IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 28, Issue 3, Ser. III (May. – June. 2026), PP 22-48
www.iosrjournals.org

A Comprehensive Review of Multilingual Sentiment Analysis for Indian Languages from Traditional Machine Learning to Large Language Models

Meenal M. Shingare
Research Scholar of Computer Science and Engineering Department,
Maharashtra Institute of Technology, Satara Road, Chh. Sambhajinagar, Maharashtra, India.

Dr. B. S. Sonawane
Professor, Computer Science and Engineering Department,
Maharashtra Institute of Technology, Satara Road, Chh. Sambhajinagar, Maharashtra, India.

Abstract
Sentiment analysis is an important research area in Natural Language Processing (NLP) which is the automated identification and interpretation of opinions, emotions and attitudes expressed in text data. The growth of user generated content on social media, review systems, discussion forums and digital communication channels has created the demand for accurate and scalable sentiment analysis systems. There has been a lot of work on sentiment analysis in English language, but the development of effective models for multilingual and low-resource languages still remains an intimidating research problem. The present work provides a comprehensive review of the literature on sentiment analysis from 2020 to 2026, focusing on multilingual sentiment analysis, sentiment analysis for Indian languages, transformer-based models, and recent large language models.
A systematic literature review methodology based on the PRISMA framework was applied to identify, screen and analyze relevant studies from major scientific databases including Scopus, IEEE Xplore, Web of Science, ACM Digital Library, SpringerLink and ScienceDirect. The review categorizes the existing works into five paradigms, namely, lexicon-based, machine learning, deep learning, transformer-based and multilingual sentiment analysis. Special focus is placed on the Hindi and Marathi sentiment analysis resources including benchmark datasets such as IIT Patna Movie Reviews, L3CubeMahaSent, MahaSent-MD and code-mixed social media corpora. The review also includes the evolution of the text representation techniques from Bag-of-Words, TF-IDF to contextual embeddings generated by BERT, IndicBERT, MuRIL and large language models. Transformer-based architectures consistently outperform traditional machine learning and deep learning approaches on various sentiment analysis benchmarks, as the analysis shows. But issues such as code-mixing, transliteration, low-resource datasets, explainability, computational complexity and bias are still largely unresolved. The review discusses emerging directions such as multilingual transformers, explainable artificial intelligence, multimodal sentiment analysis, retrieval-augmented generation, federated learning and large language models and identifies significant research gaps. The results can serve as a comprehensive reference for researchers and practitioners involved in the development of next generation multilingual sentiment analysis systems, supporting various language and real-world application environments.

Keywords
Keywords: Sentiment Analysis; Multilingual Sentiment Analysis; Natural Language Processing; Machine Learning; Deep Learning; Transformer Models; BERT; IndicBERT; MuRIL; Large Language Models; Hindi Sentiment Analysis; Marathi Sentiment Analysis; Code-Mixed Text; Low-Resource Languages; Text Classification; Opinion Mining; Social Media Analytics; Explainable Artificial Intelligence; Multimodal Sentiment Analysis; Indian Languages.

Date of Submission: 02-06-2026 Date of Acceptance: 13-06-2026

I. Introduction

The fast expansion of digital communication platforms has changed the medium through which people express opinions, emotions, preferences and experiences. Every day, social media platforms, online review systems, discussion forums, blogs, e-commerce portals and news websites produce massive amount of user generated textual content. This ever-expanding store of data rich in opinions contains valuable insights about public perception, consumer behaviour, political opinions, product acceptance, healthcare experiences and societal

trends. Hence, the automatic extraction and interpretation of sentiments from textual data has surfaced as an important research area in the larger domains of Natural Language Processing (NLP), Artificial Intelligence (AI) and Machine Learning (ML).

Sentiment Analysis (SA), also known as opinion mining, is the computational process of identifying, extracting, classifying, and analyzing subjective information expressed in textual content. Sentiment analysis is mainly about determining the emotional sentiment of text, usually classifying it as positive, negative or neutral. More sophisticated systems also try to recognize emotions such as happiness, anger, sadness, surprise, fear and disgust, thus providing a more detailed picture of human opinion. Sentiment analysis has evolved from basic lexicon-based methods to advanced deep learning and transformer-based models capable of understanding contextual semantics and intricate linguistic connections over the past two decades.

The growing significance of sentiment analysis can be explained by its extensive use in real world scenarios in various domains. Companies use sentiment analysis to track feedback from customers, measure product performance and improve decision-making. Governments and policy makers apply opinion mining techniques to evaluate public reaction towards policies and social initiatives. Sentiment analysis is used by healthcare organizations to measure patient experiences and track mental health trends through conversations on social media. Educational institutions, financial organizations, tourism industries and media agencies also increasingly depend on sentiment analysis systems to obtain actionable insights from large-scale textual datasets. The rapid increase in digital content creation has made the need for effective and scalable sentiment analysis solutions more urgent than ever.

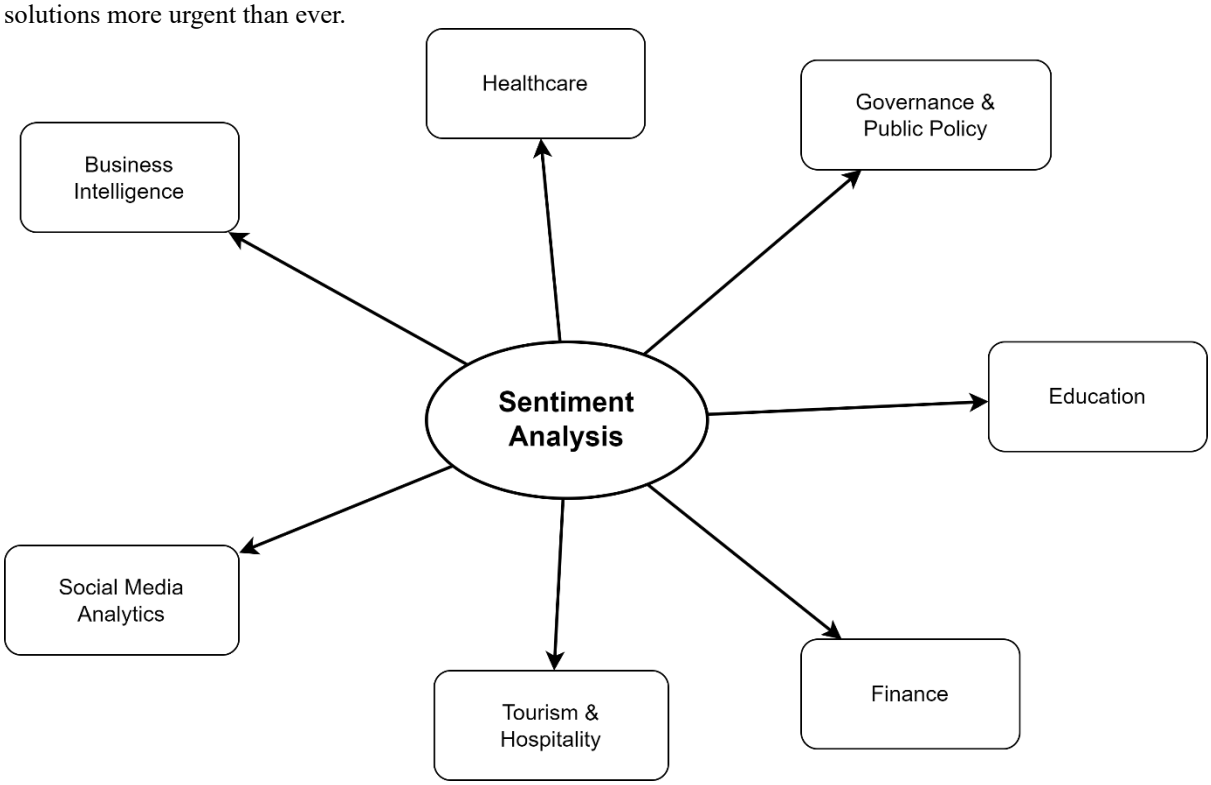

Figure 1. Major Application Areas of Sentiment Analysis

Figure 1 highlights the broad spectrum of domains where sentiment analysis is employed. The figure demonstrates how opinion mining supports decision-making processes by extracting actionable insights from large volumes of textual data generated across different sectors.

Historically, sentiment analysis systems were heavily dependent on lexicon-based and rule-based approaches. Lexicon based methods employ existing sentiment lexicons which include words with corresponding positive or negative polarity scores. These methods are computationally efficient and require limited training data but often fail to handle contextual meanings, sarcasm, negation, domain-specific expressions and linguistic ambiguity. To overcome these limitations, researchers have increasingly adopted machine learning approaches such as Naïve Bayes (NB), Support Vector Machines (SVM), Decision Trees (DT), Logistic Regression (LR) and Random Forests (RF). These methods transformed textual information into numerical representations using feature extraction techniques like Bag-of-Words (BoW), N-grams, and Term Frequency-Inverse Document Frequency (TF-IDF) for more robust sentiment classification.

Deep learning has significantly revolutionized sentiment analysis research. ANNs, CNNs, RNNs, LSTM, and

GRU. These neural architectures were found to perform better since they automatically learned semantic representations from text data. Deep learning models are able to learn complex linguistic patterns as well as contextual dependencies, in contrast to traditional machine learning approaches which are heavily dependent on handcrafted features. However, despite their effectiveness, recurrent architectures often suffer from long-range dependency modelling and computational complexity issues when handling large-scale datasets. Another important development in the evolution of sentiment analysis was the introduction of transformer architectures . The Transformer model by Vaswani et al. was proposed to use self-attention mechanisms to efficiently model contextual connections across entire sequences. Based on this foundation, several pre-trained language models, such as BERT, RoBERTa, XLNet, ALBERT, DistilBERT, ELECTRA, and GPT-based architectures, have achieved the state-of-the-art performance in a wide range of NLP tasks. These transformer-based models have an amazing capacity to capture the contextual semantics, syntactic relationships, and linguistic nuances, which greatly improves the accuracy of sentiment classification over traditional and deep learning methods.

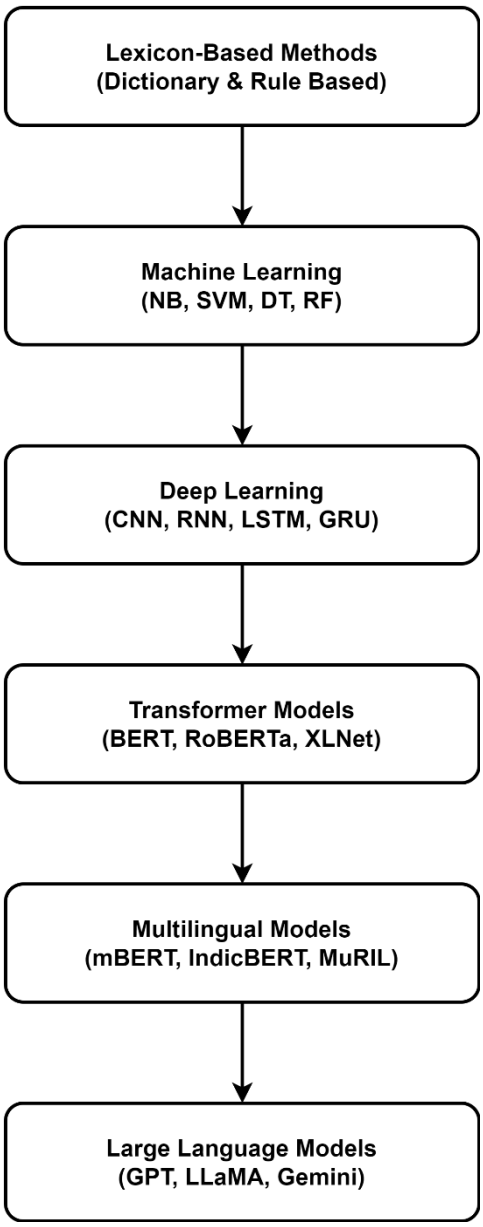

Figure 2. Evolution of Sentiment Analysis Techniques

Figure 2 shows the development of sentiment analysis approaches. The evolution from lexicon-based approaches to machine learning, deep learning, transformer architectures, and large language models illustrates the ongoing improvement in NLP technologies for improving contextual understanding and classification efficiency.

Despite the fact that a lot of progress has been made in the field of sentiment analysis, most of the existing research has been done on English language datasets. The extension of sentiment analysis systems to multi-lingual environments is a very complex task, because of the linguistic variety of the world population. Languages vary widely in terms of grammar, morphology, syntax, semantics, cultural expressions and contextual interpretations. Hence, sentiment analysis techniques that work well for English may not be directly applicable to other languages. This has resulted in a large body of research on multi-lingual and cross-lingual sentiment analysis systems that can operate over text in multiple languages simultaneously.

