iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.81139

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Fake Product Review Detection System Using Machine Learning and Natural Language Processing

Lakshya Kanaujia, Mayank Verma, Er. Ayush Pratap Singh
Stud. of Computer Science & Engineering Shri Ramswaroop Memorial College Of Engineering & Management Lucknow, India

Abstract: ***Online product reviews play a critical role in shaping consumer purchase decisions. However, the increasing prevalence of fake and deceptive reviews undermines trust in e-commerce platforms. This paper presents a Fake Product Review Detection System that leverages Natural Language Processing (NLP) and Machine Learning (ML) techniques to automatically classify reviews as genuine or fake. The system processes textual data through NLP pipelines for tokenization, lemmatization, and feature extraction, followed by model training using supervised learning algorithms such as Support Vector Machines (SVM), Random Forest, and BERT fine-tuning. Evaluation metrics including precision, recall, and F1-score demonstrate the model's effectiveness in identifying fraudulent reviews, thereby improving online trust and transparency.***
Keywords: Fake Review Detection, NLP, Machine Learning, Deep Learning, BERT, Sentiment Analysis, E-commerce, Opinion Mining, Cyber Fraud, Text Analytics

I. INTRODUCTION

The rapid growth of e-commerce platforms such as **Amazon, Flipkart, Yelp, and TripAdvisor** has revolutionized the global retail landscape by allowing consumers to purchase products with just a few clicks. A critical component of this digital ecosystem is the **online review system**, where customers share their opinions, experiences, and satisfaction levels with products and services. These reviews act as **social proof** and play an influential role in shaping consumer perception, enhancing seller credibility, and driving sales conversions.

However, this heavy reliance on reviews has also led to the emergence of a significant problem — **fake product reviews**. These deceptive reviews are often created with malicious intent: some artificially **inflate product ratings** to attract customers, while others aim to **defame competitors** by posting negative and misleading content. This manipulation damages consumer trust, misguides purchasing decisions, and poses challenges to maintaining transparency in online marketplaces. Studies have shown that **30–40% of reviews** in popular e-commerce platforms may be **partially or completely fabricated**, making automated detection essential for ensuring fair digital commerce.

Traditional review moderation methods, such as **manual verification** or **rule-based systems**, are insufficient to handle the massive volume of user-generated content. Manual screening is time-consuming, labor-intensive, and prone to human bias, while rule-based systems—relying on fixed heuristics like excessive punctuation or repeated keywords—lack adaptability against evolving spamming tactics. As fake reviews continue to evolve in structure and linguistic complexity, **static approaches fail to generalize across domains** or languages.

Figure 1. Fake Product Reviews (by the customer)

To overcome these limitations, researchers have turned to **Artificial Intelligence (AI)**, particularly **Natural Language Processing (NLP)** and **Machine Learning (ML)**, to build systems capable of detecting deceptive content automatically. NLP enables computers to analyze and understand human language, while ML allows the system to learn hidden patterns in textual and behavioral data that distinguish genuine from fake reviews. Advanced models such as **Support Vector Machines (SVM)**, **Random Forests**, and **Deep Neural Networks** have demonstrated promising results by identifying linguistic cues, sentiment inconsistencies, and unusual reviewer behaviors.

Recently, **transformer-based models** like **BERT (Bidirectional Encoder Representations from Transformers)** have revolutionized text classification by capturing deeper semantic and contextual relationships between words. These models, when fine-tuned on labeled review datasets, significantly outperform traditional algorithms in recognizing deceptive language constructs. Furthermore, **behavioral features**, such as posting frequency, average rating deviation, and reviewer history, add a powerful layer of context that enhances prediction accuracy.

The combination of **linguistic**, **semantic**, and **behavioral** analysis
makes fake review detection a multi-dimensional problem. The proposed Fake Product Review Detection System leverages this multi-faceted approach to ensure accurate classification and real-time deployment. The system aims to not only flag suspicious reviews but also explain why they are likely fake by providing interpretable insights such as key influential words or behavioral patterns.

Overall, this research contributes toward building a trustworthy and transparent digital marketplace. By automating the detection of fake product reviews, the system assists consumers in making informed decisions, supports e-commerce platforms in maintaining integrity, and ultimately enhances the reliability of online ecosystems. The following sections discuss the literature background, methodology, system design, and implementation details of this intelligent detection framework.

