Library Progress International
Vol.44 No.3, July-December 2024: P.8139-8145

Print version ISSN 0970 1052
Online version ISSN 2320 317X

Original Article

Available online at www.bpasjournals.com

A Comparative Analysis of Machine Learning Algorithms for Sentiment Analysis in Indian Social Media

Abhishek Gandhar[1], Shashi Gandhar[2]*, S B kumar [3] Arvind Rehalia [4] Prakhar Priyadarshi [5], Mohit Tiwari [6]

[1]Professor, Bharati Vidyapeeth's College of Engineering, New Delhi, India
[2]Associate Professor, Bharati Vidyapeeth's College of Engineering, New Delhi, India
[3]Associate Professor, Bharati Vidyapeeth's College of Engineering, New Delhi, India
[4]Associate Professor, Bharati Vidyapeeth's College of Engineering, New Delhi, India
[5]Professor, Bharati Vidyapeeth's College of Engineering, New Delhi, India
[6]Assistant Professor, Bharati Vidyapeeth's College of Engineering, New Delhi, India

How to cite this article: Abhishek Gandhar, Shashi Gandhar, S B kumar, Arvind Rehalia, Prakhar Priyadarshi , Mohit Tiwari (2024). A Comparative Analysis of Machine Learning Algorithms for Sentiment Analysis in Indian Social Media. *Library Progress International,* 44(3), 8139-8145.

ABSTRACT

This research paper presents a comprehensive comparative analysis of various machine learning algorithms for sentiment analysis in the diverse and multilingual context of Indian social media. The primary objective of the study was to evaluate and compare the effectiveness of different machine learning models, with a specific focus on the Support Vector Machine (SVM) algorithm, in accurately deciphering sentiments expressed in Indian languages on social media platforms, predominantly Twitter. The methodology employed a structured approach to data collection and analysis, extracting around 100,000 tweets in multiple Indian languages using the Twitter API. The SVM algorithm was then applied to this data, and its performance was assessed based on accuracy, efficiency, and adaptability in handling multilingual content.

Key findings revealed that while SVM shows reasonable accuracy in sentiment analysis, its performance varies across different languages, with English tweets exhibiting the highest accuracy. The algorithm also faced challenges in processing mixed-language tweets, indicating a need for more advanced, linguistically versatile models. Comparative analysis with other machine learning and deep learning models suggested potential for improved performance with models like BERT, especially in linguistic and contextual understanding.

The implications of these findings are significant, highlighting the necessity for more culturally and linguistically nuanced sentiment analysis tools in India's diverse digital landscape. This research contributes to the field by providing insights into the selection of appropriate algorithms for sentiment analysis, particularly in linguistically diverse settings, and underscores the need for continuous innovation in machine learning applications for social media analysis.

Keywords: Sentiment Analysis, Machine Learning, Support Vector Machine, Multilingual Natural Language Processing, Algorithm Comparison

1.1 **1. Introduction**

The burgeoning interest in sentiment analysis within Indian social media platforms presents a unique challenge and opportunity for machine learning applications. Sentiment analysis, at its core, is the computational study of opinions, sentiments, subjectivity, and emotions expressed in text. The significance of sentiment analysis in India, particularly within the context of its diverse social media landscape, stems from the country's vast, multilingual population and the increasing digital penetration.

The utilization of machine learning algorithms for sentiment analysis in Indian social media is not just a technical challenge but also a cultural and linguistic one. Different machine learning algorithms, ranging from traditional models like Support Vector Machines (SVM) to advanced deep learning techniques, have shown varied effectiveness in deciphering the nuances of sentiment in Indian languages and contexts. For instance, The application of SVM classifiers in sentiment analysis for social media, highlighting the potential of such methods in handling large and complex datasets[1].

Furthermore, the analysis of sentiment on platforms like Twitter has gained prominence[2].. These studies underscore the relevance of machine learning in passing through vast amounts of unstructured data to extract meaningful sentiment-oriented insights.

The presented research takes a comprehensive look at machine learning-based sentiment analysis for social media platforms, indicating the broad scope and applicability of these techniques in the Indian context[3]..

In summary, the exploration of machine learning algorithms for sentiment analysis in Indian social media is pivotal for understanding public opinion and consumer behaviour in one of the world's largest and most dynamic digital arenas. This research, therefore, stands at the intersection of technology, linguistics, and cultural studies, offering insights into the digital pulse of India.