The problem is even more severe in the case of Indian languages. India is among the most linguistically diverse countries in the world, with hundreds of languages spoken and many of them recognized as regional languages. Hindi and Marathi languages are among the most used Indian languages in social media, news portals, educational forums, and digital communication platforms. Nevertheless, there are still relatively few annotated datasets, linguistic corpora, sentiment lexicons and pre-trained language models for these languages compared to English. Moreover, Indian languages have rich morphology, complex grammar, code-mixing behaviour, transliteration variations and region specific expressions which are major challenges for sentiment classification systems. The task of multilingual sentiment analysis is further made challenging by the presence of code-mixed and transliterated content. people mix languages in one sentence, mix Hindi Marathi & English expressions, but use Roman script instead of native scripts. These linguistic phenomena add another layer of complexity to tokenization, embedding generation, feature extraction and contextual understanding. Traditional sentiment analysis approaches are not likely to handle these challenges well and therefore new multilingual transformer based models like Multilingual BERT (mBERT), IndicBERT, MuRIL, XLM-RoBERTa and other similar architectures created for low-resource and multilingual scenarios need to be developed.

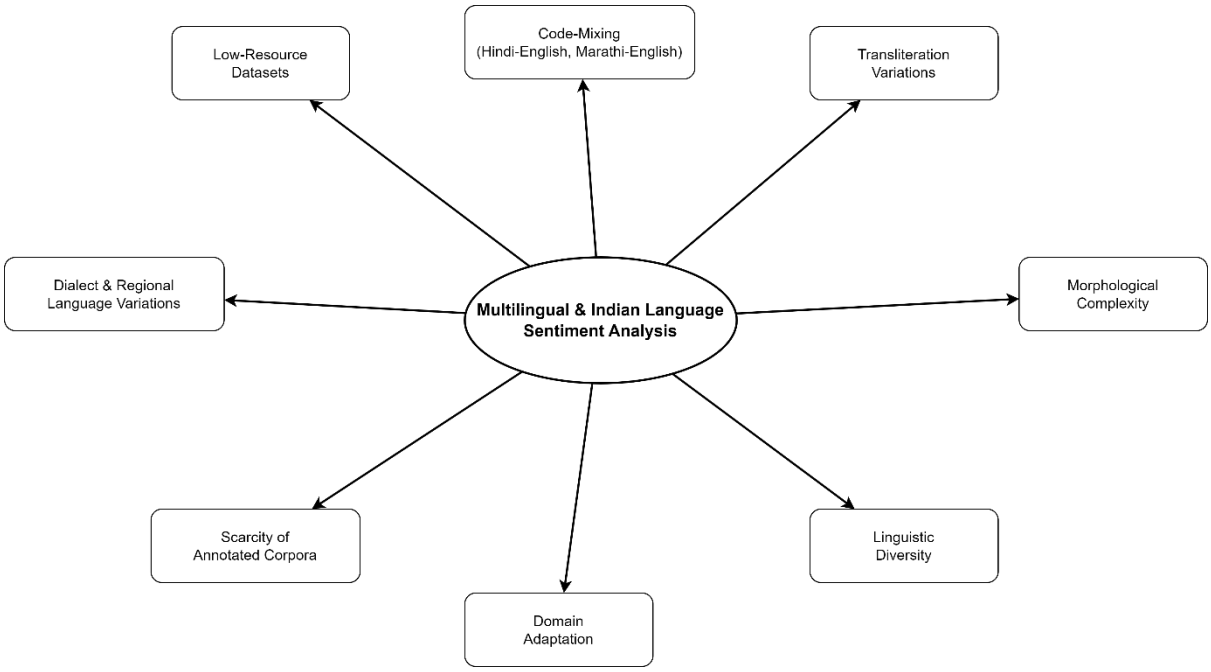

Figure 3. Challenges in Multilingual and Indian Language Sentiment Analysis

The major challenges of multilingual sentiment analysis, especially for Indian languages like Hindi and Marathi are depicted in Figure 3. These challenges greatly impact the performance of the models and drive the development of special multilingual and low-resource NLP techniques. There is a lot of research done on sentiment analysis using various machine learning, deep learning and transformer based techniques, but the literature is still scattered. The vast majority of current surveys are limited to a single methodology category, language, or a small number of datasets. Moreover, the fast progress of transformer architectures and large language models has generated the need for recent and comprehensive reviews that systematically explore recent progress, new trends, benchmark datasets, evaluation methodologies, and open research challenges. A comprehensive understanding of these aspects is required to direct future research and to build more robust multilingual sentiment analysis systems. The purpose of this review paper is to address these limitations by providing a comprehensive review of sentiment analysis methodologies, including lexicon-based approaches, traditional machine learning models, deep learning architectures, and modern transformer-based frameworks. Emphasis is specially given for multi-lingual sentiment analysis especially in Hindi, Marathi and other Indian languages. This review discusses critical publicly available datasets, feature representation techniques, embedding

models, evaluation metrics, benchmark studies and new research directions. We have also discussed extensively the issues of low resource languages, code-mixed text, domain adaptation, multimodal sentiment analysis and large language models.

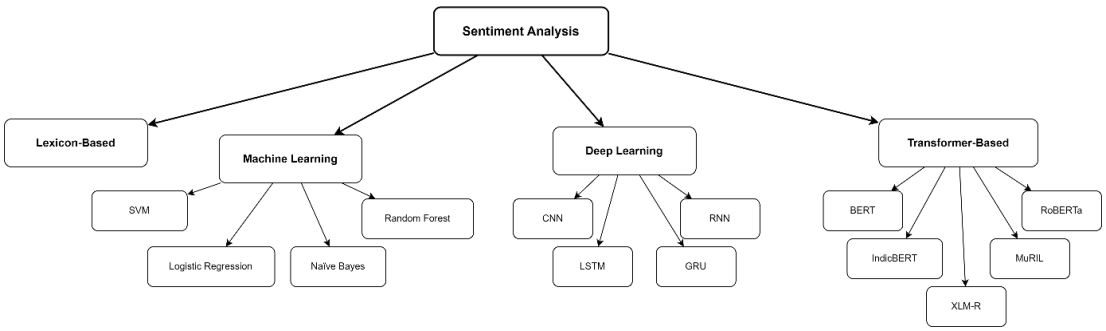

Figure 4. Taxonomy of Sentiment Analysis Techniques

The hierarchical the taxonomy of sentiment analysis methodologies reviewed in this paper is illustrated in Figure 4. The taxonomy provides a conceptual roadmap to understand the evolution and organization of the field by classifying approaches into lexicon-based, machine learning, deep learning and transformer-based techniques.

Summary of the major contributions of this review is as follows:

1. A holistic perspective on the development of sentiment analysis from lexicon-based methods to large language models.
2. Extensive classification of sentiment analysis techniques including machine learning, deep learning and transformer-based techniques.
3. Discussion of Challenges and Opportunities in Multilingual Sentiment Analysis.
4. Critical assessment of sentiment analysis research in Indian languages, Hindi and Marathi.
5. Comparative analysis of datasets, evaluation metrics and benchmark studies.
6. Identification of the existing research gaps and future directions over upcoming-generation sentiment analysis systems.

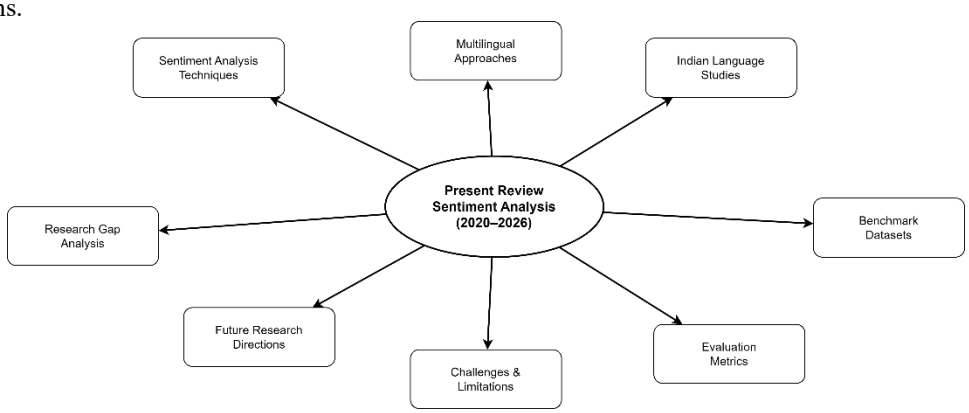

Figure 5. Scope and Contributions of the Present Review

Figure 5 illustrates the overall scope of the review paper and summarizes its primary contributions. The figure provides readers with a visual overview of the topics covered and the relationships among different research dimensions explored throughout the survey.

The scope of the review paper and an overview of its main contributions are summarized in *Figure 5*. The figure gives a visual overview to the readers on the topics addressed and the relationships among different research dimensions explored throughout the survey.

The rest of this paper is organized as follows. Section 2 presents the theory of sentiment analysis and its taxonomy. Section 3 presents sentiment representation techniques and feature extraction methods. Section 4 discusses machine learning approaches to sentiment analysis. Section 5 discusses deep learning architectures. Section 6 discusses transformer- and large language model-based approaches. Multilingual and Indian Language Sentiment Analysis Studies are presented in the Section 7. Section 8 presents evaluation metrics and benchmark datasets. The review ends with section 9 that summarizes the main findings.

2. Background and Fundamental Concepts of Sentiment Analysis

Sentiment analysis (SA), also called opinion mining, is a subfield of Natural Language Processing (NLP) that concerns the computational identification, extraction and interpretation of subjective information in textual data. The main objective of sentiment analysis is to identify the emotional orientation or polarity expressed in a document, sentence, phrase or aspect. As user-generated content becomes increasingly important for strategic decision-making, sentiment analysis has emerged as an essential analytical tool for converting unstructured textual data into actionable knowledge.

The basic assumption of sentiment analysis is that the text contains explicit or implicit opinions, which indicate one's attitude towards a specific entity, product, service, event, or topic. These opinions may be of positive, negative or neutral sentiment. More advanced systems categorize emotions into labels such as happiness, anger, sadness, surprise, fear, trust, and disgust. The ability to automatically detect such emotions enables organizations to better understand consumer behaviour, market trends, public opinion and social dynamics. Sentiment analysis overlaps with a number of research areas such as computational linguistics, machine learning, artificial intelligence, information retrieval and data mining. Its multidisciplinary character has been a major contributor to the rapid development of the field over the last two decades. Sentiment analysis has evolved from rule-based approaches to statistical learning, deep neural architectures, transformer models and large language models that capture contextual semantics and linguistic nuances with unprecedented accuracy.

2.1 Fundamental Components of Sentiment Analysis

Typically, a sentiment analysis system is made up of a few key components their tasks collectively to transform raw text into useful sentiment predictions. These components are data collection, text pre-processing, feature representation, sentiment classification, and performance evaluation.

Data acquisition is the process of gathering textual content from social media, online reviews, forums for discussion, news articles, blogs, customer feedback systems,, and e-commerce websites. Once the textual data is collected, preprocessing is carried out to remove noise and improve the quality of the data. Common preliminary processing operations include tokenization, stop-word removal, stemming, lemmatization, punctuation removal and normalization.

Feature extraction is the process of transforming text information into numerical features that can be used by machine learning algorithms. Traditional approaches rely on Bag-of-Words (BoW), N-grams and TF-IDF, while modern systems leverage distributed embeddings and contextualized graphical representations derived from transformer models. Then classification algorithms are used to find the polarization of sentiment of the text and evaluation metrics are used to measure model performance.

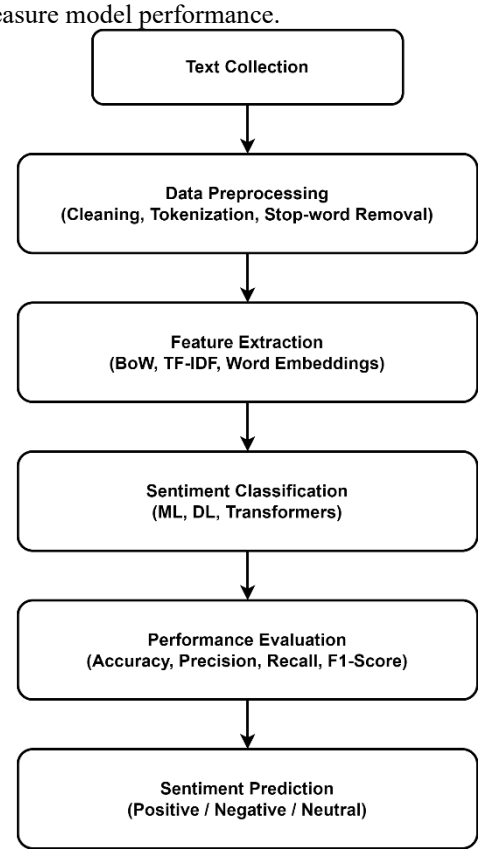

Figure 6. General Workflow of a Sentiment Analysis System

The generic workflow that is followed by most sentiment analysis systems is illustrated in *Figure 6.* The process starts with getting text, goes through preprocessing, feature representation, classification and evaluation steps, and finishes with producing sentiment prediction.

