International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

SENTIMENT ANALYSIS

[1] Farooqhussain Mohammad, [2] MD Kaif, [3] S Soundarya, [4] Venkata Srihari Bhanu [5] M Venkat Sai
[1]AssistantProfessor, [2345]Students
Department of Computer Engineering(Internet Of Things)
Siddhartha Institute of Technology & Sciences, Narapally
farooqhussain.cse@siddhartha.co.in , 23tq1a6923@siddhartha.co.in, 23tq1a6940@siddhartha.co.in, 23tq1a6950@siddhartha.co.in, 23tq1a6922@siddhartha.co.in

Abstract

This project presents an automated sentiment analysis system designed to classify textual data into positive, negative, and neutral categories. The approach follows a structured machine learning pipeline that begins with data collection from a labeled dataset and proceeds through essential preprocessing steps such as text cleaning, tokenization, and stopword removal to ensure data quality. Exploratory Data Analysis (EDA), including word clouds and distribution visualizations, is conducted to understand underlying patterns in the data. Feature extraction is performed using the Bag-of-Words model with CountVectorizer, and a Random Forest classifier is employed for sentiment classification. The model is evaluated using standard performance metrics such as accuracy and classification reports. The results demonstrate that the proposed system effectively captures sentiment in textual data, highlighting the capability of machine learning techniques to support real-time sentiment analysis applications.

I. Introduction

Sentiment analysis, also known as opinion mining, is a key application of natural language processing (NLP) that focuses on identifying and extracting subjective information from textual data. With the rapid growth of digital platforms such as social media, online reviews, and forums, vast amounts of user-generated content are produced every day. Analyzing this data manually is time-consuming and inefficient, making automated sentiment analysis systems essential for understanding public opinions, customer feedback, and trends.

This project aims to develop a machine learning-based sentiment analysis system capable of classifying text into positive, negative, and neutral categories. By leveraging techniques such as text preprocessing, feature extraction, and classification algorithms, the system transforms unstructured textual data into meaningful insights. The implementation uses the Bag-of-Words model for feature representation and applies a Random Forest classifier to perform accurate sentiment classification.

Through this project, the effectiveness of machine learning in processing and analyzing textual data is demonstrated, providing a practical solution for applications

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

such as product review analysis, social media monitoring, and decision-making support systems.

II. Literature survey

The field of sentiment analysis has evolved significantly over the years, beginning with foundational work by Bo Pang and Lillian Lee (2008), who demonstrated the effectiveness of supervised machine learning techniques for analyzing movie reviews. Their research laid the groundwork for using algorithms such as Naïve Bayes and Support Vector Machines for text classification tasks. Building on this, Bing Liu (2012) provided a comprehensive survey of opinion mining and sentiment analysis, highlighting important areas such as aspect-based sentiment analysis and the challenge of detecting fake or spam opinions.

Over time, research in this domain has expanded to include lexicon-based approaches, traditional machine learning models, and more advanced deep learning techniques. Recent studies emphasize the strong performance of models like Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and transformer-based architectures such as BERT, which capture contextual meaning more effectively. Despite these advancements, challenges such as domain adaptation, sarcasm detection, and real-time sentiment processing continue to be active research areas.

This project builds upon these foundational and modern approaches by implementing a classical machine learning pipeline on a contemporary social media dataset. It demonstrates that even with the rise of deep learning, traditional methods like Random Forest combined with effective preprocessing and feature extraction techniques remain relevant and capable of delivering reliable sentiment classification results.

III. System Analysis

System analysis involves understanding the requirements and objectives of building a sentiment analysis system. The main goal is to classify text data into different sentiment categories such as positive, negative, or neutral. This process includes identifying data sources like social media, reviews, or feedback, and analyzing the structure of textual data. It also focuses on selecting appropriate preprocessing techniques such as tokenization, stop-word removal, and stemming. The system must handle challenges like sarcasm, ambiguity, and context understanding. Various machine learning and deep learning algorithms are evaluated to determine the most suitable model. Performance metrics such as accuracy, precision, recall, and F1-score are considered. The system also requires scalability to handle large datasets efficiently. Additionally, it ensures proper data cleaning and feature extraction methods. Overall,

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X ww.ijdim.com Original Research Paper

system analysis helps in designing an effective and reliable sentiment classification system.

