Airo International Journal
Peer-Reviewed
Multidisciplinary

ISSN: 2320-3714
Volume:2 Issue:1
April 2026
Impact Factor: 10.1
Subject: IT

SENTIMENT ANALYSIS OF SOCIAL MEDIA DATA USING NLP TECHNIQUE

Subhnandani Vishwakarma / Ambuj Yadav
Student – CDOE, Mumbai University
Email – vishwakarmasubhnandani@gmail.com / ambujyadav098765@gmail.com

Abstract

Social media platforms generate a massive volume of unstructured textual data where users express opinions, emotions, and feedback on various topics. Extracting meaningful insights from this data has become crucial for organizations, policymakers, and researchers. Sentiment Analysis, a key application of Natural Language Processing (NLP), focuses on identifying the sentiment polarity of textual content. This research paper presents a sentiment analysis system for social media data using NLP techniques and machine learning algorithms. The proposed approach involves data preprocessing, feature extraction using TF-IDF, and sentiment classification using Naïve Bayes and Support Vector Machine (SVM) classifiers. Experimental analysis demonstrates that machine learning-based NLP models can effectively classify social media sentiments with satisfactory accuracy. The study also highlights challenges such as noise, sarcasm, and informal language commonly found in social media data. Future enhancements using deep learning techniques are also discussed.

The findings confirm that NLP-based machine learning models provide an efficient solution for sentiment analysis of large-scale textual data. Future work may focus on incorporating deep learning techniques, multilingual analysis, and real-time sentiment monitoring to further enhance performance and applicability.

Keywords : Sentiment Analysis, Natural Language Processing, Social Media Data, Machine Learning, Text Mining, TF-IDF

I. Introduction

In recent years, social media platforms such as Twitter, Facebook, Instagram, and YouTube have become major communication channels where users actively share opinions, reviews, and emotions. This vast amount of user-generated content contains valuable information that can be used for business intelligence, market analysis, political analysis, and public opinion monitoring.

However, social media data is highly unstructured, informal, and noisy, making manual analysis difficult and time-consuming. Sentiment Analysis, also known as opinion mining, is a computational technique used to determine the emotional tone behind a body of text. It classifies opinions into categories such as positive, negative, or neutral.

ISSN: 2320-3714
Volume:2 Issue:1
April 2026
Impact Factor: 10.1
Subject: IT

Natural Language Processing (NLP) plays a vital role in enabling machines to understand, interpret, and process human language. By combining NLP techniques with machine learning algorithms, automated sentiment analysis systems can efficiently analyze large volumes of social media data.

The purpose of this research is to develop and evaluate a sentiment analysis model that uses NLP techniques and machine learning classifiers to analyze social media text and accurately identify sentiment polarity.

With the increasing penetration of smartphones and internet connectivity, social media usage has expanded rapidly across the globe. This growth has resulted in a continuous stream of real-time user-generated content, making sentiment analysis more relevant than ever. Organizations are leveraging this data to understand customer preferences, improve products, and enhance user experience. Moreover, governments and public institutions use sentiment analysis to gauge public opinion on policies and social issues, enabling more informed decision-making.

II. Identification of Research Problem

Although social media provides valuable opinion-based data, extracting accurate sentiment information is difficult due to:

- Presence of noise such as emojis, hashtags, and URLs
- Informal and unstructured language
- Ambiguity and sarcasm in user expressions
- High dimensionality of text data

Existing sentiment analysis methods often struggle to handle these challenges efficiently while maintaining high accuracy.

The rapid growth of social media platforms such as Twitter, Facebook, and Instagram has led to an exponential increase in user-generated textual data. Users continuously express opinions, emotions, and feedback on products, services, social issues, and events. This vast amount of data contains valuable insights; however, it is largely **unstructured, noisy, and complex**, making manual analysis inefficient and impractical.

Social media text poses several challenges for automated analysis. It often includes informal language, abbreviations, slang words, emojis, hashtags, spelling errors, and grammatical inconsistencies. Additionally, users frequently express opinions implicitly, use sarcasm, or mix neutral and emotional expressions within the same sentence. These characteristics significantly reduce the effectiveness of traditional text processing and information retrieval techniques.