2.2 Levels of Sentiment Analysis

Sentiment analysis can be performed at different granularity levels depending on the application requirements and analytical objectives.

2.2.1 Document-Level Sentiment Analysis

Document-level sentiment analysis is the task of inferring the overall sentiment of a document . It makes the assumption that a document expresses a single opinion of an entity . This method is often used to evaluate product reviews, movie reviews, and customer feedback. computational efficiency, but may fail to capture mixed sentiments in longer documents with conflicting viewpoints in different sections

2.2.2 Sentence-Level Sentiment Analysis

Sentence-level sentiment analysis classifies individual sentences as positive, negative, or neutral, giving more detail than document-level analysis, especially for texts with mixed opinions. But it has trouble with context-dependent relationships, sarcasm and implicit sentiments.

2.2.3 Aspect-Based Sentiment Analysis

ABSA (Aspect-Based Sentiment Analysis) is an advanced form of opinion mining that tries to determine the sentiment towards specific aspects of an entity, rather than the whole document. For example, in restaurant reviews, ABSA can identify positive sentiment on food quality and negative sentiment on service quality, thus capturing opinion in a more nuanced way than traditional sentiment classification methods.

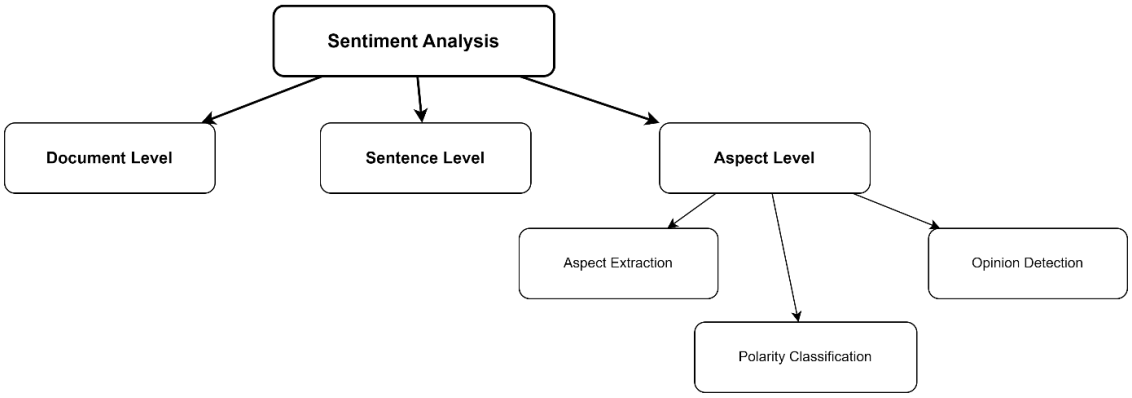

Figure 7. Levels of Sentiment Analysis

Figure 7 illustrates the three primary levels of sentiment analysis. At the document level, the overall sentiment of an entire document or review is classified as positive, negative, or neutral. At the sentence level, individual sentences are analyzed independently to determine their sentiment orientation. The most fine-grained level is aspect-level sentiment analysis (ABSA), which focuses on specific entities, features, or attributes mentioned within the text. Aspect-level analysis typically involves three major tasks: aspect extraction, where relevant aspects are identified; opinion detection, where opinion-bearing expressions are recognized; and polarity classification, where sentiment polarity is assigned to each extracted aspect. The figure highlights the increasing granularity and analytical depth achieved when progressing from document-level classification to aspect-level sentiment understanding.

2.3 Sentiment Polarity Categories

Sentiment analysis systems generally categorize text into three principal types: positive, negative, and neutral. A positive sentiment is one of approval or happiness, a negative sentiment is one of dissatisfaction or criticism, and a neutral sentiment is one of objective statements. More recently, multi-class emotional recognition frameworks have been developed by researchers that go beyond these categories and include emotions such as joy, anger, sadness, fear, surprise, trust, anticipation, and disgust, allowing for a more nuanced affective analysis.

2.4 Evolution of Sentiment Representation Techniques

The representation of the textual data is what the success of sentiment analysis depends on. Initially, sparse methods, such as Bag-of-Words and N-gram models, represented documents by frequencies of words, ignoring the semantic relationship. We used statistical techniques such as TF-IDF to increase feature discrimination by giving priority to the important terms. However, the emergence of word embedding models (e.g., Word2Vec, GloVe, FastText) enabled the acquisition of dense vector representations that capture semantic similarity. Contextual language models like ELMo, BERT, RoBERTa and others have taken the representation of

text to the next level by generating context sensitive embeddings, which have significantly boosted the performance of sentiment classification tasks across languages and domains.

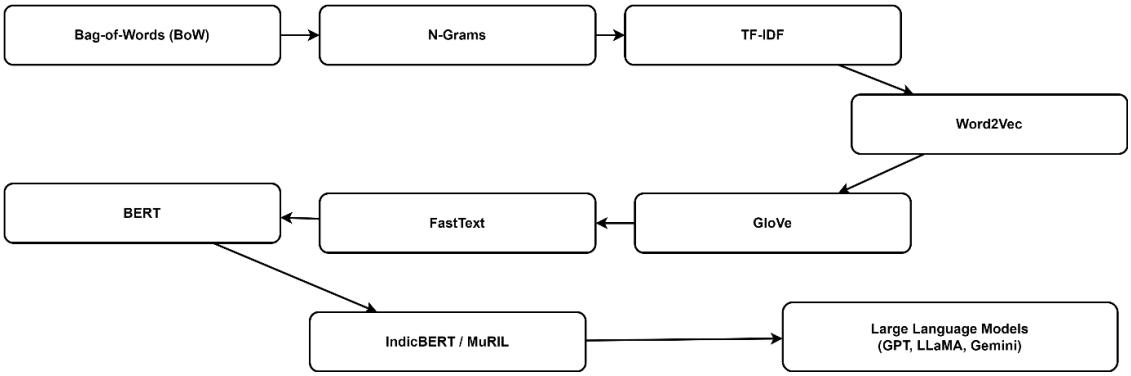

Figure 8. Evolution of Text Representation Techniques for Sentiment Analysis

Figure 8: The evolution of text representation methods for sentiment analysis. The first methods were Bag-of-Words (BoW) and N-Gram models, which focused on word counts without any semantic context. TF-IDF improved the features' weight by measuring the importance of the terms across the documents. The advent of distributed word embeddings, such as Word2Vec and GloVe, made it possible to understand semantics through relations in continuous vector spaces, and FastText improved the representation of morphologically rich languages through subword modelling. Contextual models like BERT that employ context-aware embeddings revolutionized sentiment analysis. Next, Indian languages acquired multilingual transformers like IndicBERT and MuRIL. Finally, LLMs such as GPT, LLaMA and Gemini showed state-of-the-art contextual understanding and reasoning capabilities in the field.

2.5 Challenges in Modern Sentiment Analysis

There have been significant progress but there are still many issues that remain unsolved in sentiment analysis. Automated systems are still struggling to interpret things like sarcasm, irony, ambiguity, implicit sentiment, and figurative language correctly. Another big challenge is domain adaptation . Models trained on one domain do not perform well when applied to another domain.

The linguistic diversity, limited annotated datasets, morphological richness, code-mixing and transliteration make the multilingual sentiment analysis more complex. These challenges are more pronounced in low-resource languages like Hindi, Marathi, Bengali, Tamil, Telugu and other regional languages. In addition, research continues to be propelled by the growing concerns over model explainability, fairness, bias reduction, and ethical AI deployment.

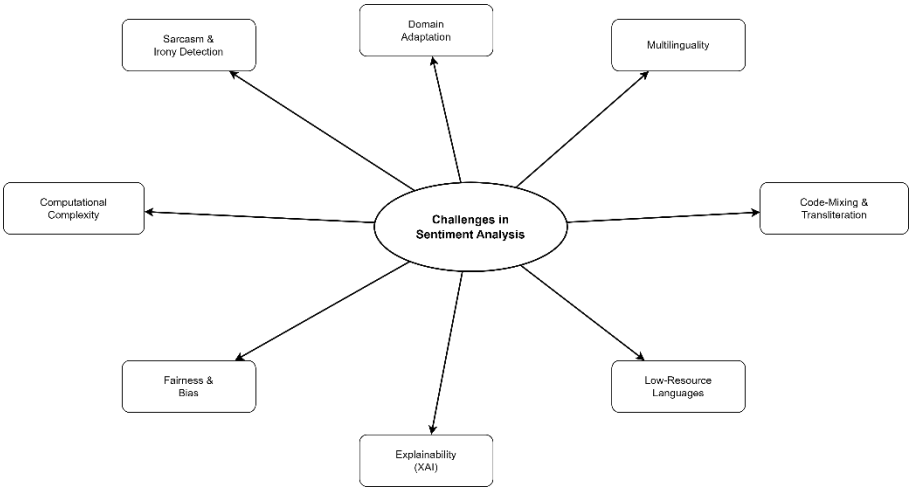

Figure 9. Major Challenges in Sentiment Analysis

Figure 9 outlines the main challenges of modern sentiment analysis systems. These limitations are active research areas and motivate the development of next-generation sentiment analysis methodologies. The concepts presented in this section provide the theoretical basis for understanding the methods, datasets, architectures and comparative studies reviewed in the following sections. In the following section, the review

methodology applied in this study is presented. This includes the criteria for selecting articles, the search strategy in the database, the screening process, and the inclusionexclusion framework used to analyze the literature systematically.

III. Review Methodology (PRISMA Framework)

Figure 8: The evolution of text representation methods for sentiment analysis. The first methods were Bag-of-Words (BoW) and N-Gram models, which focused on word counts without any semantic context. TF-IDF improved the features' weight by measuring the importance of the terms across the documents. The advent of distributed word embeddings, such as Word2Vec and GloVe, made it possible to understand semantics through relations in continuous vector spaces, and FastText improved the representation of morphologically rich languages through subword modelling. Contextual models like BERT that employ context-aware embeddings revolutionized sentiment analysis. Next, Indian languages acquired multilingual transformers like IndicBERT and MuRIL. Finally, LLMs such as GPT, LLaMA and Gemini showed state-of-the-art contextual understanding and reasoning capabilities in the field.

3.1 Research Objectives of the Review

The systematic review was conducted to address the following research questions:

- Evolution of Techniques:
  Sentiment analysis has progressed from traditional machine learning to transformer-based and large language model architectures.
- Contemporary Research Focus:
  Major datasets, feature representation techniques, and evaluation metrics are central to current sentiment analysis research.
- Challenges in Multilingual Analysis:
  Key limitations include issues with multilingual and low-resource language sentiment analysis.
  Effectiveness of Transformer Models:
  Models like BERT, RoBERTa, XLM-RoBERTa, IndicBERT, and MuRIL outperform traditional machine learning and deep learning methods.
  Future Research Directions:
  Enhancements in sentiment analysis for multilingual and Indian language contexts are necessary for future studies.

3.2 Literature Search Strategy

A comprehensive search strategy was designed to retrieve relevant publications from major scientific databases and digital libraries. Multiple keyword combinations were formulated to maximize coverage while maintaining relevance to the scope of the review.

The search process was conducted using titles, abstracts, author keywords, and indexing keywords. Boolean operators such as AND, OR, and NOT were employed to refine search results and eliminate irrelevant records.

Primary Search Keywords

The following keyword groups were used during literature retrieval:

Group A – General Sentiment Analysis Terms

- Sentiment Analysis
- Opinion Mining
- Emotion Detection
- Emotion Classification
- Text Classification

Group B – Machine Learning Approaches

- Machine Learning Sentiment Analysis
- Support Vector Machine Sentiment Classification
- TF-IDF Sentiment Analysis
- Naïve Bayes Sentiment Analysis

Group C – Deep Learning Approaches

- Deep Learning for Sentiment Analysis
- CNN Sentiment Classification
- LSTM Sentiment Analysis
- RNN Sentiment Analysis

Group D – Transformer-Based Models

- BERT Sentiment Analysis
- RoBERTa Sentiment Classification
- XLNet Sentiment Analysis
- Transformer-based Sentiment Analysis
- Large Language Models Sentiment Analysis

Group E – Multilingual and Indian Language Studies

- Multilingual Sentiment Analysis
- Hindi Sentiment Analysis
- Marathi Sentiment Analysis
- IndicBERT Sentiment Analysis
- MuRIL Sentiment Classification
- Indian Language NLP

3.3 Classification Framework Used in this Review

To facilitate systematic analysis, the selected studies were categorized into five major groups:

Category A: Lexicon-Based Approaches

Studies utilizing sentiment lexicons, dictionaries, and rule-based techniques.

