IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155155

A Survey on Machine Learning and Deep Learning Techniques for Fake Review Detection

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

Student, Department of Computer Science and Engineering, PRMIT&R, Badnera, India [1]

Professor, Department of Computer Science and Engineering, PRMIT&R, Badnera, India[2]

Assistant Professor, Department of Computer Science and Engineering, PRMIT&R, Badnera, India[3]

Abstract: Online reviews have become an essential source of information for consumers when evaluating products and services on digital platforms. However, the growing presence of deceptive or fake reviews has raised concerns about the reliability of these systems. Over the past decade, researchers have proposed various techniques to detect such reviews using natural language processing, machine learning, and deep learning methods. This paper presents a comprehensive survey of existing approaches for fake review detection, focusing on different categories of techniques including traditional machine learning models, neural network-based methods, and hybrid approaches that combine textual and behavioral features. The survey examines commonly used feature extraction techniques such as term frequency–inverse document frequency, sentiment analysis, and reviewer behavior analysis. It also discusses recent developments in deep learning models that capture contextual relationships within review text. In addition, the paper highlights key challenges faced in this domain, including limited availability of labeled datasets, data imbalance, and evolving strategies used by spammers. Finally, the study identifies important research gaps and suggests directions for future work, particularly in improving model interpretability and developing real-time detection systems. The findings provide a structured understanding of existing methods and their limitations, which can support further research in this area.

Keywords: Fake review detection, opinion spam, machine learning, deep learning, sentiment analysis, text mining, behavioral analysis, feature engineering, survey, review classification.

I. INTRODUCTION

Online reviews have become an important part of decision-making for consumers using digital platforms. Before purchasing a product or service, users often rely on feedback shared by other customers to evaluate quality and reliability. This growing dependence on user-generated content has increased the influence of online reviews on business reputation and consumer trust. However, the openness of these platforms also makes them vulnerable to misuse, where individuals or organizations post deceptive reviews to promote certain products or damage competitors.

The presence of fake reviews has created serious concerns regarding the credibility of online systems. Such reviews are intentionally written to mislead users and can significantly affect purchasing behavior. Early studies on opinion spam highlighted that deceptive reviews often follow identifiable patterns in both text and reviewer activity, making them detectable using computational methods [1], [2]. As the volume of online reviews continues to grow, manual detection has become impractical, leading to the development of automated techniques.

Researchers have explored a wide range of approaches for detecting fake reviews, including traditional machine learning models, natural language processing techniques, and more recently, deep learning methods. Machine learning algorithms such as Support Vector Machines, Logistic Regression, and Random Forest have been widely used to classify reviews based on textual features and statistical patterns [3], [7]. In addition to textual analysis, studies have shown that incorporating behavioral features such as reviewer activity and rating patterns can significantly improve detection performance [5], [8].

With recent advancements, deep learning models such as convolutional and recurrent neural networks have been applied to capture contextual relationships within review text, offering improved performance over traditional methods in many cases [9], [10]. Despite these developments, challenges such as limited labeled data, evolving spam strategies, and lack of interpretability remain significant barriers.

This paper presents a survey of existing techniques for fake review detection, focusing on machine learning and deep learning approaches, feature engineering methods, and current challenges in the field. The objective is to provide a structured understanding of the methods used by researchers and to identify gaps that can guide future work in this domain.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155155

II. BACKGROOUND OF FAKE REVIEW

Online review platforms allow users to share their experiences and opinions about products and services. These reviews are often used by potential customers to evaluate quality and make informed decisions. However, the openness of such platforms makes them susceptible to misuse, where individuals intentionally post deceptive reviews to influence public perception. These misleading reviews, commonly referred to as fake reviews or opinion spam, reduce the reliability of online systems and create challenges for both users and businesses.

Fake reviews are generally written with the intention of promoting or demoting a product or service. Based on their purpose, they can be broadly categorized into promotional reviews and defamatory reviews. Promotional reviews are designed to artificially increase the rating of a product by providing overly positive feedback, while defamatory reviews aim to harm the reputation of competitors through negative or misleading content. In some cases, these reviews are generated in large volumes by coordinated groups, making detection more difficult [1], [8].