2. Literature Review

2.1 Relevant Scholarly Works

The field of sentiment analysis using machine learning in Indian social media has been extensively explored in recent years. This literature review highlights some of the most significant works in this domain.

- Authors provided a comprehensive study on machine learning-based sentiment analysis for social media platforms. Their work emphasized the effectiveness of various machine learning models in interpreting the sentiment of social media content, a crucial step in understanding public opinion in digital spaces[3]..

- A sentiment analysis study of human thoughts using machine learning techniques. Their research is pivotal in understanding how machine learning algorithms can be adapted to analyze and interpret human emotions and opinions expressed on social media[4]..

- A noteworthy benchmark evaluation of machine learning algorithms for sentiment analysis. This study is critical for understanding the comparative effectiveness of different machine learning models in sentiment analysis tasks[5]..

- The comparison of machine learning algorithms for Twitter sentiment analysis provides insights into how different algorithms perform in analyzing sentiments on a specific platform like Twitter, which is highly relevant in the Indian context due to its growing user base[6]..

- The sentiment analysis of “Hepatitis of Unknown Origin” on social media using machine learning, studied by researchers, showcases the application of sentiment analysis in healthcare and public health monitoring, demonstrating the versatility of machine learning algorithms in various thematic areas[7]..

- The primary contribution of this work is the performance assessment of different machine learning classifiers utilizing our proposed feature set, which is created by combining the bag-of-words approach with term frequency-inverse document frequency (TF-IDF) [8]..

- This study showcases the ability of machine learning models to predict the number of future COVID-19 patients. It employs four commonly used forecasting models—linear regression (LR), least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), and exponential smoothing (ES)—to predict the critical factors related to COVID-19[9].

- This paper presents a performance comparison of various machine learning algorithms, including Support Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB), and k-Nearest Neighbours (k-NN), applied to the Wisconsin Breast Cancer (original) dataset[10].

These studies collectively represent the development and application of machine learning in sentiment analysis, particularly in the context of Indian social media. They highlight the diversity of approaches and the breadth of applications, from general sentiment analysis to specific use cases like public health monitoring and platform-specific analyses.

Despite the extensive research in machine learning for sentiment analysis on Indian social media, a notable gap exists in the comparative analysis of these algorithms' effectiveness across diverse linguistic and cultural contexts within India. Most existing studies focus on specific algorithms or platforms but do not comprehensively compare the performance of various machine learning models across the multifaceted Indian social media landscape, which is characterized by multiple languages and cultural nuances. Addressing this gap is crucial for developing more accurate and culturally sensitive sentiment analysis tools, which can provide deeper insights into consumer behaviour, public opinion, and social trends in India's diverse digital ecosystem. This research aims to fill this gap by providing a comparative analysis of different machine learning algorithms, thereby contributing to more effective and nuanced sentiment analysis in the Indian context.

3. Outlined approach to the Methodology

The methodology for this research involves a structured approach to collecting and analysing data to compare the effectiveness of various machine learning algorithms in sentiment analysis within Indian social media.

Research Design:

- **Objective**: To comparatively analyse the performance of different machine learning algorithms for sentiment analysis on Indian social media.
- **Data Collection**: Extracting social media posts and comments in multiple Indian languages.
- **Data Analysis**: Applying a specific machine learning algorithm to assess sentiment.
- **Evaluation Criteria**: Accuracy, efficiency, and adaptability of algorithms in handling multilingual data.

[Table 1]

Data Source and Analysis:

Aspect	Details
Data Source	Twitter (Indian context)
Data Type	Social media posts and comments
Language	Multiple Indian languages (Hindi, English, Bengali, etc.)
Volume	Approximately 100,000 tweets
Time Frame	Last six months
Data Collection Method	Twitter API for real-time data extraction
Machine Learning Tool	Python (libraries like NLTK, scikit-learn, TensorFlow)
Data Analysis Tool	Support Vector Machine (SVM) algorithm
Data Pre-processing	Tokenization, stemming, and removal of stopwords
Sentiment Classification	Positive, negative, neutral
Evaluation Metric	F1 Score, Precision, Recall

The chosen method focuses on the Support Vector Machine (SVM) algorithm for its known efficacy in text classification tasks. By applying this algorithm to a large and diverse dataset from Twitter, the study aims to provide insights into the algorithm's capability to accurately gauge sentiments across different Indian languages and cultural contexts.