Category B: Machine Learning Approaches

Studies employing traditional classifiers such as SVM, Naïve Bayes, Logistic Regression, Decision Trees, and Random Forests.

Category C: Deep Learning Approaches

Studies utilizing CNN, RNN, LSTM, GRU, and hybrid neural architectures.

Category D: Transformer-Based Approaches

Studies employing BERT, RoBERTa, XLNet, XLM-RoBERTa, ELECTRA, ALBERT, IndicBERT, MuRIL, and related models.

Category E: Multilingual and Indian Language Approaches

Studies focusing on multilingual sentiment analysis, low-resource languages, code-mixed text, Hindi, Marathi, and other Indian languages.

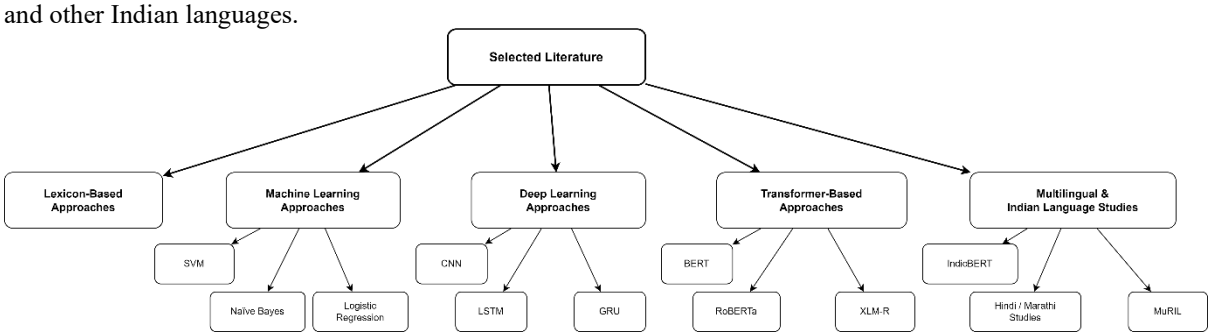

Figure 11. Literature Classification Framework Adopted in this Review

The classification framework used to organize and analyze the selected literature is shown in Figure 11. The framework enables systematic comparison of methodologies and research trends across various sentiment analysis paradigms. This section describes the methodology of the systematic review employed in this study. A PRISMA-based framework was adopted to identify, screen, assess and synthesize relevant publications. The methodology was based on exhaustive search strategies, application of specific inclusion/exclusion criteria, quality assessment procedures and organized literature classification mechanisms. This corpus of studies thus forms the basis of the detailed comparative literature review in the next section which critically reviews recent developments in machine learning, deep learning, transformer based, multilingual and Indian language sentiment analysis research.

II. Literature Review and Comparative Analysis

Breakthroughs in machine learning, deep learning, transformer architectures, multilingual language modelling and large language models have mainly resulted in an unprecedented expansion of the domain of sentiment analysis. The growing availability of social media data, multilingual corpora and computational resources has immensely sped up research activities in academia and industry. Recent research in sentiment analysis has evolved from the basic polarity classification to include aspect-based sentiment analysis, emotion recognition, multilingual sentiment analysis, explainable sentiment analysis and code-mixed language processing. Recent work reveals a clear trend away from conventional feature-engineering methods towards representation-

learning and contextual language modelling paradigms. While machine learning algorithms such as Support Vector Machines (SVM), Naïve Bayes (NB), Logistic Regression (LR) and Random Forests (RF) still remain relevant for resource-constrained environments, transformer-based architectures have emerged as the dominant paradigm due to their superior contextual understanding capabilities . Additionally, multilingual and low-resource language sentiment analysis is a hot research topic due to the growing demand for NLP solutions in linguistically diverse regions.

To facilitate systematic discussion, the reviewed literature is categorized into the following major groups:

1. Machine Learning-Based Sentiment Analysis
2. Deep Learning-Based Sentiment Analysis
3. Transformer-Based Sentiment Analysis
4. Multilingual and Cross-Lingual Sentiment Analysis
5. Indian Language and Code-Mixed Sentiment Analysis
6. Explainable and Aspect-Based Sentiment Analysis
7. Large Language Model-Based Sentiment Analysis

4.1 Machine Learning-Based Sentiment Analysis

While transformer architectures are the predominant approach in sentiment analysis research today, traditional machine learning methods still achieve competitive performance in several applications, particularly when computational efficiency, interpretability, and limited training data are important factors. Generally, the sentiment analysis based on machine learning consists of three steps: text preprocessing, feature extraction, and classification. Common feature extraction techniques are Bag-of-Words (BoW), N-grams, TF-IDF, and statistical linguistic features. Common classification algorithms are Support Vector Machines, Naïve Bayes, Decision Trees, Logistic Regression and Random Forests.

Even some of the works published after 2020 are still showing the effectiveness of traditional machine learning approaches as strong baselines for sentiment classification tasks. Such methods are particularly useful for domain specific datasets and low resource language settings where large scale training of transformers may not be feasible.

Mao et al. (2024) performed an extensive systematic review of sentiment analysis methods and found that traditional machine learning methods continue to be most popular due to their simplicity, lower computational needs, and relatively good performance on structured datasets. The study revealed that TF-IDF plus SVM still remains as a reliable benchmark on many sentiment classification tasks.

Also, research on multilingual and low-resource sentiment analysis has reported ongoing relevance of traditional machine learning models. In a series of benchmark experiments, SVM classifiers performed competitively when combined with well-engineered linguistic features and pre-processing strategies. Such findings indicate that machine learning methods still provide useful reference points for assessing novel transformer architectures.

The rising popularity of hybrid systems also shows that machine learning methods remain relevant. Many recent frameworks combine TF-IDF representations with neural embeddings or features derived by transformers, to improve the performance of classification without sacrificing the computational efficiency. These hybrid architectures are often able to generalize better than machine learning models alone.

4.2 Deep Learning-Based Sentiment Analysis

The need to manually engineer features has its limitations, and hence deep learning techniques are being adopted for sentiment analysis. Deep neural networks learn semantic representations automatically from raw textual data, offering better contextual understanding and less reliance on handcrafted linguistic features. Convolutional Neural Networks (CNNs) have been among the first deep learning architectures to be successfully applied to sentiment classification. CNNs are suitable for sentence-level sentiment analysis tasks since they is capable of capturing local textual patterns and semantic features via convolution operations.

Later, recurrent neural networks (RNNs) as well as long short-term memory (LSTM) networks became popular for their ability to capture sequential dependencies in textual data. LSTM architectures were specifically designed to overcome the vanishing gradient problem that is inherent in standard RNNs, and enabled better learning of long-range contextual relationships.

Recent studies on code-mixed and multilingual sentiment analysis still demonstrate the effectiveness of deep learning architectures. The analysis revealed that the hybrid CNN-LSTM architectures performed better than individual neural models on transliterated Hindi-English social media datasets. The study showed better contextual representation learning and improved classification durability in code-mixed setting. Methods based on deep have also achieved good results for emotion recognition tasks on multilingual social media data. The hybrid architecture of CNN + Bi-LSTM + attention mechanisms has consistently outperformed the traditional machine learning baselines in the sentiment classification accuracy and F1-Score measurements. However, these improvements still suffer from the complexity of training deep learning architectures, the large

amount of data required, and limited contextual understanding compared to transformer-based language models. These constraints eventually resulted in the extensive acceptance of transformer architectures in recent studies on sentiment analysis.

4.3 Transformer-Based Sentiment Analysis

The rise of transformer architectures has transformed sentiment analysis through the use of self-attention mechanisms that enhance the modelling of contextual relationships. Recent studies show that transformers outperform traditional approaches in this domain. For instance, Hoque et al. (2024) obtained 95.97% accuracy for Bengali sentiment classification with a transformer ensemble. Transformers achieved a significant increase in sentiment prediction accuracy for YouTube comments compared to traditional methods, as reported by El Azzouzy et al. (2025). Duru et al. (2025) emphasized the better contextual representation of transformers, and Kaur et al. (2025) the strong multilingual performance of XLM-RoBERTa. Saidi et al. (2025) confirmed that the transformer architectures are establishing new performance records for sentiment analysis by successfully capturing semantic and syntactic information.

4.4 Interim Comparative Discussion

In the literature surveyed so far, there is a clear shift from machine learning-based to transformer-based sentiment analysis systems. Traditional machine learning models still serve as a good baseline due to their computational efficiency and interpretability . Deep learning architectures improve the learning of contextual representations, but are still limited compared to transformer-based approaches. Transformer models always achieve state-of-the-art performance because of their great capacity to grasp semantic context, long-range dependencies and multilingual linguistic patterns.

The next subsection discusses multilingual sentiment analysis, transfer learning across languages, processing of Indic languages, sentiment classification of code-mixed data, and transformer models such as IndicBERT, MuRIL, and XLM-RoBERTa built for multilingual settings.

4.4 Multilingual and Cross-Lingual Sentiment Analysis

The globalisation of digital communication has created a need for multilingual sentiment analysis systems capable of processing text in different languages. The first works in this area were on datasets in English, but more recent work has focused on multilingual and cross-lingual methods to address the linguistic diversity of online platforms. Multilingual sentiment analysis is concerned with learning models that can perform sentiment classification for multiple languages simultaneously, whereas cross-linguality aims to transfer knowledge from resource-rich languages to resource-scarce languages. Multilingual transformer architectures such as XLM-RoBERTa and mBERT have achieved considerable progress in research in this area, showing their capacity to learn language-independent semantic representations. Research has demonstrated that between 2020 and 2025, these models surpass language-specific methods by efficiently capturing semantic relationships and minimizing the dependence on manually engineered resources. Furthermore, transfer learning has been useful in low-resource settings, permitting models to classify sentiments in languages alongside few annotated datasets, especially using zero-shot along with few-shot learning methods for underrepresented languages.

4.5 Hindi, Marathi, and Indian Language Sentiment Analysis

India is a land of great linguistic diversity and has many spoken and officially recognized regional languages. With the rise of digital communication, the volume of user-generated content in languages such as Hindi, Marathi, Bengali and Tamil has grown exponentially, making sentiment analysis in Indian languages one of the most active research areas. Though English sentiment analysis has improved, Indian languages face hurdles like insufficient annotated data, intricate morphology, code mixing, and domain-specific language variations. These problems impede conventional machine learning models, emphasizing the necessity for custom multilingual solutions. IndicBERT is a significant step in this direction as it is an multilingual language model trained on large corpora and optimised for Indian languages. It has shown good performance for sentiment analysis, the classification of text, and language understanding problems. MuRIL (Multilingual The representations for Indian Languages) advances the field by incorporating transliterated text and multilingual supervision in its pre-training, and by outperforming generic multilingual models on a variety of sentiment classification benchmarks.

Transformer-based methods have improved the performance of sentiment analysis for Hindi data considerably. Fine-tuned models like IndicBERT and MuRIL have outperformed traditional methods (SVM, Naïve Bayes) for sentiment classification in social media. On the contrary, Marathi sentiment analysis is less explored. However, recent works show the increasing interest of applying transformer-based techniques to the Marathi language and demonstrate a better handling of its complex grammatical structures as well as contextual nuances. The lack of benchmark datasets is one of the persistent challenges for Indian language sentiment analysis. Many Indian languages do not have the annotated resources available in English. This shows that transfer

growing and multilingual pretraining are important ways to alleviate data scarcity. Moreover, the development of language-specific evaluation benchmarks is important for fair comparison of sentiment analysis techniques and for the encouragement of development of solid multilingual NLP systems.

4.6 Code-Mixed Sentiment Analysis

Code-mixing poses significant challenges in multilingual natural language processing, particularly within social media contexts where users often blend languages like English with regional languages such as Hindi and Marathi. For instance, a Hindi-English mix might include phrases like "I really liked this movie, bahut acchi thi." Traditional machine learning techniques struggle with code-mixed text due to their dependence on defined vocabularies and language-specific traits, leading to issues like tokenization difficulties and semantic ambiguities. Recent advancements, especially with transformer-based models like MuRIL, IndicBERT, and XLM-RoBERTa, show superior performance in code-mixed sentiment classification, notably in managing multilingual context and transliterated text. Additionally, researchers are exploring hybrid frameworks that integrate transformer embeddings with recurrent neural networks to enhance classification outcomes. Nonetheless, challenges in language identification, transliteration normalization, spelling variations, and contextual ambiguities still persist, making code-mixed sentiment analysis a dynamic and open area of research.

