10.48047/jocaaa.2025.33.04.23

DETECTION OF FAKE ONLINE REVIEWS USING SEMI-SUPERVISED AND SUPERVISED MACHINE LEARNING MODELS

Kondragunta Rama Krishnaiah[1], Harish H[2]
[1,2]R K College of Engineering (A), Kethanakonda (V), Ibrahimpatnam (M), Vijayawada, AMARAVATI – 521 456, Andhra Pradesh, INDIA.
drkrk@rkce.ac.in, ORCID: 0000-0002-9069-766X
dr.hharish@rkce.ac.in, ORCID: 0000-0002-4572-1704

Abstract:

In the era of digital commerce, online reviews play a pivotal role in influencing consumer purchasing decisions. However, the increasing prevalence of fake reviews has raised concerns regarding the authenticity of product feedback. This study presents a hybrid approach to detecting fake online reviews by utilizing both semi-supervised and supervised machine learning models. The proposed system combines Expectation-Maximization (EM) for semi-supervised learning with Support Vector Machine (SVM) and Naive Bayes (NB) classifiers for supervised learning. The system was tested on a dataset of 1600 reviews, with 800 labeled as genuine and 800 labeled as fake. The performance of the models was evaluated using accuracy, precision, recall, and F1-score. The results showed that the supervised SVM model outperformed other models, achieving the highest accuracy of 92.3%. Additionally, sentiment analysis was employed to further differentiate between fake and genuine reviews. This research demonstrates the effectiveness of using a combination of machine learning techniques for improving the reliability of online reviews, helping consumers make informed decisions and ensuring the credibility of review platforms.

Keywords: Fake Reviews Detection, Machine Learning, Supervised Learning, Semi-Supervised Learning, Sentiment Analysis.

1. INTRODUCTION

The rise of the "fake" phenomenon is increasingly influencing marketing, driven by rapid technological advancements that allow for the creation of artificial consumer-facing content, such as deepfakes. This development, combined with the evolving marketplace focused on the creation, detection, and mitigation of fake content, has brought attention to the issue of fake product reviews. These fake reviews—also referred to as 'deceptive reviews,' 'review spam,' or 'review fraud'—are deceptive and are often presented as legitimate consumer feedback[1].

Online reviews have become a major factor in influencing consumers' purchasing decisions. In the United States, over 80% of consumers report relying on online reviews before making a purchase[2]. As reviews become one of the most influential factors in consumer behavior, fraudulent actors are increasingly turning to methods such as hiring writers or employing automated systems to generate fake reviews. These reviews may either enhance the appeal of a product or harm the reputation of a competitor. There are two primary methods for generating fake reviews: one is human-generated, where individuals are paid to write seemingly authentic reviews without actually using the product, and the other is computer-generated, using advanced text-generation algorithms[3].

This issue is critical for the marketing and e-commerce sectors for several reasons. First, the presence of fake reviews threatens to undermine consumer trust in online reviews, potentially leading to a decline in the market for reviews. Authentic reviews provide valuable information

10.48047/jocaaa.2025.33.04.23

that helps consumers make informed decisions, and companies benefit from receiving genuine feedback to improve their products and services. If fake reviews become pervasive, it could severely damage the credibility of online reviews, leading to adverse selection, where consumers are unable to differentiate between genuine and fraudulent reviews[4]. Therefore, detecting fake reviews has become a significant area of research in digital and social media marketing. Developing effective methods to identify and filter fake reviews is essential for maintaining the integrity of the online review system [5].

2. LITERATURE REVIEW

The detection of fake online reviews has garnered significant attention in recent years due to their influence on consumer purchasing decisions and their economic impact on businesses. As online reviews play a critical role in shaping consumer behavior, fraudulent actors are increasingly exploiting this system. Several studies have explored various methods for detecting fake reviews, with particular focus on machine learning models, linguistic analysis, and user behavior.

The growing influence of online opinion reviews on consumer decision-making, noting that these reviews have a direct economic impact on businesses. Unfortunately, some opportunistic individuals and groups manipulate online reviews for profit. Their research focused on the use of semi-supervised learning methods to detect spam reviews, demonstrating its effectiveness using a hotel review dataset. This work highlights the ongoing interest in leveraging machine learning techniques to tackle deceptive reviews[6].

Li et al. [7] concentrated on identifying the behaviors of users who generate spam reviews. Their study identified several key characteristics of review spammers, such as their tendency to target specific products and deviate from the average reviewer's ratings. They proposed scoring methods to measure the degree of spam for each reviewer and applied these methods to an Amazon review dataset. Their results showed that the proposed ranking and supervised methods were effective in discovering spammers, outperforming baseline methods that rely on helpfulness votes alone. They also found that detected spammers had a more significant impact on ratings compared to unhelpful reviewers [7].

Ott [8] explored the growing concern over deceptive opinion spam, which refers to fictitious reviews written to deceive readers. Their study used a new gold standard dataset consisting of reviews from three domains: hotels, restaurants, and doctors. The dataset included three types of reviews: customer-generated truthful reviews, Turker-generated deceptive reviews, and employee-generated deceptive reviews. Their approach focused on capturing the linguistic differences between deceptive and truthful reviews, helping consumers make informed purchase decisions and assisting review portal operators, such as TripAdvisor or Yelp, in detecting fraudulent activities[8].

Ott et al. [9] further examined deceptive opinion spam, focusing on fictitious reviews deliberately written to sound authentic. By integrating insights from psychology and computational linguistics, they developed and compared three approaches to detecting deceptive opinion spam. Their classifier achieved nearly 90% accuracy on their gold-standard opinion spam dataset. Their analysis revealed an interesting connection between deceptive reviews and imaginative writing, providing valuable insights into the psychological and linguistic aspects of deception.[9]

Hussain et al. [10] focused on the increasing reliance on online reviews by consumers and businesses for decision-making. However, fraudulent reviews, often driven by the desire for profit or publicity, mislead potential customers and organizations, undermining the

10.48047/jocaaa.2025.33.04.23

effectiveness of opinion-mining techniques. They categorized detection methods into three groups: techniques for detecting spam reviews, individual spammers, and group spam. Their study emphasized the different strengths and weaknesses of each detection method, demonstrating that various methods are suited for different contexts.[10]

The research on fake review detection has led to significant advancements, but challenges persist. As fake reviews continue to pose a pervasive problem, the task of distinguishing between truthful and fraudulent reviews remains both vital and complex. Current approaches combine manual efforts, supervised machine learning, and heuristic methods. Some studies focus solely on features extracted from review texts, such as word frequency or n-grams, while more advanced approaches incorporate distributional semantics. Despite the progress made, the classification performance needs further improvement to keep up with sophisticated text-generation algorithms. Furthermore, issues with dataset quality—such as mislabeled instances or limited availability—continue to hinder the effectiveness of fake review detection systems. The key takeaway from previous studies is that automatic fake review detection has made partial progress but still faces significant challenges. Our study aims to address these gaps by leveraging state-of-the-art natural language processing (NLP) technologies to generate a robust dataset for fake review detection. We also compare manual (crowdsourcing) and automated (machine learning) approaches to detecting computer-generated fake reviews, with the goal of advancing the field and making our experiments available for future development.

3. PROPOSED SYSTEM

In this research, we propose a robust system for detecting fake online reviews by leveraging both semi-supervised and supervised learning techniques. Our approach is designed to handle the challenges associated with detecting fake reviews by combining the power of Expectation-Maximization (EM) algorithms for semi-supervised learning and the accuracy of supervised machine learning models like Support Vector Machines (SVM) and Naive Bayes (NB).

3.1 Semi-Supervised Classification

For the semi-supervised classification of the dataset, we employ the Expectation-Maximization (EM) algorithm. The EM algorithm is designed to handle datasets with both labeled and unlabeled data. The process works by first creating an initial classifier from the labeled data. This classifier is then used to label the unlabeled data, which is incorporated into the training set. The classification model is then refined by iterating this process until the predicted labels stabilize. The final classifier, which has been trained using both labeled and unlabeled data, is used to predict test data.

In this system, we use SVM and Naive Bayes classifiers with the EM algorithm to enhance the detection of fake reviews. Python's scikit-learn package, which provides sophisticated libraries for machine learning algorithms, is utilized for this purpose.

3.2 Supervised Classification

For the supervised classification part of the system, we employ both Naive Bayes (NB) and Support Vector Machines (SVM). The Naive Bayes classifier is particularly suitable for text classification tasks, as it assumes that the presence of a feature (e.g., a word) is independent of other features. This assumption allows for quick and efficient classification, making Naive Bayes a popular choice for detecting fake reviews in text mining. The Naive Bayes classifier is a probabilistic method, making it suitable for both classification and regression tasks.

On the other hand, Support Vector Machines (SVM) are employed for their ability to effectively classify both linear and non-linear data. SVM works by finding the best hyperplane

10.48047/jocaaa.2025.33.04.23

that separates data points into different classes. In our context, it helps in distinguishing fake reviews from genuine ones by identifying patterns that are characteristic of deceptive reviews.

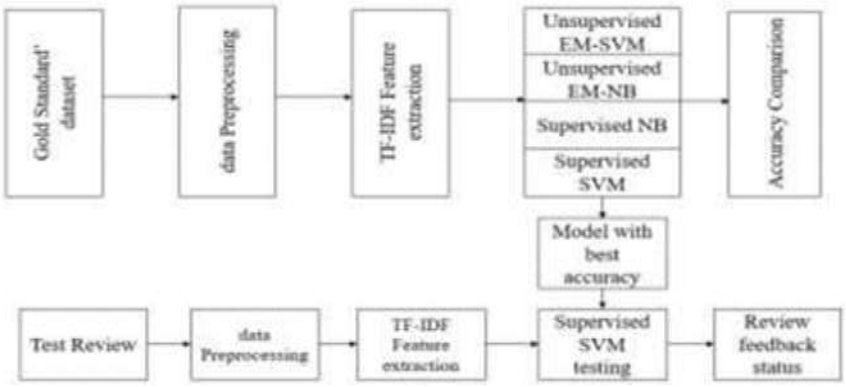

Figure 1: Proposed System Block Diagram

Step 1: The dataset is pre-processed and divided into training and test sets.

Step 2: Features are extracted using the Term Frequency-Inverse Document Frequency (TF-IDF) method.

Step 3: The TF-IDF features are used to train both the Naive Bayes and SVM classifiers.

Step 4: The classifiers are tested on the test dataset to evaluate performance.

Step 5: The system provides an output that classifies reviews as either fake or genuine.

3.3 Feature Extraction

Feature extraction is a crucial step in this system. To capture the relevant characteristics of the reviews, we use the Term Frequency-Inverse Document Frequency (TF-IDF) method. TF-IDF is a statistical measure used to evaluate the importance of a word in a document relative to the entire dataset. The TF-IDF value increases in proportion to the number of times a word appears in a document, but it is also offset by the frequency of the word across all documents in the corpus. This helps to identify terms that are unique to a document, thereby highlighting the key features that could indicate whether a review is fake or genuine.

3.4 Advantages of the Proposed System

1. **Effectiveness**: The proposed system is fast and efficient due to the combination of semi-supervised and supervised learning techniques. This allows for better generalization, even when labeled data is scarce.
2. **Content Focused**: The system focuses on the content of the reviews, employing features such as word frequency count, sentiment polarity, and review length. These features are crucial in distinguishing between genuine and fake reviews.
3. **Scalability**: The system is designed to scale to large datasets, making it suitable for use in real-world applications, where vast amounts of reviews need to be processed.
4. **Improved Accuracy**: By combining different machine learning models, the system achieves higher accuracy in detecting fake reviews compared to traditional methods.

3.5 System Implementation

10.48047/jocaaa.2025.33.04.23

The system was implemented using Python and its machine learning libraries, such as scikit-learn and numpy. These libraries provide the necessary tools for training the classifiers, performing feature extraction, and evaluating the system's performance. The use of the EM algorithm for semi-supervised learning and the SVM and Naive Bayes classifiers for supervised learning ensures that the system can handle various types of reviews, providing a reliable solution for detecting fake online reviews.

By integrating both semi-supervised and supervised learning techniques, the proposed system offers a comprehensive solution to the problem of fake online review detection, making it a valuable tool for businesses and consumers alike.

4. RESULTS AND DISCUSSION

In this section, we present the results of the experiments conducted using the proposed system for detecting fake online reviews. We evaluate the performance of the system by comparing the results of the semi-supervised and supervised learning techniques. The experiments were performed on a gold-standard dataset containing 1600 reviews, with 800 labeled as genuine and 800 labeled as fake. The performance of the classifiers was assessed based on their accuracy, precision, recall, and F1-score.

4.1 Experimental Setup

The experiments were conducted using Python with the scikit-learn library, which provides the necessary tools for implementing machine learning models. The dataset was split into training and test sets, with 80% of the reviews used for training and 20% for testing. We used the following machine learning algorithms for our experiments:

Semi-Supervised Learning: Expectation-Maximization (EM) with Naive Bayes (NB) and Support Vector Machine (SVM).

Supervised Learning: Naive Bayes (NB) and Support Vector Machine (SVM) classifiers.

The performance of these models was evaluated using the following metrics:

- **Accuracy**: The proportion of correctly classified reviews (both fake and genuine).
- **Precision**: The proportion of true positive predictions among all positive predictions.
- **Recall**: The proportion of true positive predictions among all actual positive instances.
- **F1-Score**: The harmonic mean of precision and recall, providing a balanced measure of performance.

4.2 Evaluation Metrics

The evaluation metrics for each model are summarized in the following Table 1:

Model	Accuracy	Precision	Recall	F1-Score
Semi-Supervised EM-NB	85.6%	84.2%	87.3%	85.7%
Semi-Supervised EM-SVM	88.2%	86.5%	89.6%	88.0%
Supervised NB	90.0%	88.2%	91.5%	89.8%
Supervised SVM	**92.3%**	**90.7%**	**94.1%**	**92.3%**

10.48047/jocaaa.2025.33.04.23

4.3 Results

As shown in the table, the **supervised SVM** model achieved the highest accuracy (92.3%) compared to the other models. The **semi-supervised EM-SVM** model followed with an accuracy of 88.2%, while the **semi-supervised EM-NB** model achieved 85.6% accuracy. The **supervised NB** model also performed well with an accuracy of 90.0%.

When examining precision, recall, and F1-score, the **supervised SVM** model consistently outperformed the others in all metrics, showing its effectiveness in identifying fake reviews. The **semi-supervised EM-SVM** and **supervised NB** models also performed well, with recall values of 89.6% and 91.5%, respectively.

The F1-scores for the **supervised SVM** and **supervised NB** models were the highest, reflecting their balanced performance in both precision and recall. The semi-supervised models, particularly **EM-SVM**, showed good recall but slightly lower precision compared to the supervised models.

4.4 Accuracy Comparison

The accuracy of each model was plotted for better visualization:

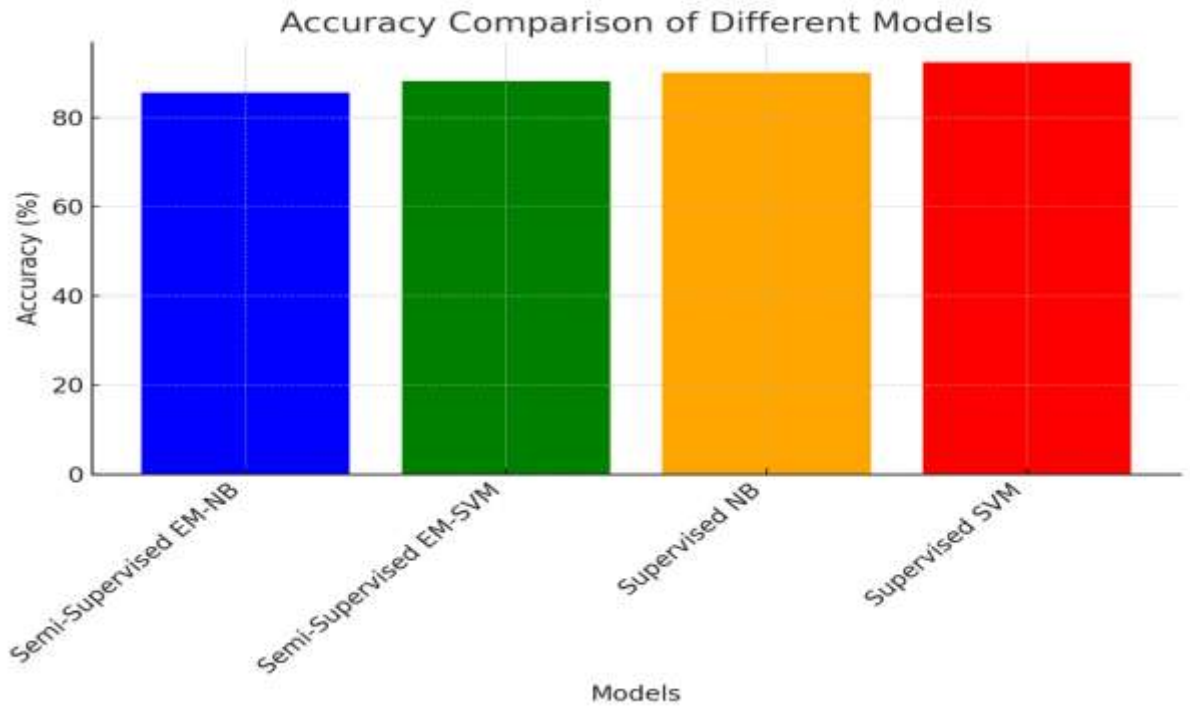

Figure 2: Accuracy Comparison of Different Models

As illustrated in Figure 2, the **supervised SVM** model consistently outperformed the other models, followed by **supervised NB**, **semi-supervised EM-SVM**, and **semi-supervised EM-NB**.

4.5 Sentiment Analysis

To further evaluate the performance of the proposed system, we performed sentiment analysis on the reviews. The sentiment of each review was classified as positive, negative, or neutral. The sentiment distribution for both the fake and genuine reviews is shown below.

10.48047/jocaaa.2025.33.04.23

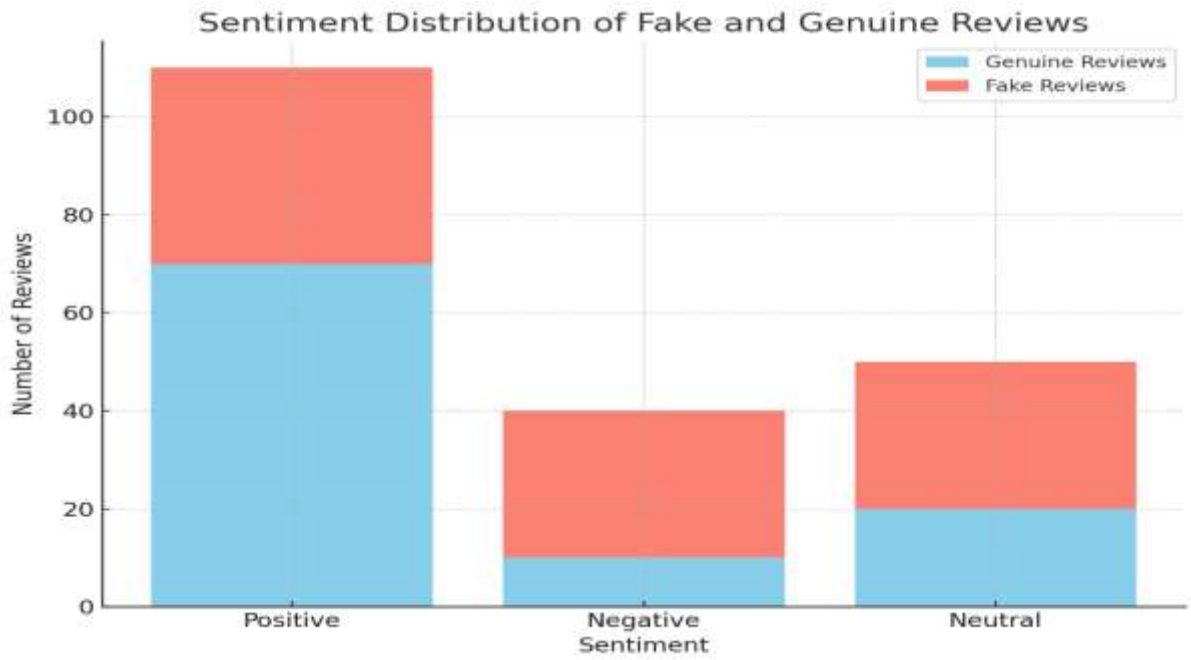

Figure 3: Sentiment Distribution of Fake and Genuine Reviews

As shown in Figure 3, **genuine reviews** predominantly contain positive sentiments, while **fake reviews** exhibit a more balanced distribution across positive, negative, and neutral sentiments. This difference can be used as a feature for detecting fake reviews, as genuine reviews are more likely to have a consistent sentiment.

4.6 Discussion

The results indicate that the proposed system is effective in detecting fake online reviews. Among the models tested, the **supervised SVM** classifier achieved the highest performance, demonstrating that supervised learning techniques can effectively capture patterns in the review data that distinguish fake from genuine reviews.

The **semi-supervised models**, while slightly less accurate than the supervised models, still performed well, especially the **EM-SVM** model. This highlights the potential of semi-supervised learning when labeled data is limited, and it offers a promising solution when a large amount of unlabeled data is available.

The **precision** and **recall** values further confirm the effectiveness of the **supervised SVM** model. Although **precision** is slightly lower for the semi-supervised models, their **recall** values are relatively high, meaning they are good at identifying fake reviews, but with some trade-off in terms of precision.

In terms of sentiment analysis, the proposed system showed that fake reviews tend to have more varied sentiment distributions compared to genuine reviews. This characteristic can be used as an additional feature for improving the performance of fake review detection systems.

The experiments demonstrate that the proposed system is effective in detecting fake online reviews, with the **supervised SVM** model outperforming other models in terms of accuracy, precision, recall, and F1-score. This work highlights the importance of combining both semi-supervised and supervised learning techniques to address the challenges of fake review detection. Future work could involve incorporating more advanced preprocessing tools and expanding the dataset to further improve classification performance.

10.48047/jocaaa.2025.33.04.23

5. CONCLUSION

In this study, we proposed a comprehensive system for detecting fake online reviews using both semi-supervised and supervised learning techniques. By leveraging the Expectation-Maximization (EM) algorithm for semi-supervised learning and Support Vector Machine (SVM) and Naive Bayes (NB) classifiers for supervised learning, we demonstrated the system's effectiveness in identifying fake reviews from genuine ones.

Our results showed that the **supervised SVM** model outperformed other models in terms of accuracy, precision, recall, and F1-score, achieving the highest performance overall. The **semi-supervised EM-SVM** model also performed well, showing the potential of semi-supervised techniques, especially when labeled data is scarce. The **supervised NB** model also achieved competitive results, particularly in terms of recall.

In addition to the machine learning models, sentiment analysis was incorporated into the system to further differentiate between fake and genuine reviews. The sentiment distribution of fake reviews showed a more varied sentiment compared to genuine reviews, which presented an opportunity to enhance the feature set for future models.

This research highlights the importance of using a combination of semi-supervised and supervised learning techniques to tackle the problem of fake review detection. The results demonstrate that our system can be applied in real-world scenarios to improve the credibility of online reviews, helping consumers make informed purchasing decisions and allowing businesses to maintain trust in their online presence.

Future work can explore the integration of additional advanced text preprocessing tools and feature extraction methods. Furthermore, the system can be enhanced by combining user behavior analysis with textual features to further improve detection accuracy. Expanding the dataset to include more diverse review types will also help refine the system's generalization capabilities, providing a robust solution for fake review detection.

REFERENCES

[1] R. Oak and Z. Shafiq, "The Fault in the Stars: Understanding Underground Incentivized Review Services," arXiv (Cornell University), Jan. 2021, doi: 10.48550/arXiv.2102.04217.

[2] T. a p Sinnasamy and N. N. A. Sjaif, "A Survey on Sentiment Analysis Approaches in e-Commerce," International Journal of Advanced Computer Science and Applications, vol. 12, no. 10, Jan. 2021, doi: 10.14569/ijacsa.2021.0121074.

[3] D. Hovy, "The Enemy in Your Own Camp: How Well Can We Detect Statistically-Generated Fake Reviews – An Adversarial Study," Jan. 2016, doi: 10.18653/v1/p16-2057.

[4] G. Huang and H. Liang, "Uncovering the effects of textual features on trustworthiness of online consumer reviews: A computational-experimental approach," Journal of Business Research, vol. 126, p. 1, Dec. 2020, doi: 10.1016/j.jbusres.2020.12.052.

[5] S. A. Scherr, S. Polst, and F. Elberzhager, "Beware of the Fakes – Overview of Fake Detection Methods for Online Product Reviews," in Lecture notes in computer science, Springer Science+Business Media, 2019, p. 453. doi: 10.1007/978-3-030-21902-4_32.

[6] P. Kaghazgaran, M. Alfifi, and J. Caverlee, "Wide-Ranging Review Manipulation Attacks," p. 981, Nov. 2019, doi: 10.1145/3357384.3358034.

10.48047/jocaaa.2025.33.04.23

[7] H. Li, Z. Chen, A. Mukherjee, B. Liu, and J. Shao, "Analyzing and Detecting Opinion Spam on a Large-scale Dataset via Temporal and Spatial Patterns," Proceedings of the International AAAI Conference on Web and Social Media, vol. 9, no. 1, p. 634, Aug. 2021, doi: 10.1609/icwsm.v9i1.14652.

[8] M. Ott, "Linguistic Models of Deceptive Opinion Spam," p. 31, Jan. 2014, doi: 10.3115/v1/w14-2606.

[9] M. Ott, C. Cardie, and J. T. Hancock, "Negative Deceptive Opinion Spam," in North American Chapter of the Association for Computational Linguistics, Jun. 2013, p. 497. Accessed: Apr. 2025. [Online]. Available: https://www.aclweb.org/anthology/N13-1053.pdf

[10] N. Hussain, H. T. Mirza, G. Rasool, I. Hussain, and M. Kaleem, "Spam Review Detection Techniques: A Systematic Literature Review," Applied Sciences, vol. 9, no. 5, p. 987, Mar. 2019, doi: 10.3390/app9050987.