International Journal of
Technology and Emerging Research
ISSN: 3068-109X (online) | ISSN: 3068-1995 (print) | https://www.ioro.org/ijter

International Conference on Intelligent Technologies for Tomorrow (ICITT-2026) · 18 August 2026

Data Mining Approaches for Sentiment Analysis in Indian Regional Languages: A Review of Challenges, Limitations, and Future Research Directions

Reshmi V

Research Scholar, United College Of Arts and Science, Periyanaickenpalayam, Coimbatore

lakshmikrishnan.v@gmail.com

Abstract

People across India increasingly use regional languages online to share opinions, reviews, and reactions, but understanding the sentiment behind this text is not easy. Indian regional languages are often used in informal ways, mixed with English, written in different scripts, and supported by only a small number of labeled datasets and language tools. This review paper looks at the main data mining methods used for sentiment analysis in Indian regional languages, including machine learning, lexicon-based, rule-based, deep learning, and transformer-based approaches. It compares how these methods have been used in previous studies and highlights what they do well and where they struggle. The review shows that simple machine learning methods can still be useful when data is limited, while deep learning and transformer models are more effective when better resources are available. Even so, many problems remain, especially code-mixed text, transliteration, sarcasm, negation, and the lack of reliable benchmarks across languages. The paper concludes that future progress depends on building richer datasets, improving language-specific preprocessing, and designing models that are both accurate and practical for real-world use.

Keywords: Sentiment analysis; Indian regional languages; data mining; low-resource NLP; code-mixing.

1. Introduction

Sentiment analysis has become an important area of natural language processing because it helps identify opinions, emotions, and attitudes expressed in text. It is widely used in social media monitoring, customer feedback analysis, political forecasting, and public opinion mining. In India, this task is especially significant because digital communication increasingly occurs in regional languages rather than only in English, reflecting the country's linguistic diversity and growing internet access. However, sentiment analysis in Indian regional languages is far more difficult than in English because these languages are often low-resource, morphologically rich, and heavily affected by code-mixing, transliteration, spelling variation, and informal writing styles.

Although many sentiment analysis techniques have been developed for resource-rich languages, their direct use in Indian regional languages often leads to poor performance due to the lack of annotated corpora, sentiment lexicons, parsers, and standardized preprocessing tools. Existing studies show that traditional machine learning methods can work reasonably well on small datasets, while deep learning and transformer-based models offer better contextual understanding, but still depend on data availability and language-

specific adaptation,. As a result, there remains a clear need for a structured review of the methods, challenges, and future directions in this area.

This paper reviews the major data mining approaches used for sentiment analysis in Indian regional languages, compares their strengths and limitations, and identifies key research gaps. It also discusses the main challenges faced by current systems and highlights promising future directions for building more robust and inclusive sentiment analysis models for India's multilingual environment.

2. Review Methodology

This review was carried out to understand the recent developments and challenges in sentiment analysis for Indian regional languages. Relevant studies were collected from major academic databases, including Scopus, IEEE Xplore, SpringerLink, ScienceDirect, ACM Digital Library, and Google Scholar. The search mainly covered publications from 2018 to 2026 and used keywords such as *Indian regional language sentiment analysis*, *Indic NLP*, *low-resource language sentiment analysis*, *machine learning*, *deep learning*, and *transformer-based sentiment analysis*.

Studies were selected when they focused on sentiment analysis or related NLP methods for Indian or Indic languages and provided relevant information about their methodology and findings. Peer-reviewed journal articles, conference papers, review papers, and significant dataset or model studies were considered. Duplicate publications and studies that were not directly related to Indian-language sentiment analysis were excluded.

The selected studies were reviewed based on the languages covered, datasets used, methods adopted, major findings, and limitations reported by the authors. The findings were then organized into key themes, including **data mining approaches, linguistic challenges, resource scarcity and dataset development, emerging trends, and research gaps**. Recent studies were given greater emphasis, while important earlier studies were included to provide the necessary background and show how the field has evolved.

