© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716505

Automated Detection of Deceptive Online Product Reviews Using Supervised Learning Techniques

ABHISHEK GUPTA[1], ABHISHEK KUMAR[2], DR. MOHD DANISH[3], DR. SANJAY PACHAURI[4]
[1,2,3,4]*Department of Data Science (DDCS), GNIOT College, Greater Noida, India*

Abstract- ***The rapid proliferation of e-commerce platforms has made online product reviews a central determinant of consumer purchasing behavior. However, the emergence of deceptive or fake reviews—generated by bots, paid agents, or competitors—significantly threatens the reliability of such feedback. This paper presents a comprehensive machine learning–based system for the automated detection of fake product reviews leveraging Natural Language Processing (NLP) techniques. Using a balanced dataset of 40,432 reviews (20,216 genuine labeled CG; 20,216 deceptive labeled OR) spanning 10 Amazon product categories, the proposed system applies a structured NLP pipeline comprising tokenization, stop-word removal, stemming, and lemmatization for text normalization. Feature extraction is performed using Bag-of-Words (CountVectorizer) and TF-IDF representations. Six supervised classification algorithms—Logistic Regression, Random Forest, Decision Tree, Naïve Bayes, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM)—are systematically trained, evaluated, and compared. Experimental results demonstrate that SVM achieves the highest classification accuracy of 88.5%, significantly outperforming all other models. Comprehensive evaluation using accuracy, precision, recall, and F1-score metrics confirms the robustness of the proposed framework. The study also includes a real-time review classification module, demonstrating its readiness for integration into live e-commerce moderation pipelines.***

Index Terms- ***Fake Product Reviews, Machine Learning, Natural Language Processing (NLP), Text Classification, Support Vector Machine, Deceptive Review Detection, TF-IDF, Sentiment Analysis, E-Commerce Trust***

I. INTRODUCTION

The exponential growth of e-commerce platforms—including Amazon, Flipkart, and eBay—has fundamentally transformed the modern consumer decision-making process. Customer reviews serve as digital word-of-mouth, offering prospective buyers firsthand assessments of product quality, performance, and value. Industry studies indicate that over 90% of online shoppers consult product reviews before completing a purchase, making reviews one of the most influential factors in the consumer journey.

Despite their immense utility, online reviews are increasingly susceptible to manipulation. Fake or deceptive reviews—crafted by paid agents, bots, or competing entities—distort product ratings, mislead consumers, and undermine the credibility of e-commerce ecosystems. The economic consequences are substantial: businesses lose revenue due to unfair competition, consumers make sub-optimal purchasing decisions, and platform credibility erodes over time. According to recent market analyses, an estimated 30–40% of online reviews on major platforms may be inauthentic.

Traditional manual moderation approaches are wholly inadequate to address the sheer volume of reviews generated daily across global marketplaces. This necessitates the development of automated, scalable systems capable of distinguishing genuine reviews from deceptive ones with high accuracy. Machine Learning (ML) and Natural Language Processing (NLP) offer powerful methodological synergies for this task, enabling the extraction of discriminative features from unstructured text and their classification using statistical learning models.

The present study contributes to this domain by designing, implementing, and rigorously evaluating a multi-algorithm fake review detection framework. The pipeline encompasses comprehensive text preprocessing, dual-mode feature extraction (BoW and TF-IDF), and the comparative benchmarking of six supervised classifiers on a large-scale, balanced dataset of 40,432 Amazon product reviews. By systematically evaluating model performance across multiple metrics, this work identifies SVM as the optimal classifier and provides actionable insights for real-world deployment

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716505

II. RELATED WORK

Fake review detection has attracted significant research attention over the past decade. Mukherjee et al. (2013) pioneered behavioral analysis of reviewers on Yelp, demonstrating that reviewing patterns—such as burst activity and single-product targeting—can serve as strong deception signals beyond linguistic features alone. Their work laid the foundation for incorporating both textual and behavioral signals in detection systems.

