© NOV 2025 | IRE Journals | Volume 9 Issue 5 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I5-1712033

Automated Detection of Deceptive Online Product Reviews Using Supervised Learning Techniques

ABHISHEK GUPTA[1], AYUSH NIGAM[2], ABHISHEK KUMAR[3], DR. ISHRAT ALI[4], PROF. (DR.) SANJAY PACHAURI[5]

[1, 2, 3, 4, 5]*Department of Data Science (DDCS), GNIOT College, Greater Noida, India*

Abstract: - This paper presents a machine learning–based system for detecting fake product reviews using Natural Language Processing (NLP) techniques. With the rapid growth of e-commerce, online reviews significantly influence consumer purchasing behavior, but the rise of deceptive or manipulated reviews has undermined their reliability. The proposed model utilizes text preprocessing methods such as tokenization, stop-word removal, and stemming, followed by feature extraction using TF-IDF and CountVectorizer. Multiple supervised learning algorithms, including Logistic Regression, Random Forest, Decision Tree, Naïve Bayes, and K-Nearest Neighbors (KNN), were implemented to classify reviews as genuine or fake. Experimental results show that the Support Vector Machine (SVM) achieved the highest accuracy of approximately 88.5%, outperforming other models. Analysis of feature importance and confusion matrices revealed that linguistic and frequency-based attributes play a key role in deception detection. The developed system also includes a real-time review classification module, demonstrating its potential for application in Deceptive Review Detection, review moderation, and consumer trust enhancement.

Keywords — Fake Product Reviews, Machine Learning, Natural Language Processing (NLP), Text Classification, Deceptive Review Detection.

I. INTRODUCTION

In recent years, the exponential growth of e-commerce platforms such as Amazon, Flipkart, and eBay has revolutionized the way consumers make purchasing decisions. A significant factor influencing these decisions is the availability of customer reviews, which provide firsthand insights into product quality, usability, and customer satisfaction. However, with the increasing dependence on online feedback, the problem of fake or deceptive reviews has emerged as a critical challenge for both consumers and businesses. These reviews, often generated by bots or paid individuals, are designed to manipulate product ratings, distort public opinion, and gain an unfair competitive advantage.

The detection of fake product reviews is not a trivial task, as deceptive reviews are typically written to appear authentic and linguistically similar to genuine ones. Traditional manual moderation techniques are insufficient to handle the vast volume of reviews generated daily. Consequently, there is a growing need for automated systems capable of identifying deceptive reviews with high accuracy and scalability. This is where the combined power of Machine Learning (ML) and Natural Language Processing (NLP) becomes indispensable.

In this project, machine learning algorithms are employed to learn linguistic and statistical patterns from review data to distinguish between genuine and deceptive content. The NLP component plays a crucial role in preprocessing and feature extraction, transforming unstructured text into meaningful numerical representations that machine learning models can analyze. Techniques such as tokenization, stop-word removal, stemming, lemmatization, and vectorization (TF-IDF and CountVectorizer) enable the conversion of textual information into feature matrices suitable for classification tasks.

The study explores and compares multiple supervised learning models, including Logistic Regression, Random Forest, Decision Tree, K-Nearest Neighbors (KNN), and Naïve Bayes, to determine the most effective approach for detecting fake reviews. Each algorithm's performance is evaluated using standard classification metrics such as accuracy, precision, recall, and F1-score.

By leveraging these methodologies, the project aims to contribute to the ongoing research in deceptive review detection and to develop a reliable, reproducible framework that can be adapted to various e-commerce datasets. Ultimately, the findings are expected to support platforms and businesses in enhancing transparency, improving

recommendation systems, and building consumer trust in digital marketplaces.

II. METHODOLOGY

The proposed system for fake product review detection employs a structured pipeline that combines Natural Language Processing (NLP) and Machine Learning (ML) techniques to accurately classify reviews as genuine or deceptive. The dataset is first preprocessed to remove null entries, duplicates, and special characters to ensure data consistency. Each review undergoes tokenization, stop-word removal, and stemming or lemmatization to normalize the text. The cleaned text data is then transformed into numerical representations using TF-IDF (Term Frequency–Inverse Document Frequency) and CountVectorizer, which capture word frequency and importance.