Existing sentiment analysis approaches struggle to maintain high accuracy when applied to real-world social media data. Many systems fail to handle linguistic ambiguity, contextual dependency, and high-dimensional feature spaces efficiently. Moreover, selecting appropriate preprocessing techniques, feature extraction methods, and classification algorithms remains a critical issue in sentiment analysis research.

Airo International Journal
Peer-Reviewed
Multidisciplinary

ISSN: 2320-3714
Volume:2 Issue:1
April 2026
Impact Factor: 10.1
Subject: IT

III. Literature Review

Sentiment analysis has been an active research area for over a decade. Pang and Lee (2008) provided one of the foundational studies on sentiment classification, highlighting challenges such as subjectivity detection and contextual polarity. Their work demonstrated the effectiveness of supervised machine learning techniques for text sentiment classification.

Go et al. (2009) introduced a distant supervision approach for Twitter sentiment analysis by using emoticons as sentiment labels. This method enabled large-scale dataset creation without manual annotation. Subsequent studies focused on improving feature representation using techniques such as Bag of Words and TF-IDF.

Liu (2012) provided a comprehensive overview of sentiment analysis and opinion mining, covering lexicon-based and machine learning-based approaches. Recent research has explored deep learning models such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, which achieved improved accuracy but require more computational resources.

Despite advancements, sentiment analysis of social media data remains challenging due to informal language, abbreviations, sarcasm, and multilingual content. This research builds upon existing studies by applying classical NLP and machine learning techniques to achieve reliable sentiment classification.

Recent advancements in natural language processing have introduced transformer-based models that significantly improve sentiment classification performance. These models are capable of capturing contextual relationships between words more effectively than traditional approaches. Furthermore, hybrid models that combine lexicon-based and machine learning techniques have also been explored to enhance accuracy and interpretability. Despite these developments, the trade-off between computational efficiency and model performance remains an important consideration in practical applications.

IV. Problem Definition

Social media platforms generate a vast amount of textual data where users express opinions, emotions, and attitudes toward various topics. While this data is valuable for understanding public sentiment, it is predominantly unstructured, informal, and noisy, which makes automated sentiment identification a complex task. The presence of slang, abbreviations, emojis, spelling errors, hashtags, and sarcastic expressions further complicates accurate sentiment classification.

Traditional text analysis techniques are insufficient for handling the linguistic variability and contextual ambiguity found in social media content. Even existing sentiment analysis systems often face difficulties in selecting appropriate preprocessing techniques, extracting meaningful features, and choosing effective classification algorithms that can operate efficiently on high-dimensional textual data.

Therefore, the problem addressed in this research is to design and implement an effective sentiment analysis system using Natural Language Processing (NLP) techniques and machine learning algorithms that can accurately classify social media text into positive, negative, and neutral sentiments despite noise and linguistic complexity. The study focuses on identifying suitable preprocessing methods, feature extraction techniques, and classifiers to improve sentiment prediction accuracy on social media data.

Airo International Journal
Peer-Reviewed
Multidisciplinary

ISSN: 2320-3714
Volume:2 Issue:1
April 2026
Impact Factor: 10.1
Subject: IT

Another significant challenge in sentiment analysis is context understanding. Words may carry different meanings depending on the context in which they are used. For instance, a word that is typically positive may convey a negative sentiment in certain situations, especially when combined with sarcasm or irony. Additionally, code-mixing and multilingual expressions, which are common in social media platforms, further complicate sentiment detection and require more advanced linguistic processing techniques.

V. Objectives and Scope

Objectives : The primary objectives of this research work are as follows:

- To study and understand the role of Natural Language Processing (NLP) techniques in analyzing textual data from social media platforms.
- To preprocess unstructured and noisy social media data by applying appropriate text cleaning and normalization techniques.
- To implement and analyze machine learning algorithms such as Naïve Bayes and Support Vector Machine (SVM) for sentiment classification.
- To perform feature extraction using TF-IDF and evaluate its effectiveness in representing textual data numerically.
- To compare the performance of different classification models based on evaluation metrics such as accuracy, precision, recall, and F1-score.
- To identify key challenges in sentiment analysis, including sarcasm detection, ambiguity, and informal language usage.
- To develop an efficient and reliable model for classifying social media text into positive, negative, and neutral sentiments.