II. LITERATURE SURVEY

The detection of fake product reviews has been an evolving research challenge within the domains of Natural Language Processing (NLP), Machine Learning (ML), and Data Mining. Early research primarily relied on linguistic and statistical methods to identify patterns of deception in textual content. Ott et al. [1] pioneered one of the first benchmark datasets for deceptive hotel reviews and demonstrated that classifiers such as Support Vector Machines (SVM) and Naïve Bayes could effectively differentiate fake reviews based on linguistic cues such as adjective frequency, pronoun usage, and syntactic structure. Similarly, Pang and Lee [2] explored sentiment polarity detection to identify emotional tone in reviews, but it was found that fake reviews often mimic genuine sentiment, limiting the reliability of sentiment-only approaches. Jindal and Liu [7] expanded on this by studying opinion spam detection and identifying unusual patterns of review duplication and rating anomalies on e-commerce platforms.

As research progressed, behavior-based and graph-based approaches gained traction for their ability to capture relationships between users, products, and temporal patterns. Mukherjee et al. [8] proposed a graph-based spam detection framework where reviewer–product interactions were modeled as networks to detect coordinated fake review groups. Wang et al. [10] and Xie et al. [9] extended this idea by introducing rating deviation and temporal behavior as key features to detect bursts of deceptive activity. Although these methods improved detection accuracy, they required rich metadata and suffered from high computational cost when applied to large-scale systems.

With the advent of ensemble and hybrid models, researchers began combining multiple feature sets to enhance generalization. Lim et al. [11] and Heydari et al. [12] demonstrated that merging textual and behavioral features in ensemble classifiers such as Random Forest and Gradient Boosting produced higher accuracy and lower false-positive rates. Mukherjee and Liu [13] further observed that combining metadata with content-based features significantly improved detection reliability across multiple domains. These studies established the foundation for hybrid models that leverage both linguistic and behavioral aspects of online reviews.

In recent years, deep learning has revolutionized fake review detection by capturing semantic and contextual relationships in text. Neural architectures such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks were used to automatically extract deep features from raw text. Cao et al. [14] utilized CNN-LSTM hybrids to simultaneously capture local and temporal dependencies, while Li et al. [5] employed Bi-LSTM models for sequential pattern learning. However, the real breakthrough came with the introduction of transformer-based architectures such as BERT by Devlin et al. [3] and the self-attention mechanism proposed by Vaswani et al. [4]. These models significantly outperformed previous systems by understanding bidirectional context and subtle semantic variations in deceptive writing. Subsequent works such as those by Sun et al. [15], Wang et al. [17], and Chen et al. [18] reported over 90% accuracy in fake review detection using fine-tuned BERT and DistilBERT models, highlighting the power of contextual embeddings and attention mechanisms.

Despite these advancements, challenges remain. Data imbalance, domain shift, and adversarial review generation continue to hinder real-world deployment. Kumar et al. [23] highlighted that GPT-based large language models can now generate highly realistic fake reviews that easily evade traditional detection algorithms. To mitigate this, researchers have introduced transfer learning [21][22], adversarially trained models, and explainable AI (XAI) frameworks to enhance model interpretability and trustworthiness [24][25][26]. Recent studies also focus on multimodal fusion, integrating textual, visual, and behavioral data to further strengthen detection accuracy [19][20]. Overall, the literature demonstrates a clear evolution from rule-based and statistical methods to advanced transformer-based hybrid architectures that combine linguistic, behavioral, and contextual understanding. Continued efforts in explainability, dataset diversity, and adversarial defense remain critical for the sustainable deployment of these systems in modern e-commerce platforms.

III. SYSTEM ARCHITECTURE AND METHODOLOGY

The proposed system architecture integrates multiple AI components to achieve high precision in fake review detection.

A. Data Collection

Data was obtained from public datasets such as the Ott Hotel Review Dataset, Amazon Product Review Corpus, and Yelp Challenge Dataset. Each review includes metadata such as reviewer ID, rating, timestamp, verified purchase status, and helpful votes. Data scraping is performed using APIs compliant with platform policies.

B. Data Preprocessing

Textual preprocessing ensures normalization and consistency across reviews. The following steps were used:

- Text normalization and lowercasing.
- Removal of HTML tags, URLs, and emojis.
- Tokenization and lemmatization using SpaCy.
- Stop-word removal.
- Sentiment tagging using TextBlob.

This process reduces noise and prepares the data for feature extraction.

C. Feature Extraction

The study employs a hybrid feature extraction framework combining:

- Lexical Features – word count, average word length, punctuation frequency.
- Syntactic Features – part-of-speech tag distributions, pronoun usage.
- Semantic Features – TF-IDF vectors, Word2Vec embeddings, and BERT sentence representations.
- Behavioral Features – reviewer rating patterns, time intervals between posts, account activity rate, and helpful vote ratio.

D. Model Training

Multiple ML models were trained for comparative evaluation:

- Logistic Regression, Naïve Bayes, Random Forest, SVM
- Deep Learning models (Bi-LSTM, CNN)
- Transformer models (BERT, DistilBERT, RoBERTa)
- Fine-tuning BERT on the combined datasets yielded the best performance.