Following preprocessing, multiple supervised learning algorithms—including Logistic Regression, Random Forest, Decision Tree, Naïve Bayes, and K-Nearest Neighbors (KNN)—are trained and evaluated. The dataset is split into training (80%) and testing (20%) subsets to assess generalization. Performance evaluation is conducted using accuracy, precision, recall, and F1-score metrics.

Among the tested models, the Random Forest classifier achieved the highest accuracy, demonstrating its robustness and effectiveness in detecting deceptive reviews. This methodology establishes a reliable framework for applying NLP-driven ML models to enhance review authenticity analysis in e-commerce platforms.

III. RESULTS AND DISCUSSION

The performance of various machine learning algorithms was evaluated using accuracy, precision, recall, and F1-score metrics to determine their effectiveness in detecting fake product reviews. Among all models tested—Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, and K-Nearest Neighbors (KNN)—the Support Vector Machine (SVM) Classifier achieved the highest accuracy of approximately 88.5%, outperforming the Naïve Bayes Classifier (85%) and Logistic Regression (86%). This indicates that ensemble-based methods like Support Vecctor Machine (SVM) Classifier can effectively capture complex linguistic patterns and non-linear relationships within textual data.

The confusion matrix analysis revealed that the Support Vector Machine (SVM) model had the lowest rate of false classifications, demonstrating its ability to generalize well on unseen reviews. Additionally, the feature importance analysis showed that term frequency–based attributes and linguistic cues significantly contribute to distinguishing deceptive reviews from genuine ones.

Overall, the results confirm that NLP-based preprocessing combined with ensemble learning provides a robust solution for identifying fake reviews. The system's high accuracy and reliability suggest its potential integration into real-world e-commerce platforms to improve transparency, reduce manipulation, and enhance consumer trust in online product feedback systems.

IV. CONCLUSION

This study presents an effective approach for detecting fake product reviews using Machine Learning (ML) and Natural Language Processing (NLP) techniques. The proposed framework successfully identifies deceptive reviews by combining advanced text preprocessing, feature extraction, and supervised learning models. Various algorithms, including Logistic Regression, Decision Tree, Naïve Bayes, K-Nearest Neighbors (KNN), and Random Forest, were implemented and evaluated to determine the most efficient model for classification.

The experimental analysis demonstrated that the Support Vector Machine (SVM) classifier achieved the best overall performance with an accuracy of approximately 88.5%, outperforming other models in terms of precision and recall. The findings highlight the importance of linguistic and frequency-based features in distinguishing genuine reviews from deceptive ones.

This research establishes a solid foundation for applying machine learning to e-commerce review verification. The developed system can be integrated into online platforms to automatically detect suspicious reviews, helping consumers make informed decisions and enhancing marketplace credibility. Future work may focus on expanding the dataset, incorporating deep learning models such as

LSTM or BERT, and improving model generalization for multilingual and cross-domain review detection.

REFERENCES

[1] Ahmad, A., & Siddiqui, M. F. (2022). *Detecting deceptive online reviews using machine learning and NLP techniques*. Journal of Information and Computational Science, 12(5), 45–53.

[2] Banerjee, S., & Choudhary, A. (2021). *Fake review detection using natural language processing and supervised learning approaches*. International Journal of Data Science and Analytics, 9(4), 315–327.

[3] Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. (2013). *What Yelp fake review filter might be doing?* Proceedings of the International AAAI Conference on Web and Social Media, 409–418.

[4] Scikit-learn Documentation: https://scikit-learn.org/stable/ 2. NLTK Toolkit: https://www.nltk.org/ 3. Kaggle Datasets: Fake Reviews Detection Dataset 4. BERT for Text Classification — Devlin et al., Google AI (2018)

[5] Ray, S., & Chakraborty, M. (2020). *Fake review detection using ensemble learning and text analytics*. International Journal of Advanced Computer Science and Applications, 11(8), 110–118.

[6] Zhou, L., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5), 1–40.

[7] Sharma, R., & Gupta, D. (2021). Detection of spam product reviews using machine learning and linguistic features. *Journal of Big Data*, 8(1), 1–15.

[8] Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.