International Journal of Web of Multidisciplinary Studies

(Peer-Reviewed, Open Access, Fully Refereed International Journal)

website: *https://ijwos.com*
Volume 3 Issue 3 , March 2026

OPEN ACCESS

E-ISSN : 3049-2424

DOI: https://doi.org/10.71366/ijwos03032624949

BIG DATA SENTIMENT ANALYSIS ON PRODUCT REVIEWS

G.SOUNDARYA DEVI[1], S.KESAVAN[2], G.S.SAMPATH KUMAR[3], A.MURALIDHARAN[4], S.GOKUL[5]

[1] Assistant Professor, Department of Information Technology , M P Nachimuthu M Jaganthan Engineering College

[2,3,4,5] Final Year B.Tech (IT), Department of Information Technology , M P Nachimuthu M Jaganthan Engineering College.

Article Info

Article History:

Published:19 March 2026

Publication Issue:
Volume 3, Issue 3
March-2026

Page Number:
392-396

Corresponding Author:
G.SOUNDARYA DEVI

Abstract:

The rapid growth of e-commerce platforms has resulted in a massive amount of customer product reviews being generated every day. These reviews contain valuable insights regarding customer satisfaction, product quality, and user experience. However, manually analyzing large volumes of reviews is difficult and time-consuming. This research proposes a Big Data Sentiment Analysis framework for product reviews using machine learning techniques. The system collects product review data and processes it using Natural Language Processing (NLP) methods including tokenization, stop- word removal, and text normalization. Feature extraction is performed using TF-IDF representation, and classification models such as Logistic Regression, Naïve Bayes, and. Random Forest are used to identify the sentiment of reviews as positive, negative, or neutral. Experimental results show that the proposed system effectively analyzes large-scale review data and achieves high classification accuracy. The framework helps businesses understand customer opinions and improve product quality and customer satisfaction.

Keywords: Sentiment Analysis, Big Data, Product Reviews, Machine Learning, Natural Language Processing

1. INTRODUCTION

In recent years, online shopping platforms such as Amazon, Flipkart, and other e-commerce websites have become an important part of modern consumer behavior. Customers often share their experiences and opinions about products by writing reviews and giving ratings. These reviews influence purchasing decisions and provide valuable feedback to manufacturers and sellers.The number of product reviews generated daily on these platforms is extremely large. . As a result, analyzing this huge amount of textual information manually becomes impractical. This challenge has led to the development of automated sentiment analysis systems that can process and analyze large volumes of textual data efficiently . Sentiment analysis, also known as opinion mining, is a field of study within Natural Language Processing that focuses on identifying and extracting subjective information from text. It helps determine whether a particular piece of text expresses a positive, negative, or neutral opinion . By applying sentiment analysis to product reviews, companies can gain insights into customer preferences, identify product issues, and improve overall customer satisfaction. Big Data technologies further enhance this process by enabling the analysis of massive datasets quickly and efficiently . This project aims to develop a Big Data-based sentiment analysis system that can automatically analyze product reviews and classify them according to their sentiment polarity. The proposed system combines text preprocessing, feature extraction, and machine learning algorithms to perform accurate sentiment classification.

LITERATURE REVIEW AND RELATED WORK

Sentiment analysis has become an important research area in the fields of Natural Language Processing, machine learning, and data mining. Many researchers have proposed various techniques and models to analyze customer opinions from textual data.

Pang and Lee (2008) conducted one of the earliest studies on sentiment classification and demonstrated the effectiveness of machine learning algorithms such as Naïve Bayes, Maximum Entropy, and Support Vector Machines for opinion mining tasks. Their work laid the foundation for modern sentiment analysis research.

Liu (2012) explored sentiment analysis in the context of e-commerce and highlighted the importance of analyzing customer reviews to understand product quality and user satisfaction. The study emphasized that customer feedback can help businesses improve products and services . Medhat et al. (2014) provided a comprehensive survey of sentiment analysis techniques, including machine learning methods, lexicon-based approaches, and hybrid met.

Despite significant advancements, several challenges still exist in sentiment analysis, such as handling sarcasm, detecting fake reviews, processing multilingual text, and managing extremely large datasets. These challenges continue to motivate further
research in this area.

2. METHODOLOGY

2.1 Data Collection and Preprocessing

The first step in the proposed system is the collection of product review data
from available datasets such as Amazon product review datasets or other e- commerce platforms. The dataset typically includes information such as product name, review text, rating, and user feedback.Raw datasets often contain noisy and unstructured data. Therefore, preprocessing is necessary to clean and prepare the data before applying machine learning models. Several preprocessing techniques are applied to improve data quality.These preprocessing steps include removing duplicate records, handling missing values, removing punctuation marks and special characters, converting text into lowercase, and eliminating common stop-words such as “the”, “is”, and “and”. Tokenization is also performed to split sentences into individual words or tokens. Additionally, stemming or lemmatization techniques are applied to reduce words to their root forms. These preprocessing steps help improve the accuracy and efficiency of the sentiment analysis model.