3. Literature Review

Sentiment analysis has emerged as a key research area within Natural Language Processing (NLP), driven by the rapid growth of user-generated content on social media platforms, online review sites, blogs, and discussion forums. As people increasingly express their opinions online, the need to automatically identify and interpret sentiment has become more important than ever. Although significant progress has been made in English sentiment analysis, applying these techniques to Indian regional languages remains a challenging task. India's rich linguistic diversity, coupled with the limited availability of annotated datasets and language-specific resources, has slowed the development of robust sentiment analysis systems for many regional languages. Consequently, researchers have focused on designing methods that address these language-specific challenges while improving sentiment classification for low-resource languages.

3.1. Evolution of Sentiment Analysis Techniques

Early research on sentiment analysis in Indian regional languages primarily relied on lexicon-based methods and traditional machine learning algorithms such as Naïve Bayes (NB), Support Vector Machines (SVM), and Decision Trees. These techniques were widely adopted because they performed reasonably well on small datasets and required relatively fewer computational resources. However, their effectiveness was often limited by the linguistic complexity of Indian languages and the scarcity of annotated training data.

Shah and Kaushik, in their review *Sentiment Analysis on Indian Indigenous Languages*, highlighted that the performance of traditional machine learning models depends heavily on the availability of annotated corpora and language-specific lexical resources. They observed that many Indian languages still lack these essential resources, making it difficult to develop accurate and reliable sentiment classification systems.

A similar perspective is presented by Kale et al. (2023) in *A Comprehensive Review of Sentiment Analysis on Indian Regional Languages*. Their survey, covering languages such as Hindi, Malayalam, Tamil, Telugu, Marathi, Bengali, Gujarati, and Urdu, shows that the field has gradually moved from feature-based machine learning methods toward deep learning techniques. However, the authors emphasize that improvements in learning algorithms alone cannot fully address the challenges posed by limited language resources.

This shift is further discussed in *A Journey of Indian Languages over Sentiment Analysis: A Systematic Review*, where the authors describe the growing adoption of deep learning architectures, including CNNs, LSTMs, and transformer-based models. These approaches have improved sentiment classification by learning contextual representations directly from text, reducing the dependence on handcrafted features. At the same time, the review points out that the success of these models largely depends on the availability of sufficient high-quality labelled data, which remains a major challenge for many Indian languages.

Recent work on low-resource languages has also demonstrated the potential of transformer-based models. In *Sentiment Analysis for Low-Resource Languages: Insights from Tamil and Tulu Using Deep Learning and Machine Learning Models*, the authors compared conventional machine learning algorithms with multilingual transformer models and found that multilingual BERT consistently achieved better performance by capturing contextual information more effectively. Nevertheless, the study also reported that issues such as limited labelled datasets and class imbalance continue to affect the performance of these advanced models.

Overall, the literature reflects a clear evolution in sentiment analysis techniques for Indian regional languages. While research has progressed from traditional machine learning approaches to deep learning and transformer-based models, most studies agree that the availability of high-quality datasets and language-specific resources remains a key factor influencing the success of sentiment analysis systems.

3.2. Linguistic Challenges in Indian Regional Languages

A common finding across the literature is that the linguistic diversity of Indian regional languages presents one of the greatest challenges for sentiment analysis. Unlike English, Indian languages differ significantly in their scripts, grammatical structures, morphology, and vocabulary. In addition, the widespread use of multiple dialects and informal writing styles on social media makes sentiment classification even more challenging.

Soman et al., in A Comparative Review of the Challenges Encountered in Sentiment Analysis of Indian Regional Language Tweets vs English Language Tweets, highlighted that social media posts in Indian languages frequently contain transliterated text, code-mixed content, and non-standard spellings. These characteristics complicate fundamental NLP tasks such as tokenization, feature extraction, and sentiment classification, making them more difficult than their English counterparts.

Similar observations were reported by Kale et al. (2023), who identified code-mixing, spelling inconsistencies, dialectal variations, and the absence of standardized writing conventions as major factors affecting classification performance across Indian regional languages. Their review suggests that these linguistic variations often reduce the effectiveness of conventional text preprocessing techniques.

Shah and Kaushik also emphasized that preprocessing methods developed for English cannot be directly applied to Indian languages because of their rich morphology and limited availability of language-specific linguistic resources. As a result, tasks such as stemming, lemmatization, and part-of-speech tagging remain challenging for many regional languages.

