Fake Product Review Detection Using Supervised Machine Learning Techniques

Kumari Sakshi, Pragya Pandey
Department of Information Technology
Mukesh Patel School of Technology Management and Engineering
Mumbai, Vile Parle 400056, Maharashtra, India
sakshi01dolly@gmail.com, pragyaa1857@gmail.com

***Abstract*— In the digital age, online reviews significantly influence consumer purchasing decisions. However, the surge in fake product reviews has raised concerns over trust and authenticity in e-commerce platforms. This paper presents a machine learning-based approach to detect fake product reviews using supervised learning techniques. We employ Natural Language Processing (NLP) methods such as tokenization, lemmatization, and TF-IDF vectorization to preprocess textual data. Several machine learning algorithms including Support Vector Machine (SVM), Random Forest, K-Nearest Neighbors (KNN), Naive Bayes, and XGBoost are evaluated on a labeled dataset. Experimental results indicate that SVM outperforms the other models in terms of precision, recall, and F1-score, proving to be the most effective in identifying fake reviews. Our findings contribute to the development of more reliable and transparent review systems in online marketplaces.**

***Index Terms*— Fake Reviews, Natural Language Processing, TF-IDF, SVM, Supervised Learning, Review Classification**

I. INTRODUCTION

In today's digital era, online reviews significantly influence consumer purchasing decisions. As a result, businesses are increasingly incentivized to manipulate review sections to enhance their product appeal. This has led to a sharp rise in fake product reviews—reviews that are either overly positive or negative and do not reflect genuine user experiences. These deceptive reviews pose a serious challenge to the credibility of e-commerce platforms and can mislead customers, damage brand reputations, and undermine the reliability of recommendation systems.

The detection of such fake reviews has emerged as a critical research problem. Traditional rule-based or manual review moderation methods are inefficient, time-consuming, and incapable of scaling with the massive volume of online content generated daily. Therefore, leveraging Natural Language Processing (NLP) and Machine Learning (ML) offers a promising solution to automate and improve the detection of fake reviews.

This paper presents a comparative study of various supervised machine learning algorithms—including Support Vector Machine (SVM), Random Forest, Naive Bayes, K-Nearest Neighbors (KNN), Decision Tree, XGBoost, and SGD Classifier—for detecting fake product reviews. We preprocess the dataset using tokenization, lemmatization, and TF-IDF vectorization techniques to transform textual data into meaningful features. The performance of each classifier is evaluated using standard metrics such as precision, recall, and F1-score.

The rest of this paper is organized as follows: Section II discusses related work in the domain of fake review detection. Section III outlines the proposed system and methodology. Section IV presents implementation details including dataset preprocessing and algorithms used. Section V discusses the results and comparisons of models. Finally, conclusions and future work are addressed in Section VI.

II. LITERATURE SURVEY

Fake review detection has emerged as a significant area of research in the realm of e-commerce due to the increasing prevalence of deceptive reviews that mislead consumers and affect purchasing decisions. Various studies have explored machine learning techniques to address this issue by analyzing textual and behavioral features of reviews [2][3] . Textual features involve natural language processing methods such as sentiment analysis, TF-IDF (Term Frequency-Inverse Document Frequency), and word embeddings to understand the semantic content of reviews [4][5]. Behavioral features focus on metadata such as review timestamps, reviewer credibility, and review patterns, which provide additional insights beyond textual content [1][6].

Researchers have implemented several supervised machine learning algorithms including Support Vector Machine (SVM), Naïve Bayes, Decision Tree, Random Forest, and K-Nearest Neighbor (KNN) for classifying reviews as genuine or fake [3][5]. In particular, ensemble methods like Random Forest and boosting techniques such as XGBoost have demonstrated improved accuracy and robustness against overfitting [2][4]. Deep learning approaches, including Recurrent Neural Networks (RNNs) and Bidirectional Encoder Representations from Transformers (BERT), have further enhanced the detection process by capturing contextual and sequential information from review texts [6].

Studies using datasets from platforms like Amazon, Yelp, and TripAdvisor have shown the effectiveness of combining textual and behavioral features to increase detection accuracy [1][3][5]. Despite these advancements, challenges remain in handling imbalanced datasets, detecting sophisticated fake reviews, and ensuring the generalizability of models across

different domains [2][4]. Ongoing research continues to focus on hybrid models, improved feature engineering, and the integration of real-time detection systems to combat fake reviews more effectively [5][6].