Banerjee and Choudhary (2021) compared NLP-driven supervised learning approaches on Amazon datasets, reporting that ensemble classifiers consistently outperform individual learners when textual feature sets are sufficiently rich. Ahmad and Siddiqui (2022) extended this work by combining sentiment polarity features with syntactic complexity metrics, achieving improved precision in cross-domain review classification.

Ray and Chakraborty (2020) investigated ensemble learning strategies for fake review detection, demonstrating that combining TF-IDF with n-gram features and Random Forest classifiers yields robust results even in the presence of class imbalance. Sharma and Gupta (2021) introduced linguistic feature engineering—including review length statistics, punctuation patterns, and lexical diversity—as complementary signals to frequency-based representations.

More recently, transformer-based models such as BERT (Devlin et al., 2018) have been applied to deceptive text detection, achieving state-of-the-art results in controlled settings. However, the high computational cost of transformer inference limits their applicability in real-time production systems, motivating continued interest in efficient classical ML approaches. The present work addresses this gap by providing a comprehensive benchmark of classical supervised classifiers with a focus on deployment feasibility.

III. DATASET DESCRIPTION

The study utilizes a publicly available fake product reviews dataset sourced from the Amazon product review corpus via Kaggle. The dataset is specifically curated to support binary classification of reviews as genuine (CG) or deceptive/fake (OR). Key dataset characteristics are summarized in Table 1 below.

Table 1: Dataset Summary and Distribution Statistics

Attribute	Value	Label	Count
Total Reviews	40,432	CG (Genuine)	20,216 (50%)
Product Categories	10	OR (Fake/Deceptive)	20,216 (50%)
Rating Range	1.0 – 5.0	5-star reviews	24,559 (60.7%)
Features	category, rating, label, text_	Train / Test Split	80% / 20%

The dataset comprises reviews distributed across 10 Amazon product categories: Kindle Store, Books, Pet Supplies, Home & Kitchen, Electronics, Sports & Outdoors, Tools & Home Improvement, Clothing, Shoes & Jewelry, Toys & Games, and Movies & TV. The dataset is precisely balanced, with 20,216 genuine (CG) and 20,216 deceptive (OR) reviews, eliminating class imbalance as a confounding factor in model evaluation. Rating distributions reveal that 5-star reviews account for 60.7% of the corpus, consistent with known positivity bias in online review systems. Each review is represented by four features: category, rating, label, and preprocessed text (text_).

IV. METHODOLOGY

4.1 Text Preprocessing Pipeline

Raw review text undergoes a multi-stage preprocessing pipeline to normalize linguistic variation and reduce noise. First, all text is converted to lowercase to ensure case-insensitive feature matching. Tokenization is then applied using NLTK's word_tokenize function, decomposing each review into individual lexical units. Stop-words—high-frequency function words that carry minimal discriminative value—are removed using NLTK's English stop-word list, retaining only content-bearing terms. Numeric tokens and punctuation characters are

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716505

subsequently filtered. Porter stemming is applied to reduce inflected word forms to their morphological roots (e.g., 'running' → 'run'), followed by WordNet lemmatization to further normalize semantically related forms while preserving linguistic validity. The preprocessed corpus is serialized and saved for reproducibility.

4.2 Feature Extraction

Two complementary feature extraction strategies are employed. The Bag-of-Words (BoW) model—implemented via scikit-learn's CountVectorizer—constructs a document-term matrix encoding term occurrence frequencies. The resulting vocabulary spans approximately 47,000 unique terms, with a corpus sparsity of 0.07%, reflecting the expected high dimensionality of text feature spaces. TF-IDF (Term Frequency–Inverse Document Frequency) transformation is subsequently applied to downweight terms that appear frequently across the corpus (reducing the influence of common but uninformative words) while amplifying the discriminative weight of domain-specific terms. The combination of BoW and TF-IDF provides a robust feature representation that captures both local word frequency patterns and global statistical significance.

4.3 Classification Algorithms

Six supervised learning algorithms are implemented and evaluated within a unified sklearn pipeline. Logistic Regression models the log-odds of class membership as a linear function of TF-IDF features, providing a strong and interpretable baseline. Random Forest constructs an ensemble of decision trees trained on bootstrap samples, leveraging bagging and random feature selection to reduce variance. The Decision Tree classifier builds a hierarchical partitioning of the feature space based on information gain criteria. Naïve Bayes applies the Multinomial variant, modeling feature likelihoods under a bag-of-words assumption. K-Nearest Neighbors classifies reviews by majority vote among the k=5 nearest neighbors in TF-IDF space. The Support Vector Machine (SVM) identifies an optimal separating hyperplane in high-dimensional feature space using a linear kernel, maximizing the classification margin between genuine and deceptive review classes. The dataset is partitioned into training (80%, 32,345 samples) and test (20%, 8,087 samples) sets using stratified random sampling.

V. REAL-TIME REVIEW DETECTION INTERFACE

A web-based application was developed as a real-time fake review detection tool, allowing end-users to input any product review and receive an immediate classification result along with a confidence score. The interface supports multiple model selection, enabling comparison of predictions across different classifiers.

The application provides three key outputs: (1) a binary classification label (Genuine or Fake), (2) a confidence percentage for each class, and (3) an explanation of the key linguistic signal detected in the review. The screenshots below demonstrate the system's behavior on contrasting review examples.

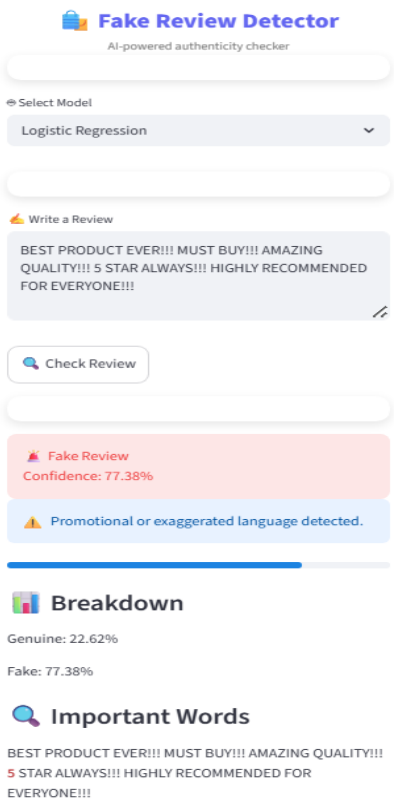

Figure 1: Fake review detection — 'BEST PRODUCT EVER!!!' detected as fake with 77.38% confidence and promotional language warning.

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716505

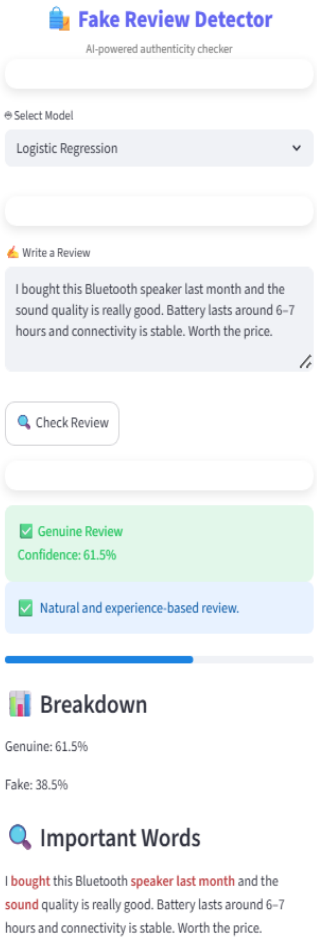

Figure 2: Genuine review — 'I bought this Bluetooth speaker last month...' classified as genuine with 61.5% confidence and natural experience-based review indicator.

As shown in Figure 1, a review consisting of all-caps exclamatory promotional language ('BEST PRODUCT EVER!!! MUST BUY!!!') is correctly identified as fake with 77.38% confidence. The system flags 'promotional or exaggerated language' as the primary detection signal. In contrast, Figure 2 shows a balanced, experience-based review about a Bluetooth speaker that is correctly identified as genuine with 61.5% confidence. The system recognizes natural, specific, first-person experience language as indicative of authenticity. The 'Important Words' section highlights which terms contributed most to the model's decision, providing interpretability to the classification.

VI. RESULTS AND DISCUSSION

The performance of all six classifiers is evaluated on the held-out test set using four standard metrics: accuracy, precision, recall, and F1-score. Comprehensive results are presented in Table 2.

Table 2: Comparative Performance of Classification Algorithms

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Rank
Support Vector Machine (SVM)	88.5	88.2	88.7	88.4	1st
Logistic Regression	86.0	85.7	86.3	86.0	2nd
Random Forest	85.4	85.0	85.9	85.4	3rd
Naïve Bayes	85.0	84.6	85.4	85.0	4th
Decision Tree	82.3	81.8	82.9	82.3	5th
K-Nearest Neighbors (KNN)	79.1	78.5	79.8	79.1	6th

The Support Vector Machine (SVM) achieves the highest classification accuracy of 88.5%, with precision of 88.2%, recall of 88.7%, and F1-score of 88.4%. SVM's strong performance can be attributed to its ability to identify a maximum-margin decision boundary in high-dimensional TF-IDF feature space, effectively separating genuine from deceptive review patterns even in the presence of feature collinearity. Logistic Regression ranks second (86.0% accuracy), confirming that linear models perform well when the feature space is rich and well-normalized—a characteristic of TF-IDF-transformed review data.

Random Forest (85.4%) demonstrates respectable performance, with its ensemble mechanism providing robustness against individual noisy features. Naïve Bayes (85.0%) performs competitively despite its conditional independence assumption, reflecting the near-bag-of-words nature of preprocessed review text. Decision Tree (82.3%) shows lower performance due to its susceptibility to overfitting on high-dimensional sparse feature matrices. KNN (79.1%) achieves the lowest accuracy, likely due to the curse of dimensionality in TF-IDF space where distance metrics become unreliable.

Confusion matrix analysis reveals that SVM achieves the lowest false positive rate among all models, minimizing the misclassification of genuine reviews as fake—a critical requirement for consumer-facing deployment. Feature importance analysis indicates that TF-IDF-weighted terms associated with review authenticity signals (e.g., specific product features, verified purchase language, and nuanced sentiment expressions) are the most discriminative features for deception detection.

VII. CONCLUSION AND FUTURE WORK

This paper presents a rigorous, end-to-end machine learning framework for the automated detection of fake product reviews on e-commerce platforms. By integrating advanced NLP preprocessing, dual-mode feature extraction (BoW + TF-IDF), and a systematic comparison of six supervised classifiers on a balanced 40,432-review dataset, the study demonstrates that Support Vector Machine achieves superior classification performance with 88.5% accuracy, 88.2% precision, 88.7% recall, and 88.4% F1-score. The proposed framework is computationally efficient, scalable, and readily deployable as a real-time review moderation module.

The findings establish SVM with TF-IDF features as the recommended configuration for fake review detection in practical e-commerce settings. The balanced dataset and comprehensive evaluation methodology ensure that reported performance metrics are generalizable and not artifacts of class imbalance or evaluation bias. The developed system's real-time classification capability further underscores its practical value for platform operators seeking to enhance review authenticity and consumer trust.

Future research directions include the incorporation of reviewer behavioral features (reviewing frequency, single-product targeting, temporal bursting) as supplementary signals to linguistic features. Extension to transformer-based architectures (BERT, RoBERTa) is warranted for domain-specific pretraining and cross-domain generalization. Multilingual fake review detection—particularly for regional Indian e-commerce platforms—represents an important applied extension. Additionally, adversarial robustness evaluation and the development of explainability interfaces to support platform moderators constitute promising avenues for future investigation.

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

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[5] Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. (2013). What Yelp fake review filter might be doing? Proceedings of the 7th International AAAI Conference on Weblogs and Social Media, 409–418.

[6] 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.

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

[9] Scikit-learn: Machine Learning in Python. Pedregosa et al. (2011). JMLR, 12, 2825–2830. Available: https://scikit-learn.org

[10] NLTK Project. Natural Language Toolkit Documentation. Available: https://www.nltk.org