More recent studies have further highlighted the influence of social media language on sentiment analysis. The increasing use of emojis, abbreviations, slang, sarcasm, irony, and implicit expressions has introduced additional complexity, as these elements often convey sentiment beyond the literal meaning of words. Although transformer-based models have improved the ability to capture contextual information, several studies report that accurately interpreting such informal and context-dependent expressions remains an open research problem.

Overall, the reviewed literature indicates that linguistic diversity continues to be a major barrier to developing robust sentiment analysis systems for Indian regional languages. Most researchers agree that improving language-specific preprocessing techniques, expanding linguistic resources, and developing models capable of handling code-mixed and informal text are essential for achieving better sentiment classification performance.

3.3. Resource Scarcity and Dataset Development

A recurring theme across the reviewed literature is the lack of language resources, which continues to be one of the biggest obstacles to sentiment analysis in Indian regional languages. The performance of both traditional and modern sentiment analysis models depends heavily on the availability of high-quality datasets and supporting linguistic resources. Unfortunately, many Indian languages still have limited annotated corpora, sentiment lexicons, and other essential NLP resources.

Shah and Kaushik highlighted that the absence of annotated datasets, sentiment dictionaries, stop-word lists, and comprehensive WordNet resources significantly restricts the development of reliable supervised sentiment analysis models. Without these foundational resources, training accurate language-specific models becomes a challenging task.

Similar observations were made by Kale et al. (2023), who reported that resource scarcity remains a common issue across most Indian regional languages, despite the growing interest in this research area. Their review suggests that the lack of large, publicly available benchmark datasets continues to limit both model development and fair comparison of different approaches.

An important step toward addressing this challenge was the introduction of **IndiSentiment140** by Kumar et al. This multilingual sentiment dataset was developed to support sentiment analysis across several Indian languages and demonstrated the benefits of combining multilingual datasets with transfer learning techniques. Their work showed that knowledge learned from resource-rich languages can be effectively transferred to languages with fewer annotated samples, leading to improved sentiment classification performance.

Researchers have also explored data augmentation as a way to overcome the shortage of labelled data. For example, Pingle et al. (2023) investigated augmentation techniques such as back-translation, paraphrasing, BERT-based token replacement, and GPT-assisted text generation for Marathi sentiment analysis. Their findings indicate that these methods can improve model robustness and generalization, particularly in low-resource settings.

Overall, the literature suggests that expanding high-quality multilingual datasets, developing language-specific lexical resources, and establishing standardized benchmark datasets remain essential for advancing sentiment analysis in Indian regional languages.

3.4. Emerging Trends and Research Gaps

Recent research reflects a clear shift toward transformer-based models and multilingual learning frameworks for sentiment analysis in Indian regional languages. Compared with earlier machine learning approaches, these models are better equipped to capture contextual meaning and support knowledge transfer across related languages.

Models such as mBERT, IndicBERT, and MuRIL have become widely adopted because they learn multilingual representations that can be applied to several Indian languages simultaneously. This capability has proved particularly useful for low-resource languages, where large annotated datasets are not readily available. In addition to transformer-based approaches, researchers have begun exploring federated learning as a means of training sentiment analysis models collaboratively while preserving data privacy and making use of distributed datasets.

Another notable development is the creation of multilingual benchmark datasets, such as **IndiSentiment140**, which facilitate cross-lingual learning and provide a common platform for evaluating sentiment analysis models. These resources have contributed to more consistent performance across multiple Indian languages and have encouraged further research in multilingual sentiment analysis.

Despite these advances, the literature identifies several areas that require further investigation. Many existing datasets remain relatively small, domain-specific, and imbalanced, limiting the generalizability of trained models. The absence of standardized evaluation protocols also makes it difficult to compare the performance of different approaches across studies. Furthermore, relatively few studies have explored emerging topics such as explainable artificial intelligence (XAI), multimodal sentiment analysis, and the application of large language models (LLMs) to Indian regional languages. These areas present promising opportunities for future research.