From a textual perspective, fake reviews often exhibit certain patterns such as exaggerated language, repetitive phrases, or lack of specific details. However, these patterns are not always consistent, as spammers continuously adapt their writing styles to avoid detection. In addition to textual characteristics, reviewer behavior also provides useful signals. Abnormal patterns such as frequent posting, duplicate content, and unusual rating distributions have been identified as indicators of deceptive activity [5], [16].

The impact of fake reviews extends beyond individual purchasing decisions. They can distort product rankings, mislead consumers, and reduce trust in online platforms. Studies have shown that manipulated reviews can significantly influence user perception and business performance, highlighting the need for effective detection mechanisms [23], [24].

Understanding the nature and characteristics of fake reviews is essential for developing reliable detection techniques. This background provides a foundation for analyzing the different approaches proposed by researchers, which are discussed in the following sections.

III.MACHINE LEARNING APPROACHES

Machine learning techniques have been widely used for detecting fake reviews due to their ability to learn patterns from labeled data and classify reviews based on extracted features. Early approaches focused on supervised learning methods, where models are trained using datasets containing both genuine and deceptive reviews. These methods rely heavily on feature engineering; particularly textual and statistical features derived from review content.

Among the commonly used algorithms, Support Vector Machines (SVM), Logistic Regression, Naïve Bayes, and Decision Trees have shown effective performance in classification tasks. SVM is widely preferred because of its ability to handle high-dimensional textual data and identify optimal decision boundaries between classes. Logistic Regression provides a simple yet efficient approach for binary classification, while Naïve Bayes is effective in handling probabilistic relationships between words in a review [3], [7]. These models are typically trained using features such as term frequency–inverse document frequency (TF-IDF), n-grams, and linguistic patterns.

Several studies have demonstrated that combining multiple features improves the performance of machine learning models. In addition to textual features, statistical attributes such as review length, word distribution, and frequency of specific terms are often used to enhance classification accuracy. However, relying only on textual data has limitations, as deceptive reviews can be written to closely resemble genuine ones.

To address this issue, researchers have incorporated behavioral features into machine learning models. These features include reviewer activity patterns, posting frequency, rating deviations, and duplicate reviews. Mukherjee et al. showed that analyzing reviewer behavior can significantly improve detection accuracy by identifying patterns that are difficult for spammers to disguise [8]. Similarly, Elmogy et al. reported improved performance when combining textual and behavioral features compared to using textual features alone [5].

Ensemble learning techniques such as Random Forest and boosting methods have also been applied to fake review detection. These methods combine multiple classifiers to improve robustness and reduce overfitting. Studies have shown that ensemble approaches can achieve better performance compared to individual models, especially when dealing with complex datasets [15].

Despite their effectiveness, machine learning approaches face certain challenges. Their performance depends heavily on the quality of feature extraction and availability of labeled datasets. In addition, these models may struggle to adapt to new types of deceptive behavior, as they rely on patterns learned from historical data. Nevertheless, machine learning remains a strong baseline for fake review detection and is often used in combination with more advanced techniques.

IV.DEEP LEARNING APPROACHES

In recent years, deep learning techniques have gained significant attention in fake review detection due to their ability to automatically learn complex patterns from large textual data. Unlike traditional machine learning methods, which

DOI: 10.17148/IJARCCE.2026.155155

depend on manually engineered features, deep learning models can capture contextual and semantic relationships within review text, making them more effective in many classification tasks.

Convolutional Neural Networks (CNNs) are widely used for text classification because they can identify local patterns and important phrases within a review. CNN-based models apply convolution operations over word embeddings to extract meaningful features, which helps in distinguishing deceptive content from genuine reviews. Studies have shown that CNN models perform well in capturing key textual patterns that are difficult to identify using traditional approaches [10].

Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, are also commonly used in this domain. LSTM models are designed to capture sequential dependencies in text, allowing them to understand the order and context of words within a review. This makes them suitable for detecting subtle linguistic cues present in deceptive reviews. Ren and Ji demonstrated that neural network models such as LSTM can effectively identify deceptive opinion spam by learning contextual relationships in textual data [9].

More advanced approaches combine CNN and LSTM architectures to leverage both local feature extraction and sequential learning. These hybrid models first extract important features using convolutional layers and then analyze their relationships using recurrent layers. Such combinations have shown improved performance compared to using a single deep learning model, particularly when dealing with complex review data.

