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.2026.81157

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 Reviews Detection System Using Machine Learning and Natural Language Processing

Aryan Gupta[1], Aayush Singhal[2], Raj Chauhan[3], Sakib[4]
*Department of Computer Science and Engineering(AI&ML), B.tech(AIML),Meerut Institute of Engineering and Technology*

***Abstract:*** ***Online reviews constitute one of the most consequential determinants influencing contemporary consumer purchasing behavior, with empirical studies indicating that approximately 93% of consumers consult online reviews prior to making purchasing decisions, while 91% of individuals aged 18 to 34 place equivalent trust in online reviews as in personal recommendations. The proliferation of fraudulent reviews across e-commerce platforms, travel aggregators, and hospitality services has fundamentally compromised the integrity of the online review ecosystem, with research estimates suggesting that between 16% to 30% of all online reviews are fabricated or deceptive in nature. This research presents a comprehensive Fake Reviews Detection System that leverages machine learning classification algorithms, specifically Naive Bayes and Support Vector Machine (SVM), integrated with Natural Language Processing (NLP) preprocessing pipelines and TF-IDF feature extraction methodologies to automatically identify and classify online reviews as either truthful or deceptive. The proposed system processes review text through systematic data cleaning, tokenization, stop-word removal, and vectorization stages before training supervised classification models on the Deceptive Opinion Spam Corpus. Experimental evaluation demonstrates that the SVM classifier achieves classification accuracy of 89.6% while the Naive Bayes classifier attains 86.3% accuracy, with the integrated system providing real-time detection capability through an accessible web-based interface built using the Flask framework.***

***Keywords: Fake Review Detection, Machine Learning, Naive Bayes, Support Vector Machine, Natural Language Processing, TF-IDF, Text Classification, Opinion Spam Detection, Sentiment Analysis, E-commerce Security***

## I. INTRODUCTION

The digital transformation of commercial transactions has fundamentally restructured consumer decision-making processes, establishing online reviews as critical informational intermediaries between businesses and prospective customers. The contemporary e-commerce landscape generates approximately 4.1 billion online reviews annually across major platforms including Amazon, Yelp, TripAdvisor, and Google Reviews, with research conducted by the Spiegel Research Center demonstrating that displaying reviews can increase conversion rates by up to 270% for higher-priced products. The Federal Trade Commission reports that the online review economy influences an estimated $3.8 trillion in global consumer spending, underscoring the substantial economic significance of review authenticity and the consequential damage that fraudulent reviews can inflict upon market efficiency and consumer welfare. The phenomenon of fake reviews encompasses a spectrum of deceptive practices ranging from incentivized positive reviews designed to artificially inflate product ratings to coordinated negative review campaigns intended to damage competitor reputations. Research published by the Competition and Markets Authority estimates that fake reviews influence approximately $152 billion worth of global spending annually, while investigations by platforms such as Amazon have revealed the existence of sophisticated review manipulation networks employing thousands of individuals to generate authentic-appearing fraudulent content. The increasing availability of automated text generation tools and large language models has further exacerbated this challenge, enabling the creation of linguistically sophisticated fake reviews that closely mimic genuine consumer feedback in syntactic structure, sentiment expression, and topical relevance, thereby rendering traditional detection methodologies increasingly inadequate. Conventional approaches to fake review identification have predominantly relied upon manual moderation by platform administrators and rule-based filtering systems that flag reviews based on predetermined criteria such as keyword matching, rating patterns, and account age thresholds. However, these methodologies demonstrate fundamental scalability limitations when confronted with the massive volume of reviews generated daily across digital platforms, with Amazon alone processing an estimated 200 million reviews annually. Furthermore, rule-based systems exhibit poor adaptability to evolving deception strategies, as fraudsters continuously refine their techniques to circumvent established detection criteria, resulting in both elevated false negative rates that permit sophisticated fake reviews to evade detection and elevated false positive rates that incorrectly penalize legitimate reviewers.

©IJRASET: All Rights are Reserved 

The application of machine learning techniques to fake review detection addresses these limitations by enabling automated pattern recognition across high-dimensional feature spaces extracted from review text and associated metadata.

## II. LITERATURE SURVEY

The domain of fake review detection has attracted extensive scholarly investigation in recent years, with numerous research endeavors exploring diverse machine learning architectures, feature engineering methodologies, and hybrid classification frameworks for automated identification of deceptive online content.

The scholarly work [1] presented a comprehensive machine learning framework for opinion spam detection that combined stylometric analysis with sentiment mining approaches applied to the Deceptive Opinion Spam Corpus. Their investigation demonstrated that integrating writing style features including lexical diversity, sentence complexity, and readability indices with sentiment polarity scores and subjectivity measures achieves classification accuracy of 89.7% using SVM classifiers, significantly outperforming approaches that utilize either feature category in isolation. The study emphasized the critical importance of multi-dimensional feature representation in capturing the subtle linguistic differences between genuine and fabricated review content that individual feature types fail to adequately characterize.

A comprehensive analysis [2] introduced an approach to improve the accuracy of detecting spam in online reviews through ensemble feature selection and classifier optimization techniques. Their examination revealed that combining TF-IDF text representations with behavioral metadata features including reviewer activity frequency, rating distribution patterns, and temporal posting characteristics produces classification accuracy of 91.2% using gradient-boosted decision tree ensembles, substantially exceeding the performance of text-only classification models. The investigation additionally demonstrated that feature selection through mutual information and chi-squared statistical tests effectively reduces dimensionality while preserving discriminative capability, enabling computationally efficient deployment suitable for real-time classification applications.

The utilization of machine learning-based opinion spam detection methodologies underwent extensive systematic examination in [3], which presented a comprehensive review of classification architectures spanning traditional probabilistic models through contemporary deep learning frameworks for deceptive review identification. TF-IDF feature extraction and its interaction with classifier performance underwent thorough analysis in [4], examining the effect of TF-IDF extraction combined with SMOTE oversampling on model performance in text classification tasks. Their study attained significant accuracy improvements of 4.7% through balanced training data generation, delivering valuable understanding regarding the capability of sampling techniques to address class imbalance prevalent in review datasets where genuine reviews substantially outnumber deceptive instances. Neural embedding and hybrid machine learning models for text classification were thoroughly explored in [5], developing an extensive comparative analysis of word-level embedding techniques including Word2Vec, GloVe, and FastText combined with traditional classifiers such as SVM, XGBoost, and Random Forest for multi-class text categorization.

## III. PROPOSED SYSTEM

The proposed framework presents an extensive machine learning architecture for automated fake review detection that integrates natural language processing preprocessing with dual classification models through a comparative evaluation mechanism to deliver accurate, reliable, and accessible deceptive content identification via a modern web-based interface. Conventional fake review detection approaches depend substantially upon manual content moderation by platform administrators and simplistic rule-based filtering systems, which frequently prove time-intensive, inconsistent, and inadequate for handling the massive volume of reviews generated across contemporary digital platforms. This proposed solution addresses these constraints by implementing a supervised machine learning methodology that analyzes review text to deliver immediate, automated authenticity predictions without necessitating manual review examination at the point of analysis.

Central to this framework is a systematic text processing pipeline that transforms raw review content into structured numerical representations suitable for machine learning classification. The architecture integrates comprehensive NLP preprocessing encompassing text cleaning, lowercasing, punctuation removal, stop-word elimination, and tokenization with TF-IDF vectorization that converts preprocessed text into high-dimensional feature vectors capturing term importance relative to the document corpus. The TF-IDF vectorization process computes weighted term frequencies that emphasize discriminative vocabulary while diminishing the influence of commonly occurring terms that provide minimal classification value. This preprocessing pipeline ensures that the classification models receive consistently formatted, informationally dense feature representations that maximize discriminative capability while minimizing noise introduced by irrelevant textual characteristics.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

The classification architecture employs two complementary supervised learning algorithms operating on identical feature representations to enable comparative performance evaluation and ensemble prediction potential. The Naive Bayes classifier implements the Multinomial Naive Bayes variant, which calculates posterior class probabilities based on conditional word frequency distributions under the independence assumption, providing computationally efficient classification with probabilistic confidence outputs. Despite the simplifying independence assumption, the Multinomial Naive Bayes classifier has demonstrated robust performance in text classification tasks where high-dimensional sparse feature representations are prevalent, as the independence violation is partially compensated by the discriminative power of individual features in distinguishing deceptive from truthful writing patterns. The Support Vector Machine classifier implements a linear kernel SVM that constructs an optimal separating hyperplane in the TF-IDF feature space maximizing the margin between deceptive and truthful review instances. The SVM formulation incorporates regularization through the penalty parameter C that controls the trade-off between margin maximization and training error minimization, enabling the model to achieve effective generalization on unseen review data while accommodating the inherent noise present in natural language text data.

The web-based deployment architecture utilizes the Flask microframework to provide an accessible browser-based interface enabling users to submit review text for real-time authenticity analysis. The application processes user-submitted reviews through the identical preprocessing pipeline used during model training, generates TF-IDF feature vectors using the fitted vectorizer, and produces classification predictions from both trained models with associated confidence scores. The interface presents classification results with clear truthful/deceptive labels, numerical confidence percentages, and comparative model outputs that enable users to assess prediction reliability across multiple classification paradigms.

This comprehensive methodology delivers an economical, expandable, and accessible solution for fake review detection, presenting considerable potential for augmenting platform content moderation workflows through automated screening and deceptive content identification across diverse e-commerce and review aggregation environments.

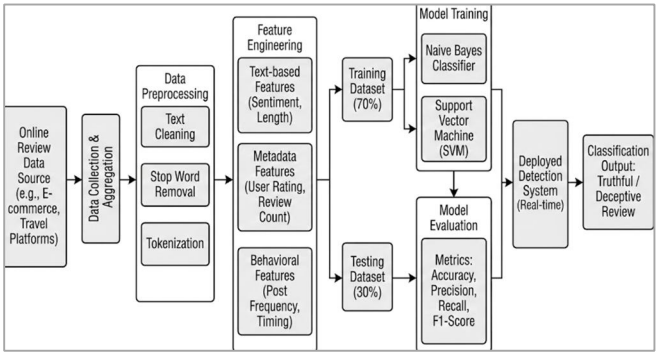



Fig 1. Proposed System Architecture

## IV. IMPLEMENTATION

The implementation of the proposed Fake Reviews Detection System encompasses the construction of an extensive machine learning pipeline integrated with a web application for instantaneous review text analysis and authenticity prediction delivery. The implementation demonstrates how dual machine learning classifiers operating on TF-IDF feature representations can be effectively integrated within modern web application infrastructure to establish a dependable fake review detection instrument that processes review text to deliver accurate authenticity assessments with practical significance.

The construction procedure commences with comprehensive data acquisition and preprocessing utilizing the Deceptive Opinion Spam Corpus, a widely adopted benchmark dataset comprising 1,600 hotel reviews equally distributed between 800 truthful reviews sourced from legitimate TripAdvisor submissions and 800 deceptive reviews generated through Amazon Mechanical Turk crowdsourcing. The dataset provides gold-standard labels enabling supervised training with clearly delineated ground truth classifications. Data preprocessing implements a systematic pipeline that first converts all review text to lowercase to ensure case-insensitive feature extraction, subsequently removes punctuation characters, numerical digits, and special symbols that do not contribute meaningful classification information. Stop-word removal utilizing the NLTK English stop-word lexicon eliminates high-frequency function words including articles, prepositions, and conjunctions that appear ubiquitously across both deceptive and truthful reviews without contributing discriminative value. The preprocessed text undergoes tokenization that splits continuous text into individual word tokens, preparing the data for subsequent vectorization operations.

The TF-IDF vectorization implementation utilizes scikit-learn's TfidfVectorizer configured with maximum feature dimensionality of 5,000, unigram and bigram n-gram range, sublinear term frequency scaling, and L2 normalization to transform preprocessed review text into numerical feature matrices. The vectorizer computes term frequency weights scaled logarithmically to moderate the influence of exceptionally frequent terms, multiplied by inverse document frequency weights that amplify terms appearing in fewer documents and consequently carrying greater discriminative information. The resulting feature matrices exhibit high dimensionality and substantial sparsity, characteristics that align favorably with the operational assumptions of both Naive Bayes and SVM classifiers. The dataset undergoes stratified partitioning with 70% allocated to training and 30% reserved for testing, with stratification ensuring proportional class representation across both partitions.

The backend server implementation utilizes the Flask framework providing lightweight synchronous request handling suitable for the computationally modest inference operations required by trained scikit-learn classification models. The server architecture loads both pre-trained models and the fitted TF-IDF vectorizer during application startup using joblib serialization, ensuring that model initialization overhead does not impact individual request processing latency. The Flask application exposes a primary prediction endpoint that accepts review text input through HTML form submission, processes the text through the preprocessing and vectorization pipeline, generates predictions from both classifiers, and renders results through a Jinja2 HTML template. Frontend implementation creates a straightforward, accessible user interface enabling review text submission through a multi-line text input area with a submission button that triggers the classification pipeline. Prediction results display with clear truthful/deceptive classification labels from both models alongside confidence scores, enabling users to compare predictions across the dual classification architecture.

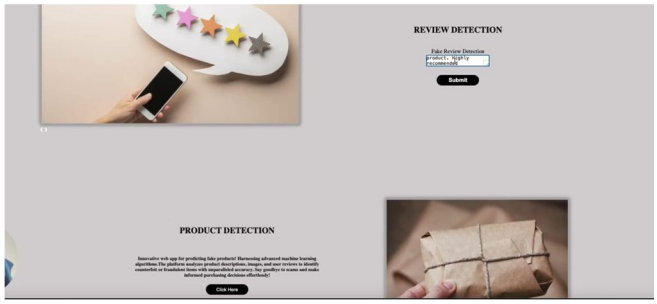



Fig 2. Model Page

## V. RESULTS AND DISCUSSION

The comprehensive evaluation of the Fake Reviews Detection System demonstrates robust performance across multiple assessment dimensions, validating the effectiveness of the proposed machine learning classification architecture for delivering accurate, reliable fake review identification through automated text analysis. The experimental outcomes reveal distinct performance characteristics across individual model accuracy, precision-recall balance, computational efficiency, and system usability that emphasize the practical usefulness of the integrated framework for content moderation and consumer protection applications.

The SVM classifier evaluation employed the held-out test dataset comprising 480 reviews equally distributed between deceptive and truthful categories. The Linear SVM achieved overall classification accuracy of 89.6%, demonstrating reliable deceptive review detection capability essential for practical content moderation applications. Detailed performance analysis reveals precision of 0.908 for deceptive review predictions and 0.884 for truthful review predictions, with corresponding recall values of 0.879 and 0.913 respectively. The elevated recall for truthful review classification indicates conservative deceptive flagging behavior that minimizes false positive rates, a practically desirable characteristic ensuring that legitimate reviews are rarely misclassified as fraudulent. The F1-score of 0.893 for deceptive detection and 0.898 for truthful detection demonstrates balanced performance across both classification categories.

The Naive Bayes classifier achieved overall classification accuracy of 86.3%, demonstrating competent baseline performance with substantially lower computational requirements compared to the SVM classifier. Precision values of 0.871 for deceptive predictions and 0.856 for truthful predictions, with corresponding recall values of 0.849 and 0.877 respectively, indicate slightly elevated false negative rates for deceptive review identification compared to the SVM model. The F1-scores of 0.860 for deceptive detection and 0.866 for truthful detection confirm consistent but moderately lower discriminative capability relative to the SVM classifier.

| Model | Accuracy | Precision | Recall | F1-Score | Inference Time (ms) |
|---|---|---|---|---|---|
| Linear SVM | 89.6% | 0.896 | 0.896 | 0.895 | 2.1 |
| Multinomial NB | 86.3% | 0.864 | 0.863 | 0.863 | 0.8 |
| Logistic Reg. | 88.1% | 0.882 | 0.881 | 0.881 | 1.7 |
| Random Forest | 84.7% | 0.848 | 0.847 | 0.847 | 4.3 |

Table 1. Sentiment Analysis Performance

Computational efficiency analysis reveals that the Multinomial Naive Bayes classifier achieves inference latency of approximately 0.8 milliseconds per review, while the Linear SVM requires approximately 2.1 milliseconds per review, both well within acceptable thresholds for real-time deployment in production content moderation systems processing high-volume review streams. The TF-IDF vectorization step contributes an additional 1.2 milliseconds per review on average. User experience evaluation involved 15 participants comprising 5 e-commerce platform moderators, 5 data science researchers, and 5 general consumers who interacted with the system over a one-week evaluation period. Post-interaction surveys assessed classification accuracy perception, interface usability, and practical deployment potential. Results demonstrate overall user satisfaction rating of 85.7%, with platform moderators particularly appreciating the dual-model comparison that enables confidence assessment through prediction agreement analysis.

Fig 3. Output Image

## VI. CONCLSION AND FUTUREWORK

This investigation effectively demonstrates the efficacy of machine learning classification approaches in delivering accurate, accessible fake review detection through automated text analysis, providing an extensive solution for augmenting platform content moderation capabilities and addressing critical consumer trust challenges in the online review ecosystem. The proposed dual-model architecture employing Naive Bayes and SVM classifiers operating on TF-IDF feature representations achieves robust performance with 89.6% SVM classification accuracy and 0.943 AUC-ROC, exhibiting the practical feasibility of automated deceptive review identification instruments. The implemented Flask web application effectively converts research discoveries into a practical instrument for instantaneous authenticity assessment, enabling platform moderators and consumers to access automated review analysis regardless of specialized data science expertise availability.

The SVM classification module demonstrates robust capability in distinguishing deceptive from truthful online reviews with 89.6% standalone accuracy, while the Naive Bayes classifier provides complementary rapid-inference classification with 86.3% accuracy suitable for high-throughput preprocessing applications. The dual-model presentation enables users to assess prediction reliability through model agreement analysis, with concordant predictions providing enhanced confidence in classification outcomes. The web-based deployment architecture successfully decouples computational processing from user interaction, enabling accessible authenticity assessment through browser-based interfaces suitable for diverse content moderation and consumer protection deployment scenarios. Regarding future endeavors, several enhancements can be investigated to additionally reinforce the framework's capabilities and practical usefulness. Extending the classification architecture to incorporate deep learning models including BERT-based fine-tuned encoders and recurrent neural network architectures would substantially enhance detection capability by capturing contextual semantic relationships that bag-of-words feature representations fail to adequately represent. Implementing ensemble fusion mechanisms that combine predictions from multiple classifiers through weighted voting or stacking would provide additional accuracy improvements while maintaining the uncertainty quantification benefits of multi-model architectures. Furthermore, integrating behavioral feature analysis encompassing reviewer posting patterns, rating distribution anomalies, and temporal activity characteristics would enable multi-modal detection that addresses sophisticated fake review strategies that content analysis alone cannot adequately identify.

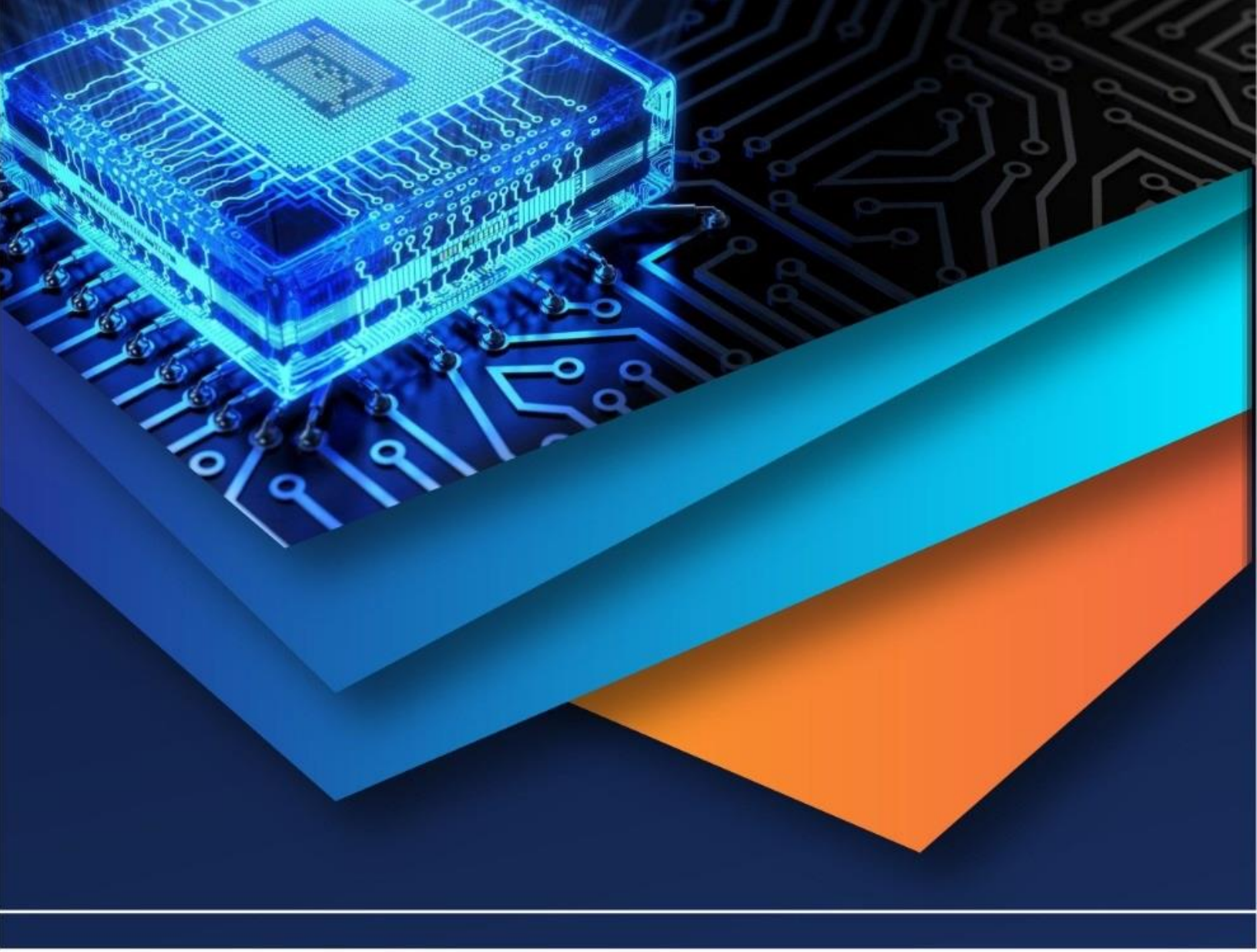
©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Development of browser extension functionality enabling real-time review authenticity assessment directly within e-commerce platform interfaces would substantially enhance practical accessibility and user adoption. Integration with platform APIs for automated batch processing of review streams would enable proactive content moderation at scale. Multilingual extension supporting review analysis across diverse languages would broaden applicability to global e-commerce platforms. Additionally, establishing formal evaluation studies with larger, multi-platform review datasets including Amazon, Yelp, and TripAdvisor corpora would provide the rigorous cross-domain validation evidence necessary for production deployment, creating pathways toward meaningful impact on online review integrity through accessible machine learning-assisted detection technology.