3.5. Summary of the Literature

The literature reviewed in this study demonstrates the significant progress made in sentiment analysis for Indian regional languages over the past decade. Research has evolved from lexicon-based and traditional machine learning methods to more sophisticated deep learning and transformer-based approaches that provide improved contextual understanding and higher classification accuracy.

At the same time, the reviewed studies consistently identify several challenges that continue to affect progress, including limited annotated datasets, resource scarcity, code-mixing, transliteration, and the linguistic diversity of Indian languages. Recent developments in multilingual transformers, transfer learning, data augmentation, and multilingual dataset construction have helped address some of these issues, particularly for low-resource languages. However, the literature also makes it clear that further work is needed to develop larger benchmark datasets, richer linguistic resources, and standardized evaluation frameworks. This review brings together these findings to provide a comprehensive understanding of the current state of research while highlighting key directions for future investigation.

4. Data Mining Approaches for Sentiment Analysis in Indian Regional Languages

Researchers have investigated a wide range of data mining techniques to improve sentiment analysis in Indian regional languages. The selection of an appropriate approach often depends on factors such as the availability of annotated datasets, language-specific resources, and the linguistic characteristics of the target language. Earlier studies mainly focused on machine learning and lexicon-based methods because they were well suited to low-resource settings. As larger datasets and improved computational resources have become available, research has gradually shifted toward deep learning and transformer-based models, which offer better contextual understanding and have shown promising results across several Indian languages.

4.1. Machine Learning Approaches

Machine learning has been one of the most commonly adopted approaches for sentiment analysis in Indian regional languages, particularly in situations where annotated datasets are limited. Algorithms such as Naïve Bayes (NB), Support Vector Machines (SVM), Logistic Regression (LR), Decision Trees, and k-Nearest Neighbours (KNN) have been successfully applied to languages including Hindi, Malayalam, Tamil, Telugu, and Marathi. Their popularity stems from their relatively simple implementation and their ability to produce reliable results even with modest amounts of training data.

Shah and Kaushik, in their review of sentiment analysis for Indian indigenous languages, noted that these algorithms rely heavily on manually engineered features such as Bag-of-Words, n-grams, and TF-IDF representations. Likewise, Kale et al. (2023) observed that SVM has been one of the most widely used classifiers because of its effectiveness in handling high-dimensional textual data. However, both studies emphasized that the performance of these models is closely tied to the quality of feature engineering and the effectiveness of language-specific preprocessing techniques.

Despite their widespread use, traditional machine learning models have several limitations. While they are computationally efficient and suitable for small datasets, they often fail to capture the contextual meaning of words and phrases. This limitation becomes more apparent when analysing social media content, where sarcasm, negation, code-mixing, and informal expressions are common. As a result, many recent studies suggest that although machine learning methods continue to provide a strong baseline, more advanced models are better suited to handling the linguistic complexity of Indian regional languages.

4.2. Lexicon-Based Approaches

Lexicon-based approaches have been widely explored for sentiment analysis in Indian regional languages, especially in cases where annotated training data are limited or unavailable. Unlike supervised machine learning methods, these techniques determine sentiment by using predefined sentiment lexicons or opinion dictionaries, making them a practical option for low-resource languages.

Shah and Kaushik pointed out that lexicon-based methods are particularly useful when labelled datasets are scarce, as they can classify sentiment without extensive training data. However, they also emphasized that the performance of these approaches depends heavily on the availability and quality of language-specific lexical resources. Similarly, Kale et al. (2023) and Yadav et al. (2024) reported that many Indian regional languages still lack comprehensive sentiment lexicons, domain-specific vocabularies, and well-developed WordNet resources. This limitation often reduces the ability of lexicon-based systems to accurately interpret context-dependent expressions, idioms, and newly emerging words commonly found in social media.

Recognizing these challenges, recent studies have suggested that lexicon-based methods are more effective when integrated with machine learning or deep learning models. Such hybrid approaches combine the

linguistic knowledge provided by sentiment lexicons with the contextual learning capability of data-driven models, leading to improved sentiment classification, particularly in low-resource settings.

4.3. Rule-Based Approaches

Rule-based approaches were among the first methods used for sentiment analysis in Indian regional languages. These systems determine sentiment by applying manually defined linguistic rules, such as grammatical structures, negation patterns, and predefined sentiment expressions. Because their decision-making process is based on explicit rules, they are relatively easy to understand and interpret.

