*The Knowledge Engineering Review* (2024), **39**, e8, pp. 1–47
doi:10.1017/S0269888924000067

The Knowledge Engineering Review

CAMBRIDGE UNIVERSITY PRESS

REVIEW

# Recent state-of-the-art of fake review detection: a comprehensive review

Richa Gupta[1,2], Vinita Jindal[1], and Indu Kashyap[2]

[1]Keshav Mahavidyalaya, University of Delhi, New Delhi, India
[2]Faculty of Engineering & Technology, Manav Rachna International Institute of Research and Studies, Faridabad, India
**Corresponding author:** Richa Gupta; Email: richa.gupta@keshav.du.ac.in

**Received:** 9 November 2023; **Revised:** 1 August 2024; **Accepted:** 13 August 2024

**Abstract**

Online reviews have a significant impact on the purchasing decisions of potential consumers. Positive reviews often sway buyers, even when faced with higher prices. This phenomenon has given rise to a deceptive industry dedicated to crafting counterfeit reviews. Companies frequently indulge in procuring bulk fake reviews, employing them to tarnish their rivals' reputations or artificially bolster their credibility. These spurious reviews materialize through automated systems or compensated individuals. Thus, detecting fake reviews is becoming increasingly important due to their deceptive nature, as they are extremely difficult for humans to identify. To address this issue, current work has focused on machine learning and deep learning techniques to identify fake reviews. However, they have several limitations, including a lack of sufficient training data, inconsistency in providing accurate solutions across different datasets, concept drift, and inability to address new methods that evolved to create fake reviews over time. The objective of this review paper is to find the gaps in the existing research in the field of fake review detection and provide future directions. This paper provides the latest, comprehensive overview and analysis of research efforts focusing on various techniques employed so far, distinguishing characteristics utilized, and the existing datasets used.

## 1. Introduction

Online reviews have become a dominant force in the realm of e-commerce websites. Whether it's purchasing products on platforms like Amazon and Myntra, exploring hotel options on TripAdvisor, checking out restaurant reviews on Yelp, or assessing services on Google, consumer/user reviews play a pivotal role in shaping the e-commerce landscape. Potential customers get attracted to these services and products, after checking the positive reviews of other people about that particular service or product. Hence, these reviews help in the consumer's decision-making process. However, the presence of fake reviews creates a misleading environment that not only deceives but also misguides potential customers who rely on these reviews. The dissemination of false information, whether it be through news articles, reviews, or social media posts, has a disruptive effect on society and leads to social issues. For instance, the spread of inaccurate COVID-19-related news regarding vaccinations and lockdowns on social media resulted in widespread negative behaviour (Marco-Franco *et al*., 2021).

Fake reviews manifest as either disinformation, where fraudulent reviews are intentionally generated to harm others, or misinformation, where false information is spread without malicious intent. Regardless of the intentions behind them, fake reviews pose a significant threat to consumer behaviour and fair business practices. The internet has facilitated the dissemination of false narratives and biased opinions that serve self-interests at the expense of others (Amos, 2022). The inability to identify fake reviews poses significant disadvantages for all parties involved, including consumers, business service

**Cite this article:** Richa Gupta, Vinita Jindal and Indu Kashyap. Recent state-of-the-art of fake review detection: a comprehensive review. *The Knowledge Engineering Review* 39(e8): 1–47. https://doi.org/10.1017/S0269888924000067

providers, and e-commerce platforms. Consumers must be protected against the influence of these deceptive reviews. Businesses and service providers need defence against artificially generated competition for the preservation of their reputation, as reviews directly impact product acceptability (He *et al*., 2022). E-commerce and online travel platforms have a responsibility to ensure fair review channels. Although industry giants like Google and Yelp have implemented fake review detection algorithms, these algorithms are not publicly accessible. Amazon, on the other hand, removes fake reviews over time, but there is typically a delay of around 100 days. However, by the time these reviews are deleted, a considerable number of consumers may have already been exposed to them, causing potential damage.

