Sampath, K. K., & Supriya, M. (2024). Transformer based sentiment analysis on code mixed data. *Procedia Computer Science*, *233*, 682-691

[4] Perera, A., & Caldera, A. (2024). Sentiment analysis of code-mixed text: a comprehensive review. *Journal of Universal Computer Science*, *30*(2), 242.

[5] Maurya, C. G., & Jha, S. K. (2024). Sentiment analysis: a hybrid approach on Twitter data. *Procedia Computer Science*, *235*, 990-999.

[6] Kanakkahewa, K. H. S. L., Mohotti, W. A., & Subhashini, L. D. C. S. (2023). PoS tag-based Attention for Feature Selection in Sentiment Analysis. Preprint at Researchsquare https://doi.org/10.21203/rs.3.rs-3151544/v1

[7] Garg, N., & Sharma, K. (2022). Text pre-processing of multilingual for sentiment analysis based on social network data. *International Journal of Electrical & Computer Engineering (2088-8708)*, *12*(1).

[8] Chiche, A., & Yitagesu, B. (2022). Part of speech tagging: a systematic review of deep learning and machine learning approaches. *Journal of Big Data*, *9*(1), 10.

[9] Ahmad, G. I., Singla, J., Ali, A., Reshi, A. A., & Salameh, A. A. (2022). Machine learning techniques for sentiment analysis of code-mixed and switched indian social media text corpus-a comprehensive review. *International Journal of Advanced Computer Science and Applications*, *13*(2).

[10] Mandalam, A. V., & Sharma, Y. (2021, April). Sentiment analysis of Dravidian code mixed data. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages* (pp. 46-54)

[11] Mahadzir, N. H., Omar, M. F., Nawi, M. N. M., Salameh, A. A., & Hussin, K. C. (2021). Sentiment analysis of code-mixed text: a review. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(3), 2469-2478.

[12] Singh, M., & Goyal, V. (2020, December). Sentiment analysis of English-Punjabi code-mixed social media content. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON): System Demonstrations* (pp. 24-25).
[13] Bansal, N., Goyal, V., & Rani, S. (2020). Experimenting language identification for sentiment analysis of english punjabi code mixed social media text. *International Journal of E-Adoption (IJEA)*, *12*(1), 52-62.
[14] Chakravarthi, B. R., Muralidaran, V., Priyadharshini, R., & McCrae, J. P. (2020). Corpus creation for sentiment analysis in code-mixed Tamil-English text. *arXiv preprint arXiv:2006.00206*.
[15] Alshamsi, A., Bayari, R., & Salloum, S. (2020). Sentiment analysis in English texts. *Advances in Science, Technology and Engineering Systems Journal*, *5*(6), 1683-1689.
[16] Tripathi, N., Singh, M., & Yadav, N. (2024, November). A Comprehensive Review on Emotion Detection in Code-Mixed Social Media Posts. In *2024 International Conference on Advances in Computing, Communication and Materials (ICACCM)* (pp. 1-6). IEEE.
[17] Singh, M., Goyal, V., & Raj, S. (2021). Sentiment analysis of social media Tweets on Farmer Bills 2020. *Journal of Scientific Research*, *65*(3), 156-162.
[18] Sunita, S., Kumar, A., & Neetika, N. (2023). A Comprehensive Survey of Techniques Used for Part-of-Speech Tagging of Code-Mixed Social Media Text. Preprint at ResearchSquare https://doi.org/10.21203/rs.3.rs-3274325/v1
[19] Maity, K., Jha, P., Saha, S., & Bhattacharyya, P. (2022, July). A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval* (pp. 1739-1749).
[20] Yadav, K., Lamba, A., Gupta, D., Gupta, A., Karmakar, P., & Saini, S. (2020, October). Bilingual sentiment analysis for a code-mixed Punjabi English social media text. In *2020 5th International conference on computing, communication and security (ICCCS)* (pp. 1-5). IEEE.
[21] Chen, M. H., Chen, W. F., & Ku, L. W. (2018). Application of sentiment analysis to language learning. *IEEE Access*, *6*, 24433-24442.
[22] Neetika, Goyal, V., & Rani, S. (2020). Automatic understanding of code mixed social media text: A state of the art. *Advances in Information Communication Technology and Computing: Proceedings of AICTC 2019*, 91-100.
[23] Pasupa, K., & Ayutthaya, T. S. N. (2019). Thai sentiment analysis with deep learning techniques: A comparative study based on word embedding, POS-tag, and sentic features. *Sustainable Cities and Society*, *50*, 101615.
[24] Srividya, K., & Sowjanya, A. M. (2019). Aspect based sentiment analysis using POS tagging and TFIDF. *International Journal of Engineering and Advanced Technology*, *8*(6), 1960-1963.
[25] Nandakumar, R., & Pallavi, M. S. (2022, October). Sentimental analysis on student feedback using NLP & POS tagging. In *2022 International conference on edge computing and applications (ICECAA)* (pp. 309-313). IEEE.
[26] Kalarani, P., & Selva Brunda, S. (2019). Sentiment analysis by POS and joint sentiment topic features using SVM and ANN. *Soft computing*, *23*(16), 7067-7079.
[27] Jhanwar, M. G., & Das, A. (2018). An ensemble model for sentiment analysis of Hindi-English code-mixed data. *arXiv preprint arXiv:1806.04450*.
[28] Pravalika, A., Oza, V., Meghana, N. P., & Kamath, S. S. (2017, July). Domain-specific sentiment analysis approaches for code-mixed social network data. In *2017 8th international conference on computing, communication and networking technologies (ICCCNT)* (pp. 1-6). IEEE.
[29] Giatsoglou, M., Vozalis, M. G., Diamantaras, K., Vakali, A., Sarigiannidis, G., & Chatzisavvas, K. C. (2017). Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications*, *69*, 214-224.
[30] Dinesh Kumar, M., & Josan, G. S. (2016). Prediction of part-of-speech tags for Punjabi using Support Vector Machines. *International Journal of Advanced Research in Computer Science and Software Engineering*, *6*(4), 123–129.
[31] Kaur, H., & Singh, P. (2021). Machine learning approaches for sentiment analysis of Punjabi social media text. *International Journal of Computational Linguistics and Applications*, *12*(3), 87–95.
[32] Lehal, G. S., & Gupta, V. (2012). Automatic Punjabi text extractive summarization system. In *Proceedings of COLING 2012: Demonstration Papers* (pp. 191–198). Association for Computational Linguistics.