Several review studies recognize the interpretability of rule-based systems as one of their main strengths. However, they also point out the practical challenges involved in developing and maintaining these systems. Kale et al. (2023) observed that the grammatical diversity, rich morphology, and regional variations found across Indian languages make it difficult to create a single set of rules that can be applied universally. As a result, separate rule sets often need to be developed for individual languages, making the process both time-consuming and labour-intensive.

For these reasons, recent studies rarely use rule-based methods as standalone solutions. Instead, they are often combined with machine learning or deep learning techniques, where linguistic rules complement data-driven models and help improve performance in specific domains or language-specific applications.

4.4. Deep Learning Approaches

With the increasing availability of computational resources and larger datasets, deep learning has become a popular approach for sentiment analysis in Indian regional languages. Unlike traditional machine learning methods, deep learning models can automatically learn meaningful semantic and contextual features from text, reducing the reliance on manual feature engineering and improving the overall classification process.

The systematic review *A Journey of Indian Languages over Sentiment Analysis* reported that deep learning architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Gated Recurrent Units (GRUs) consistently achieve better performance than conventional machine learning models across several Indian languages. Similar observations were made in *Sentiment Analysis for Low-Resource Languages: Insights from Tamil and Tulu Using Deep Learning and Machine Learning Models*, where deep learning models demonstrated higher classification accuracy by effectively capturing contextual relationships within the text.

Although these models have significantly improved sentiment classification, the literature also highlights their limitations. Most deep learning techniques require large volumes of labelled training data to perform well. Since many Indian regional languages still lack sufficient annotated datasets, the advantages of deep learning cannot always be fully realized. This has led researchers to emphasize the need for developing larger, high-quality corpora and better language resources to support deep learning-based sentiment analysis.

4.5. Transformer-Based Approaches

Transformer-based models have emerged as the most promising approach for sentiment analysis in Indian regional languages and have become a major focus of recent research. Models such as multilingual BERT (mBERT), IndicBERT, MuRIL, and XLM-R have shown remarkable improvements in sentiment classification by learning rich contextual representations from multilingual text. Their ability to transfer knowledge across related languages makes them particularly suitable for the linguistically diverse and low-resource nature of many Indian languages.

Kale et al. (2023) and Yadav et al. (2024) reported that transformer-based models consistently achieve better performance than both traditional machine learning algorithms and earlier deep learning models on a variety of sentiment analysis tasks. Similar findings were presented in studies involving Tamil and Tulu, where multilingual transformers demonstrated superior accuracy by leveraging knowledge learned from resource-rich languages and applying it to languages with limited training data.

Despite these encouraging results, the literature also highlights a few practical challenges. Transformer models require considerable computational resources for training and fine-tuning, and their performance is still influenced by the availability of high-quality annotated datasets. To address these limitations, recent research has increasingly focused on parameter-efficient fine-tuning techniques, multilingual transfer learning, and lightweight transformer architectures that aim to deliver high performance while reducing computational cost.

Table 1. Comparison of Data Mining Approaches Used in Indian Regional Language Sentiment Analysis

Approach	Representative Algorithms	Advantages	Limitations	Suitable Scenarios
Machine Learning	SVM, Naïve Bayes, Logistic Regression, KNN	Simple, computationally efficient, performs well on small datasets	Requires manual feature engineering; poor contextual understanding	Small annotated datasets
Lexicon-Based	Sentiment Lexicons, WordNet	No labelled data required; interpretable	Depends on high-quality lexicons; weak contextual understanding	Low-resource languages
Rule-Based	Grammar Rules, Negation Rules	Transparent and explainable	Labour-intensive; difficult to scale across languages	Domain-specific applications
Deep Learning	CNN, RNN, LSTM, GRU	Learns semantic features automatically; higher accuracy	Requires large annotated datasets and computational resources	Medium to large datasets
Transformer-Based	mBERT, IndicBERT, MuRIL, XLM-R	State-of-the-art performance; captures contextual s emantics; supports multilingual transfer learning		

5. Discussion and Future Research Directions

The studies reviewed in this paper show that sentiment analysis in Indian regional languages has developed considerably over the years. The field has gradually moved from traditional machine learning methods to deep learning and, more recently, transformer-based models. However, this progress has not been equal across all Indian languages. Languages with larger datasets and better-developed linguistic resources have generally benefited more from these advances, while low-resource languages continue to face basic challenges.