Existing System

The existing system for sentiment analysis primarily relies on traditional methods such as manual analysis or basic lexicon-based approaches. In these systems, predefined dictionaries of positive and negative words are used to determine sentiment. Some systems also use simple machine learning algorithms like Naïve Bayes or Support Vector Machines with limited feature extraction. These methods often depend heavily on keyword matching rather than understanding the context of the text. They may not effectively handle complex language patterns, sarcasm, or mixed sentiments. The preprocessing techniques used are often basic and may not capture deeper semantic meanings. Additionally, existing systems may struggle with domain-specific data and require frequent updates to the lexicon.

Disadvantages of Existing System

· Limited understanding of context and semantics
· Poor handling of sarcasm and irony
· Dependence on predefined lexicons
· Low accuracy for complex or mixed sentiments
· Not adaptable to new language trends
· Requires frequent manual updates
· Inefficient for large-scale data processing

Proposed System

The proposed system aims to improve sentiment analysis by using advanced machine learning and deep learning techniques. It incorporates algorithms such as Logistic Regression, Random Forest, or deep learning models like LSTM and transformer-based models such as BERT. The system uses advanced preprocessing techniques including tokenization, lemmatization, and vectorization methods like TF-IDF or word embeddings. It focuses on capturing contextual meaning rather than relying only on keywords. The proposed system can handle large datasets efficiently and is scalable for real-time applications. It also improves accuracy by using better feature engineering and model tuning techniques. The system is capable of handling domain-specific data with minimal adjustments.

Advantages of Proposed System

· Higher accuracy and better performance
· Improved context understanding

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

- Handles sarcasm and complex sentences better
- Scalable for large datasets
- Supports real-time analysis
- Less dependency on manual rules
- Adaptable to different domains

IV. Methodology

The methodology of the sentiment analysis system follows a structured pipeline to ensure accurate and efficient classification of text data. Initially, data is collected from various sources such as social media platforms, customer reviews, or datasets. This data is then preprocessed to remove noise, including stop words, punctuation, and irrelevant characters. Techniques like tokenization, stemming, and lemmatization are applied to standardize the text. After preprocessing, feature extraction methods such as TF-IDF or word embeddings are used to convert textual data into numerical form. The processed data is then split into training and testing datasets. Machine learning or deep learning models such as Logistic Regression, Naïve Bayes, or LSTM are trained on the dataset. The model is evaluated using metrics like accuracy, precision, recall, and F1-score. Once optimized, the model is deployed to classify new incoming text data. Finally, the system continuously improves by retraining with updated data to enhance performance and adaptability.