This methodology is designed to assess the effectiveness of SVM in the dynamic and linguistically diverse environment of Indian social media, providing a focused approach to understanding the nuances of sentiment analysis in this context.

4. Result and Analysis:

This section outlines the results obtained from applying the Support Vector Machine (SVM) algorithm for sentiment analysis on Indian social media data, specifically Twitter. The results are presented in a series of tables, each followed by an explanation.

[Table 2]
Overall Accuracy of SVM Algorithm

Language	**Accuracy (%)**
Hindi	78%
English	82%
Bengali	74%
Tamil	77%
Telugu	75%

shows the accuracy of the SVM algorithm in classifying sentiments in different Indian languages. English tweets showed the highest accuracy, possibly due to a more extensive dataset and better-developed natural language processing tools.

[Table 3]
Precision, Recall, and F1-Score for Hindi Sentiment Analysis

Metric	**Score**
Precision	0.79
Recall	0.76
F1-Score	0.77

The table 2 provides detailed metrics for sentiment analysis in Hindi. The balance between precision and recall suggests a relatively even performance of the SVM in identifying true positives and minimizing false negatives in table 3.

[Table 4]
Sentiment Distribution in English Tweets

Sentiment	**Percentage**
Positive	40%
Negative	35%
Neutral	25%

[Table 4] shows **t**he distribution of sentiments in English tweets is fairly balanced, with a slight leaning towards positive sentiments. This could indicate general trends in public opinion or the nature of the dataset.

[Table 5]
Comparison of SVM with Other Algorithms in Bengali

Algorithm	**Accuracy (%)**
SVM	74%
Naïve Bayes	68%
Random Forest	70%

[Table 5] presents a comparison with other machine learning algorithms, SVM shows superior performance in analyzing Bengali tweets, highlighting its effectiveness in this linguistic context.

[Table 6]
Efficiency Analysis - Time Taken for Analysis

Language	**Time (seconds)**
Hindi	120
English	110
Bengali	130

[Table 6] This table shows the time efficiency of the SVM algorithm in processing and analyzing tweets in different languages. English tweets were processed the fastest, which may be due to better optimization of tools for English language processing.

[Table 7]
Error Analysis - Common Misclassifications

Sentiment (Actual → Predicted)	Count
Positive → Negative	150
Negative → Positive	130
Neutral → Positive	110

[Table 7] provides an insight into the common misclassifications made by the SVM algorithm. The majority of errors were in wrongly classifying positive sentiments as negative, which could indicate a bias in the algorithm or the complexity of understanding contextual nuances.

[Table 8]
Sentiment Classification in Mixed Language Tweets

Language Combination	Accuracy (%)
Hindi-English	72%
Bengali-English	70%

[Table 8] shows the accuracy in classifying sentiments in mixed-language tweets is slightly lower, reflecting the increased complexity in processing tweets with code-switching, a common phenomenon in Indian social media.

[Table 9]
Sentiment Classification Accuracy by Region

Region	Hindi (%)	English (%)	Bengali (%)	Tamil (%)	Telugu (%)
Northern India	79	81	-	-	-
Southern India	-	80	-	76	75
Eastern India	-	83	75	-	-
Western India	78	82	-	-	-
Central India	77	80	-	-	-

[Table 9] This table illustrates the accuracy of sentiment analysis across different regions of India. The variation in accuracy can be attributed to regional linguistic nuances and the prevalence of certain languages in specific areas.

[Table 10]
Algorithm Performance by Tweet Length

Tweet Length	Accuracy (%)
Short (< 50 chars)	80
Medium (50-100 chars)	78
Long (> 100 chars)	76

[Table 10] The table 10 shows how tweet length affects the accuracy of the SVM algorithm. Shorter tweets tend to have higher accuracy, possibly due to less complexity and clearer sentiment expression.

[Table 11]
Comparison of SVM with Deep Learning Models

Algorithm	Accuracy (%)
SVM	78
CNN	80
LSTM	82
BERT	85

[Table 11] This table compares the SVM algorithm with various deep learning models like CNN, LSTM, and BERT. Deep learning models, especially BERT, show higher accuracy, indicating their advanced capability in understanding context and nuances in language.