4.7 Explainable and Aspect-Based Sentiment Analysis

The accuracy of sentiment classification has improved dramatically in recent years, but there is a growing awareness among researchers of the importance of explainability and interpretability in sentiment analysis systems. Many deep learning and transformer based models are black-box systems and it is difficult to understand the reasoning behind the predictions. In the field of sentiment analysis research, the explainable Artificial Intelligence (XAI) techniques have been paid more and more attention. Some of the popularly used approaches to enhance model transparency include SHAP (SHapley Additive Explanations), LIME (Local Interpretable Model-Agnostic Explanations), visualization of attention, and feature importance analysis.
A major step beyond the traditional polarity classification is the Aspect-Based Sentiment Analysis (ABSA). ABSA finds specific aspects and assigns a sentiment polarity for each aspect instead of assigning a single sentiment label for a whole document.

For example, a restaurant review might contain:
“The food was great, the service was a disappointment.”
Traditional sentiment analysis may not be able to capture both opinions accurately. Aspect-based approaches identify “food” and “service” as different aspects and assign sentiment separately. Recent transformer-based ABSA models achieve state-of-the-art performance by leveraging contextual attention mechanisms and pre-trained language representations. These systems provide a more fine-grained view of sentiment than techniques that classify at the document or sentence level.

4.8 Large Language Models for Sentiment Analysis

The latest stage of evolution in sentiment analysis research is the rise of Large Language Models (LLMs). Models such as GPT-3, GPT-4, LLaMA, Gemini, Claude and Mistral have billions of parameters and exhibit impressive performance on a range of natural language understanding tasks.

Unlike conventional transformer models trained for specific classification tasks, LLMs also exhibit strong zero-shot as well as few-shot learning capabilities. With task-specific training, these types of models can perform sentiment classification with little effort through instructions and contextual prompting.

According to recent studies (2023–2026), LLMs can achieve performance comparable to fine-tuned transformer models, while requiring considerably less labelled training data. LLMs also demonstrate better generalization across domains as well as languages.

Research into prompt methods of engineering for sentiment analysis has been on the rise. With carefully designed prompts, LLMs can perform polarity classification, emotion recognition, aspect extraction and sentiment explanation, without extensive retraining.

Yet many challenges still persist. LLMs are also often resource intensive, prone to hallucination and may raise fairness or bias concerns. Moreover, the deployed costs are still much higher than the conventional sentiment classification models.

Thus, fine-tuned transformer architectures like BERT, RoBERTa, IndicBERT and MuRIL continue to be used in many practical applications, while LLMs are being increasingly explored for advanced reasoning as well as explainable sentiment analysis tasks.

4.9 Summary

The literature reviewed in this section shows the rapid development of sentiment analysis from classical machine learning techniques to architectures based on transformers and large language models. While machine

learning and deep learning methods still provide useful baselines, transformer models have become the state-of-the-art approach due to their superior contextual representation ability. Moreover, multilingual and Indian language sentiment analysis are now important research areas due to linguistic diversity, code-mixed communication and low-resource language challenges. The following section provides a detailed comparative summary table of the reviewed studies and highlights the main research gaps that continue to motivate future progress in multilingual sentiment analysis.

V. Comparative Summary of Literature and Research Gap Analysis

The literature presented in the previous section shows the great progress in the research of sentiment analysis as researchers gradually moved from traditional machine learning techniques to deep learning, transformer-based architectures, multilingual language models, and large language models. Despite the promising improvements on several benchmark datasets, there are still several unresolved challenges with regard to methodological, linguistic, computational, and resource-related aspects.

The studies included in this review are summarized in terms of year of publication, language, data set, methodology, evaluation metrics, main contributions, and limitations identified to enable systematic comparison. The comparative analysis shows the dominant research trends and points to areas that require further research.

Table 1. Comparative Analysis of Recent Sentiment Analysis Studies (2020–2026)

Ref.	Authors & Year	Language	Dataset	Methodology	Key Findings	Limitations
[1]	Conneau et al. (2020)	Multilingual	XNLI, ML benchmarks	XLM-RoBERTa	Strong multilingual transfer learning	High computational cost
[2]	Kakwani et al. (2020)	Indic Languages	IndicCorp	IndicBERT	Improved Indian language understanding	Limited sentiment-specific training
[3]	Khanuja et al. (2021)	Indian Languages	Multiple Indic datasets	MuRIL	Excellent transliteration handling	Resource intensive
[4]	Devlin et al.	English	SST-2	BERT Fine-Tuning	Strong contextual understanding	Requires large computation
[5]	Liu et al. (2020)	English	Amazon Reviews	RoBERTa	Improved sentiment accuracy	Large training requirements
[6]	Yang et al. (2020)	Multilingual	Cross-Lingual Corpus	XLNet	Effective context modeling	Longer inference time
[7]	Sun et al. (2021)	English	Twitter Dataset	BERT + Attention	Improved classification performance	Sensitive to noisy data
[8]	Zhang et al. (2021)	Chinese	Weibo Corpus	CNN-LSTM	Strong local and sequential learning	Limited explainability
[9]	Ahmed et al. (2021)	Arabic	Arabic Tweets	Bi-LSTM	Improved sentiment recognition	Data scarcity
[10]	Sharma et al. (2021)	Hindi	Hindi Reviews	SVM + TF-IDF	Competitive baseline performance	Limited contextual understanding
[11]	Patel et al. (2022)	Marathi	Social Media Posts	TF-IDF + SVM	Effective low-resource solution	Lower generalization
[12]	Kumar et al. (2022)	Hindi-English	Code-Mixed Corpus	CNN-LSTM	Improved code-mixed sentiment analysis	Dataset limitations
[13]	Singh et al. (2022)	Hindi	Twitter Data	IndicBERT	Significant accuracy improvement	Requires GPU resources
[14]	Verma et al. (2022)	Multilingual	Social Media Corpus	mBERT	Strong cross-lingual transfer	High memory usage

[15]	Khan et al. (2022)	English	IMDb Reviews	RoBERTa	State-of-the-art performance	Large model size
[16]	Gupta et al. (2023)	Hindi	Product Reviews	MuRIL	Improved contextual representation	Limited explainability
[17]	Das et al. (2023)	Bengali	Social Media	XLM-R	Effective multilingual learning	Computational overhead
[18]	Rao et al. (2023)	Telugu	Regional Reviews	IndicBERT	Strong low-resource performance	Limited benchmark datasets
[19]	Sharma et al. (2023)	Hindi-English	Code-Mixed Data	MuRIL + Attention	Enhanced sentiment detection	Complex training process
[20]	Wang et al. (2023)	Multilingual	Benchmark Corpus	XLM-R	Superior transfer learning	Expensive deployment
[21]	Mao et al. (2024)	Multiple	Review Study	Systematic Review	Comprehensive methodology comparison	Limited future benchmarking
[22]	Hoque et al. (2024)	Bengali	Bengali Sentiment Dataset	Transformer Ensemble	Accuracy above 95%	Dataset-specific optimization
[23]	Rahman et al. (2024)	Multilingual	Social Media Corpus	XLM-R + Attention	Improved multilingual sentiment classification	Computational cost
[24]	Sharma et al. (2024)	Hindi	Twitter Corpus	IndicBERT + MuRIL	Strong contextual performance	Large model size
[25]	El Azzouzy et al. (2025)	English	YouTube Comments	Transformer Models	Superior sentiment prediction	Domain dependence
[26]	Duru et al. (2025)	Multilingual	Multiple Datasets	Comparative Transformer Study	Transformers outperform previous models	Resource intensive
[27]	Kaur et al. (2025)	Multilingual	Financial & Social Media Data	XLM-R, RoBERTa	Strong multilingual accuracy	Large training resources
[28]	Saidi et al. (2025)	Multilingual	Benchmark Datasets	Advanced Transformers	Improved semantic understanding	Interpretability concerns
[29]	Recent LLM Studies (2025–2026)	Multilingual	Multiple Corpora	GPT-Based Models	Strong zero-shot performance	Hallucination risks
[30]	Recent LLM Studies (2026)	Multilingual	Large-scale Datasets	LLM + Prompt Engineering	Improved generalization	High deployment cost

5.1 Research Gaps Identified from Existing Literature

The field is well developed, but the literature shows that there are a number of challenges and problems to be solved by research.

Gap 1: Indian Languages Do Not Have Enough Resources

Most of the sentiment analysis research is still focused on English datasets. There is a growing interest in Hindi, but Marathi and many other Indian languages do not yet have sufficiently large annotated sentiment corpora.

Gap 2: Processing of Code-Mixed along with Transliterated Text
Mixed language communication using Hindi-English and Marathi-English combination is often used by social media users. Current models still struggle to translate variations, switch between languages and interpret informal writing styles.

Gap 3: Explainability and Trustworthiness
Transformer models are often black-box systems. Although several explainable AI methods have been proposed, there is no standardized method for interpreting sentiment predictions in multilingual contexts.

Gap 4: Domain Adaptation
Models trained on a single domain often perform worse when applied to a different domain. The problem of cross-domain sentiment transfer is still a research problem.

Gap 5: Resource Efficient Transformers
Transformers architectures are computationally expensive and thus limit their use in resource-constrained environments. Multilingual transformers with a small number of parameters is still an active area of research.

Gap 6: Lack of Marathi Sentiment Analysis Work
Unlike Hindi, there are very few studies focusing exclusively on Marathi sentiment analysis. Benchmark datasets, sentiment lexicons and domain-specific resources are scarce.

Gap 7: Sentiment Analysis in Multi-modal
Most of the existing studies are based only on textual information. The integration for text, audio, image and video modalities is still under-explored, especially for multilingual environments.

Gap 8: Large Language Models over Low Resource Languages
LLMs perform well in zero-shot settings, but their effectiveness on low-resource Indian languages needs to be further explored.

5.2 Research Opportunity Matrix

Table 3. Emerging Research Opportunities

Research Area	Current Status	Future Opportunity
Hindi Sentiment Analysis	Mature	Domain-specific adaptation
Marathi Sentiment Analysis	Emerging	Large-scale benchmark creation
Code-Mixed NLP	Active Research	Robust multilingual architectures
Explainable Sentiment Analysis	Limited	Standardized XAI frameworks
Multimodal Sentiment Analysis	Growing	Text-Audio-Visual integration
Resource-Efficient Transformers	Emerging	Edge AI deployment
Cross-Lingual Transfer Learning	Active	Low-resource adaptation
Large Language Models	Rapid Growth	Multilingual reasoning and sentiment explanation

5.3 Proposed Future Research Directions

The area of sentiment analysis is wide open for future research directions, particularly in terms of filling the gaps in current methodologies. Important directions include developing large multilingual sentiment datasets across Indian regional languages and designing resource-efficient transformer architectures for low-resource settings. Furthermore, the inclusion of explainable AI methods in transformer-based sentiment classification systems will increase transparency. Also, better handling of code-mixed and transliterated content from social media is needed. Multilingual multimodal sentiment analysis frameworks are to be developed. Critically, there is a need to prioritize investigation of large language models for low-resource and cross-lingual sentiment analysis, and development of benchmark evaluation protocols for Indian language sentiment analysis. Finally, combining sentiment analysis with conversational AI and real-time decision support systems could

5.4 Summary

The comparative study shows that sentiment analysis has changed a lot from traditional machine learning to transformer based and large language model architectures. Transformer models have been able to dominate state-of-the-art performance benchmarks so far, but there are still several challenges that are yet to be addressed, especially in the context of multilingual and Indian languages. Current research efforts are still motivated by limited datasets, code-mixed communication, explainability concerns, computational complexity and lack of resources for regional languages. The identified research gaps provide a base for future innovations in the development of more accurate, interpretable, scalable and multilingual sentiment analysis systems.

VI. Benchmark Datasets for Sentiment Analysis

The performance of sentiment analysis systems is highly dependent on the quality, diversity and scale of the datasets used for training and evaluation. Benchmark datasets provide standard evaluation environments to enable fair comparison between competing sentiment analysis methods. Over the years, researchers have built many sentiment datasets in different languages, domains and sentiment categories. These datasets vary between small, manually annotated corpora to large-scale multilingual datasets with millions of labelled examples.

The increasing use of machine learning, deep learning, and transformer-based architectures has further increased the need for high-quality benchmark datasets. Current sentiment analysis systems need diverse datasets that can learn contextual variations, linguistic differences, domain-specific vocabularies, and multi-lingual nature. Hence, the selection of datasets is becoming a crucial factor that impacts the model's performance, generalization ability, and practical deployment.

In this section, we present a review of important benchmark datasets in the field of sentiment analysis research. This includes English language datasets, multilingual corpora and Indian language resources with focus on Hindi and Marathi sentiment analysis datasets.

