Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 4 | April-2026
ISSN (O) 3107-6696

A Comparative Review of Fake Review Detection Techniques Using Machine Learning and Transformer Models

Humaid Ahmad Kidwai[1], Nihal Gupta[2], Abdullah Suhail[3], Ms. Hina Parveen[4]

Department of Computer Science, Integral University, Lucknow, India
kidwaihumaid@gmail.com, nihalgupta.in@gmail.com, abdullahcodes1812@gmail.com, hparveen@iul.ac.in

----------***----------

Abstract – Online review systems play a crucial role in shaping consumer decisions in modern e-commerce environments. However, the increasing prevalence of deceptive or fake reviews has raised serious concerns regarding the reliability of such platforms. Over the years, a wide range of techniques have been proposed to address this issue, spanning traditional machine learning methods, deep learning architectures, and transformer-based models. This paper presents a comprehensive comparative review of major approaches used for fake review detection. The analysis covers feature-based classification methods, network-oriented models, neural architectures such as CNN and LSTM, and advanced transformer models including BERT and RoBERTa. Each approach is evaluated in terms of model design, dataset usage, performance metrics, advantages, and limitations. The study highlights a clear shift from manually engineered feature-based systems to context-aware deep learning frameworks. Although recent transformer-based models demonstrate strong performance, challenges such as cross-domain adaptability, computational complexity, interpretability, and evolving spam tactics remain unresolved. This review aims to provide a structured understanding of existing techniques and identify future research directions for building efficient and scalable fake review detection systems.

Keywords—*Fake review detection, opinion spam, machine learning, deep learning, BERT, RoBERTa, transformer models, text classification.*

Introduction

The expansion of digital commerce platforms has reshaped consumer decision-making by increasing reliance on user-generated feedback. In this environment, user-generated reviews now act as a primary information channel, allowing users to evaluate products based on the experiences of others. As a result, review systems have become a critical component of digital marketplaces, directly influencing customer trust and product reputation.

However, the increasing reliance on user-generated reviews has also led to the emergence of deceptive practices, commonly known as fake reviews or opinion spam. These reviews are intentionally created to manipulate product ratings by promoting certain products or damaging competitors. Such activities not only mislead consumers but also undermine the credibility of online platforms.

To address this issue, researchers have explored various computational techniques over the past decade. Early approaches primarily relied on traditional machine learning models that utilized manually designed features derived from textual content, user behavior, and metadata. Algorithms such as logistic regression and support vector machines were widely applied to identify suspicious patterns in reviews.

With advancements in artificial intelligence, deep learning models have been introduced to automatically learn representations from textual data. Techniques such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks improved detection performance by capturing semantic and sequential dependencies. More recently, transformer-based models like BERT and RoBERTa have further enhanced text understanding by leveraging attention mechanisms to capture contextual relationships.

Despite these advancements, several challenges remain unresolved. Issues such as poor generalization across datasets, high computational requirements, class imbalance, and continuously evolving spam strategies continue to affect detection systems. Therefore, there is a need for more robust and scalable solutions.

In this paper, a comparative review of fake review detection techniques is presented, covering traditional machine learning methods, deep learning models, and transformer-based approaches. The study systematically evaluates existing research and highlights key limitations and future research opportunities.

Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 4 | April-2026
ISSN (O) 3107-6696

Literature Survey

A wide range of studies have investigated methods for identifying deceptive review content across online platforms, leading to the development of diverse computational techniques. One of the earliest contributions in this domain was made by Jindal and Liu [1], who formally introduced the concept of opinion spam. Their work categorized different types of spam reviews and applied similarity-based techniques to detect duplicate or near-duplicate content. By combining textual and behavioral features with logistic regression, their study demonstrated the potential of automated systems in identifying spam patterns, although detecting subtle deception remained challenging.

Subsequently, Ott et al. [2] developed one of the first benchmark datasets for fake review detection, consisting of manually curated hotel reviews. Using linguistic and psycholinguistic features such as n-grams and LIWC, they evaluated multiple classifiers and found that Support Vector Machines achieved strong performance. Their findings indicated that machine learning models could effectively distinguish deceptive content, often outperforming human judgment.

To overcome the limitations of purely text-based approaches, Mukherjee et al. [3] proposed a graph-based framework that models relationships among users, reviews, and products. Their method, known as SpEagle, utilized probabilistic graphical models to capture dependencies within review networks. Experimental results on datasets such as YelpChi and YelpNYC showed that incorporating relational and behavioral information significantly improves detection robustness.

With the advancement of deep learning, researchers began exploring neural network-based models for automatic feature extraction. Jain et al. [4] investigated architectures including CNN, LSTM, and multilayer perceptrons. Their experiments demonstrated that LSTM models achieved superior performance by effectively capturing sequential dependencies in text, highlighting the advantage of deep learning over traditional approaches.

More recent studies have focused on transformer-based models, which have shown remarkable success in natural language processing tasks. A hybrid approach combining BERT with LSTM and Monte Carlo Dropout was proposed in [7], aiming to improve prediction reliability. The model achieved strong performance on Yelp datasets, demonstrating the effectiveness of contextual embeddings. Similarly, Boobalan [5] applied transfer learning using BERT and reported competitive results, reinforcing the applicability of transformer architectures.

Further improvements have been achieved through hybrid models that integrate transformers with sequential learning mechanisms. Mohawesh et al. [6] introduced a framework combining RoBERTa with LSTM to capture both contextual and temporal features. Their model achieved high accuracy and F1-scores across benchmark datasets, indicating that hybrid architectures can enhance detection performance.

Overall, the progression of research in fake review detection reflects a transition from feature-based machine learning methods to advanced deep learning and transformer-based techniques. While modern models achieve high accuracy, challenges related to generalization, computational cost, and interpretability continue to require attention.

In addition to the discussed approaches, several studies have explored alternative perspectives for detecting deceptive reviews. For instance, Li et al. [9] and Feng et al. [10] explored language-based characteristics and writing patterns associated with deceptive reviews, highlighting the role of syntactic and semantic cues in identifying fake reviews. Similarly, Jindal et al. [11] focused on unusual review patterns, while Li et al. [12] proposed collective learning methods to improve detection performance. These studies further demonstrate the diversity of approaches in opinion spam detection.

The evolution of fake review detection techniques from traditional machine learning methods to modern transformer-based models is illustrated in Fig. 1.

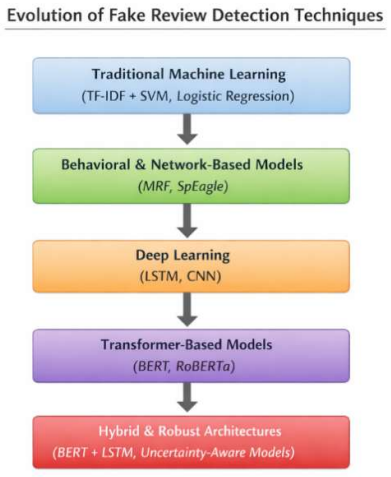

Fig. 1. Evolution of Fake Review Detection Techniques

Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 4 | April-2026
ISSN (O) 3107-6696

COMPARATIVE ANALYSIS OF EXISTING APPROACHES

In order to better understand the differences between existing fake review detection techniques, a comparative analysis of representative studies is presented in this section. The selected research works are evaluated based on their methodological approach, dataset characteristics, and reported performance metrics. The comparison highlights the evolution of detection strategies from traditional machine learning models to modern deep learning and transformer-based architectures. A summary of the analyzed studies is provided in Table I.

TABLE I COMPARATIVE SUMMARY OF FAKE REVIEW DETECTION TECHNIQUES

Ref	Year	Approach	Model	Dataset	Performance
[1]	2008	Feature-based ML	Logistic Regression	Amazon Reviews	AUC 98.7%
[2]	2011	Text Classification	SVM	OpSpam Dataset	Accuracy 89.8%
[3]	2015	Graph-based Detection	SpEagle (MRF)	YelpChi / YelpNYC	AUC ≈ 0.79
[4]	2019	Deep Learning	LSTM / CNN	Ott + Yelp	Accuracy 96.75%
[5]	2023	Transformer-based	BERT + MCD	Yelp Dataset	Accuracy 91.75%
[6]	2024	Hybrid Transformer	RoBERTa + LSTM	OpSpam / Deception	Accuracy 96.03%

The comparison presented in Table I highlights the gradual evolution of fake review detection techniques from traditional feature-based machine learning approaches to advanced deep learning and transformer-based architectures. Early studies relied primarily on handcrafted textual and behavioral features combined with classical classifiers such as logistic regression and support vector machines. In contrast, more recent approaches leverage contextual language representations and neural architectures to capture deeper semantic relationships in review text, leading to improved detection performance.

A visual comparison of the reported performance of the analyzed models is illustrated in Fig. 2.

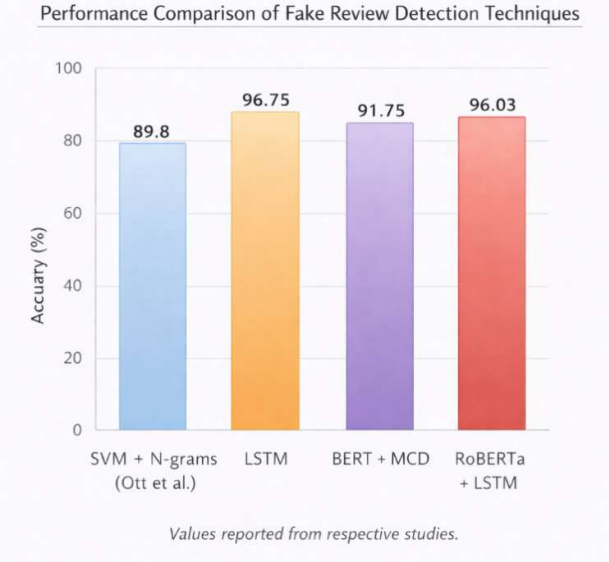

Fig. 2. Performance comparison of fake review detection models

The scale of datasets used in different studies is illustrated in Fig. 3.

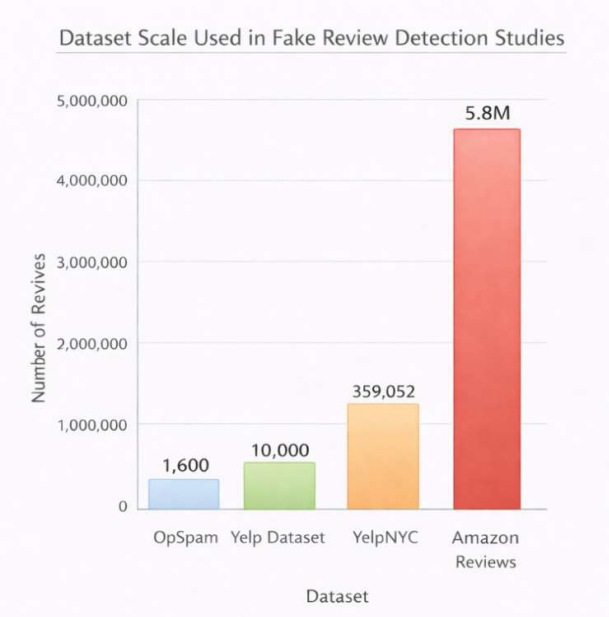

Fig. 3. Dataset scale used in fake review detection studies

IV. TREND ANALYSIS AND OBSERVATIONS

From the analysis of the selected studies, several clear trends can be identified in the evolution of fake review detection techniques. A key observation from the reviewed studies is the progressive shift from traditional feature-based machine learning approaches to more advanced deep learning and transformer-based models. Earlier methods primarily relied on manually engineered features, such as linguistic patterns, review metadata, and user behavior, combined with classifiers like logistic regression and support vector machines. While these approaches provided foundational insights, their dependence on handcrafted features limited their adaptability and scalability.

With the emergence of deep learning, models such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks enabled automatic feature extraction from textual data. These models significantly improved performance by capturing semantic and sequential relationships within reviews, reducing the reliance on manual feature engineering.

More recently, transformer-based architectures, including BERT and RoBERTa, have demonstrated superior performance due to their ability to model contextual dependencies using attention mechanisms. These models consider bidirectional context, allowing for a deeper understanding of linguistic nuances in review text. As observed in the reviewed studies, transformer-based approaches generally achieve higher accuracy and improved generalization compared to earlier methods.

The emergence of transformer-based architectures has significantly influenced recent research in natural language processing. Models such as BERT [13] and its optimized variant RoBERTa [14] have demonstrated strong capabilities in capturing contextual representations, making them highly effective for text classification tasks including fake review detection.

Another emerging trend is the development of hybrid models that integrate multiple techniques. For example, combining transformer-based embeddings with sequential models such as LSTM allows systems to capture both contextual and temporal patterns. This hybridization has shown promising improvements in detection performance.

In addition to methodological advancements, there has been a noticeable increase in dataset size and diversity. However, challenges such as class imbalance, noisy labels, and dataset reliability continue to affect model performance. Furthermore, despite achieving high accuracy, modern models often introduce increased computational complexity, which may limit their deployment in real-time environments.

Overall, the progression of research indicates a shift toward more data-driven and context-aware systems, but achieving a balance between performance, efficiency, and interpretability remains a key challenge.

V. METHODOLOGY

In this study, a machine learning-based approach is adopted for detecting fake reviews using textual data. The dataset consists of labeled reviews categorized as genuine and fake, obtained from a publicly available dataset on Kaggle [8]. The labels are converted into binary format, and the dataset is divided into training and testing sets using an 80:20 split while preserving class distribution.

In the preprocessing stage, the text data is normalized by converting it to lowercase and removing stopwords. The processed text is then transformed into numerical features using the Term Frequency–Inverse Document Frequency (TF-IDF) technique. Both unigrams and bi-grams are considered to capture contextual relationships between words and improve semantic understanding.

A Support Vector Machine (SVM) classifier is used for the classification task due to its effectiveness in text-based problems, as also demonstrated in earlier foundational work on text categorization [15]. Class imbalance is handled using appropriate weighting, and probability calibration is applied to obtain confidence scores for predictions. The model is evaluated using accuracy and F1-score to ensure balanced performance.

Once trained, the system can classify new input reviews in real time by predicting whether a review is genuine or fake, along with an associated confidence score.

The overall workflow of the proposed system is illustrated in Fig. 4.

Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 4 | April-2026
ISSN (O) 3107-6696

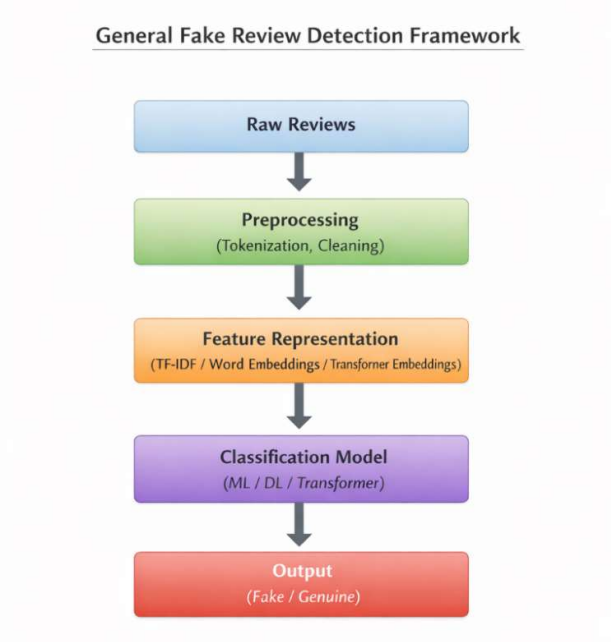

Fig. 4. General fake review detection framework

In addition to traditional approaches, recent advancements in fake review detection incorporate deep learning and transformer-based models. These models utilize contextual embeddings to capture semantic meaning and relationships within text. A hybrid architecture combining transformer-based embeddings with sequential models is shown in Fig. 5, which highlights the integration of contextual and sequential learning for improved detection performance.

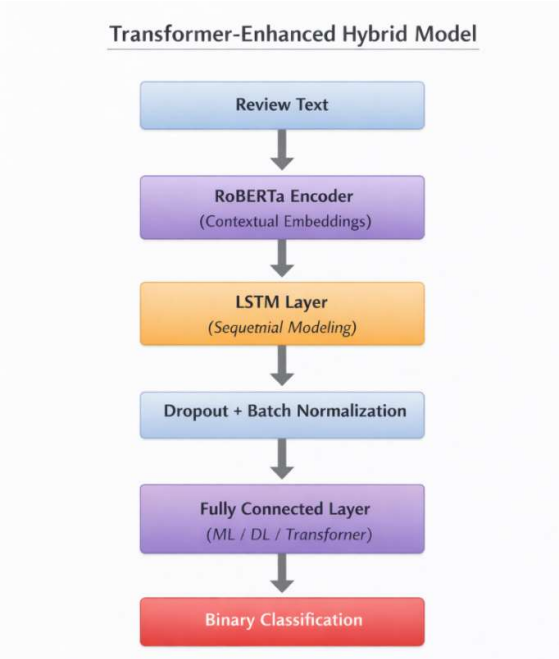

Fig. 5. Transformer-based hybrid architecture for fake review detection

To ensure robustness, the model performance is evaluated using multiple metrics, and efforts are made to minimize overfitting through appropriate validation strategies.

VI. RESEARCH GAPS

Despite significant advancements in fake review detection, several critical research gaps remain unresolved. One of the primary challenges is the limited ability of existing models to generalize across domains. Most approaches are trained and evaluated on specific datasets, such as Yelp or Amazon, which restricts their applicability to real-world scenarios involving diverse platforms.

Another major limitation is the heavy reliance on labeled datasets. Supervised learning methods require large volumes of annotated data, which are often expensive and time-consuming to obtain. Moreover, inconsistencies and noise in labeling can negatively impact model performance and reliability.