Recent developments in natural language processing have introduced transformer-based models, such as BERT, which provide deeper contextual understanding of text. These models use attention mechanisms to capture relationships between words across the entire sentence, rather than relying only on local or sequential patterns. Transformer-based models have demonstrated strong performance in various text classification tasks, including fake review detection, due to their ability to understand context more effectively.

In addition to text representation, deep learning models have also been applied to aspect-based sentiment analysis, where specific aspects of a review are analyzed instead of the entire text. This approach improves detection accuracy by focusing on relevant parts of the review and reducing noise in the data.

Despite their advantages, deep learning approaches have certain limitations. They require large amounts of labeled data for training and involve higher computational costs compared to traditional machine learning models. Moreover, these models often lack interpretability, making it difficult to understand the reasoning behind their predictions. As a result, researchers are increasingly exploring hybrid approaches that combine deep learning with feature-based methods to balance performance and interpretability.

V. FETURE ENGINEERING TECHNIQUES

Feature engineering plays a central role in fake review detection, as the performance of both machine learning and deep learning models largely depends on how effectively the input data is represented. Researchers have explored different types of features to capture the characteristics of deceptive reviews, including textual, sentiment-based, and behavioral features. Combining these features has been shown to improve detection accuracy compared to relying on a single type of information.

A. Textual Features

Textual features are the most commonly used in fake review detection. These features are extracted directly from review content and include representations such as term frequency–inverse document frequency (TF-IDF), bag-of-words, and n-grams. TF-IDF assigns importance to words based on their frequency within a document and across the dataset, helping to highlight distinguishing terms [3]. N-grams capture sequences of words, which can reveal common phrases used in deceptive reviews.

Linguistic patterns such as part-of-speech tags, writing style, and readability have also been used to identify deceptive content. However, purely textual features may not always be sufficient, as fake reviews can be written to closely resemble genuine ones.

B. Sentiment –Based Feature

Sentiment analysis is widely used to capture the emotional tone of a review. These features include polarity scores such as positive, negative, and neutral sentiments, as well as compound scores that represent overall sentiment. Deceptive reviews often exhibit extreme sentiment, either overly positive or overly negative, which can act as an indicator of manipulation.

Lexicon-based approaches, such as those using sentiment dictionaries, and machine learning-based sentiment models have been applied to extract these features. Resources such as SentiWordNet and concept-based sentiment analysis frameworks have been used to improve the quality of sentiment representation [19], [20]. Although sentiment features provide useful insights, they are more effective when combined with other feature types.

DOI: 10.17148/IJARCCE.2026.155155

C. Behavioral Features

Behavioral features focus on patterns related to reviewer activity rather than the content of the review. These features include reviewer posting frequency, review timing, rating distribution, and duplication of content. Such patterns are difficult for spammers to disguise and therefore provide strong indicators of deceptive behavior.

Mukherjee et al. demonstrated that analyzing reviewer behavior can significantly improve detection accuracy by identifying unusual activity patterns [8]. Similarly, studies have shown that combining behavioral features with textual features leads to better classification performance compared to using either approach alone [5], [15].

D. Hybrid Feature Approaches

Recent research has emphasized the importance of combining multiple feature types to improve detection performance. Hybrid approaches integrate textual, sentiment, and behavioral features to capture different aspects of deceptive reviews. These approaches provide a more comprehensive representation of review data and help overcome the limitations of individual feature types.

Studies have shown that hybrid feature models achieve higher accuracy and robustness, especially in complex datasets where deceptive reviews are designed to mimic genuine ones [11], [22]. As a result, feature combination has become a standard practice in modern fake review detection systems.

VI. CHALLENGES AND LIMITATIONS

Despite significant progress in fake review detection, several challenges continue to affect the performance and reliability of existing approaches. One of the primary issues is the limited availability of high-quality labeled datasets. Many datasets used in research are either small, domain-specific, or artificially generated, which makes it difficult for models to generalize to real-world scenarios. In addition, labeling reviews as genuine or fake often requires manual effort, which is time-consuming and may introduce subjectivity [11], [22].

Another major challenge is data imbalance. In real-world platforms, genuine reviews significantly outnumber fake ones, leading to skewed datasets. Machine learning models trained on such data may become biased toward the majority class, reducing their ability to accurately detect deceptive reviews. Handling this imbalance requires careful data preprocessing or specialized learning techniques.

