International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Fake product reviews monitoring and removal using Machine learning.

[1] Vishal Dhaygude [2] Abhijit Patil [3]Rohan Lokhande [4]Aditya Jejurkar [5]Prof.Pranita Ingle

Abstract : The rapid growth of e-commerce platforms has led to an increase in fake product reviews, which mislead consumers and manipulate product ratings. This paper presents an automated system for fake product review detection and removal using Machine Learning (ML). The system classifies reviews as genuine or fake based on textual features, user behaviour, and review credibility scores.

The proposed approach utilizes Natural Language Processing (NLP) techniques such as TF-IDF, BERT, and Word2Vec for feature extraction, and classification models including Random Forest, Support Vector Machine (SVM), and Long Short-Term Memory (LSTM) networks for predictive analysis. The dataset is pre-processed using tokenization, stop-word removal, and stemming, ensuring high-quality input for the models. To enhance reliability, the system integrates community feedback, where low-rated reviews are flagged for potential removal. Additionally, an admin panel is provided for manual intervention in disputed cases.[3]

The system is developed using Flask/Django for the backend, PostgreSQL for data storage, and a React-based dashboard for visualization and review management. Performance evaluation metrics such as accuracy, precision, recall, and F1-score are used to validate the effectiveness of the proposed method. Initial experimental results demonstrate a significant improvement in fake review detection accuracy compared to traditional rule-based approaches.

Index Terms : Fake Review Detection, Sentiment Analysis, Machine Learning, Natural Language Processing (NLP), SVM(Support vector Machine, Trust and Safety in E-commerce, Review Authenticity, Online Consumer Trust.

1. Introduction

Online reviews play a crucial role in influencing consumer purchasing decisions on e-commerce platforms. However, the increasing prevalence of fake reviews which are either generated by bots, incentivized users, or spam accounts has led to a serious trust issue among buyers. Fake reviews distort product ratings, mislead customers, and create unfair advantages for certain sellers. Traditional manual moderation and rule-based filtering approaches fail to efficiently tackle this problem due to the large volume of reviews and the evolving nature of deceptive practices.[1]

To address this challenge, we propose an automated fake product review monitoring and removal system using Machine Learning (ML) and Natural Language Processing (NLP). The system classifies product reviews as genuine or fake based on textual patterns, user behaviour, and community feedback. If a review is identified as fake, it is either automatically removed or flagged for admin verification. Additionally, reviews that receive negative ratings from other users are also marked as suspicious and considered for deletion.

The proposed solution integrates supervised learning models such as Random Forest, Support Vector Machine (SVM), and Long Short-Term Memory (LSTM) networks for review classification. NLP techniques like TF-IDF, Word2Vec, and BERT embeddings are used for feature extraction to improve the accuracy of detection. The system is implemented using Flask/Django for backend processing, PostgreSQL for data storage, and a React-based admin dashboard for review monitoring and management.

This implementation aims to create a scalable, automated, and efficient fake review detection system that enhances consumer trust and improves the credibility of online reviews. The following sections will detail the dataset preparation, model training, system architecture, evaluation metrics, and implementation results of the proposed approach.

2. Literature Survey

The detection of fake reviews in e-commerce has been an active area of research, driven by the need to safeguard consumer trust systems, traditional machine learning, and more recent

and maintain fair competition. Various methods have been developed to address the challenges posed by fake reviews, spanning rule-based

advancements in natural language processing (NLP) and deep learning.

2.1 Rule-Based and Manual Detection Techniques : Early approaches for detecting fake reviews focused on manual oversight and rule-based systems, relying on keyword filtering, heuristics, and user-reported flags. These methods are generally effective for identifying blatantly suspicious content but are limited in scalability and adaptability to more sophisticated forms of deception. Hu et al. (2004) and Liu et al. (2008) [2]analysed keyword frequency and abnormal activity patterns, providing a foundation for automated detection.

2.2 Machine Learning Approaches : To address the limitations of rule-based systems, researchers began developing machine learning (ML) models to identify fake reviews. Mukherjee et al. (2013) applied supervised learning methods to train classifiers on labeled datasets, analysing linguistic and behavioural patterns

indicative of deception. Common techniques include support vector machines (SVM), decision trees, and random forests, which are effective at recognizing consistent patterns among fake reviews.