Traditional methods such as SVM, Naïve Bayes, and Logistic Regression are still useful when only small datasets are available. They are relatively simple, require less computational power, and can provide reasonable results. Their main weakness, however, is their dependence on manually selected features. This makes them less effective when sentiment depends on context, as is often the case with sarcasm, negation, code-mixing, transliteration, and informal social media language.

Deep learning models such as CNNs, LSTMs, and GRUs have helped overcome some of these limitations by learning useful features directly from text. Even so, their performance is closely tied to the amount and quality of labelled data available for training. This creates a continuing problem for Indian languages: the languages that need better computational resources often have the least amount of training data.

Transformer-based models such as mBERT, IndicBERT, MuRIL, and XLM-R have brought another important improvement. Their ability to understand context and transfer knowledge across languages makes them

particularly useful for low-resource settings. At the same time, they are not a complete solution. They can be computationally expensive and may still perform poorly when the available data are small, imbalanced, or poorly annotated. Future research should therefore look not only at improving model accuracy but also at making these models more practical through lightweight architectures, efficient fine-tuning, and other resource-saving techniques.

One of the strongest conclusions emerging from the literature is that **better models alone cannot solve the problem**. The quality of the dataset plays an equally important role. Future research should give greater attention to developing larger and better-balanced datasets for Indian languages. These datasets should represent the way people actually communicate online, including code-mixed sentences, transliterated words, spelling variations, emojis, and informal expressions. Consistent annotation guidelines would also help improve the reliability of these resources.

Code-mixing and transliteration deserve particular attention. They are not unusual exceptions in Indian social media; they are part of everyday digital communication. Future sentiment analysis systems should therefore be designed to handle them naturally rather than simply removing them during preprocessing. Better transliteration normalization, word-level language identification, and multilingual representation learning could help address this issue.

Another area that needs more attention is the interpretation of sarcasm, irony, and implicit sentiment. A sentence may appear positive on the surface while expressing a negative opinion in context. Even advanced transformer models can struggle with such cases. Building datasets specifically designed to capture these forms of communication and incorporating conversational or cultural context could be useful directions for future research.

The literature also suggests that future evaluation should go beyond reporting accuracy or F1-score. A model that performs well on a carefully prepared dataset may not necessarily work equally well on real-world social media data. Future studies should therefore examine robustness, explainability, fairness, and performance across different domains such as education, healthcare, tourism, public services, and social media. Multimodal sentiment analysis and the use of large language models also deserve further investigation, particularly for handling the combination of text, emojis, images, and other forms of online expression.

6. Conclusion

Sentiment analysis in Indian regional languages has attracted growing attention in recent years, reflecting the increasing need to understand opinions expressed in diverse Indian languages across digital platforms. The literature reviewed in this study shows a clear progression from traditional machine learning techniques to deep learning and, more recently, transformer-based models, each contributing to improved sentiment classification performance.

Despite these advancements, several challenges continue to hinder the development of robust sentiment analysis systems. Limited annotated datasets, code-mixed and transliterated text, linguistic diversity, and the lack of comprehensive language resources remain common issues across most Indian regional languages. These challenges are particularly evident in low-resource languages, where the effectiveness of even advanced models is constrained by insufficient training data.

Overall, the reviewed studies suggest that future research should focus on developing larger multilingual datasets, improving language-specific preprocessing techniques, and designing resource-efficient transformer models. Strengthening collaboration between researchers in natural language processing, data

mining, and linguistics will also play an important role in building more accurate, scalable, and inclusive sentiment analysis systems for the diverse linguistic landscape of India.