2.2 Feature Extraction and Engineering

After preprocessing, textual data must be converted into numerical features so that machine learning algorithms can process it. Feature extraction is a crucial step in sentiment analysis because it determines how text data is represented in the model. One of the most commonly used feature extraction techniques is TF-IDF (Term Frequency – Inverse Document Frequency). This method measures the importance of words in a document relative to a collection of documents.

The TF-IDF formula is defined as:

TF-IDF(t,d) = TF(t,d) × log(N / DF(t))

Where:

- TF represents the frequency of a term in a document
- DF represents the number of documents containing that term
- N represents the total number of documents in the dataset

This technique assigns higher weights to words that appear frequently in a specific document but rarely in other documents, helping the model identify meaningful patterns in the text.

2.3 Machine Learning Classification Once the features are extracted, machine learning algorithms are trained to classify product reviews into different sentiment categories.Several classification algorithms are used in this project. **Logistic Regression** is a widely used statistical model for binary classification problems. It estimates the probability that a given input belongs to a particular class.

Naïve Bayes is a probabilistic classifier based on Bayes' theorem. It assumes independence between features and is particularly effective for text classification tasks.

Random Forest is an ensemble learning algorithm that constructs multiple decision trees and combines their predictions to improve accuracy and reduce overfitting.

The performance of these models is evaluated using common evaluation metrics such as **accuracy, precision, recall, and F1-score**.

3. IMPLEMENTATION AND RESULTS

The proposed sentiment analysis system is implemented using Python programming language

due to its powerful libraries for data analysis and machine learning. Several libraries and tools are used in the implementation process, including: Pandas for data manipulation and analysis NumPy for numerical computations Scikit-learn for machine learning algorithms NLTK for Natural Language Processing tasks Matplotlib for visualization of results The implementation process involves several stages including loading the dataset, preprocessing text data, extracting features using TF-IDF, training machine learning models, and evaluating their performance. The dataset is divided into training and testing sets to evaluate the model's predictive performance. After training the models, predictions are generated for the testing data and compared with actual labels.

Experimental results indicate that the **Random Forest classifier achieves the highest accuracy**, followed by Logistic Regression and Naïve Bayes

4. DISCUSSION

The experimental results demonstrate that machine learning techniques can effectively classify product reviews into sentiment categories. The use of TF-IDF feature extraction significantly improves classification performance by identifying the most important words in the review text.

Among the tested models, Random Forest performed better due to its ability to combine multiple decision trees and capture complex relationships in the data. Logistic Regression also produced reliable results due to its efficiency in handling binary classification tasks.

However, the system still faces certain limitations. For example, detecting sarcasm or ironic statements in text remains challenging. Additionally, customer reviews often contain informal language, spelling errors, and slang words that may affect model accuracy.

Future research can focus on applying advanced deep learning models such as **LSTM, CNN, or transformer-based models like BERT** to further improve sentiment classification performance.

5. CONCLUSION

This research presents a Big Data sentiment analysis system for product reviews using machine learning techniques. The proposed system processes large volumes of customer reviews and automatically classifies them into positive, negative, and neutral sentiment categories.

The system uses Natural Language Processing techniques for text preprocessing and TF-IDF for feature extraction. Machine learning algorithms such as Logistic Regression, Naïve Bayes, and Random Forest are applied for sentiment classification.

Experimental results demonstrate that the proposed approach can efficiently analyze large datasets and provide accurate sentiment predictions. The system can help businesses gain valuable insights from customer feedback and improve product quality and customer satisfaction.

Future work may focus on integrating deep learning models, handling multilingual datasets, and performing real-time sentiment analysis for large- scale e-commerce platforms.

References

1. B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.
2. B. Liu, *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, 2012.
3. W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, 2014.
4. E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 15–21, 2013.
5. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification using Distant Supervision," *Stanford University Technical Report*, 2009.
6. M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004, pp. 168–177.
7. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of WordRepresentations in Vector Space," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2013.
8. Maas et al., "Learning Word Vectors for Sentiment Analysis," *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 2011.
9. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of NAACL- HLT*, 2019.
10. Y. Kim, "Convolutional Neural Networks for Sentence Classification," *Proceedings of EMNLP*, 2014.
11. S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, O'Reilly Media, 2009.
12. J. Leskovec, A. Rajaraman, and J. Ullman, *Mining of Massive Datasets*, Cambridge University Press, 2014.
13. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.
14. J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.
15. M. Zaharia et al., "Apache Spark: A Unified Engine for Big Data Processing," *Communications of the ACM*, vol. 59, no. 11, pp. 56–65, 2016