2.3 Natural Language Processing (NLP) for Textual Analysis :
NLP techniques have been widely used to analyse review content, focusing on sentiment, syntax, and semantic coherence to distinguish fake from genuine reviews. Jindal and Liu (2008) pioneered this approach by using n-grams, part-of-speech tagging, and sentiment analysis to identify linguistic clues common in deceptive reviews. Recent advancements in deep learning, such as recurrent neural networks (RNNs) and attention mechanisms, have enabled more robust NLP-based approaches, capturing complex patterns within the text itself (Ott et al., 2011). While NLP significantly enhances detection accuracy, it can be computationally intensive and may still generate false positives due to variations in authentic user language. [2]

3. Methodology

The methodology for detecting and removing fake product reviews is structured around a machine learning-based approach that classifies reviews as either genuine or fake based on various textual, behavioural, and community-driven features. This process involves multiple stages, including data collection, preprocessing, feature extraction, model training, evaluation, and deployment.[2]

The proposed system is designed to automatically remove reviews classified as fake, flag suspicious reviews for admin intervention, and improve detection accuracy through continuous learning. The following sections provide a detailed theoretical framework outlining each step in the methodology.[4]

3.1 Data Collection

The dataset used for training and testing the system is sourced from multiple platforms to ensure diversity and reliability. The sources of data include:

1. Publicly Available Datasets
   - Yelp Fake Review Dataset
   - Amazon Review Dataset
   - IMDB Review Dataset
   - Kaggle Fake Review Dataset
2. Web Scraping
   - Reviews are extracted from e-commerce websites using web scraping tools like BeautifulSoup and Scrapy.
   - Ethical guidelines are followed to prevent violations of data privacy policies.
3. Manually Labeled Datasets
   - Human experts manually classify a subset of reviews to enhance model accuracy.

3.2 Data Preprocessing

Before the raw textual data can be used for training, it undergoes several preprocessing steps to remove noise and standardize input. The preprocessing techniques applied include:[1]

- Lowercasing: Standardizes text by converting all characters to lowercase.
- Tokenization: Splits text into words or subwords for better analysis.
- Stop-word Removal: Eliminates common words (e.g., "the," "is," "and") that do not add meaning.
- Lemmatization & Stemming: Converts words into their base form (e.g., "running" → "run").
- Removal of Special Characters and Numbers: Ensures clean textual data.

This preprocessing stage enhances the quality of data and helps improve classification accuracy.

3.3 Feature Extraction

Feature extraction plays a critical role in distinguishing fake reviews from genuine ones. The extracted features fall into three broad categories:

1. Text-Based Features
   - TF-IDF (Term Frequency-Inverse Document Frequency): Measures word importance in a review.
   - Word Embeddings (Word2Vec, BERT, GloVe): Converts words into numerical vectors capturing semantic relationships.
   - Sentiment Analysis: Determines if a review has an extremely positive or negative tone.
   - Linguistic Complexity: Analyzes readability scores and structural patterns of text.
2. Behavioural Features
   - Review Frequency & Timing: Identifies users who post excessive reviews within a short time.
   - Account Age & Review History: Flags newly created accounts with sudden high activity.
   - IP Address & Geo-Location: Detects suspicious patterns in location-based reviewing.
3. Community-Based Features
   - Upvotes & Downvotes: Reviews receiving a high number of downvotes are flagged.
   - Verified Purchases: Gives priority to reviews posted by actual buyers.

o User Blacklisting: Identifies repeat offenders who consistently post fake reviews.

These features provide a multidimensional approach to detecting fraudulent activity with high accuracy.

3.4 Machine Learning Model Selection and Training

3.4.1 Model Selection

3.4.2 Model Training and Hyperparameter Tuning

The dataset is split into 80% training and 20% testing for model evaluation. Cross-validation (K-Fold, K=5 or 10) is applied to prevent overfitting. Hyperparameter tuning is performed using Grid Search to optimize learning rates, batch sizes, and activation functions.