References

[1] S. D. Kale, R. Prasad, G. P. Potdar, P. N. Mahalle, D. T. Mane, and G. D. Upadhye, "A Comprehensive Review of Sentiment Analysis on Indian Regional Languages: Techniques, Challenges, and Trends," International Journal on Recent and Innovation Trends in Computing and Communication, vol. 11, no. 9s, pp. 93–110, 2023, doi:10.17762/ijritcc.v11i9s.7401.

[2] S. R. Shah and A. Kaushik, "Sentiment Analysis on Indian Indigenous Languages: A Review on Multilingual Opinion Mining," arXiv preprint arXiv:1911.12848, 2019.

[3] S. Kumar, R. Sanasam, and S. Nandi, "IndiSentiment140: Sentiment Analysis Dataset for Indian Languages with Emphasis on Low-Resource Languages using Machine Translation," in Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), Mexico City, Mexico, 2024, pp. 7689–7698, doi:10.18653/v1/2024.naacl-long.425.

[4] S. Khanuja, D. Bansal, S. Mehtani, et al., "MuRIL: Multilingual Representations for Indian Languages," arXiv preprint arXiv:2103.10730, 2021.

[5] B. G. Patra, D. Das, and A. Das, "Sentiment Analysis of Code-Mixed Indian Languages: An Overview of SAIL_Code-Mixed Shared Task @ ICON-2017," arXiv preprint arXiv:1803.06745, 2018.

[6] S. D. Kale, "Sentiment Analysis on Indian Regional Languages: A Comprehensive Review," International Journal of Computer Sciences and Engineering, vol. 7, no. 1, pp. 966–974, 2019, doi:10.26438/ijcse/v7i1.966974.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of NAACL-HLT, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[8] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), 2017.

[9] Y. Goldberg, "A Primer on Neural Network Models for Natural Language Processing," Journal of Artificial Intelligence Research, vol. 57, pp. 345–420, 2016.

[10] S. D. Kale, R. Prasad, G. P. Potdar, P. N. Mahalle, D. T. Mane, and G. D. Upadhye, "Recent Trends and Future Directions in Sentiment Analysis for Indian Regional Languages," 2023

Appendix A. Supplementary Material

This appendix provides supplementary information supporting the review of sentiment analysis in Indian regional languages. It includes the scope of the review, the languages covered, and a summary of the principal data mining approaches discussed in the article.

A.1 Scope of the Review

The review focuses on sentiment analysis techniques developed for Indian regional languages, with particular emphasis on data mining approaches, linguistic challenges, resource limitations, and future research directions. The literature considered spans traditional machine learning methods, lexicon-based and rule-based techniques, deep learning models, and transformer-based architectures. The review primarily includes studies published between 2019 and 2025, together with a few foundational works that have significantly influenced research in multilingual sentiment analysis.

A.2 Indian Regional Languages Covered in the Review

The reviewed studies address sentiment analysis in several Indian languages, including:

- Hindi
- Malayalam

- Tamil
- Telugu
- Kannada
- Marathi
- Bengali
- Gujarati
- Punjabi
- Urdu
- Tulu
- Other low-resource Indic languages

Although the amount of research varies across these languages, the reviewed literature consistently identifies common challenges such as limited annotated datasets, code-mixed text, transliteration, dialectal variation, and insufficient language resources.

A.3 Summary of Data Mining Approaches

Table A.1 summarizes the major data mining approaches discussed in this review.

Table A.1. Overview of Data Mining Approaches for Sentiment Analysis

Approach	Representative Methods	Strengths	Major Limitations
Machine Learning	Naïve Bayes, SVM, Logistic Regression, KNN	Simple, efficient, suitable for small datasets	Requires manual feature engineering and struggles with contextual understanding
Lexicon-Based	Sentiment dictionaries, Opinion lexicons	Useful for low-resource languages and interpretable	Depends on high-quality lexical resources and cannot effectively capture contextual meaning
Rule-Based	Grammar rules, Negation rules	Transparent and easy to interpret	Difficult to develop and maintain across multiple languages
Deep Learning	CNN, RNN, LSTM, GRU	Learns semantic features automatically and improves classification accuracy	Requires large annotated datasets and high computational resources
Transformer-Based	mBERT, IndicBERT, MuRIL, XLM-R	Captures contextual semantics and supports multilingual transfer learning	Computationally intensive and dependent on quality training data