E. Evaluation

The system's performance was assessed using precision, recall, F1-score, and ROC-AUC metrics. Class imbalance was addressed using SMOTE oversampling. Cross-validation with 10 folds was used for robustness.

F. Deployment

The best-performing model was deployed using Flask REST API. The system features a ReactJS-based frontend for user interaction and integrates with cloud services via Docker for scalability.

IV. ANALYSIS

Model evaluation demonstrated that transformer-based architectures outperformed traditional ML models. Logistic Regression and SVM achieved average F1-scores of 0.84 and 0.88 respectively, while fine-tuned BERT reached 0.94.

Random Forest performed well in terms of interpretability but struggled with high-dimensional embeddings. LSTM captured temporal dependencies effectively but required more training time. The inclusion of behavioral features increased overall accuracy by 6–8%.

The confusion matrix revealed that false positives (genuine reviews misclassified as fake) were significantly reduced by integrating semantic embeddings. The ROC-AUC score of 0.96 confirms the robustness of the model.

Attention visualization (via transformer layers) highlighted that fake reviews often contained vague adjectives (“great”, “nice”, “best”) without specific product features, whereas genuine reviews included concrete details (“battery lasted 12 hours”, “screen resolution is crisp”).

The model’s performance across domains showed some degradation in unseen categories, indicating the need for transfer learning strategies for cross-domain adaptability.

V. MAJOR FINDING

Hybrid Features Are Crucial: Combining linguistic, semantic, and behavioral features results in the highest detection accuracy.

Transformer Models Dominate: Fine-tuned BERT and RoBERTa outperform traditional classifiers by a large margin.

Explainability Matters: XAI tools like SHAP improve user trust by explaining model predictions.

Behavioral Analytics Enhance Detection: Review frequency and account metadata serve as reliable deception signals.

Multilingual and Multimodal Potential: Cross-lingual and visual-text fusion models enhance adaptability.

Scalable Deployment Feasible: The Flask-based API supports integration into commercial review systems.

VI. CONCLUSION

This study presents a comprehensive, AI-powered framework for detecting fake product reviews using NLP and ML techniques. By incorporating deep contextual embeddings from BERT and hybrid feature engineering, the proposed system achieves high precision and interpretability.

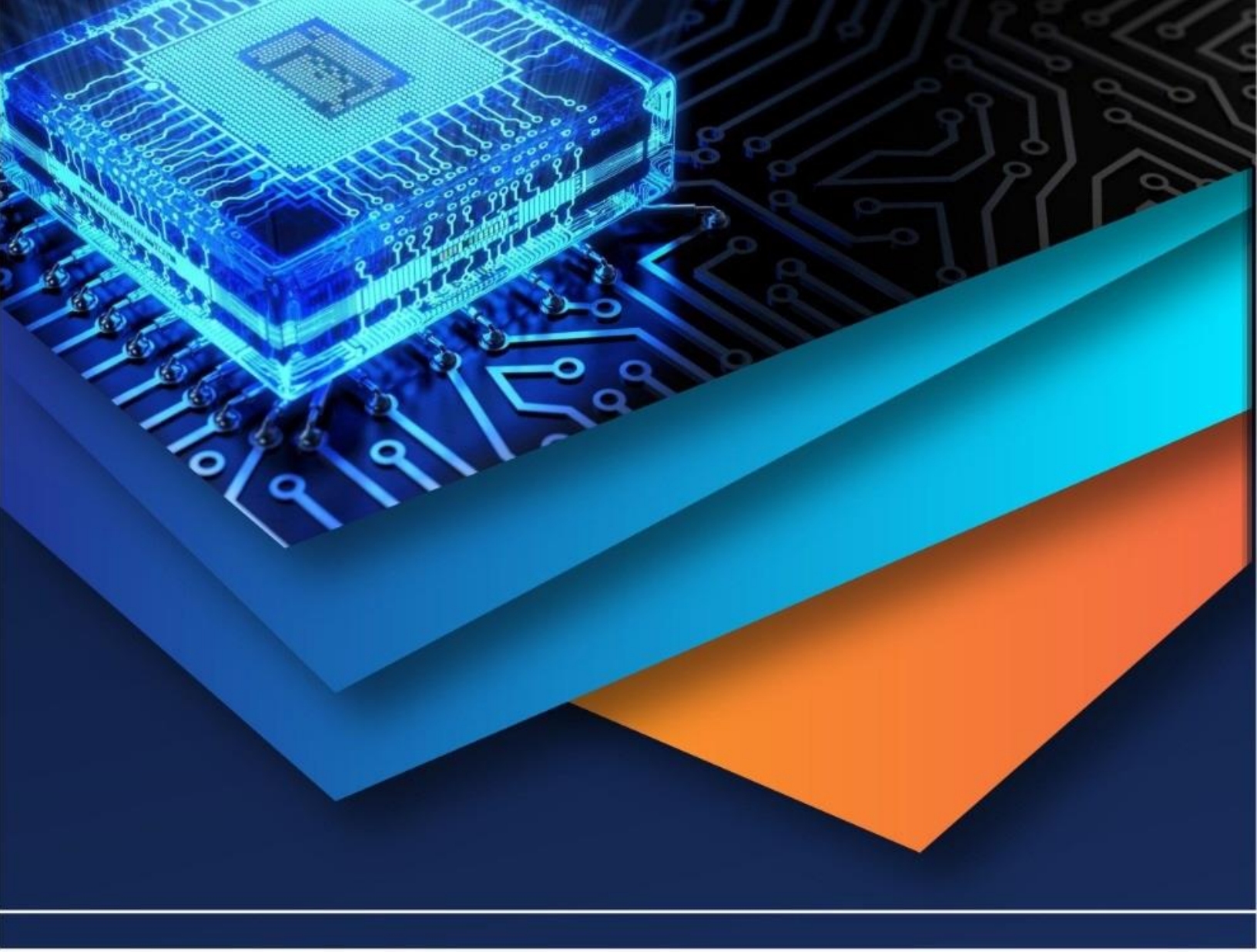
The findings demonstrate that automated detection of deceptive reviews is feasible and essential for maintaining the integrity of online marketplaces. Future work will focus on developing cross-domain transfer learning, adversarial resistance, and multimodal detection systems that analyze both text and images.