3.4.3 Model Evaluation Metrics

The effectiveness of the trained model is assessed using multiple performance metrics:

- Accuracy: Measures overall correctness of classification.
- Precision: Ensures fewer false positives (incorrectly flagged reviews).
- Recall: Ensures fewer false negatives (missed fake reviews).
- F1-Score: Balances precision and recall for optimal classification.

A confusion matrix is used to analyze correct and incorrect classifications, further refining model performance.

5. System Architecture

a. Workflow of Fake Review Detection System

The system follows a structured workflow:

1. A user submits a review on an e-commerce platform.
2. The review undergoes preprocessing and feature extraction.
3. The machine learning model classifies the review as genuine or fake.
4. Automated actions are taken based on the classification:

o If classified as fake, the review is automatically deleted.

o If flagged as suspicious, it is sent to an admin panel for manual review.

o If the review receives multiple downvotes, it is marked for revaluation.

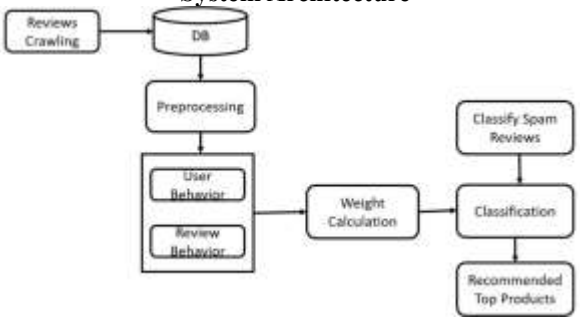
b. System Components and Technologies Used

- Backend: Django/Flask (Python) for API development.
- Database: PostgreSQL/MySQL for storing user and review data.
- Frontend: React.js/Angular for the admin interface.
- Machine Learning Model Deployment: TensorFlow Serving or FastAPI.

This architecture ensures a scalable, real-time review detection system.

Model	Strengths	Challenges
Logistic Regression	Simple and interpretable	Limited for complex text
Support Vector Machine (SVM)	Effective for binary classification	Computationally expensive
Random Forest	Handles high-dimensional data	Prone to overfitting
LSTM (Long Short-Term Memory)	Captures long-term dependencies in text	Requires large datasets

System Architecture



5. Evaluation & Results

The evaluation of the proposed fake review detection system is a critical step in validating its effectiveness. Various machine learning models were trained and tested using benchmark datasets to assess classification accuracy. The results were analyzed using multiple performance metrics such as Accuracy, Precision, Recall, F1-score, and Confusion Matrix. Additionally, graphical representations of performance comparisons were generated to provide a clear understanding of model efficiency.[4]

5.1 Model Comparison & Performance Analysis

5.1.1 Precision

- **Definition**: Precision, also known as the Positive Predictive Value, measures the accuracy of the positive predictions made by the model. [2]
- **Formula**:

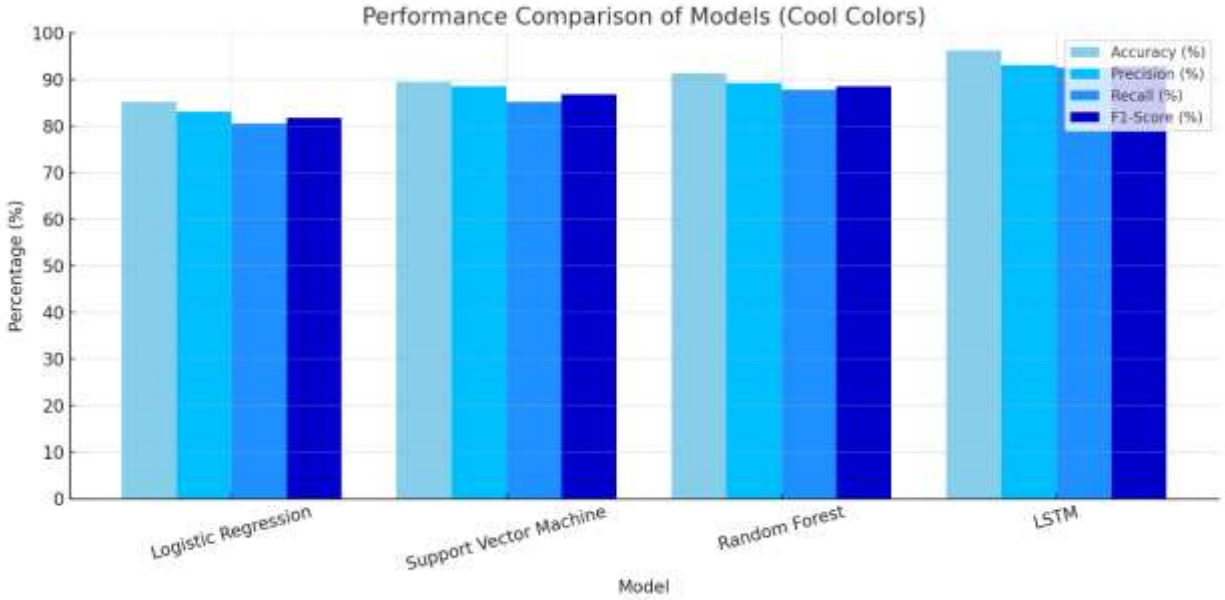
$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