[33] Lehal, G. S., & Singh, S. (2012). Punjabi text-to-speech synthesis system. In *Proceedings of COLING 2012: Demonstration Papers* (pp. 409–416). Association for Computational Linguistics.
[34] Lehal, G. S., & Saini, T. S. (2011). Using Hidden Markov Models to improve the accuracy of Punjabi POS tagger. *IEEE International Conference on Computer Science and Automation Engineering*, 2011, 234–238.
[35] Sharma, R., & Singh, H. (2022). Deep learning approaches for sentiment analysis of Punjabi social media text. *International Journal of Data Science and Analytics, 10*(1), 45–57.
[36] Tehseen, S., Singh, M., & Kaur, H. (2023). POS tagging for Punjabi using BiLSTM with contextual embeddings. *Journal of South Asian Language Technologies, 8*(2), 112–130.
[37] Gill, H., & Singh, J. (2024). Multilingual sentiment analysis for Punjabi–English social media data. *Journal of Multilingual Information Processing, 17*(1), 88–102.
[38] Kaur, P., & Kaur, M. (2023). Transformer-based sentiment analysis for Punjabi language text. *International Journal of Artificial Intelligence Research, 15*(4), 62–77.
[39] Advani, L., Patel, V., & Mehta, S. (2020). Feature engineering approaches for sentiment analysis of code-mixed social media text. *Proceedings of the International Conference on Computational Linguistics*, 152–160.
[40] Aguilar, G., Kar, S., Solorio, T., & Diab, M. (2020). LinCE: A centralized benchmark for linguistic code-switching evaluation. *Proceedings of the 12th Language Resources and Evaluation Conference*, 1803–1813.
[41] Banerjee, S., Chakraborty, T., & Das, A. (2019). FastText embeddings for sentiment analysis of code-mixed social media text. *Proceedings of the Workshop on Computational Approaches to Linguistic Code Switching*, 1–8.
[42] Baroi, S. J., Chakraborty, A., & Bandyopadhyay, S. (2020). NITS-Hinglish-SentiMix at SemEval-2020 task 9: Sentiment analysis for Hinglish code-mixed tweets using ensemble learning. *Proceedings of the 14th International Workshop on Semantic Evaluation*, 1135–1140.
[43] Bhat, G., Choudhury, M., & Bali, K. (2020). PHINC: A parallel Hinglish corpus for machine translation and sentiment analysis. *Proceedings of the Language Resources and Evaluation Conference*, 4944–4950.
[44] Bohra, A., Vijay, D., Singh, V., & Akhtar, S. S. (2018). A dataset of Hindi-English code-mixed social media text for sentiment analysis. *Proceedings of the Workshop on Computational Approaches to Linguistic Code Switching*, 36–41.
[45] Chakravarthi, B. R., Jose, N., Suryawanshi, S., Sherly, E., & McCrae, J. (2020). DravidianCodeMix: Sentiment analysis and offensive language identification dataset for code-mixed languages. *Proceedings of the Workshop on Computational Approaches to Linguistic Code Switching*, 113–118.
[46] Choudhary, N., Singh, R., & Gupta, S. (2022). Multilingual transformer models for sentiment analysis of code-mixed social media text. *Journal of Artificial Intelligence Research*, 73, 123–138.
[47] Das, A., Gambäck, B., & Das, A. (2015). Overview of the FIRE shared task on mixed script information retrieval. *Proceedings of the Forum for Information Retrieval Evaluation*, 19–25.
[48] Ghosh, S., Saha, S., & Bandyopadhyay, S. (2021). Attention-based deep neural networks for sentiment analysis of code-mixed social media text. *IEEE Access*, 9, 84534–84545.
[49] Gupta, R., Sharma, P., & Singh, K. (2021). Hybrid deep learning models for sentiment analysis of code-mixed social media text. *Expert Systems with Applications*, 168, 114329.
[50] Jamatia, A., Gambäck, B., & Das, A. (2020). Deep learning based sentiment analysis of code-mixed social media text. *Proceedings of the International Conference on Natural Language Processing*, 258–267.
[51] Khanuja, S., Dandapat, S., Srinivasan, A., & Varma, V. (2020). GLUECoS: An evaluation benchmark for code-switched NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3575–3585.
[52] Kumar, A., Verma, R., & Patel, S. (2022). Ensemble learning for sentiment classification of code-mixed social media text. *Applied Soft Computing*, 114, 108105.
[53] Kaur, P., Singh, M., & Kaur, H. (2024). Transformer-based sentiment analysis for code-mixed social media text. *Information Processing & Management*, 61(2), 102976.
[54] Lal, Y. K., Kumar, V., & Joshi, A. (2019). De-mixing sentiment from code-mixed text. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 352–357.