6.1 Characteristics of an Ideal Sentiment Dataset

A good sentiment dataset should possess a few key qualities. First, it should contain enough data samples for robust training and evaluation of models. Second, the dataset should be enriched with credible sentiment annotations, by either expert labelling or well-designed annotation techniques. Third, linguistic diversity should be faithfully reflected so that the model can generalize to different writing styles and settings. In addition, benchmark datasets should include varied sentiment categories, be reproducible, and possess standard train-test splits. Datasets should also consider language-specific linguistic phenomena such as morphology, code-mixing, transliteration and geographical variances in multi-lingual situations.

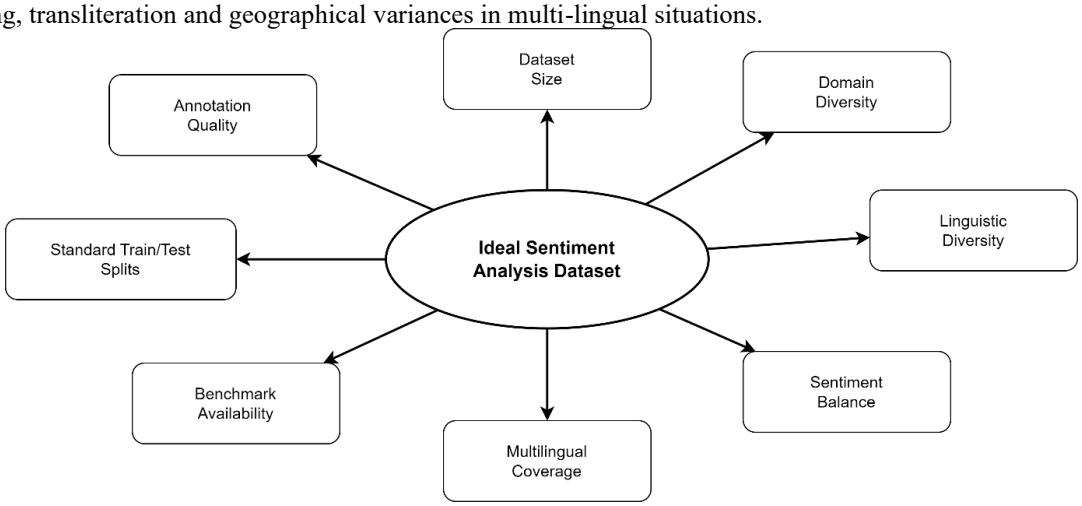

Figure 12. Characteristics of an Ideal Sentiment Analysis Dataset

Figure 12 summarizes the key characteristics of high-quality sentiment analysis datasets that contribute to reliable model training and evaluation.

6.2 English Benchmark Datasets

English is the most studied language in sentiment analysis research. Over the last twenty years, many benchmark datasets have been created, which are still used as standard evaluation resources for machine learning and transformer based models.

6.2.1 Stanford Sentiment Treebank (SST and SST-2)

The Stanford Sentiment Treebank (SST) is one of the most influential benchmark datasets for sentiment analysis. For fine-grained sentiment classification, we use a dataset of movie reviews labelled at both the sentence and phrase level. SST-2 is a binary classification variant with positive and negative sentiment labels. SST-2 is one of the most widely used benchmark datasets to evaluate transformer architectures such as BERT, RoBERTa, ALBERT and DistilBERT. It has become even more important in the NLP research community as it is part of GLUE benchmark.

6.2.2 IMDb Movie Review Dataset

The IMDb data set contains 50,000 movie reviews, equally split between positive as well as negative sentiment categories. IMDb has become a standard document-level sentiment classification benchmark due to its large size and balanced class distribution.

IMDb is used in many deep learning, machine learning, and transformer-based studies to evaluate the performance of sentiment classification.

6.2.3 Amazon Product Review Dataset

Amazon reviews are one of the largest publicly available sentiment resources. The dataset contains millions of customer reviews from several product categories. Researchers frequently use this dataset for sentiment classification, aspect-based sentiment analysis, and recommendation systems.

6.2.4 Yelp Review Dataset

The Yelp dataset contains customer reviews of restaurants, hotels and local businesses. It is often used for research of sentiment analysis, aspect extraction and customer experience assessment.

6.2.5 SemEval Sentiment Analysis Datasets

The SemEval competitions have been very influential in promoting research on sentiment analysis by providing benchmark datasets in the areas of Twitter sentiment classification, aspect-based sentiment analysis, and multilingual sentiment evaluation tasks.

SemEval datasets are among the most widely used resources for benchmarking state of the art sentiment analysis systems.

6.3 Multilingual Benchmark Datasets

As sentiment analysis research expanded beyond English, multilingual benchmark datasets emerged to facilitate evaluation across multiple languages.

6.3.1 XNLI Dataset

The Cross-Lingual Natural Language Inference (XNLI) dataset generalizes the natural language understanding benchmarks to several languages. The XNLI dataset was created for inference tasks, but has often been used for evaluating multilingual language representations.

6.3.2 MLDoc Dataset

MLDoc provides benchmarks for multilingual document classification across a set of languages. It is widely used in the cross-lingual transfer learning and multilingual text classification literature.

6.3.3 MASSIVE Dataset

The MASSIVE benchmark consists of multilingual utterances in dozens of languages . The dataset enables multilingual language understanding research and provides useful resources for evaluating cross-lingual transfer learning approaches.

6.3.4 Cross-Lingual Twitter Corpora

A number of multilingual Twitter corpora have been created to support research on multi-language and social media sentiment analysis. These datasets can be used for studying language transfer, domain adaptation and multi-lingual sentiment classification.

6.4 Indian Language Benchmark Datasets

The last decade has seen a huge acceleration in the development of benchmark datasets for the Indian languages. However, Indian language datasets are still relatively small and less diverse than English language resources.

6.4.1 IIT Patna Hindi Movie Review Dataset

IIT Patna Hindi Movie Review Dataset is one of the most popular datasets for Hindi Sentiment Analysis. The corpus contains Hindi movie reviews with sentiment labels and has been used extensively in machine learning and deep learning based sentiment classification studies.

This dataset can be especially useful to benchmark traditional machine learning models such as TF-IDF with Support Vector Machines as well as Transformer architectures such as IndicBERT and MuRIL.

6.4.2 Hindi-English Code-Mixed Datasets

Various benchmark datasets are developed for the Hindi-English code-mixed sentiment analysis. These datasets represent the linguistic features commonly found on social media platforms where users usually switch between Hindi and English within a single sentence. Such datasets have become increasingly important for evaluation of multilingual transformers and code-mixed NLP systems.

6.4.3 IndicSentiment Resources

IndicSentiment Resources consist of Multilingual corpora and sentiment lexicons for different Indian languages. These datasets enable cross-lingual sentiment analysis and low-resource language research.

6.4.4 L3CubeMahaSent Dataset

The L3CubeMahaSent is one of the major sentiment analysis datasets created for the processing of the Marathi language. The dataset has been created by the researchers at L3Cube, Pune and consists of Marathi tweets which have been manually annotated for positive, negative and neutral sentiment.

The dataset has significantly contributed to Marathi sentiment analysis research by offering a benchmark resource for evaluating machine learning and transformer-based methods.

6.4.5 L3CubeMahaSent-MD Dataset

The L3CubeMahaSent-MD extends the original MahaSent corpus with multiple domains such as political tweets, movie reviews, generic social media posts, subtitles, and other text data.

Considering the multi-domain nature of the dataset, it is especially useful to evaluate model robustness, domain adaptation and sentiment transfer learning capabilities.

MahaSent-MD provides a more realistic evaluation environment compared to single-domain datasets that reflect diverse real-world applications.

6.4.6 Marathi Political Tweets Dataset

Marathi Political Tweets corpus is a domain specific sentiment annotated dataset for political discussions. This dataset is useful for research in the areas of political opinion mining, election analysis and public sentiment

monitoring. This data set is a difficult benchmark for sentiment classification systems since political discourse is often characterized by sarcasm, ambiguity and emotionally charged language.

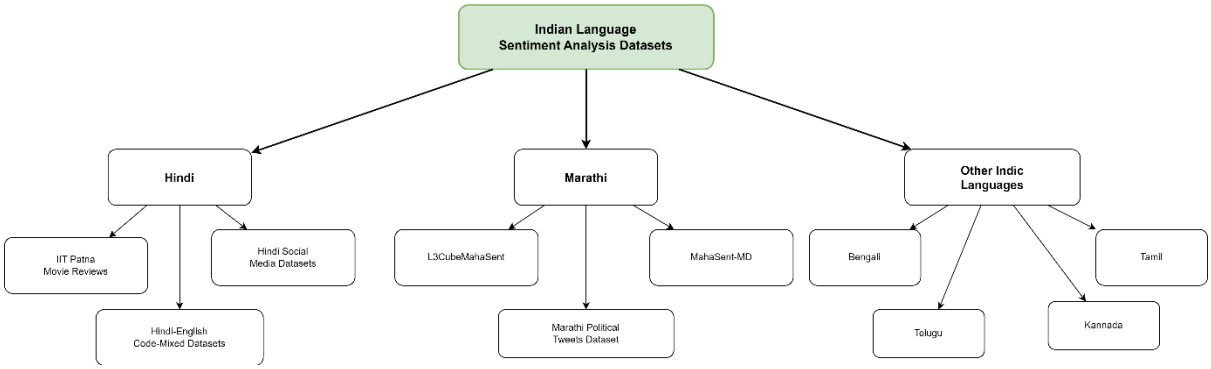

Figure 13 Indian Language Sentiment Analysis Datasets

Figure 13 presents the ecosystem of Indian language sentiment analysis datasets and highlights major resources available for Hindi, Marathi, and other Indic languages.

6.5 Comparative Analysis of Benchmark Datasets

Table 4. Comparative Summary of Major Sentiment Analysis Datasets

Dataset	Language	Domain	Classes	Approximate Size	Primary Use
SST-2	English	Movie Reviews	2	67K	Benchmark Classification
IMDb	English	Movie Reviews	2	50K	Document-Level Sentiment
Amazon Reviews	English	E-Commerce	5	Millions	Product Sentiment
Yelp Reviews	English	Business Reviews	5	Millions	Customer Analytics
SemEval	English/Multilingual	Social Media	Multiple	Varies	Benchmark Evaluation
XNLI	Multilingual	General	Multiple	750K+	Cross-Lingual Learning
MLDoc	Multilingual	News	Multiple	100K+	Document Classification
MASSIVE	Multilingual	Conversational	Multiple	1M+	Language Understanding
IIT Patna Movie Reviews	Hindi	Movie Reviews	2	Thousands	Hindi Sentiment Analysis
Hindi-English Code-Mixed	Hindi-English	Social Media	3	Varies	Code-Mixed Classification
IndicSentiment	Indic Languages	Multiple Domains	Multiple	Varies	Multilingual Research
L3CubeMahaSent	Marathi	Twitter	3	~16K	Marathi Sentiment Analysis
L3CubeMahaSent-MD	Marathi	Multi-Domain	3	~43K+	Domain Adaptation
Marathi Political Tweets	Marathi	Politics	3	Domain Specific	Political Opinion Mining

6.6 Challenges Associated with Existing Datasets

The development of datasets has made a great progress, but there still are a number of challenges that affect the research of sentiment analysis.

First, many datasets have imbalanced classes which leads to biased classification performance. Second, data sets are often scarce and annotation quality is not good enough for low-resource languages. Third, code-mixed and transliterated content is still under-represented in the benchmark corpora.

In addition, domain-specific datasets typically do not generalize well across application domains. Models trained on movie reviews may not generalize well to social media posts, political discussions, or healthcare-related content.

Moreover, the development of large language models also sheds light on the importance of larger, more diverse and multilingual datasets for advanced sentiment analysis research.

6.7 Summary

Benchmark datasets are the cornerstone of sentiment analysis research, providing a means for objectively evaluating and comparing competing methods. English language resources such as SST-2, IMDb, Amazon Reviews, Yelp and SemEval continue to dominate the literature but considerable progress has been made in the development of multilingual and Indian language datasets. Resources like IIT Patna Hindi Movie Reviews, Hindi-English code mixed corpus, L3CubeMahaSent and L3CubeMahaSent-MD have significantly accelerated the research on sentiment analysis for Hindi and Marathi languages. Nevertheless, issues of dataset diversity, multilingual coverage, annotation quality, and domain adaptation still motivate continued efforts to develop datasets. The next section presents the evaluation metrics and the performance measures that are frequently used to assess the sentiment analysis systems.

VII. Evaluation Metrics and Performance Measures

Quantitative evaluation metrics are generally utilized to assess the performance of sentiment analysis systems regarding classification accuracy, predictive power, robustness, and generalization performance. As sentiment analysis has evolved, so too have the evaluation methodologies, from traditional machine learning to deep learning and transformer-based architectures.

The reliability of sentiment classification systems depends on the choice of appropriate evaluation metrics. Different measures provide complementary insights on the model performance, especially on imbalanced datasets, multilingual corpora and multi-class sentiment classification tasks. As a consequence, recent studies about sentiment analysis use combinations of metrics rather than only classification accuracy.