Scope : The scope of this research is defined as follows:

- The study focuses on sentiment analysis of English-language social media data only.
- The classification is limited to three sentiment categories: positive, negative, and neutral.
- The research uses traditional machine learning techniques rather than advanced deep learning models.
- The dataset used for analysis is pre-collected and labeled, and real-time data streaming is not considered.
- The research emphasizes text-based analysis and does not include multimedia content such as images, videos, or audio.
- The system is designed for offline analysis, making it suitable for academic and research purposes rather than real-time industrial deployment.
- The study provides a foundation that can be extended in the future to include multilingual analysis, real-time processing, and deep learning approaches.

VI. Research Methodology

The proposed research methodology consists of the following phases:

Data Collection: A publicly available social media dataset containing labeled text samples is used for experimentation. The dataset includes user comments categorized into sentiment classes.

Airo International Journal
Peer-Reviewed
Multidisciplinary

ISSN: 2320-3714
Volume:2 Issue:1
April 2026
Impact Factor: 10.1
Subject: IT

Data Preprocessing: To clean and normalize the text data, the following preprocessing steps are applied:

- Removal of URLs, mentions, hashtags, and special characters
- Conversion of text to lowercase
- Tokenization of text into words
- Removal of stop words
- Stemming or lemmatization

Feature Extraction: TF-IDF (Term Frequency–Inverse Document Frequency) is used to convert textual data into numerical feature vectors. This method emphasizes important words while reducing the influence of commonly occurring terms.

Model Implementation: Two machine learning classifiers are implemented:

- Naïve Bayes Classifier
- Support Vector Machine (SVM)

The dataset is divided into training and testing sets for model evaluation.

Evaluation Metrics: The models are evaluated using:

- Accuracy
- Precision
- Recall
- F1-score

In addition to TF-IDF, other feature representation techniques such as word embeddings can be considered to improve semantic understanding. Word embeddings capture contextual meaning by representing words in continuous vector space, allowing models to recognize relationships between similar words. Incorporating n-grams alongside unigram features can also help in capturing phrase-level sentiment, which is particularly useful in detecting negations and short expressions commonly found in social media text.

VII. Analysis and Findings

The experimental results indicate that both classifiers are capable of performing sentiment classification effectively. However, SVM achieved higher accuracy due to its ability to handle high-dimensional text data.

Classifier	Accuracy
Naïve Bayes	81%
SVM	88%

The analysis shows that preprocessing and feature extraction significantly influence model performance. Misclassification mainly occurs in sarcastic and ambiguous sentences.

Airo International Journal
Peer-Reviewed
Multidisciplinary

ISSN: 2320-3714
Volume:2 Issue:1
April 2026
Impact Factor: 10.1
Subject: IT

The experimental results also indicate that feature engineering plays a crucial role in determining model performance. Proper handling of negation words, punctuation, and repeated characters can significantly improve classification accuracy. It was observed that shorter texts, such as tweets, are more challenging to classify due to limited contextual information. On the other hand, longer texts provide better context but may introduce noise, requiring careful preprocessing and feature selection.

VIII. Limitations and Future Scope

Limitations

- Sarcasm and irony detection is limited
- Analysis is restricted to English language
- Dataset size impacts accuracy
- Real-time sentiment analysis is not implemented

Future Scope: Future enhancements may include:

- Use of deep learning models such as LSTM and BERT
- Multilingual sentiment analysis
- Emotion-based sentiment classification
- Real-time social media sentiment monitoring

Another limitation of the current study is the inability to capture fine-grained emotions such as happiness, anger, fear, and surprise. The classification into only three categories (positive, negative, neutral) may not fully represent the complexity of human emotions. Future research can explore emotion detection models that provide deeper insights into user sentiments. Additionally, integrating real-time data streams using APIs can enable dynamic sentiment tracking and trend analysis.