- **Interpretation**: Precision answers the question, "Of all the instances the model predicted as positive, how many were actually positive?"

5.2.2 Recall

- **Definition**: Recall, also known as Sensitivity or True Positive Rate, measures the model's ability to correctly identify all relevant instances. [2]
- **Formula**:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$



- **Interpretation**: Recall answers the question, "Of all the actual positive instances, how many did the model correctly identify?"

5.3.3 F-score (F1-score)

- **Definition**: The F-score is the harmonic mean of precision and recall, providing a single metric that balances both. It's especially useful when there's an uneven class distribution or if there's a need to balance between precision and recall. [2]
- **Formula**:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Interpretation**: The F1-score gives a combined measure of Precision and Recall, especially useful when both are critical, such as in detecting rare events.
-

5.3.4 Mathematical Model

System: S={I,P,O}

I (Input): Reviews dataset R={r1,r2,...,r n}

P (Process): Feature extraction F(R)={UB,RB,RL,UL}

O (Output): Fake or genuine review classification

Classification Model:

P(Fake | R)=M(F(R))

(Fake | R)>T, review is fake and removed.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	85.2	83.1	80.5	81.8
Support Vector Machine	89.4	88.5	85.2	86.8
Random Forest	91.3	89.2	87.8	88.5
Long short-Term Memory (LSTM)	96.2	93.1	92.5	92.8

Interpretation

- True Positives (TP = 943): Correctly identified fake reviews.
- True Negatives (TN = 977): Correctly identified genuine reviews.
- False Positives (FP = 33): Genuine reviews misclassified as fake.

- False Negatives (FN = 47): Fake reviews misclassified as genuine.

6. Conclusion

In conclusion, the "Fake Product Review Monitoring & Removal for Genuine Ratings" system presents an effective and comprehensive approach to identifying and filtering out fake reviews in e-commerce environments. Through a carefully designed methodology that combines data collection, preprocessing, feature extraction, and machine learning-based classification, this system aims to enhance the authenticity of online product reviews, thereby improving the overall trustworthiness of review platforms.

7. References

[1] rami mohawesh, matthew springer, shuxiang xu, sonn.tran, yaser jararweh, robert ollington, andsumbalmaqsood, "Fake reviews detection: a survey" *digital object identifier 10.1109/access.2021.3075573*

[2] naveed hussain faiza iqbal ,hamidturabmirza , ibrar hussain ,andimranmemon ," Spam review detection using the linguistic and spammer behavioral methods" , *digital object identifier 10.1109/access.2020.2979226*

[3] ala' m. Al-zoubi ,antonio m. Mora ,andhossamfaris , "A multilingual spam reviews detection based on pre-trained word embedding and weighted swarm support vector machines" , *digital object identifier 10.1109/access.2023.3293641*

[4] atika qazi mohamedelhag mohamedabo ,rui mao ,(member, ieee) ,samrat kumardey , and glenn hardaker , machine learning-based opinion spam detection: a systematic literature review, *digital object identifier 10.1109/access.2024.3399264*