By ensuring authenticity and transparency in online feedback, this research contributes to improving consumer trust and ethical digital commerce.

REFERENCES

[1] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, “Findingdeceptive opinion spam by any stretch of the imagination,” Proc.ACL,2011.
[2] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” Foundations and Trends in Information Retrieval, 2008.
[3] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” NAACL, 2019.
[4] A. Vaswani et al., “Attention is all you need,” NeurIPS, 2017.
[5] J. Li, T. Zhu, and H. Hu, “Deep learning for fake review detection,” IEEE Conf., 2017.
[6] Y. Zhang and B. Varadarajan, “Utility scoring of product reviews,” Proc. ACM CIKM, 2006.
[7] N. Jindal and B. Liu, “Opinion spam and analysis,” Proc. WSDM, 2008.
[8] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, “What Yelp fake review filter might be doing,” Proc. ICWSM, 2013.
[9] S. Xie et al., “Review spam detection via temporal pattern discovery,” Proc. WWW, 2012.
[10] G. Wang et al., “Review graph-based online spam detection,” Proc. ICDM, 2011.
[11] E.-P. Lim et al., “Detecting product review spammers using rating behaviors,” Proc. CIKM, 2010.
[12] A. Heydari et al., “Detection of fake opinions using time series,” Expert Systems with Applications, 2016.
[13] A. Mukherjee and B. Liu, “Spotting fake reviewer groups in consumer reviews,” Proc. WWW Companion, 2013.
[14] Y. Cao et al., “Detecting deceptive reviews using CNN and LSTM,” IEEE Access, 2020.
[15] J. Sun et al., “Fake review detection using BERT and attention mechanism,” Knowledge-Based Systems, 2021.
[16] X. Sun, L. Xu, and M. Wang, “ALBERT for deceptive review detection,” Applied Intelligence, 2022.
[17] C. Wang et al., “An XLNet-based approach for online fake review detection,” IEEE Trans. Comput. Soc. Syst., 2022.
[18] Z. Chen et al., “DistilBERT with attention layer for detecting deceptive content,” Future Generation Computer Systems, 2023.
[19] A. Banerjee et al., “Multimodal fake review detection using visual and textual cues,” Information Fusion, 2024.
[20] T. Haque et al., “Hybrid models for fake review detection using BERT and metadata fusion,” IEEE Access, 2023.
[21] H. Li et al., “Domain adaptation for opinion spam detection using transfer learning,” Expert Systems with Applications, 2022.

[22] D. Zhou et al., "Cross-domain deceptive review detection with adversarial learning," Neural Networks, 2024.
[23] P. Kumar et al., "The threat of AI-generated fake reviews: A study of GPT-based systems," Computers & Security, 2024.
[24] C. Sun and D. Wu, "Explainable reinforcement learning for safety-critical applications," Pattern Recognition, 2023.
[25] A. Roy et al., "Trustworthy and transparent AI in intelligent review systems," AI Ethics, 2023.
[26] F. Zhang et al., "Attention-based interpretability for fake review classification," IEEE Trans. Affective Computing, 2025.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY