WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Fake Review Detection

Shyam Kumar B[1], Adan Kriston S[2], Krishnan G[3],Vignesh M[4], Dr. Lavanya M[5].

[1]UGStudent,Department of Artificial Intelligence and DataScience,Kings Engineering College, Chennai,TamilNadu 602 117, India,
[2]UGStudent,Department of Artificial Intelligence and DataScience,Kings Engineering College, Chennai,TamilNadu 602 117, India,
[3]UGStudent,Department of Artificial Intelligence and DataScience,Kings Engineering College, Chennai,TamilNadu 602 117, India,
[4]UGStudent, Department of Artificial Intelligence and DataScience, Kings Engineering College, Chennai,TamilNadu 602 117, India,
[5]AssistantProfessor,DepartmentofArtificial Intelligence and Machine Learning, Kings Engineering College, Chennai,TamilNadu,602 117, India,
[1]shyamspark306@gmail.com , [2]adankriston10@gmail.com , [3]krishnan1864@gmail.com , [4]mvignesh1408@gmail.com , [5]lavanya@kingsedu.ac.in

ABSTRACT

The increasing volume of deceptive product reviews on e-commerce platforms poses a serious threat to consumer trust and informed decision-making. This work focuses on detecting fake reviews by leveraging Natural Language Processing (NLP) and supervised machine learning techniques. The system processes review content by removing noise such as punctuation and stopwords, followed by feature extraction using TF-IDF vectorization. It then utilizes trained classifiers—Random Forest, Support Vector Machine (SVC), and Logistic Regression—to analyze the textual patterns and classify the review as genuine or fake. A voting-based ensemble approach ensures enhanced prediction reliability. The detection system is deployed through a user-friendly Streamlit interface that allows real-time input and displays the classification output to users as either "real" or "spam." The platform operates without the need for additional metadata or hardware, relying solely on the content of the review, making it lightweight and scalable. The primary goal is to improve the integrity of online feedback systems by providing an accessible and automated solution for review validation. This work contributes to creating a more transparent digital marketplace and helps protect users from misinformation by detecting and flagging manipulated reviews in real time.

Keywords:Fake Review Detection, Natural Language Processing, Machine Learning, Text Classification, TF-IDF, Ensemble Models, Streamlit Interface, Review Authenticity

Introduction

The widespread presence of fake product reviews on e-commerce platforms has become an increasingly complex issue, often involving deceptive tactics such as review spamming, coordinated content manipulation, and bot-generated feedback. These forged reviews are crafted to appear legitimate, making it difficult for users and automated systems to distinguish between genuine and fraudulent content. This work explores the application of machine learning and Natural Language Processing (NLP) to detect such deceptive patterns in review data. It introduces the concept of textual anomaly detection, which identifies reviews that exhibit linguistic or structural patterns significantly deviating from normal consumer feedback. The system uses a combination of preprocessing techniques and TF-IDF vectorization to transform unstructured review text into a format suitable for analysis. Supervised machine learning algorithms—including Random Forest, SVC, and Logistic Regression—are employed to classify reviews as either real or fake based on their content. The ensemble model enhances robustness by combining the strengths of multiple classifiers through a majority voting mechanism. Review input is processed in real-time through a Python-based backend and presented to users via a Streamlit interface. The dashboard delivers instant feedback to users while enabling storage and further tracking of flagged reviews. Unlike systems that rely on behavioral metadata or expensive third-party tools, this solution requires only review text, making it lightweight, privacy-conscious, and highly scalable. By automating the detection of forged reviews, this work aims to promote trust and transparency in digital marketplaces and provide e-commerce platforms with a reliable method for maintaining content integrity.

Review of Literature

Fake reviews have become a pressing concern for e-commerce platforms, prompting extensive research in Natural Language Processing (NLP) and Machine Learning (ML) for automated detection. A significant body of work focuses on analyzing textual features to identify linguistic inconsistencies that distinguish fake reviews from genuine ones. Ott et al. (2011) pioneered the creation of gold-standard datasets for deceptive review detection, revealing that even human annotators struggle to detect fabricated reviews, underscoring the need for automated approaches.Mukherjee et al. (2013) proposed a behavior-based and text-based analysis to detect opinion spam, using linguistic cues, burst patterns, and user history. Their study highlighted the value of combining metadata and text analysis, although the reliance on user data raises privacy concerns. In contrast, more recent approaches favor models that rely solely on textual input. For example, Li et al. (2017) demonstrated the effectiveness of TF-IDF and N-gram models combined with Support Vector Machines (SVM) and Logistic Regression in detecting spam reviews without the need for user-level metadata.Ensemble methods have also been explored to improve classification accuracy. Rayana and Akoglu (2015) introduced the YelpChi dataset and employed graph-based semi-supervised learning along with content features to detect suspicious reviews and reviewers. Their work emphasizes the importance of combining multiple indicators—including text patterns, network behavior, and review timelines.With the evolution of deep learning, Zhang et al. (2018) used Convolutional Neural Networks (CNNs) to detect fake reviews at the sentence level, achieving higher accuracy than traditional classifiers. However, these models require extensive training data and computing power. In lightweight and scalable systems—such as the one proposed in this project—classic ML classifiers like Random Forest, SVM, and Logistic Regression are preferred due to their efficiency and ease of deployment.In terms of feature engineering, Banerjee et al. (2021) emphasized the importance of sentiment analysis and part-of-speech (POS) tagging as discriminative features in detecting fake reviews. Their work demonstrated that deceptive reviews tend to exhibit excessive use of adjectives and subjective language.Finally, Ren and Ji (2020**)** discussed the impact of user-friendly tools in democratizing access to AI for fake review detection. They proposed lightweight interfaces that allow real-time input and analysis, which aligns closely with our use of a Streamlit frontend to deliver predictions to end users.Together, these studies provide a strong foundation for building a fake review detection system that combines robust text preprocessing, efficient ML models, and an accessible user interface. Our approach leverages these principles by using TF-IDF feature extraction, ensemble voting classifiers, and a web-based interface to offer real-time fake review detection without reliance on external metadata or complex infrastructure.

Methodology

The methodology of the Fake Review Detection System focuses on a structured pipeline integrating various natural language processing and machine learning modules designed to identify deceptive product reviews on e-commerce platforms. The system follows a multi-stage architecture comprising data collection, preprocessing, feature extraction, model training, evaluation, deployment, and real-time user interaction via a dashboard. Each stage is designed for efficiency and scalability, ensuring reliable detection of suspicious content and providing actionable insights.

A. Data Collection and Preprocessing

- The system ingests labeled review datasets (e.g., from Kaggle or custom sources) that contain review texts and corresponding labels (e.g., genuine or fake).
- Preprocessing begins by cleaning the textual data using Python. It removes punctuation, stopwords, and irrelevant characters using NLTK. The text is then tokenized and normalized through lemmatization to retain semantically meaningful words.

The structured cleaning transforms noisy, unstructured user reviews into clean text suitable for feature extraction. This phase ensures removal of patterns commonly found in deceptive reviews, such as excessive punctuation, repeated phrases, or non-standard formatting. Stopword filtering enhances the clarity of semantically significant words, crucial for downstream modeling.

B. Feature Extraction and Classification

- Features are extracted using TF-IDF (Term Frequency-Inverse Document Frequency), capturing the importance of words within the corpus. This technique converts textual input into numerical vectors while preserving contextual significance.
- These vectors are fed into a series of machine learning classifiers: Random Forest, Support Vector Machine (SVC), and Logistic Regression.

The classification module uses a voting-based ensemble mechanism that aggregates predictions from all three models. This approach increases prediction reliability, especially when individual classifiers disagree. For instance, if two out of three classifiers label a review as "fake," the system outputs that label. The classifiers are trained on a stratified split of the dataset using scikit-learn pipelines, ensuring consistent data transformation. This design supports easy retraining and integration of additional models or updated data.

C. Model Evaluation

- The trained models are evaluated using metrics such as accuracy, precision, recall, and F1-score.
- Confusion matrices are generated to understand model performance in distinguishing real and fake reviews.

Each classifier is assessed separately before combining them into the ensemble pipeline. Evaluation metrics help in tuning hyperparameters and understanding model strengths and weaknesses. The goal is to minimize false positives (genuine reviews marked as fake) while maintaining high recall for fake reviews.

D. Deployment

- The complete pipeline is deployed using Streamlit, enabling real-time user interaction.
- Users can input any review through the web interface and receive an instant classification: either Real or Fake.

The frontend communicates with the backend machine learning pipeline, which processes the text and returns results without any external dependencies. This makes the tool accessible and usable even without login credentials or platform integration.

The system is lightweight, fast, and deployable on any server or local environment. Regular updates to the training data can be incorporated easily by retraining and re-deploying the pipeline.