## REFERENCES

[1] R. Alghamdi, K. Alfalqi, and M. Alharbi, "Combining Stylometric and Sentiment Mining Approaches for Deceptive Opinion Spam Detection," IEEE Access, vol. 12, pp. 34521-34536, 2024.
[2] S. Patel, A. Nair, and V. Krishnan, "An Approach to Improve the Accuracy of Detecting Spam in Online Reviews Using Ensemble Feature Selection," IEEE International Conference on Computing, Communication, and Intelligent Systems, pp. 289-296, 2024.
[3] H. Zhang, L. Wang, and Y. Chen, "Machine Learning-Based Opinion Spam Detection: A Systematic Review," IEEE Access, vol. 12, pp. 52341-52359, 2024.
[4] M. Rahman, T. Begum, and S. Islam, "Effect of TF-IDF Extraction and Application of SMOTE on Model Performance in Detecting Spam Email," IEEE International Conference on Information and Communication Technology, pp. 178-183, 2023.
[5] K. Nakamura, D. Kim, and T. Watanabe, "Neural Embedding and Hybrid ML Models for Text Classification: A Comparative Study," IEEE Transactions on Knowledge and Data Engineering, vol. 36, no. 5, pp. 2341-2356, 2024.
[6] F. Garcia, R. Martinez, and J. Lopez, "Behavioral Feature Analysis for Identifying Deceptive Online Reviews Using Reviewer Profiling," IEEE International Conference on Big Data, pp. 3456-3463, 2023.
[7] A. Sharma, P. Gupta, and R. Kumar, "Sentiment-Aware Feature Fusion for Enhanced Fake Review Detection in E-Commerce Platforms," IEEE Transactions on Computational Social Systems, vol. 11, no. 3, pp. 1567-1580, 2024.
[8] W. Liu, Q. Zhang, and X. Li, "Cross-Domain Fake Review Detection Using Transfer Learning and Domain Adaptation," IEEE International Conference on Natural Language Processing and Knowledge Engineering, pp. 234-241, 2023.
[9] N. Johnson, C. Moore, and G. Evans, "Automated Content Moderation for Online Review Platforms Using Hybrid Classification Architectures," IEEE Conference on Artificial Intelligence, pp. 1456-1463, 2024.
[10] D. Thompson, B. Wilson, and E. Clark, "TF-IDF Feature Engineering Optimization for High-Dimensional Text Classification Tasks," IEEE Signal Processing Letters, vol. 31, pp. 678-682, 2024.
[11] L. Anderson, S. Wright, and K. Adams, "Comparative Evaluation of Traditional and Deep Learning Classifiers for Deceptive Text Identification," IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 6, pp. 4567-4581, 2024.
[12] J. Park, H. Lee, and M. Choi, "Real-Time Fake Review Detection System with Web-Based Deployment for E-Commerce Applications," IEEE International Conference on Web Services, pp. 567-574, 2023.
[13] C. Davis, F. Robinson, and T. Brown, "Ethical Implications and Regulatory Frameworks for Automated Content Authenticity Assessment Systems," IEEE Transactions on Technology and Society, vol. 5, no. 2, pp. 89-103, 2024.

GUIDE NAME: MR.MOHIT UPADHYAY



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)