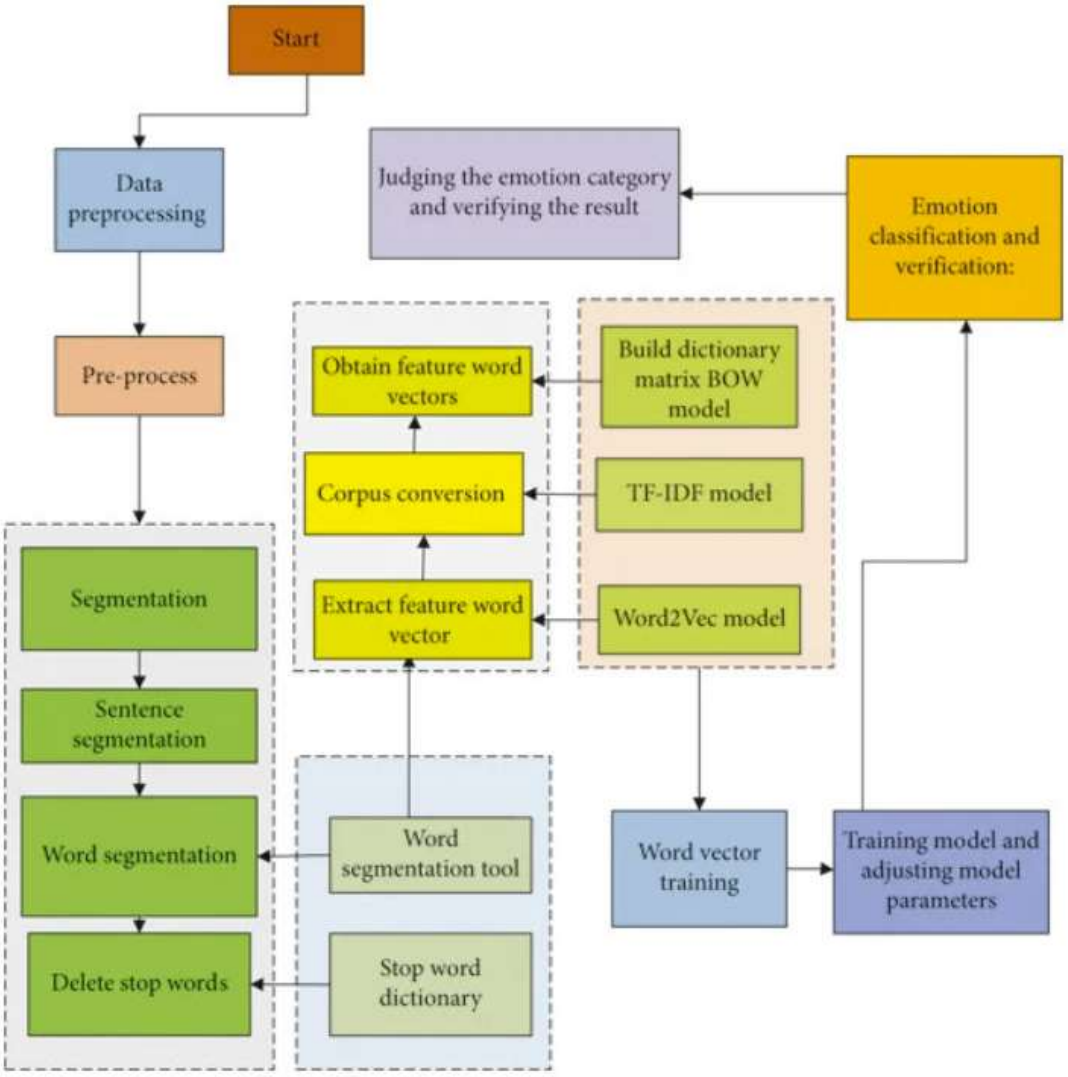
System Architecture

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper



The system architecture of the sentiment analysis system is designed as a multi-layered pipeline that processes text data efficiently from input to output. It begins with the data collection layer, where raw textual data is gathered from sources such as social media, reviews, or datasets. This data is then passed to the preprocessing layer, where noise such as stop words, punctuation, and irrelevant characters is removed, and techniques like tokenization, stemming, and lemmatization are applied. Next, the feature extraction layer converts the cleaned text into numerical representations using methods such as TF-IDF or word embeddings. The processed data is then fed into the model training layer, where machine learning or deep learning algorithms are used to train the sentiment classification model. After training, the model evaluation layer assesses performance using metrics like accuracy, precision, recall, and F1-score to ensure reliability. Once validated, the system moves to the prediction layer, where

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

new input text is classified into sentiment categories such as positive, negative, or neutral. Finally, the deployment layer integrates the model into real-world applications like web platforms or APIs, enabling real-time sentiment analysis and continuous system improvement.

V. Result and Output

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

VI. Conclusion

Sentiment analysis is a powerful tool for extracting actionable insights from unstructured text data. This project successfully demonstrated the development of an end-to-end sentiment analysis pipeline, from data loading and exploratory analysis to model training and evaluation. Using a dataset of social media posts, we performed thorough preprocessing, visualized data insights with distribution plots and word clouds, and trained a *Random Forest* machine learning classifier. The results showed that the Random Forest model can effectively classify sentiment, achieving an accuracy of around 68%. The project successfully addressed the core objectives of automating sentiment classification and visualizing key data patterns. Despite inherent challenges like handling sarcasm and ambiguity, this classical machine learning approach provides a scalable and effective foundation for real-world applications in brand monitoring, customer service, and public opinion analysis. With the future integration of advanced deep learning models and real-time processing capabilities, the potential and accuracy of such systems will only continue to grow, solidifying sentiment analysis's critical role in data-driven decision-making.

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine

learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.

[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.

[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm

[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.

[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, "Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks," in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.

[7] R. D. Kumar, V. N. S.Manaswini, "Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology," in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.

[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie recommender using collaborative filtering and sentiment analysis," International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.

[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.

[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.

[11] Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.