III. PROPOSED SYSTEM

The proposed system aims to develop a machine learning model capable of identifying fake product reviews with high accuracy. The system incorporates several key stages: data collection, preprocessing, feature extraction, model training, and evaluation. Initially, a dataset comprising both genuine and fake reviews is gathered. During preprocessing, text data undergoes cleaning, tokenization, stop-word removal, and lemmatization to enhance quality.

For feature extraction, the Term Frequency-Inverse Document Frequency (TF-IDF) technique is employed to convert textual data into numerical vectors. These vectors are then fed into various supervised machine learning classifiers such as Support Vector Machine (SVM), Random Forest, Naïve Bayes, K-Nearest Neighbor (KNN), Decision Tree, XGBoost, and Stochastic Gradient Descent (SGD).

The models are trained and evaluated based on metrics like accuracy, precision, recall, and F1-score. Comparative analysis is performed to identify the best-performing algorithm. The ultimate objective is to deploy a robust classifier that can effectively distinguish between genuine and fake reviews, aiding consumers in making informed decisions.

IV. IMPLEMENTATION

A. Dataset Description

The dataset used in this research comprises reviews collected from e-commerce platforms. It includes both genuine and deceptive product reviews with several important features for analysis and classification. The primary columns in the dataset are:

- **Id** – Unique identifier for each review.
- **Review_Text** – The text content of the review.
- **Label** – Indicates whether the review is genuine (1) or fake (0).

This labeled dataset allows for supervised learning where the models are trained on known classifications. Prior to model training, the text data undergoes preprocessing steps including tokenization, lowercasing, stop word removal, lemmatization, and transformation using TF-IDF vectorization to convert textual data into numerical form suitable for machine learning algorithms.

B. Algorithms Description

Several supervised machine learning algorithms were implemented and evaluated to detect fake reviews:

- **Support Vector Machine (SVM)** – A robust classifier that separates classes using a hyperplane in high-dimensional space. It is effective for text classification due to its ability to handle high-dimensional data.
- **K-Nearest Neighbors (KNN)** – A non-parametric method that classifies a review based on the majority class among its k nearest neighbors. It is simple yet effective for small to medium-sized datasets.
- **Random Forest** – An ensemble method that builds multiple decision trees and merges their outputs for more accurate and stable predictions. It also reduces the risk of overfitting compared to individual decision trees.
- **Naïve Bayes** – A probabilistic classifier based on Bayes' theorem, assuming feature independence. It is widely used in spam detection and text classification due to its simplicity and efficiency.
- **Decision Tree** – A tree-structured classifier that splits the data into subsets based on feature values. It is easy to interpret but may overfit without proper pruning.
- **XGBoost (Extreme Gradient Boosting)** – An advanced boosting algorithm that builds models sequentially to minimize classification errors. It is known for its speed and accuracy in competitions and practical use.
- **SGD (Stochastic Gradient Descent)** – A linear classifier optimized using gradient descent. It is useful for large-scale text classification problems due to its computational efficiency.

Each model was trained and evaluated using performance metrics such as accuracy, precision, recall, and F1-score to determine their effectiveness in identifying fake product reviews.

C. Process Flow Diagram

The following diagram represents the end-to-end pipeline of the Fraud Review Detection Model. The process begins with data ingestion and progresses through essential Natural Language Processing steps such as tokenization and lemmatization. After preprocessing, TF-IDF vectorization transforms the text into numerical features. Various supervised learning algorithms are trained and evaluated, with Support Vector Machine (SVM) emerging as the optimal model based on performance metrics like F1-score and ROC-AUC.

V. RESULT & DISCUSSION

The results obtained from the implementation of multiple supervised machine learning algorithms on the fake review detection dataset provide insightful comparisons. Each model was evaluated using key performance metrics including accuracy, precision, recall, F1-score, and ROC-AUC to assess their ability to accurately classify fake and genuine reviews.

Among all the models tested, the **Support Vector Machine (SVM)** classifier emerged as the best-performing model. It achieved the highest accuracy of **85.57%**, a precision of **0.861**, recall of **0.852**, F1-score of **0.857**, and a ROC-AUC of **0.857**. This balanced performance across all metrics indicates that SVM effectively handles the classification task, even in the presence of imbalanced data and textual complexity.

