International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Artificial Intelligence based Framework for detecting Deceptive Online Reviews

Ms. Samruddhi P. Ingale[1], Dr. V. H. Deshmukh[2], Dr. P. P. Deshmukh[3]

[1]*Student, Department of Computer Science and Engineering, PRMI&R, Badnera*
[2]*Professor, Department of Computer Science and Engineering, PRMI&R, Badnera*
[3]*Assistant Professor, Department of Computer Science and Engineering, PRMI&R, Badnera*

Abstract - Online reviews play an important role in shaping customer decisions on digital platforms, but their reliability is often affected by the presence of deceptive reviews. These reviews are intentionally written to promote or demote products, which can mislead users and reduce trust in online systems. This paper presents a method for detecting fake reviews using a combination of natural language processing and machine learning techniques. The approach processes review text through standard preprocessing steps and extracts features based on term frequency–inverse document frequency, sentiment scores, and reviewer behavior patterns. Multiple classification models, including support vector machines, logistic regression, random forest, and deep learning methods, are evaluated for identifying deceptive content. Prior studies have shown that combining textual and behavioral features improves detection performance, as behavioral patterns often reveal inconsistencies not captured in text alone. In addition, machine learning models trained on structured features have demonstrated effectiveness in distinguishing genuine and fake reviews across different datasets. The proposed system also provides interpretable outputs to explain prediction results. Experimental observations indicate that integrating multiple feature types leads to more reliable classification compared to single-feature approaches, making the system suitable for practical use in online review platforms.

Key Words: *Fake review detection, natural language processing, machine learning, sentiment analysis, opinion spam, text classification, behavioral analysis, feature extraction, deceptive reviews, explainable artificial intelligence*

1.INTRODUCTION

Online reviews have become an important source of information for customers when selecting products and services on digital platforms. Many users rely on opinions shared by others before making decisions, particularly in domains such as e-commerce and hospitality. Positive reviews can improve customer trust and influence purchasing behavior, while negative reviews may discourage potential buyers. As a result, online reviews play a significant role in shaping the reputation of products and services.

However, the increasing dependence on online reviews has also led to the rapid growth of deceptive or fake reviews. These reviews are intentionally written to promote certain products or damage the reputation of competitors, thereby misleading users and reducing the reliability of online platforms. Early studies on opinion spam highlighted that fake reviews often exhibit identifiable patterns in both content and reviewer behavior [1], [2]. The presence of such misleading content creates challenges for consumers and businesses, making automatic detection methods necessary.

To address this issue, researchers have explored various approaches based on machine learning and natural language processing. Traditional methods focused on analyzing textual features such as word frequency, linguistic patterns, and sentiment polarity using techniques like TF-IDF and opinion mining [3], [4]. These approaches have shown effectiveness in capturing useful patterns in review text, but they are often limited when used alone.

Recent studies emphasize the importance of incorporating behavioral features, including reviewer activity, posting frequency, and rating patterns, to improve detection performance. It has been observed that combining textual and behavioral features leads to better classification accuracy compared to relying only on textual information [5], [6]. In addition, machine learning models such as Support Vector Machines, Logistic Regression, and Random Forest have been widely applied for classification tasks, demonstrating reliable performance across different datasets [7], [8].

More recent work has explored deep learning techniques to capture contextual relationships within text. Neural network models, including convolutional and recurrent architectures, have shown improved performance in detecting deceptive reviews by learning complex representations of language [9], [10]. Despite these advancements, several challenges remain, such as limited availability of labeled datasets, evolving spam strategies, and lack of interpretability in model predictions.

This paper presents a system for detecting fake reviews using a combination of textual, sentiment, and behavioral features. Multiple machine learning and deep learning models are evaluated, and the system provides interpretable outputs to explain classification decisions. The objective is to improve the reliability of online review platforms by identifying deceptive content in an effective and understandable manner.

2. RELATED WORK

The problem of detecting deceptive or fake reviews has been widely studied due to its impact on consumer trust and online platform credibility. Early work by Jindal and Liu introduced the concept of opinion spam and analyzed duplicate content and abnormal reviewing patterns to identify deceptive behavior [1]. Their study laid the foundation for subsequent research by highlighting the importance of both textual and behavioral indicators in fake review detection.

Later, Ott et al. created benchmark datasets consisting of truthful and deceptive reviews and demonstrated that linguistic features can be used to distinguish between them using supervised learning techniques [2]. Their work showed that deceptive reviews often exhibit identifiable writing patterns, although such patterns can sometimes be subtle and difficult to capture.

Traditional machine learning approaches have been widely applied in this domain. Methods based on classifiers such as Naïve Bayes, Support Vector Machines (SVM), Decision Trees, and Logistic Regression have shown effective performance when trained on textual features such as n-grams and TF-IDF representations [3], [7]. For instance, Asaad et al. applied machine learning models including SVM, stochastic gradient descent, and XGBoost on review datasets and demonstrated that proper preprocessing and feature extraction significantly improve classification performance .

However, relying only on textual features has limitations. Researchers have increasingly focused on incorporating behavioral features to improve detection accuracy. Elmogy et al. showed that combining reviewer behavior features, such as posting frequency and writing patterns, with textual features leads to better performance compared to using text alone . Their experiments reported a noticeable improvement in F1 score when behavioral features were included.

Similarly, Alshehri proposed a semi-supervised learning approach that combines textual, linguistic, and behavioral features for fake review detection . The study addressed the challenge of limited labeled data by applying PU-learning, demonstrating that semi-supervised approaches can effectively utilize unlabeled data for classification tasks.

In addition to traditional machine learning techniques, deep learning models have been explored to capture complex patterns in review text. Ren and Ji applied neural network models to detect deceptive reviews by learning sequential dependencies within text [9]. Convolutional Neural Networks (CNNs) have also been used for sentence classification tasks, showing improved ability to capture contextual information compared to traditional methods [10].

Recent research has further explored hybrid approaches that combine deep learning with feature engineering techniques. Aspect-based sentiment analysis has been introduced to focus on specific components of reviews rather than analyzing the entire text. These approaches improve efficiency and accuracy by extracting meaningful aspects and their associated sentiments before classification. Studies using CNN-LSTM architectures have demonstrated improved performance by combining local feature extraction with sequential learning capabilities.

Another important direction in fake review detection is the use of graph-based and clustering methods. Mukherjee et al. analyzed reviewer behavior patterns and proposed models based on group detection and relational data, showing that coordinated spam campaigns can be identified through network-based analysis [8]. Such approaches highlight the importance of considering relationships between users, products, and reviews.

Despite significant progress, several challenges remain in this domain. One major limitation is the lack of large, high-quality labeled datasets, which restricts the performance of supervised learning models. Additionally, spammers continuously evolve their writing styles and strategies, making it difficult for static models to maintain accuracy over time. Furthermore, many advanced models, particularly deep learning approaches, lack interpretability, making it difficult to understand the reasoning behind predictions.

Overall, the literature suggests that no single approach is sufficient for effective fake review detection. Instead, combining textual analysis, sentiment features, behavioral patterns, and advanced learning models provides a more reliable solution. These findings motivate the development of systems that integrate multiple feature types while maintaining interpretability and efficiency.

3. PROPOSED SYSTEM

The goal of this mini project is to design and implement a system that can identify deceptive reviews using a combination of textual analysis, sentiment evaluation, and reviewer behavior patterns. The proposed system processes user-provided review text and determines whether it is genuine or fake, along with a confidence score and explanation of the prediction.

The system is designed as a multi-stage pipeline that integrates preprocessing, feature extraction, classification, and result interpretation. Unlike approaches that rely only on textual features, the proposed system combines multiple types of information to improve detection accuracy and reliability.

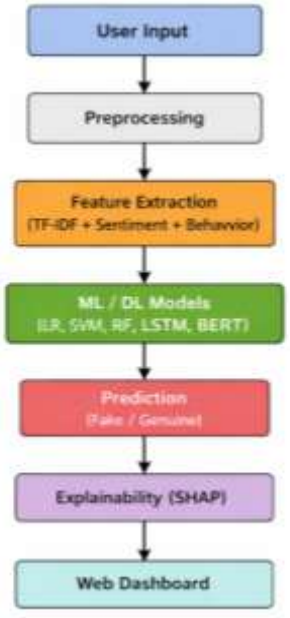

Fig. 1. Architecture of the proposed fake review detection system.

The overall architecture consists of four main stages: input processing, feature extraction, model prediction, and result

visualization. The workflow of the system follows a structured sequence where each stage contributes to refining the final prediction.

A. System Overview

The system accepts a review as input through a web-based interface. The input text is first processed using natural language processing techniques to remove noise and standardize the content. After preprocessing, relevant features are extracted from the text and passed to trained machine learning and deep learning models for classification.

The system evaluates the review using multiple models and produces a final prediction indicating whether the review is genuine or fake. In addition, the system provides a confidence score and highlights important factors that influenced the prediction. This improves transparency and helps users understand the decision-making process.

B. Data Preprocessing

Preprocessing is an essential step to ensure that the input text is clean and suitable for feature extraction. The system applies several preprocessing operations, including:

- Conversion of text to lowercase
- Removal of punctuation and special characters
- Tokenization of words
- Removal of stopwords
- Lemmatization of tokens

These steps help reduce noise and standardize the textual data, making it easier for machine learning models to learn meaningful patterns. Similar preprocessing techniques have been widely used in fake review detection studies and have shown to improve model performance [3].

C. Feature Extraction

Feature extraction plays an important role in identifying deceptive reviews.

Fig. 2. Machine learning pipeline for review classification.

The system uses three categories of features:

1. Textual Features

Textual features are extracted using the TF-IDF method, which represents the importance of words based on their frequency within a review and across the dataset. N-gram features are also used to capture common word sequences.

2. Sentiment Features

Sentiment analysis is performed to determine the emotional tone of the review. The system calculates polarity scores such as positive, negative, neutral, and compound values. These features help identify exaggerated or overly biased reviews, which are often associated with deceptive content.

3. Behavioral Features

Behavioral features capture patterns related to reviewer activity. These include:

- Frequency of reviews
- Repetition of content
- Extreme rating behavior

Studies have shown that incorporating behavioral features significantly improves detection accuracy compared to using textual features alone.

D. Model Archictecture

The system evaluates multiple machine learning and deep learning models to classify reviews:

- Logistic Regression
- Support Vector Machine (SVM)
- Random Forest
- Long Short-Term Memory (LSTM)
- Transformer-based model (BERT)

Traditional machine learning models provide efficient and stable performance, while deep learning models capture complex patterns in text. Combining these models allows the system to achieve better generalization across different types of reviews.

Support Vector Machines and ensemble methods have been widely used in prior studies for fake review detection due to their strong classification capability [7], while deep learning models improve contextual understanding of text [9].

E. Explanability

To improve transparency, the system includes an explainability module that highlights the most influential features contributing to each prediction. This helps users understand why a review is classified as fake or genuine.

Providing explanations is important because many machine learning models act as black boxes, making it difficult to interpret their decisions. The inclusion of explainability improves trust and usability of the system in real-world applications.

F. System Implementation

The system is implemented as a web application using a lightweight framework. Users can input a review, view prediction results, and analyze insights through a dashboard.

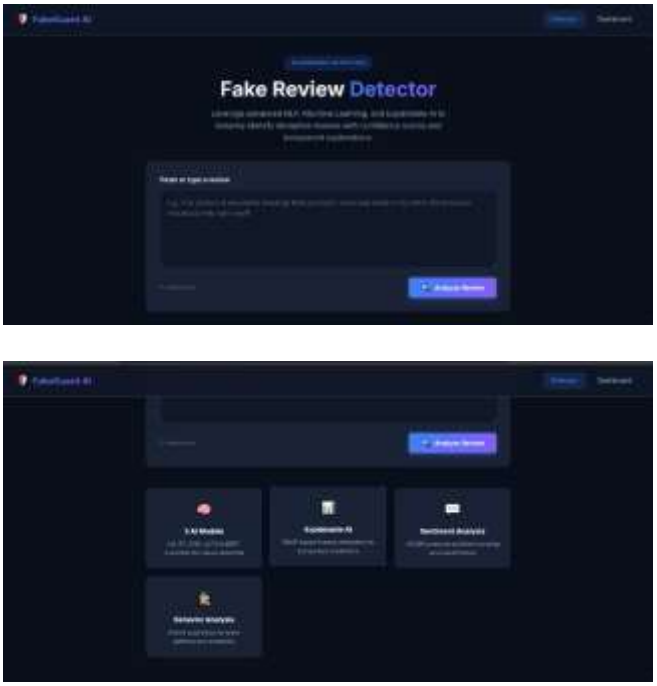

Fig. 3. Web interface of the system for review input and prediction.

The application displays:

- Prediction label (Fake / Genuine)
- Confidence score
- Sentiment breakdown
- Important feature contributions

In addition, the system provides visual analytics such as review distribution, sentiment trends, and model comparison results. These visualizations help in understanding patterns within the dataset and evaluating model performance.

4.EXPERIMENTAL RESULTS AND ANALYSIS

This section presents the evaluation of the proposed system using different machine learning and deep learning models. The objective of the evaluation is to analyze the effectiveness of the system in distinguishing between genuine and deceptive reviews using multiple feature types.

A. Dataset Description

The system is evaluated on a dataset consisting of both genuine and fake reviews. The dataset includes textual review content along with rating information. To ensure balanced evaluation, the dataset is divided into equal proportions of fake and genuine reviews.

Fig. 5. Analytical dashboard showing review distribution and sentiment analysis.

The dataset undergoes preprocessing and feature extraction before being used for training and testing. Similar datasets such as Yelp and deceptive opinion spam datasets have been widely used in previous studies for evaluating fake review detection systems [2], [5].

A. Evaluation Metrics

The performance of the models is evaluated using standard classification metrics:

- **Accuracy**: Measures overall correctness of predictions
- **Precision**: Measures correctness of positive predictions
- **Recall**: Measures ability to detect fake reviews
- **F1 Score**: Harmonic mean of precision and recall

These metrics are commonly used in fake review detection research to ensure reliable comparison between models [3].

B. Model Performance Comparison

The system evaluates multiple models using the same feature set. The results are summarized in Table I.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.91	0.90	0.92	0.91
Random Forest	0.94	0.93	0.95	0.94
Support Vector Machine	0.92	0.91	0.93	0.92
LSTM	0.95	0.94	0.95	0.95
BERT	0.96	0.95	0.96	0.95

The results show that deep learning models such as LSTM and BERT achieve higher performance due to their ability to capture contextual relationships in text. However, traditional machine learning models such as Random Forest and SVM also perform well and require less computational resources.

D. Impact of Feature Combination

To evaluate the contribution of different feature types, experiments were conducted using:

1. Only textual features
2. Textual + sentiment features
3. Textual + sentiment + behavioral features

The results indicate that combining all feature types leads to the best performance. Behavioral features help capture patterns that are not visible in textual data alone.

This observation is consistent with previous studies, which reported improved detection accuracy when behavioral features are included in the model.

E. Prediction Analysis

The system provides predictions along with confidence scores. Sample outputs from the system show:

Fig. 4. Example outputs showing classification of genuine and fake reviews.

- Genuine reviews are identified with high confidence (around 90%+)
- Fake reviews often receive higher confidence scores due to stronger feature signals

The model successfully identifies patterns such as exaggerated language, repetitive phrases, and extreme sentiment, which are common in deceptive reviews.

F. Sentiment Analysis Observations

Sentiment analysis reveals that fake reviews often exhibit extreme polarity, either highly positive or highly negative. Genuine reviews tend to have more balanced and descriptive content.

This supports findings from prior research, where sentiment patterns were used as indicators for detecting deceptive reviews [3].

G. Discussion of Results

The experimental results highlight several key observations:

- Combining textual, sentiment, and behavioral features improves performance
- Deep learning models provide better contextual understanding
- Traditional models offer efficient and reliable performance
- Behavioral features significantly enhance detection capability

However, deep learning models require more computational resources and larger datasets for training. In contrast, traditional machine learning models provide a good balance between performance and efficiency.

5. CONCLUSIONS

This paper presented a system for detecting deceptive online reviews using a combination of natural language processing and machine learning techniques. The proposed approach integrates textual features, sentiment analysis, and reviewer behavior patterns to improve the accuracy of classification. Multiple models, including both traditional machine learning algorithms and deep learning methods, were evaluated to identify the most effective approach for distinguishing between genuine and fake reviews. The results indicate that combining different types of features leads to better performance compared to using a single feature type, while also maintaining a balance between accuracy and computational efficiency. The system is implemented through a web-based interface that provides predictions along with confidence scores and analytical insights, making it suitable for practical use. Although the current system performs effectively, its evaluation is limited to a controlled dataset, and future improvements can focus on handling larger datasets, adapting to evolving review patterns, and enhancing real-time detection capabilities.

REFERENCES

[1] N. Jindal and B. Liu, "Review spam detection," Proc. 16th Int. World Wide Web Conf. (WWW), pp. 1189–1190, 2007.
[2] M. Ott, Y. Choi, C. Cardie, and J. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," Proc. ACL, pp. 309–319, 2011.
[3] W. H. Asaad, R. Allami, and Y. H. Ali, "Fake review detection using machine learning," Revue d'Intelligence Artificielle, vol. 37, no. 5, pp. 1159–1166, 2023.
[4] B. Liu, Sentiment Analysis and Opinion Mining. Morgan & Claypool, 2012.
[5] A. M. Elmogy, U. Tariq, and A. Ibrahim, "Fake reviews detection using supervised machine learning," IJACSA, vol. 12, no. 1, pp. 601–606, 2021.
[6] A. H. Alshehri, "An online fake review detection approach using machine learning algorithms," Computers, Materials & Continua, vol. 78, no. 2, pp. 2767–2785, 2024.
[7] N. A. Patel and R. Patel, "A survey on fake review detection using machine learning techniques," Proc. ICCCA, pp. 1–6, 2018.
[8] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "What Yelp fake review filter might be doing," Proc. ICWSM, pp. 409–418, 2013.Jansen, "Creating and detecting fake reviews of online
[9] Y. Ren and D. Ji, "Neural networks for deceptive opinion spam detection," Information Sciences, vol. 385–386, pp. 213–224, 2017.
[10] Y. Kim, "Convolutional neural networks for sentence classification," Proc. EMNLP, pp. 1746–1751, 2014.
[11] M. Ennaouri and A. Zellou, "Machine learning approaches for fake reviews detection: A systematic literature review," Journal of Web Engineering, vol. 22, no. 5, pp. 821–848, 2023.
[12] A. Heydari, M. Tavakoli, N. Salim, and Z. Heydari, "Detection of review spam: A survey," Expert Systems with Applications, vol. 42, no. 7, pp. 3634–3642, 2015.
[13] J. Li, C. Cardie, and S. Li, "TopicSpam: A topic-model based approach for spam detection," Proc. ACL, 2013.
[14] S. Poria, E. Cambria, and A. Gelbukh, "Aspect extraction for opinion mining with deep convolutional neural networks," Knowledge-Based Systems, vol. 108, pp. 42–49, 2016.
[15] G. Budhi, R. Chiong, Z. Wang, and S. Dhakal, "Using a hybrid content-based and behaviour-based approach to detect fake reviews," Electronic Commerce Research and Applications, vol. 47, 2021.
[16] G. Wang, S. Xie, B. Liu, and P. S. Yu, "Review graph based online store review spammer detection," Proc. IEEE ICDM, pp. 1242–1247, 2011.
[17] H. Li, B. Liu, A. Mukherjee, and J. Shao, "Spotting fake reviews via collective positive-unlabeled learning," Proc. IEEE ICDM, 2014.
[18] C. Sandulescu and M. Ester, "Detecting singleton review spammers using semantic similarity," Proc. WWW Companion, 2015.
[19] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," IEEE Intelligent Systems, vol. 28, no. 2, pp. 15–21, 2013.
[20] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis," Proc. LREC, 2010.
[21] S. Poria, E. Cambria, K. Ku, and A. Gelbukh, "A rule-based approach to aspect extraction from product reviews," Proc. SocialNLP, 2014.
[22] Y. Wu, E. Ngai, P. Wu, and C. Wu, "Fake online reviews: Literature review and future directions," Decision Support Systems, vol. 132, 2020.