IX. Conclusion

This research successfully demonstrated the application of NLP techniques and machine learning algorithms for sentiment analysis of social media data. The proposed system effectively classified user opinions into positive, negative, and neutral categories. Among the implemented models, SVM showed superior performance compared to Naïve Bayes. The study confirms that NLP-based sentiment analysis is a powerful tool for extracting insights from unstructured social media data. Further improvements using advanced deep learning techniques can enhance accuracy and scalability.

Overall, the study emphasizes the importance of selecting appropriate preprocessing techniques and machine learning models for effective sentiment analysis. While traditional methods provide satisfactory results, the integration of advanced techniques can further enhance system performance. As social media continues to evolve, sentiment analysis systems must adapt to new linguistic patterns and user behaviors to remain accurate and reliable.

ISSN: 2320-3714
Volume:2 Issue:1
April 2026
Impact Factor: 10.1
Subject: IT

X. References

1. Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*.
2. Go, A., Bhayani, R., & Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. *Stanford University*.
3. Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Morgan & Claypool Publishers*.
4. Jurafsky, D., & Martin, J. H. (2020). Speech and Language Processing. *Pearson*.
5. Aggarwal, C. C. (2018). Machine Learning for Text. *Springer*.
6. Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems*.
7. Go, A., Bhayani, R., & Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. Stanford.
8. Maas, A. L., et al. (2011). Learning Word Vectors for Sentiment Analysis (IMDB). *ACL*.
9. Socher, R., et al. (2013). Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank (SST). *EMNLP*.
10. Rosenthal, S., et al. (2017). SemEval-2017 Task 4: Sentiment Analysis in Twitter. *SemEval*.
11. Joachims, T. (1998). Text Categorization with Support Vector Machines. *ECML*.
12. McCallum, A., & Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Text Classification. *AAAI Workshop*.
13. Mikolov, T., et al. (2013). Efficient Estimation of Word Representations in Vector Space (Word2Vec). *arXiv*.
14. Pennington, J., Socher, R., & Manning, C. (2014). GloVe. *EMNLP*.
15. Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *ICWSM*.

Airo International Journal
Peer-Reviewed
Multidisciplinary

ISSN: 2320-3714
Volume:2 Issue:1
April 2026
Impact Factor: 10.1
Subject: IT

Author's Declaration

As an author of the above research paper/article, here by, declare that the content of this paper is prepared by me and if any person having copyright issue or patent or anything otherwise related to the content, I shall always be legally responsible for any issue. For the reason of invisibility of my research paper on the website /amendments /updates, I have resubmitted my paper for publication on the same date. If any data or information given by me is not correct, I shall always be legally responsible. With my hole responsibility legally and formally have intimated the publisher (Publisher) that my paper has been checked by my guide (if any) or expert to make it sure that paper is technically right and there is no unaccepted plagiarism and hentriacontane is genuinely mine. If any issue arises related to Plagiarism/ Guide Name/ Educational Qualification /Designation /Address of my university/ college/institution/ Structure or Formatting/ Resubmission /Submission /Copyright /Patent /Submission for any higher degree or Job/Primary Data/Secondary Data Issues. I will be solely/entirely responsible for any legal issues. I have been informed that the most of the data from the website is invisible, shuffled, or vanished from the database due to some technical fault or hacking and therefore the process of resubmission is there for the scholars/students who find trouble in getting their paper on the website. At the time of resubmission of my paper I take all the legal and formal responsibilities, If I hide or do not submit the copy of my original documents (Andhra/Driving License/Any Identity Proof and Photo) in spite of demand from the publisher, then my paper may be rejected or removed from the website anytime and may not be consider for verification. I accept the fact that as the content of this paper and the resubmission legal responsibilities and reasons are only mine then the Publisher (Airo International Journal/Airo National Research Journal) is never responsible. I also declare that if publisher finds any complication or error or anything hidden or implemented otherwise, my paper may be removed from the website, or the watermark of remark/actuality may be mentioned on my paper. Even if anything is found illegal publisher may also take legal action against me.

Subhnandani Vishwakarma
Ambuj Yadav

******