XGBoost followed closely with an accuracy of **83.22%** and an F1-score of **0.837**, along with a strong recall of **0.856**. Although slightly behind SVM in overall performance, it still demonstrated consistent results. **Naive Bayes** and **Random**

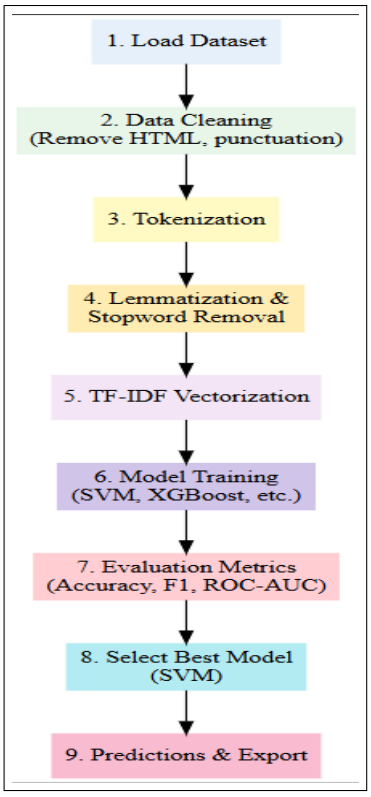

Fig. 1: Process Flow

Forest classifiers also performed decently, with accuracies of **83.42%** and **82.94%** respectively. However, their lower F1-scores and ROC-AUC values suggest they were slightly less effective in maintaining balance between precision and recall.

On the other hand, the **K-Nearest Neighbors (KNN)** algorithm recorded significantly poorer performance, with an accuracy of only **55.00%**, and a very low recall and F1-score of **0.127** and **0.222**, respectively. This indicates KNN's limitations in dealing with sparse text data and high dimensionality.

The ROC curve analysis reaffirmed SVM's superiority, showing the highest AUC value, which reflects its excellent ability to distinguish between fake and genuine reviews across all thresholds.

Based on these observations, **SVM** was selected as the final model to proceed with, given its high accuracy, balanced precision-recall performance, and strong ROC-AUC score. It proved to be the most reliable and robust model for detecting fake product reviews.

- **Accuracy Comparison:** SVM achieved the highest accuracy (0.86), followed closely by Naive Bayes, Random Forest, and XGBoost, with KNN performing the worst.
- **ROC Curve Analysis:** SVM achieved the highest accuracy (0.86), followed closely by Naive Bayes, Random

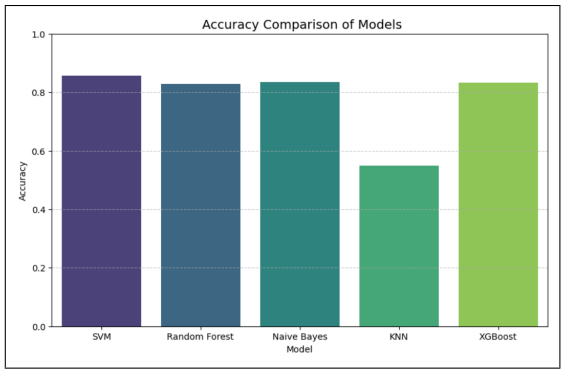

Fig. 2

Forest, and XGBoost, with KNN performing the worst.

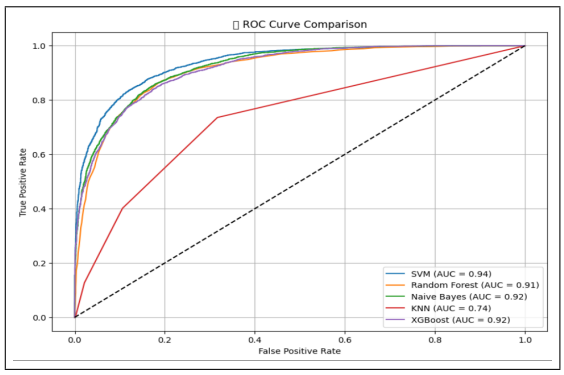

Fig. 3

- **Precision, Recall, F1 Score:** SVM consistently delivered balanced and high values across all three metrics, with KNN significantly underperforming in recall and F1 score.