[55] Patwa, P., Sharma, S., Pykl, S., Guptha, V., Kumari, G., Akhtar, M. S., Ekbal, A., Das, A., & Chakraborty, T. (2020). SemEval-2020 task 9: Sentiment analysis of code-mixed tweets. *Proceedings of the 14th International Workshop on Semantic Evaluation*, 774–790.

[56] Ranjan, P., Singh, A., & Kumar, R. (2019). Feature-based machine learning approaches for sentiment analysis of code-mixed tweets. *International Journal of Computer Applications*, 178(9), 25–31.

[57] Reddy, S., Narayan, A., & Joshi, P. (2023). Context-aware sentiment analysis for code-mixed social media text using deep neural networks. *Neural Computing and Applications*, 35, 14215–14228.

[58] Sahoo, S., Mishra, A., & Das, D. (2022). Transformer-based architectures for sentiment analysis of code-mixed social media text. *Pattern Recognition Letters*, *158*, 45–52.

[59] Sarkar, A., Chakraborty, T., & Bandyopadhyay, S. (2019). Convolutional neural networks for sentiment analysis of code-mixed tweets. *Proceedings of the International Conference on Natural Language Processing*, 193–202.

[60] Sharma, A., Bali, K., & Choudhury, M. (2018). Machine learning approaches for sentiment analysis of Hindi-English code-mixed social media text. *International Journal of Computer Applications*, 182(44), 20–25.

[61] Singh, G. (2021). Sentiment analysis of code-mixed social media text using machine learning approaches. *Journal of Computational Linguistics Research*, 5(2), 45–56.

[62] Singh, M., & Goyal, V. (2020). Sentiment analysis of English–Punjabi code-mixed social media content using lexicon-based approaches. In *Proceedings of the International Conference on Natural Language Processing* (pp. 125–132).

[63] Singh, P., & Lefever, E. (2020). Cross-lingual word embeddings for sentiment analysis of Hinglish tweets. *Proceedings of the Workshop on Computational Approaches to Linguistic Code Switching*, 140–145.

[64] Sunita, S., Kumar, A., & Bansal, N. (2026). Part-of-speech tagging of English–Punjabi code-mixed social media text using hidden Markov models. *International Journal of Latest Technology in Engineering, Management & Applied Science*, *15*(2), 112–118.

[65] Sitaram, S., Choudhury, M., Bali, K., & Black, A. W. (2019). A survey of code-switched speech and language processing. *Computer Speech & Language*, 67, 101–132.

[66] Solorio, T., Aguilar, G., & Diab, M. (2020). Computational approaches to code-switching: Overview and challenges. *Proceedings of the Workshop on Computational Approaches to Code Switching*, 1–10.

[67] Verma, R., Kumar, A., & Singh, P. (2023). Fine-tuning multilingual BERT for sentiment analysis of Hinglish social media text. *Applied Intelligence*, 53, 11245–11258.