[Table 12]
Sentiment Analysis in Mixed Sentiment Tweets

Sentiment Combination	Accuracy (%)
Positive-Negative	70
Positive-Neutral	72
Negative-Neutral	74

[Table 12] presents the accuracy of sentiment classification in tweets with mixed sentiments. The algorithm shows varied performance, with the highest accuracy in distinguishing between negative and neutral sentiments.

[Table 13]
Impact of Pre-processing on Accuracy

Pre-processing Step	Accuracy Improvement (%)
Tokenization	+2
Stemming	+1
Stopword Removal	+3

[Table 13] The table 13 highlights the impact of different pre-processing steps on the accuracy of the SVM algorithm. Stop word removal contributes the most to accuracy improvement, emphasizing the importance of pre-processing in text analysis.

[Table 14]
Algorithm Performance across Different Topics

Topic	Accuracy (%)
Politics	75
Entertainment	79
Sports	77
Technology	78
Health	76

[Table 14]: This table shows the SVM algorithm's performance in sentiment analysis across different topics. The highest accuracy is observed in entertainment-related tweets, possibly due to more explicit expressions of sentiment.

5. Result Discussion

The analysis and interpretation of the results obtained in Section 4 offer insightful revelations on the use of Support Vector Machine (SVM) algorithms for sentiment analysis in Indian social media. These findings significantly contribute to addressing the identified literature gap, specifically the need for a comparative analysis of machine learning algorithms across diverse linguistic and cultural contexts within India.

5.1 Analysis and Interpretation of Results

5.1.1 Accuracy across Languages: The varying levels of accuracy in different languages (Table 1) highlight the challenges in sentiment analysis across India's linguistic diversity. The highest accuracy in English suggests better-developed tools and resources for English, while the lower accuracy in regional languages like Bengali and Tamil indicates the need for more refined algorithms tailored to these languages.

5.1.2 Performance in Mixed-Language Tweets: The reduced accuracy in mixed-language tweets (Table 7) underscores the complexity of sentiment analysis in code-switched language contexts, common in Indian social media. This emphasizes the need for algorithms that can better handle linguistic nuances and mixed-language data.

5.1.3. Algorithm Comparison: The comparison of SVM with other machine learning and deep learning algorithms (Tables 4 and 10) suggests that while SVM performs adequately, there is potential for improved accuracy with advanced deep learning models like BERT. This comparison is crucial in determining the most effective algorithm for sentiment analysis in the Indian context.

5.1.4. Impact of Tweet Characteristics: The influence of tweet length (Table 9) and topic (Table 13) on accuracy points towards the contextual dependency of sentiment analysis. Shorter tweets and certain topics like entertainment yield higher accuracy, suggesting that sentiment expression varies with the content's nature and length.

The comprehensive analysis across multiple Indian languages and contexts, as well as the comparison with various algorithms, directly addresses the literature gap. This research provides a clearer understanding of how different machine learning algorithms, especially SVM, perform in the multifaceted environment of Indian social media. It highlights the need for algorithms that are not only linguistically competent but also culturally sensitive to effectively interpret sentiments in a country as diverse as India.

5.2 Implications and Significance:

5.2.1 Enhanced Cultural Relevance: The findings advocate for the development of more culturally and linguistically nuanced sentiment analysis tools, which can significantly improve the accuracy of consumer behavior analysis, public opinion mining, and social trend analysis in India.

5.2.2 Guidance for Algorithm Selection: This research serves as a guide for practitioners and researchers in selecting appropriate algorithms for sentiment analysis, especially in linguistically diverse settings.

5.2.3 Potential for Cross-Disciplinary Applications: The insights from this study can be applied in various domains, including marketing, political science, public health, and social research, where understanding public sentiment is crucial.

In summary, this study not only fills a critical gap in existing literature by providing a comparative analysis of machine learning algorithms for sentiment analysis in Indian social media but also lays the groundwork for future research and development in this field.

6. Conclusion

The study embarked on an insightful journey to explore the effectiveness of machine learning algorithms, particularly the Support Vector Machine (SVM), in the realm of sentiment analysis within the diverse and multilingual landscape of Indian social media. The findings revealed that while the SVM algorithm demonstrates reasonable accuracy in sentiment classification, its performance varies significantly across different Indian languages, with English tweets showing the highest accuracy. This variation underscores the challenges inherent in sentiment analysis in a linguistically diverse environment like India.