In addition, although deep learning and transformer-based models achieve high accuracy, they often come with substantial computational costs. These models require significant processing power and memory, making them less suitable for deployment in resource-constrained or real-time environments.

The issue of interpretability also remains a concern. Advanced models, particularly transformer-based architectures, function as black boxes, providing limited insight into how predictions are made. This lack of transparency can reduce trust in automated detection systems.

Furthermore, existing research predominantly focuses on textual features, while underutilizing complementary signals such as temporal patterns, reviewer credibility, and network-level interactions, which could enhance detection robustness.

Therefore, there is a strong need for future models that are not only accurate but also generalizable, efficient, interpretable, and capable of leveraging multi-dimensional data.

VII. FUTURE WORK

Based on the identified research gaps, several directions can be explored to improve fake review detection systems. Future work can focus on developing models that generalize well across different platforms and domains. This can be achieved

by using domain adaptation techniques and training on more diverse datasets.

Another promising direction is the use of semi-supervised and unsupervised learning approaches to reduce dependency on labeled data. Techniques such as active learning and self-supervised learning can help in efficiently utilizing available data.

Additionally, efforts can be made to design lightweight and computationally efficient models that can be deployed in real-time systems. Optimizing transformer-based models and reducing their complexity can significantly improve their practical applicability.

Improving model interpretability is also an important area for future research. Developing explainable AI techniques can help users understand how predictions are made, thereby increasing trust in automated systems.

Finally, integrating multiple sources of information such as textual content, user behavior, and network relationships can lead to more robust and accurate fake review detection systems.

VIII. CONCLUSION

This study provided a structured synthesis of existing techniques for detecting deceptive reviews across multiple computational paradigms, covering traditional machine learning approaches, deep learning models, and transformer-based architectures. A detailed comparative analysis of representative studies was conducted based on methodology, dataset, and performance metrics.

The study highlighted a clear evolution of detection techniques from feature-based models to advanced context-aware systems. While recent approaches demonstrate improved performance, several challenges such as generalization, computational complexity, and interpretability still remain.

The analysis offers consolidated insights into current methodologies and their practical limitations, while highlighting the current state of fake review detection and key research gaps. of fake review detection and identify key research gaps for future exploration. Overall, the study emphasizes the need for developing more robust, efficient, and scalable solutions for real-world applications.

This study not only consolidates existing approaches but also provides a structured perspective that can support the development of more practical and real-world applicable detection systems.

References

- B. Jindal and B. Liu, “Opinion spam and analysis,” in Proc. Int. Conf. Web Search Data Mining (WSDM), 2008, pp. 219–230.
- M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, “Finding deceptive opinion spam by any stretch of the imagination,” in Proc. 49th Annu. Meeting Assoc. Comput. Linguistics (ACL), 2011, pp. 309–319.
- Mukherjee, V. Venkataraman, B. Liu, and N. Glance, “Collective opinion spam detection: Bridging review networks and metadata,” in Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), 2015, pp. 985–994.
- Jain, A. Gupta, and R. Katarya, “Spam review detection using deep learning,” in Proc. IEEE Conf., 2019, pp. 1–6.
- P. Boobalan, “Fake review detection using BERT transfer learning algorithm,” Int. J. Res. Appl. Sci. Eng. Technol., vol. 12, no. 3, pp. 3360–3365, 2024.
- R. Mohawesh, H. B. Salameh, and others, “Fake review detection using transformer-based enhanced LSTM and RoBERTa,” Int. J. Cogn. Comput. Eng., 2024.
- Author et al., “Robust fake review detection using uncertainty-aware LSTM and BERT,” in Proc. IEEE Int. Conf. Comput. Intell. Commun. Netw. (CICN), 2023.
- M. Mexwell, “Fake reviews dataset,” Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/mexwell/fake-reviews-dataset
- J. Li, M. Ott, C. Cardie, and E. Hovy, “Towards a general rule for identifying deceptive opinion spam,” in Proc. ACL, 2014.
- S. Feng, R. Banerjee, and Y. Choi, “Syntactic stylometry for deception detection,” in Proc. ACL, 2012.
- N. Jindal, B. Liu, and E.-P. Lim, “Finding unusual review patterns using unexpected rules,” in Proc. CIKM, 2010.

- H. Li, Z. Chen, B. Liu, X. Wei, and J. Shao, "Spotting fake reviews via collective positive-unlabeled learning," in Proc. ICDM, 2014.
- J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019.
- Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," 2019.
- T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in Proc. ECML, 1998.