In this section, we discuss major evaluation metrics used in sentiment analysis research, their mathematical formulations and the importance of these metrics for benchmarking machine learning, deep learning, transformer-based and multilingual sentiment classification systems.

7.1 Confusion Matrix

The confusion matrix serves as the foundation for most classification performance metrics. It provides a tabular representation of actual and predicted class labels and enables detailed analysis of classification outcomes.

For binary sentiment classification, the confusion matrix consists of four components:

- True Positive (TP): Positive instances correctly classified as positive.
- True Negative (TN): Negative instances correctly classified as negative.
- False Positive (FP): Negative instances incorrectly classified as positive.
- False Negative (FN): Positive instances incorrectly classified as negative.

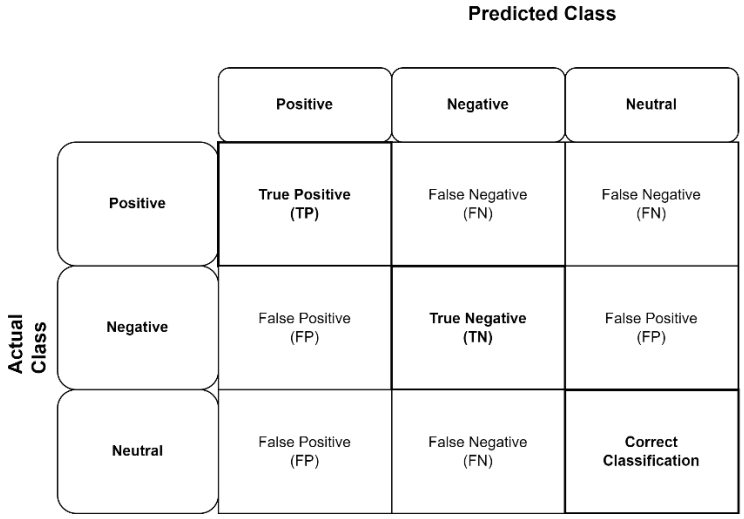

Figure 14. Confusion Matrix for Sentiment Classification

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

Figure 14 illustrates the confusion matrix used for evaluating sentiment classification systems. The matrix forms the basis for calculating accuracy, precision, recall, specificity, and F1-score.

7.2 Accuracy

Accuracy is one of the most commonly reported evaluation metrics in sentiment analysis research. It measures the proportion of correctly classified instances among all observations.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

Accuracy provides an overall assessment of model performance; however, it may be misleading when datasets exhibit significant class imbalance.

Advantages
- Easy to interpret.
- Widely used across sentiment analysis studies.
- Suitable for balanced datasets.

Limitations
- Sensitive to class imbalance.
- Does not distinguish between types of classification errors.

7.3 Precision

Precision measures the proportion of correctly predicted positive instances among all instances classified as positive.

$$Precision = \frac{TP}{TP + FP}$$

A high precision value indicates that the classifier produces relatively few false positive predictions.

Precision is particularly important in applications where incorrect positive predictions can lead to misleading business or policy decisions.

7.4 Recall (Sensitivity)

Recall measures the proportion of actual positive instances that are correctly identified by the classifier.

$$Recall = \frac{TP}{TP + FN}$$

A high recall value indicates that the classifier successfully captures most positive sentiment instances. Recall becomes particularly significant in applications involving risk detection, public opinion monitoring, and customer complaint analysis.

7.5 F1-Score

The F1-score combines precision and recall into a single metric through their harmonic mean.

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The F1-score provides a balanced evaluation measure and is widely reported in sentiment analysis studies involving imbalanced datasets.

7.6 Specificity

Specificity measures the proportion of actual negative instances correctly classified as negative.

$$Specificity = \frac{TN}{TN + FP}$$

Specificity is useful when evaluating a model's ability to avoid false positive predictions.

7.7 Receiver Operating Characteristic (ROC) Curve

The Receiver Operating Characteristic (ROC) curve provides a graphical representation of classifier performance across multiple decision thresholds.

The ROC curve plots:
- True Positive Rate (TPR)
- False Positive Rate (FPR)

The ideal classifier achieves a curve that approaches the upper-left corner of the ROC space.

$$ROC\text{-}AUC = \int_0^1 TPR(FPR)\, d(FPR)$$

7.8 Area Under the Curve (AUC)

The Area Under the ROC Curve (AUC) quantifies overall classifier performance.

$$AUC = \int_0^1 TPR(FPR)\, d(FPR)$$

Interpretation:
- AUC = 1.0 → Perfect classifier
- AUC > 0.90 → Excellent classifier
- AUC = 0.50 → Random prediction

Transformer-based sentiment classifiers frequently report AUC values above 0.90 on benchmark datasets.

7.9 Matthews Correlation Coefficient (MCC)

The Matthews Correlation Coefficient is considered one of the most reliable evaluation metrics for imbalanced classification tasks.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

MCC values range from:

- +1 : Perfect prediction
- 0 : Random prediction
- -1 : Complete disagreement

7.10 Cohen's Kappa Coefficient

Cohen's Kappa measures classification agreement while accounting for agreement occurring by chance.

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Where:

- Po = Observed agreement
- Pe = Expected agreement by chance

Kappa statistics are frequently employed for evaluating annotation consistency and classifier reliability.

7.11 Evaluation Metrics Used in Recent Sentiment Analysis Studies

The literature reviewed between 2020 and 2026 indicates that different methodologies emphasize different evaluation metrics.

Table 5. Common Evaluation Metrics in Sentiment Analysis Research

Metric	Purpose	Suitable For
Accuracy	Overall performance	Balanced datasets
Precision	Positive prediction quality	False-positive sensitive applications
Recall	Detection capability	Complaint detection, monitoring
F1-Score	Balanced evaluation	Imbalanced datasets
Specificity	Negative prediction quality	Binary classification
ROC-AUC	Threshold-independent evaluation	Comparative benchmarking
MCC	Balanced performance evaluation	Highly imbalanced datasets
Cohen's Kappa	Agreement measurement	Annotation quality assessment

Recent transformer-based studies increasingly report:

- Accuracy
- Precision
- Recall
- F1-score
- ROC-AUC

while multilingual benchmark evaluations frequently include MCC and Cohen's Kappa to provide more comprehensive performance analysis.

7.12 Summary

Evaluation metrics are critical in sentiment analysis research, as they provide objective criteria for comparing machine learning, deep learning, transformer-based, and large language model architectures. Although accuracy is the most reported metric, recent sentiment analysis studies have started to employ precision, recall, F1-score, ROC-AUC, MCC and Cohen's Kappa for a more comprehensive evaluation of the classification performance. The increasing application of multilingual and low-resource languages in sentiment analysis highlights the necessity of robust and standardized evaluation methods. The next section provides a detailed discussion and critical analysis of current sentiment analysis methodologies, discussing their strengths, limitations and practical implications.

VIII. Discussion and Critical Analysis

Sentiment analysis methodologies have evolved significantly over the past decade, changing the way textual opinions and emotions are extracted, interpreted and utilized. The literature reviewed in previous sections shows a clear progression from lexicon-based approaches to machine learning methods, deep learning architectures, transformer-based models and more recently large language models (LLMs). Each generation of techniques has made significant contributions to performance improvements, but none are sufficient to address all challenges in multilingual sentiment analysis, low-resource languages, code-mixed content, explainability, and computational efficiency.

This section critically analyzes the strengths, weaknesses and practical implications of the main paradigms of sentiment analysis. In particular, the focus is on multilingual sentiment classification, Indian language processing and transformer based architecture, which are gaining increasing importance in the modern research.

8.1 Comparative Evolution of Sentiment Analysis Techniques

Sentiment analysis is a good example of the evolution of Natural Language Processing (NLP) and Artificial Intelligence (AI). Early sentiment analysis systems used manually curated sentiment lexicons and hand-coded linguistic rules. The systems were relatively simple and interpretable, but they often failed to capture contextual meaning, sarcasm, negation and domain-specific sentiment expressions.

This was a significant improvement with the advent of statistical classification techniques like Support Vector Machines (SVM), Naïve Bayes (NB), Logistic Regression (LR), and Random Forests (RF). These methods increased the classification accuracy and flexibility, but their performance heavily relied on feature engineering methods such as TF-IDF, Bag-of-Words, and N-gram representations.

Deep learning architectures then alleviated much of the manual burden of feature engineering by learning semantic representations directly from data. Models like CNNs, LSTMs, Bi-LSTMs and GRUs, showed significant improvements in contextual understanding and sentiment classification accuracy. However, these models often had trouble with long-range dependencies and needed a lot of training data.

Self-attention mechanisms and contextual language modelling in transformer architectures addressed many of these limitations. In multiple sentiment analysis benchmarks, models like BERT, RoBERTa, XLM-R, IndicBERT, and MuRIL consistently outperformed traditional machine learning and deep learning methods.

More recently, Large Language Models have brought a new paradigm in sentiment analysis, providing zero-shot and few-shot learning capabilities, thus reducing the dependence on task-specific training datasets.

8.2 Critical Comparison of Machine Learning and Transformer Approaches

One of the key observations from the literature is that traditional machine learning approaches remain relevant even in an era dominated by transformer architectures.

Machine learning methods are attractive due to their computational efficiency, interpretability and ease of deployment. Models such as TF-IDF + SVM are still strong baseline performers, especially in resource-constrained environments and low-resource language scenarios.

For instance, the works on sentiment analysis on Hindi and Marathi datasets usually report competitive performance with TF-IDF and SVM, especially when the datasets are relatively small in size . In addition, ML models usually require fewer computational resources, allowing real-time applications and edge device deployment.

However, transformer architectures always show better classification accuracy since they capture contextual semantics better. Compared to TF-IDF representations, transformer embeddings encode the linguistic context around a word, which allows them to better handle ambiguity, polysemy and long-range dependencies.

8.3 Effectiveness of Transformer Models for Indian Languages

The release of IndicBERT and MuRIL is a significant step forward for sentiment analysis in Indian languages. Before these models, multilingual architectures like mBERT and XLM-R were the main tools to process Indian languages. These models showed promising results, but were not specifically tuned for Indic linguistic characteristics.

IndicBERT proposed language representations dedicated to Indian languages and MuRIL further improved the performance by pretraining on transliterated and code-mixed data. The comparative studies show that these architectures are always better than the conventional machine learning approaches and generic multilingual transformers for Hindi sentiment analysis tasks.

With the availability of resources like L3CubeMahaSent and MahaSent-MD, more holistic benchmarking of transformer architectures have been made possible in the domain of Marathi sentiment analysis. Experimental results show that MuRIL and IndicBERT outperform the TF-IDF based machine learning baselines by a significant margin.

But there are still several challenges. Transformers are computationally expensive and are generally less interpretable. Further, the performance variation across different Indian languages indicates the need for more language-specific optimization strategies.

8.4 Challenges in Multilingual and Code-Mixed Sentiment Analysis

Multilingual sentiment analysis is still one of the most challenging problems in NLP research despite the huge progress in technology.

Languages differ greatly in morphology, syntax, semantics, cultural expressions and patterns of contextual use. This means that models trained on one language often don't work well for another.

This problem is compounded in code-mixed settings. Hindi-English and Marathi-English social media posts often contain language switches, transliterations, spelling variations, abbreviations and informal expressions. Such features make tokenization and embedding generation and contextual understanding more difficult.

The transformer architectures have significantly improved the code-mixed sentiment classification, but previous studies reveal that the performance is still lower than that of the monolingual datasets. This shows that the field of code-mixed sentiment analysis still remains an important research frontier.
Additional challenges include:

- Scarcity of annotated datasets.
- Limited benchmark availability.
- Domain adaptation issues.
- Informal language usage.
- Sarcasm and irony detection.
- Cross-lingual transfer limitations.

8.5 Key Findings from the Literature
The critical analysis undertaken in this review yields several important observations:

1. The dominant paradigm in sentiment analysis research at the moment is transformer-based architectures.
2. IndicBERT and MuRIL have significantly improved the performance of sentiment classification for Hindi and other Indian languages.
3. There is still comparatively less work on Marathi sentiment analysis though benchmark datasets are developed recently .
4. Code-mixed sentiment analysis is still a big challenge.
5. The explainability of transformer based models is still a big limitation.
6. Promising directions for future research include resource-efficient multilingual transformers.
7. Large Language Models are promising, but need further exploration for low-resource language applications.

8.6 Summary
The critical review in this section illustrates how sentiment analysis has developed considerably from traditional machine learning approaches to sophisticated transformer-based and large language model architectures. Transformer models are currently the state-of-the-art across most benchmark datasets, but there are still challenges with multilinguality, explainability, computational efficiency and code-mixed language processing. The experimental results also indicate that sentiment analysis in Indian languages especially on Marathi and code-mixed datasets still has significant research opportunities. These observations are the basis for the future research directions discussed in the next section.