One of the key revelations of the study was the algorithm's reduced accuracy in handling mixed-language tweets, a common feature in Indian social media. This points towards the need for more advanced algorithms that can effectively navigate the complexities of code-switching. Moreover, the comparison of SVM with other machine learning and deep learning models highlighted that while SVM is a competent tool for sentiment analysis, there are opportunities for enhanced performance with more sophisticated models like BERT, particularly in terms of contextual and linguistic understanding.

The implications of these findings are far-reaching. Firstly, they call for the development of more advanced sentiment analysis tools that are not only linguistically adept but also culturally attuned, especially for a country as diverse as India. This is crucial for accurately gauging public opinion, consumer behavior, and social trends. Secondly, the study provides valuable insights for researchers and practitioners in selecting the most suitable algorithms for sentiment analysis tasks, particularly in settings marked by linguistic and cultural diversity.

Furthermore, the study's findings have significant applications across various fields, including marketing, political analysis, public health, and social research. In these domains, understanding public sentiment is key, and the enhanced accuracy in sentiment analysis tools, as suggested by the study, can lead to more informed decision-making and strategy development.

In conclusion, this research contributes substantially to the field of sentiment analysis in Indian social media, bridging a critical gap in existing literature and paving the way for future advancements in machine learning applications for sentiment analysis. The study not only enhances our understanding of the performance of different algorithms in varied linguistic contexts but also highlights the need for continuous innovation in this rapidly evolving field.

7. References:

1. Huang, Q. Sentiment analysis for social media using SVM classifier of machine learning. *Applied and Computational Engineering*.2023; 4:86-90 DOI: 10.54254/2755-2721/4/20230354
2. Jagan, L., et al. (2023). Twitter Sentiment Analysis using Machine Learning. *International journal of scientific research in science, engineering and technology*. 2023; Volume 10(2), 665-669, DOI: 10.32628/ijsrset2310281
3. P. Upadhyay, S. Saifi, R. Rani, A. Sharma and P. Bansal, "Machine Learning-Based Sentiment Analysis for the Social Media Platforms," *2023 6th International Conference on Information Systems and Computer Networks (ISCON)*, Mathura, India, 2023, pp. 1-5, doi: 10.1109/ISCON57294.2023.10112120.
4. Singh and G. Sharma, "Sentiment Analysis Study of Human Thoughts using Machine Learning Techniques," *2023 International Conference on Disruptive Technologies (ICDT)*, Greater Noida, India, 2023, pp. 776-785, doi: 10.1109/ICDT57929.2023.10150917.
5. Anuradha Vishwajit Yenkikar, C. Narendra Babu. SentiML Benchmark Evaluation of Machine Learning Algorithms for Sentiment Analysis. *Indonesian Journal of Electrical Engineering and Informatics*. Vol 11, No 1 DOI: 10.52549/ijeei.v11i1.4381
6. Parikh, A., et al. Comparison of Machine Learning Algorithms for Twitter Sentiment Analysis. *2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, Trichy, India, 2022, pp. 209-215, DOI: 10.1109/ICAISS55157.2022.10010781
7. N. Agustina, H. Gusdevi, D. Wijayati, I. Ismawati and C. N. Ihsan, "Sentiment Analysis of “Hepatitis of Unknown Origin” on Social Media Using Machine Learning," *2022 Seventh International Conference on Informatics and Computing (ICIC)*, Denpasar, Bali, Indonesia, 2022, pp. 01-06, doi: 10.1109/ICIC56845.2022.10006985.
8. Rustam F, Khalid M, Aslam W, Rupapara V, Mehmood A, Choi GS. A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. *PLoS ONE* 16(2): 2021, e0245909. https://doi.org/10.1371/journal.pone.0245909.
9. F. Rustam *et al*. COVID-19 Future Forecasting Using Supervised Machine Learning Models, *IEEE Access*, vol. 8, pp. 101489-101499, 2020, DOI: 10.1109/ACCESS.2020.2997311
10. Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, Thomas Noel, Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis, *Procedia Computer Science*, Volume 83,2016, 1064-1069, https://doi.org/10.1016/j.procs.2016.04.224