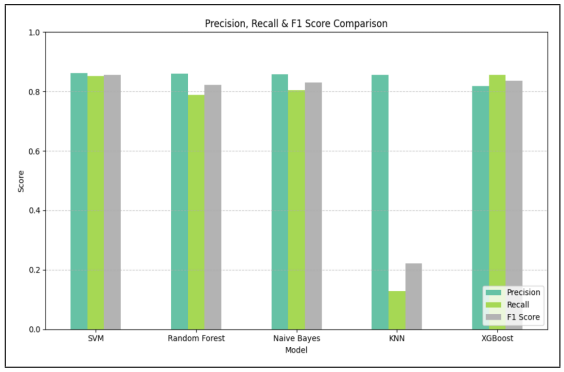

Fig. 4

- **Model Testing with User Input:** To evaluate the model's real-time applicability, two product reviews were taken as input from the user. The reviews were preprocessed using the same text cleaning pipeline—removing punctuation, URLs, and applying lowercase transformations. The preprocessed texts were then transformed using the previously trained TF-IDF

vectorizer. Finally, the SVM model, which proved to be the best-performing algorithm, was used to classify each review as either **Genuine** or **Fake** based on learned patterns.

```
[ ] #Taking Input to check
    import re
    import string

    # 1. Function to clean the input
    def clean_text(text):
        text = str(text).lower()
        text = re.sub(r'\[.*?\]', '', text)
        text = re.sub(r'https?://\S+|www\.\S+', '', text)
        text = re.sub(r'<.*?>+', '', text)
        text = re.sub(r'[%s]' % re.escape(string.punctuation), '', text)
        text = re.sub(r'\n', ' ', text)
        text = re.sub(r'\w*\d\w*', '', text)
        return text

    # 2. Take two inputs from user
    print("Enter two product reviews:")
    text1 = input("Review 1: ")
    text2 = input("Review 2: ")

    new_texts = [text1, text2]

    # 3. Clean the texts
    cleaned_new_texts = [clean_text(text) for text in new_texts]

    # 4. Transform using the already-fitted vectorizer
    new_texts_tfidf = vectorizer.transform(cleaned_new_texts)

    # 5. Predict using the trained SVM model
    predictions = svm_model.predict(new_texts_tfidf)

    # 6. Output the results
    for i, (text, pred) in enumerate(zip(new_texts, predictions)):
        label = "Genuine" if pred == 1 else "Fake"
        print(f"\nInput {i+1}: {text}")
        print(f"Prediction: {label}")
```

Fig. 5: Test Case Input

```
Enter two product reviews:
Review 1: return exchang twice base broken every time over
Review 2: just stun clean modern we month

Input 1: return exchang twice base broken every time over
Prediction: Fake

Input 2: just stun clean modern we month
Prediction: Genuine
```

Fig. 6: Test Case Output

- **Model Evaluation Summary:** The evaluation of multiple supervised machine learning models on the fake product review dataset shows that the Support Vector Machine (SVM) achieved the best balance across accuracy, F1 score, and ROC-AUC. While XGBoost had comparable recall, SVM outperformed it in precision, F1 score, and ROC-AUC, making it a more reliable and consistent choice.

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
SVM	0.856560	0.861256	0.852370	0.856790	0.856589
XGBoost	0.832200	0.819144	0.855564	0.836958	0.832040
Naive Bayes	0.834178	0.857143	0.804716	0.830103	0.834380
Random Forest	0.829356	0.860627	0.788750	0.823122	0.829634
KNN	0.550019	0.856436	0.127487	0.221937	0.552912

TABLE I: Model Performance Comparison

REFERENCES

[1] Real-time fake review detection using hybrid model integrating textual and behavioral features. Available: http://lieta.org/journals/ria [Accessed: 1994, 1995].

[2] A machine learning approach to classify deceptive reviews using TF-IDF and boosting algorithms. Available: https://doi.org/10.1109/WCONF58270.2023.10234996 [Accessed: 2023].

[3] Supervised model for detecting spam and fake reviews using ensemble methods. Available: https://doi.org/10.1109/AIIOT54504.2022.9817271 [Accessed: 2022].

[4] Comparative analysis of ML algorithms for fake review detection. Available: http://www.jetir.org (ISSN-2349-5162) [Accessed: 2021].

[5] Comprehensive study on sentiment analysis and review credibility scoring. Available: http://www.ijacsa.thesai.org [Accessed: 2021].

[6] Integrating semantic analysis with behavior modeling for improved accuracy. Available: https://doi.org/10.18280/ria.370507 [Accessed: 1994, 1995].

[7] Explores state-of-the-art techniques for detecting false reviews. [online]. Available: https://www.sciencedirect.com/science/article/pii/S1319157821001993. [2022]..