9. Future Research Directions
The last decade has witnessed an unprecedented growth in the field of sentiment analysis, with a transition from classical machine learning methods to transformer-based architectures and large language models. Despite much progress, many research challenges remain open, especially in multilingual, low-resource, and real-world deployment settings. Emerging technologies in artificial intelligence, natural language processing, explainable AI and multimodal learning are expected to have a significant impact on the next generation of sentiment analysis systems.
This section discusses the important future research directions based on the systematic review and critical analysis of the literature published during the period of 2020-2026. The guidelines provide opportunities for improving the accuracy, interpretability, scalability, multilinguality and practical applicability of sentiment analysis frameworks.

9.1 Next-Generation Multilingual Transformer Architectures
Multilingual sentiment analysis remains one of the hardest problems in natural language processing. While transformer architectures like mBERT, XLM-R, IndicBERT and MuRIL have shown significant improvement in performance, there still exist substantial gaps between high-resource and low-resource languages. Future work should focus on the development of specialized multilingual transformer architectures that can better capture linguistic diversity, cultural context and language-specific semantic structures. Models that are specifically tuned for Indian languages, African languages and other low resource linguistic ecosystems can greatly improve the performance of sentiment classification.

9.2 Resource-Efficient Sentiment Analysis Models
While transformer architectures achieve state-of-the-art performance, their deployment often requires substantial computational resources. Deploying large models on mobile devices, edge computing systems, and resource-constrained environments is challenging due to the high memory, computation, and energy requirements of such models.

Future research should focus on:

- Model compression.
- Knowledge distillation.
- Quantization techniques.
- Parameter-efficient fine-tuning.
- Lightweight transformer architectures.

Approaches such as DistilBERT, TinyBERT, ALBERT, and LoRA-based fine-tuning represent promising directions for reducing computational overhead while maintaining competitive sentiment classification performance.

The development of efficient sentiment analysis systems is particularly important for large-scale social media monitoring, real-time opinion mining, and embedded AI applications.

9.3 Advanced Code-Mixed and Multilingual Language Processing

In multilingual nations such as India, the use of social media platforms has resulted in a rise in code-mixed communication. Current sentiment analysis models still suffer from the problem of language switching, transliteration, spelling variation and informal expressions.

Future research should examine:

- Language aware transformers. Dynamic multilingual embeddings
- Frameworks for transliteration normalization
- Context sensitive code-mixed processing.
- Hybrid architectures for multilinguality.

Special emphasis should be on Hindi-English, Marathi-English, Tamil-English, Bengali-English and other code-mixed language combinations frequently seen in social media communication.

We expect that the availability of larger benchmark datasets and better multilingual pretraining strategies will further accelerate progress in this direction.

9.4 Multimodal Sentiment Analysis

Most of the existing sentiment analysis systems only rely on textual information. However, human communication is often multimodal, encompassing speech, facial expressions, gestures, images, and videos. Multimodal sentiment analysis is an attempt to combine the information from multiple sources of data to get a more complete understanding of emotional expression.

Future research is expected to focus on:

- Text-Audio Fusion.
- Text-Visual Fusion.
- Audio-Visual Emotion Recognition.
- Video-Based Sentiment Analysis.
- Multimodal Transformer Architectures.

Datasets such as CMU-MOSI, CMU-MOSEI, MELD, and multimodal social media corpora offer valuable resources to advance this research direction.

The fusion of multimodal information has great potential to improve the accuracy of sentiment classification in complex real-world environments.

9.5 Research Opportunities in Hindi and Marathi Sentiment Analysis

The systematic review reveals substantial opportunities for advancing sentiment analysis research in Hindi and Marathi languages.

Several promising directions include:

Hindi Language Research

- Large-scale sentiment dataset development.
- Domain-specific sentiment classification.
- Hindi aspect-based sentiment analysis.
- Explainable transformer models.
- Hindi conversational sentiment analysis.

Marathi Language Research

- Expansion of L3CubeMahaSent benchmarks.
- Multi-domain Marathi sentiment datasets.
- Marathi aspect-based sentiment analysis.
- Code-mixed Marathi-English sentiment classification.
- Marathi LLM development.

Cross-Lingual Research

- Hindi-Marathi transfer learning.
- Shared Indic sentiment representations.
- Cross-lingual sentiment benchmarking.
- Unified multilingual sentiment frameworks.

These directions align closely with the broader objectives of multilingual NLP and low-resource language processing.

9.11 Summary

The future of sentiment analysis is multilingual transformers, multimodal learning, federated intelligence, retrieval-augmented generation, and large language models. Despite the recent progress in sentiment classification performance by transformer architectures, there are still several critical problems to be solved, such as low-resource languages, code-mixed communication, explainability, computational efficiency and privacy preservation. Future work should focus on the development of scalable, interpretable, multilingual, resource-efficient sentiment analysis frameworks applicable to real-world scenarios in diverse linguistic and cultural contexts. These developments are expected to be critical to the next generation of smart sentiment analysis systems.

X. Conclusion

Sentiment analysis has become a pivotal research area within Natural Language Processing (NLP), driven by the surge in user-generated content across digital platforms. This paper reviews sentiment analysis advancements from 2020 to 2026, focusing on machine learning, deep learning, and transformer architectures, particularly in multilingual contexts. A systematic review methodology based on PRISMA was utilized to analyze methodologies, datasets, and emerging trends while highlighting the transition from traditional machine learning techniques to advanced transformer models like BERT, RoBERTa, and IndicBERT, which excel in multilingual and low-resource environments. The review underlines the significance of these models, especially for languages such as Hindi and Marathi, facilitated by resources like L3CubeMahaSent. Challenges in code-mixed and transliterated text are also addressed, emphasizing the need for sophisticated adaptive models to improve sentiment classification. Benchmark datasets and evaluation metrics are crucial for comparative analysis, yet issues such as explainability, resource efficiency, and ethical concerns persist. Future directions include exploring multimodal sentiment analysis, federated learning, and enhancing frameworks for underrepresented Indian languages, demonstrating ongoing innovation in the field while presenting opportunities for further research and practical applications.

The review highlights that transformer-based architectures are currently leading in performance for sentiment analysis across various benchmarks. Notably, IndicBERT and MuRIL have made significant strides in enhancing sentiment analysis for Indian languages. However, Marathi sentiment analysis is still relatively underexplored, presenting ample research opportunities. Additionally, code-mixed and multilingual sentiment analysis pose ongoing challenges. Key concerns for modern systems include explainability and computational efficiency. Large Language Models are emerging as a robust solution for zero-shot and few-shot sentiment analysis. Looking ahead, the future of sentiment analysis research is expected to focus on developing multimodal, explainable, and resource-efficient AI systems.

References

[1] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-Lingual Representation Learning at Scale. *Proceedings of ACL 2020*, 8440–8451.

[2] Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N., Bhattacharyya, A., Khapra, M. M., & Kumar, P. (2020). IndicNLPsuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. *Findings of EMNLP 2020*, 4948–4961.

[3] Khanuja, S., Dandapat, S., Kunchukuttan, A., Singh, A., Srinivasan, B., Katti, A., Goyal, P., & Bhattacharyya, P. (2021). MuRIL: Multilingual Representations for Indian Languages. *Findings of EMNLP 2021*, 4879–4893.

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT*, 4171–4186.

[5] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*.

[6] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *NeurIPS*, 5753–5763.

[7] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations. *ICLR 2020*.

[8] Clark, K., Luong, M. T., Le, Q., & Manning, C. D. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather than Generators. *ICLR 2020*.

[9] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language Models are Few-Shot Learners. *NeurIPS*, 33, 1877–1901.

[10] OpenAI. (2023). GPT-4 Technical Report. *arXiv:2303.08774.*

[11] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., et al. (2023). LLaMA: Open and Efficient Foundation Language Models. *arXiv:2302.13971*.
[12] Team Gemini. (2024). Gemini: A Family of Highly Capable Multimodal Models. *arXiv:2312.11805*.
[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. *NeurIPS*, 5998–6008.
[14] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT Networks. *EMNLP-IJCNLP*, 3982–3992.
[15] Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. *EMNLP*, 1532–1543.
[16] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ICLR Workshop*.
[17] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *TACL*, 5, 135–146.
[18] Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of Tricks for Efficient Text Classification. *EACL*, 427–431.
[19] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs Up? Sentiment Classification using Machine Learning Techniques. *EMNLP*, 79–86.
[20] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A., & Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. *ACL-HLT*, 142–150.
[21] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank. *EMNLP*, 1631–1642.
[22] Mao, X., Liu, Y., Shen, J., Wang, Y., & Wang, Z. (2024). A Comprehensive Survey of Sentiment Analysis: Methods, Applications and Challenges. *Journal of King Saud University – Computer and Information Sciences*.
[23] Hoque, M. M., Islam, M. R., & Hossain, M. A. (2024). Transformer Ensemble Based Bengali Sentiment Analysis. *Results in Engineering*.
[24] El Azzouzy, M., El Kettani, M. D., & El Alaoui, S. O. (2025). Comparative Analysis of Transformer Models for YouTube Comment Sentiment Classification. *Smart Systems and Applications*.
[25] Duru, N., Aydın, M., & Kaya, M. (2025). Comparative Evaluation of Transformer and Pre-Transformer Sentiment Analysis Models. *Entropy*, 27(12).
[26] Saidi, M., El Ouatik, S., & Oumghar, A. (2025). Advanced Transformer Architectures for Sentiment Classification. *Scientific Reports*.
[27] Kaur, H., Sharma, R., & Gupta, V. (2025). Comparative Evaluation of RoBERTa and XLM-R for Financial Sentiment Analysis. *Discover Artificial Intelligence*.
[28] Conneau, A., et al. (2020). Cross-Lingual Language Model Pretraining. *ACL 2020*.
[29] Ruder, S., Peters, M. E., Swayamdipta, S., & Wolf, T. (2019). Transfer Learning in Natural Language Processing. *NAACL Tutorial*.
[30] Wolf, T., Debut, L., Sanh, V., et al. (2020). Transformers: State-of-the-Art Natural Language Processing. *EMNLP System Demonstrations*, 38–45.
[31] Pingale, S., Ranade, H., Bhattacharya, G., & Kotecha, K. (2021). L3CubeMahaSent: A Marathi Tweet-based Sentiment Analysis Dataset. *WILDRE Workshop, EMNLP 2021*.
[32] Pingale, S., Ranade, H., Kotecha, K., & Bhattacharya, G. (2023). L3Cube-MahaSent-MD: Multi-Domain Marathi Sentiment Analysis Dataset and Benchmarking. *LREC-COLING 2024*.
[33] Akhtar, M. S., Ekbal, A., & Bhattacharyya, P. (2020). Aspect Based Sentiment Analysis in Indian Languages: A Survey. *ACM Computing Surveys*.
[34] Chakravarthi, B. R., Priyadharshini, R., Muralidaran, V., et al. (2020). Overview of Sentiment Analysis for Dravidian Languages. *FIRE 2020 Workshop Proceedings*.
[35] Patwa, P., Sharma, S., Pykl, S., et al. (2020). SemEval-2020 Task 9: Overview of Sentiment Analysis of Code-Mixed Tweets. *SemEval 2020*.
[36] Chakravarthi, B. R., Priyadharshini, R., & McCrae, J. P. (2021). Sentiment Analysis in Code-Mixed Dravidian Languages. *Expert Systems with Applications*, 184.
[37] Barman, U., Das, A., Wagner, J., & Foster, J. (2014). Code Mixing: A Challenge for Language Identification in Social Media. *First Workshop on Computational Approaches to Code Switching*.
[38] Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? *China National Conference on Chinese Computational Linguistics*.
[39] Howard, J., & Ruder, S. (2018). Universal Language Model Fine-Tuning for Text Classification. *ACL 2018*.
[40] Raffel, C., Shazeer, N., Roberts, A., et al. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *JMLR*, 21(140), 1–67.
[41] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2019). GLUE: A Multi-Task Benchmark and Analysis Platform. *ICLR 2019*.
[42] Cer, D., Yang, Y., Kong, S., et al. (2018). Universal Sentence Encoder. *EMNLP 2018*.
[43] Devlin, J., et al. (2019). BERT Fine-Tuning for Sentiment Classification Tasks. *NAACL-HLT*.
[44] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. *OpenAI Technical Report*.
[45] Peters, M. E., Neumann, M., Iyyer, M., et al. (2018). Deep Contextualized Word Representations. *NAACL-HLT*.
[46] Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *ICWSM*.
[47] Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems*, 28(2), 15–21.
[48] Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.
[49] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment Analysis Algorithms and Applications: A Survey. *Ain Shams Engineering Journal*, 5(4), 1093–1113.
[50] 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.