The evolving nature of spam behavior also poses a serious limitation. Spammers continuously adapt their strategies to avoid detection by mimicking genuine writing styles and using more sophisticated language. As a result, models trained on older datasets may fail to detect newly generated fake reviews. This makes it necessary to design adaptive systems that can handle changing patterns over time [12].

Another limitation is the dependence on feature engineering in traditional machine learning approaches. The effectiveness of these models relies heavily on the selection of appropriate features, which may not capture all aspects of deceptive behavior. Although deep learning methods reduce the need for manual feature extraction, they introduce other challenges such as high computational cost and the requirement for large training datasets [9].

Lack of interpretability is also an important concern, particularly for complex models. Many deep learning models act as black boxes, making it difficult to understand why a review is classified as fake or genuine. This limits their practical adoption in real-world applications where transparency is required.

Finally, scalability and real-time detection remain open challenges. Online platforms generate a large volume of reviews continuously, and processing this data efficiently requires scalable systems. Many existing approaches are not optimized for real-time analysis, which limits their usability in dynamic environments.

Studies have shown that hybrid feature models achieve higher accuracy and robustness, especially in complex datasets where deceptive reviews are designed to mimic genuine ones [11], [22]. As a result, feature combination has become a standard practice in modern fake review detection.

VII. RESEARCH GAPS

Although significant progress has been made in fake review detection, several research gaps still exist that limit the effectiveness of current approaches. One major gap is the lack of integration between different types of features. Many studies focus either on textual features or behavioral patterns, but only a limited number of approaches effectively combine these features into a unified model. Since deceptive reviews are designed to mimic genuine content, relying on a single feature type often leads to reduced detection accuracy. This highlights the need for more comprehensive models that integrate textual, sentiment, and behavioral features in a balanced manner [5], [11].

Another important gap is the limited focus on model interpretability. While deep learning models have improved detection performance, they often operate as black-box systems, making it difficult to understand the reasoning behind their predictions. This lack of transparency reduces trust and limits practical adoption, especially in real-world

DOI: 10.17148/IJARCCE.2026.155155

applications where explanations are necessary for decision-making. There is a clear need for methods that provide interpretable outputs while maintaining high performance.

The availability and quality of datasets also remain a critical issue. Many existing datasets are either small, domain-specific, or synthetically generated, which does not fully represent real-world scenarios. In addition, there is a lack of standardized benchmark datasets that allow fair comparison between different approaches. This makes it difficult to evaluate the true effectiveness of proposed models across diverse platforms [22].

Another gap is the limited development of real-time detection systems. Most existing approaches are designed for offline analysis and do not address the challenges of processing large volumes of streaming data. As online platforms continue to grow, there is a need for scalable systems capable of detecting fake reviews in real time without compromising accuracy.

Furthermore, current models often struggle to adapt to evolving spam strategies. Spammers continuously modify their writing style and behavior to bypass detection systems, which reduces the effectiveness of static models trained on historical data. This creates a need for adaptive learning approaches that can update models dynamically as new patterns emerge [12].

Finally, there is limited exploration of hybrid approaches that combine machine learning, deep learning, and domain knowledge. While some studies have attempted to integrate different techniques, there is still scope for developing more robust frameworks that balance performance, interpretability, and computational efficiency.

Studies have shown that hybrid feature models achieve higher accuracy and robustness, especially in complex datasets where deceptive reviews are designed to mimic genuine ones [11], [22]. As a result, feature combination has become a standard practice in modern fake review detection systems.

VIII. CONCLUSION

This paper presented a comprehensive survey of existing approaches for detecting fake reviews on online platforms. The study examined various techniques including traditional machine learning models, deep learning methods, and hybrid approaches that combine textual, sentiment, and behavioral features. It was observed that while machine learning models provide efficient and reliable performance, deep learning techniques offer improved capability in capturing contextual relationships within review text. Feature engineering plays a critical role in enhancing detection accuracy, especially when multiple feature types are combined. The survey also highlighted key challenges such as limited availability of high-quality datasets, data imbalance, evolving spam strategies, and lack of interpretability in complex models. In addition, several research gaps were identified, including the need for better feature integration, explainable models, and real-time detection systems. Overall, this study provides a structured understanding of the current state of fake review detection and outlines directions for future research aimed at improving the reliability and effectiveness of detection systems.

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]. I 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.

[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.

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155155

[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