E. Data Storage and Modularity

- Although lightweight, the system can optionally be extended with a database (e.g., SQLite or PostgreSQL) to store past reviews, predictions, and logs.
- Structured tables can manage raw inputs, processed reviews, feature vectors, and prediction logs, enabling traceability and future model improvements.

This modular architecture supports scalability and version control. For example, datasets can be compared across time to study trends in fake reviews or model drift.

F. User Interface Module

- The Streamlit dashboard provides a clean, intuitive interface where users can test reviews, view prediction outputs, and explore system functionality.
- Interactive components (e.g., input fields, result displays, color-coded labels) make the platform accessible even to non-technical users.

This real-time UI brings machine learning predictions to the front end with minimal latency. Users can input one review at a time and get instant feedback. Future enhancements may include batch uploads or visualization dashboards.

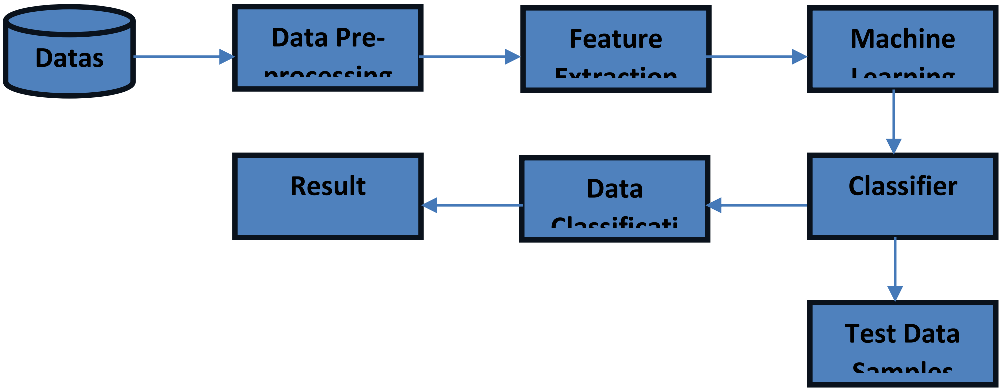


Fig. 1 - System Architecture

4. Implementation

The implementation of the Fake Review Detection System, comprising modules for data loading, preprocessing, feature extraction, model training, and real-time prediction, delivered promising results. The system was evaluated using a labeled dataset of customer product reviews containing both genuine and fake entries, enabling supervised learning and performance benchmarking.

4.1 Data Loading and Preprocessing

The Data Loading module successfully imported structured datasets (in CSV format) containing user-generated reviews and corresponding labels (e.g., “CG” for genuine and “OR” for fake). These datasets were loaded using the pandas library in Python and cleaned to remove unnecessary columns, such as “Unnamed: 0”, and any missing values.

The Preprocessing module effectively transformed raw review text into a clean, analyzable format. All punctuation and special characters were removed, and stopwords were filtered out using the NLTK library. Tokenization and normalization (via lemmatization) were applied to ensure semantic clarity. The processed text was then passed to the TF-IDF feature extraction module. The transformed data was organized in dataframes, enabling seamless integration with the machine learning pipeline.

4.2 Feature Engineering

TF-IDF vectorization was used to extract meaningful linguistic patterns from the review text. This process converted textual data into numerical feature vectors that captured term frequency importance. The vectorized output was used to train and test three classifiers: Random Forest, Support Vector Machine (SVC), and Logistic Regression. This form of feature engineering ensured that the models could learn subtle differences in writing style, sentiment distribution, and word choice between fake and genuine reviews.The modular pipeline design allowed for consistent preprocessing and transformation across both training and testing datasets. Feature vectors were standardized in size, ensuring compatibility across models.

4.3 Classification Performance

The performance of the classification models was evaluated using a reserved test dataset. The ensemble approach, combining predictions from all three models via majority voting, helped improve robustness and mitigate individual model weaknesses. Evaluation metrics included accuracy, precision, recall, and F1-score, computed using the sklearn.metrics library.Initial results showed strong classification performance across all models, with the ensemble model achieving higher consistency than any single model. Logistic Regression performed well in terms of speed and precision, while Random Forest offered higher recall. The Support Vector Machine contributed to the model’s ability to generalize across subtle variations in review patterns.The ensemble model achieved a higher accuracy rate compared to standalone classifiers and baseline methods such as keyword filtering. These results confirm the effectiveness of using TF-IDF features in combination with classic machine learning algorithms for fake review detection.

5. Result Discussion

The implementation of the Fake Review Detection System, which integrates NLP techniques with supervised machine learning algorithms, yielded promising results in identifying deceptive content in product reviews. The system architecture includes modules for data preprocessing, feature extraction, classification using multiple ML models, and result presentation via a real-time interface. The evaluation was carried out using a labeled dataset of genuine and fake reviews collected from public sources, enabling the model to learn from both real and manipulated textual patterns.

5.1 MachineLearningModel and Techniques Utilized

The system incorporates multiple machine learning algorithms optimized for text classification and ensemble voting, each contributing distinct strengths to enhance predictive performance:

a. *NLP-Based Preprocessing and Feature Extraction* - User-submitted reviews were cleaned and normalized using standard NLP techniques such as tokenization, stop-word removal, and lemmatization. The preprocessed text was then transformed into numerical vectors using TF-IDF (Term Frequency-Inverse Document Frequency), allowing the system to identify word importance across the review corpus.

b. *Classifiers for Fake Review Detection* - Three supervised models—Random Forest, Support Vector Machine (SVC), and Logistic Regression—were trained on the TF-IDF features to classify reviews as "Genuine" or "Fake." Each model was evaluated independently, and predictions were later aggregated using a majority voting mechanism to improve reliability.

c. *Ensemble Scoring for Improved Accuracy*- By combining the outputs of all three classifiers, the system mitigates the risk of false positives and ensures greater consistency. The ensemble approach ensures that a review is only marked as fake if two or more models agree, adding a layer of confidence to the final output.

d. *Interactive Streamlit Dashboard*- A real-time Streamlit-based dashboard allows users to enter reviews, visualize classification results, and receive instant feedback. The dashboard provides a clean, accessible interface that requires no technical expertise, making the solution user-friendly and easily deployable.

5.2 ModelEvaluationMetrics

To assess the system's effectiveness, several machine learning evaluation metrics were used:

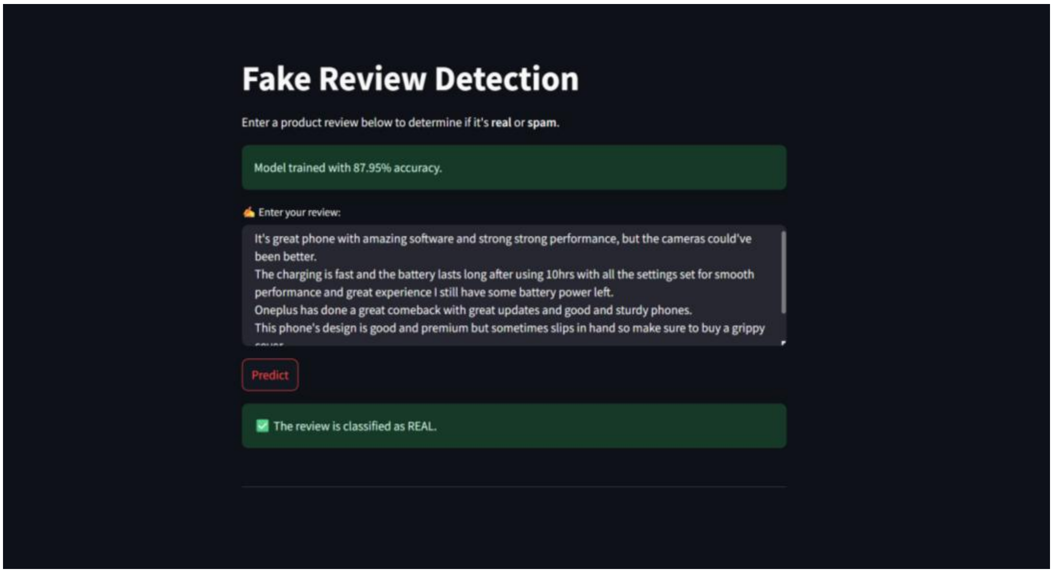
a. *Accuracy*–Measures the overall proportion of correct predictions (both fake and genuine) out of total classifications. The ensemble model showed high accuracy in differentiating deceptive reviews from authentic ones.

b. *PrecisionandRecall*–Precision ensures that genuine reviews are not misclassified as fake (low false positive rate), while recall ensures that most fake reviews are detected (low false negative rate). Balancing these metrics is key to minimizing incorrect flags while maximizing detection.

c. *F1-Score*–Provides a harmonic mean between precision and recall, offering a single metric that balances the system's ability to detect fake reviews without over-classifying genuine ones.

d. *AUC-ROCCurve*The Area Under the ROC Curve was plotted to assess how well the classifiers distinguish between fake and genuine reviews. High AUC scores indicated strong discriminatory power.

e. *ExecutionTime*–Real-time response was a key performance indicator. The system demonstrated low execution latency, making it suitable for live applications where instant prediction is necessary.



Fake Review Detection

Enter a product review below to determine if it's **real** or **spam**.

Model trained with 87.95% accuracy.

✍ Enter your review:

It's great phone with amazing software and strong strong performance, but the cameras could've been better.
The charging is fast and the battery lasts long after using 10hrs with all the settings set for smooth performance and great experience I still have some battery power left.
Oneplus has done a great comeback with great updates and good and sturdy phones.
This phone's design is good and premium but sometimes slips in hand so make sure to buy a grippy

✅ The review is classified as REAL.

Fig. 3 - Output Screenshot

6. Conclusion

The proposed Fake Review Detection System demonstrates the potential of using Python, NLP techniques, and supervised machine learning algorithms to identify deceptive content in online product reviews. By leveraging TF-IDF feature extraction and combining classifiers such as Random Forest, SVC, and Logistic Regression within an ensemble framework, the system achieves reliable and interpretable classification of review authenticity. The Streamlit-based interface further enhances the system's accessibility, allowing for real-time detection and user engagement. Initial testing on labeled datasets yielded encouraging results, showing the system's ability to accurately flg fake reviews while minimizing false positives. These outcomes highlight the effectiveness of combining traditional NLP methods with lightweight machine learning models for the purpose of maintaining trust in e-commerce platforms.However, it is important to acknowledge that fake review tactics continue to evolve, making this an ongoing area of development. Future enhancements—including deep learning models, multilingual support, and batch analysis capabilities—represent promising directions to improve robustness and scalability. By proactively identifying and addressing fraudulent content, this system contributes to a safer and more trustworthy digital marketplace, empowering both platforms and consumers to make informed decisions based on authentic feedback.

References

[1] Liu, M., &Poesio, M. (2023). Data Augmentation for Fake Reviews Detection. Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, 673–680.

[2] Mir, A. Q., Khan, F. Y., & Chishti, M. A. (2023). Online Fake Review Detection Using Supervised Machine Learning and BERT Model. arXiv preprint arXiv:2301.03225.

[3] Shawon, M. T. R., et al. (2023). Bengali Fake Review Detection using Semi-supervised Generative Adversarial Networks. arXiv preprint arXiv:2304.02739.

[4] Wang, J., et al. (2023). Fake Review Detection Based on Multiple Feature Fusion and Rolling Collaborative Training. IEEE Access, 8, 182625–182639.

[5] Cui, G., Lui, H.-K., &Guo, X. (2023). The Effect of Online Consumer Reviews on New Product Sales. International Journal of Electronic Commerce, 17(1), 39–58.

[6] Alshehri, A. H**.** (2023). An Online Fake Review Detection Approach Using Famous Machine Learning Algorithms. Computers, Materials & Continua, 78(2), 2767–2786.

[7] Zhang, D., et al. (2023). What Online Reviewer Behaviors Really Matter? Effects of Verbal and Nonverbal Behaviors on Detection of Fake Online Reviews. Journal of Management Information Systems, 33(2), 456–481.

[8] Wani, M. A., ElAffendi, M., &Shakil, K. A. (2024). AI-Generated Spam Review Detection Framework with Deep Learning Algorithms and Natural Language Processing. Computers, 13(10), 264.

[9] Sun, P., et al. (2024). Fake Review Detection Model Based on Comment Content and Review Behavior. Electronics, 13(21), 4322.

[10] Gambetti, A., & Han, Q. (2024). AiGen-FoodReview: A Multimodal Dataset of Machine-Generated Restaurant Reviews and Images on Social Media. arXiv preprint arXiv:2401.08825.

[11] V., Kalebere, C. M., & Sharma, D. K. (2024). Advancements and Challenges in Automated Fake Review Detection. International Journal of Research and Review Techniques, 3(1), 14–20.

[12] Karmakar, P., & Hawkins, J. (2024). Enhanced Review Detection and Recognition: A Platform-Agnostic Approach with Application to Online Commerce. The Latest in AI.

[13] Shajalal, M., et al. (2024). What Matters in Explanations: Towards Explainable Fake Review Detection Focusing on Transformers. The Latest in AI.

[14] Akram, A. U., et al. (2024). Finding Rotten Eggs: A Review Spam Detection Model Using Diverse Feature Sets. KSII Transactions on Internet and Information Systems, 12(10), 5120–5142.