Genuine reviews provide valuable feedback to service providers and sellers, enabling them to enhance their products or services. Unfortunately, numerous brand marketers and service providers engage in unethical practices by purchasing fake positive reviews from both known and unknown reviewers. They resort to such tactics to give a boost to their new products or artificially enhance the popularity of their brand through inflated positive reviews. In some cases, they may even employ fake negative reviews to maliciously damage the reputation of their competitors. An alternative perspective could argue that choosing not to purchase fake reviews while competitors do, might result in the potential loss of customers. Hence, even ethical players in the industry may be tempted to resort to purchasing fake reviews as a desperate measure to stay competitive in a market. Through their research, He, Hollenbeck, and Prosperpio (He *et al*., 2022) have established a direct correlation between product reviews, average ratings, prices, and sales rank on Amazon.com, providing concrete evidence of the significant impact that buying fake reviews has on business outcomes. The widespread implementation of fake review detection (FRD) algorithms can bring about several benefits to society, encompassing three key aspects. Firstly, consumers will have access to authentic feedback, enabling them to make informed decisions and develop trust in the products or services they are considering. Secondly, businesses and service providers will be incentivized to improve their offerings to meet consumer expectations. Additionally, they will be deterred from engaging in the purchase of fake reviews, as it could hamper their reputation. Lastly, e-commerce websites that effectively tackle fake reviews will be regarded as trustworthy and fair, leading to increased user engagement and utilization of their platforms.

### *1.1. Contributions of this paper*

Numerous studies have delved into the complex issue of detecting fake reviews, and this paper offers a survey and summary of these research endeavours. By exploring the application of machine learning, deep learning, and swarm intelligence techniques, researchers have aimed to identify the most effective methods for further investigation in fake review detection. However, the existing literature surveys did not encompass recent advancements in fake review detection utilizing new methodologies such as swarm intelligence and transformers, nor did they delve into the motivations behind the exponential growth of fake reviews. Furthermore, earlier surveys failed to provide a thorough examination of identifying features and used datasets. As a result, this paper incorporates relevant publications from esteemed journals and conferences between 2019 and 2023 to shed light on the challenges within this field. The contributions of this paper are as follows:

- This review paper provides an up-to-date analysis that encompasses the various state-of-the-art techniques employed in the realm of fake review detection.
- This paper conducts a comparative analysis of different methods used for fake review detection, evaluating their efficiency, and summarizing their respective strengths and weaknesses. By systematically comparing these methods, this paper aims to provide insights into their performance and help researchers and practitioners make informed decisions regarding the most effective approaches in this domain.
- This review also presents the various features extracted from the existing works and categorizes them as review-centric or reviewer-centric.

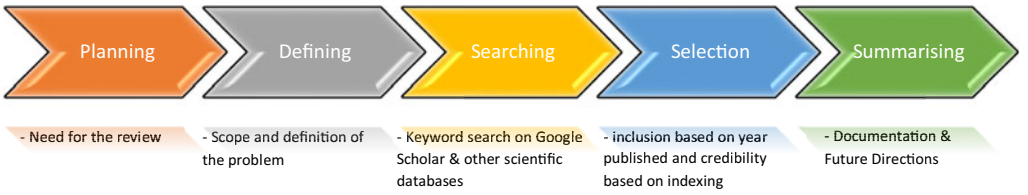


***Figure 1.*** *Research methodology steps.*

- This review paper summarizes the various datasets used in the existing research and discusses their limitations.
- Challenges and shortcomings in the current research are discovered and future research directions are introduced.

### *1.2. Methodology*

The literature survey in this paper follows a stepwise methodology, as shown in Figure 1. These steps include planning, defining, searching keywords, selecting, and summarizing the results. The detailed explanation is given below:

- Planning: The planning phase involved outlining the objective and structure of the review paper. Inclusion and exclusion criteria were devised for selecting relevant studies.
- Defining: In this phase, the methodology for the review was clearly articulated. Criteria for selecting the research papers to be included in this study were defined. A timeline with milestones was also created to guide the progress of the review.
- Searching: For this research, various search terms such as 'fake review detection', 'deceptive review detection', and 'fraudulent reviews' were employed to gather material for this systematic literature review. The search was performed on Google Scholar, ResearchGate, and ScienceDirect. The search scope was from the year 2019 to 2023 and the primary sources consulted were scholarly journals and conference proceedings.
- Selection: A total of 98 papers were selected considering their relevance, novelty, and credibility based on indexing in academic databases. Of these 98, 71 reputed journal articles were identified with fake review text, 9 were old review papers, and 13 articles were from international Scopus-indexed conferences. The rest were articles that provided valuable insights for this review paper.
- Summarizing: After a thorough reading of these papers, the strengths and limitations of various techniques and features employed in FRD, were documented and various challenges were identified and listed in the paper.

### *1.3. Comparison with previous literature surveys*

Many literature surveys have been published in the past to comprehensively analyse the methodologies for fake review detection (FRD). But they suffer from one or more limitations. Table 1 concisely summarizes their strength and limitations.

Review papers published between 2019 and 2023, specifically referenced as Ren and Ji (2019), Rodrigues *et al*. (2020), Tang and Cao (2020), Wu *et al*. (2020), Mohawesh *et al*. (2021) (in Table 1) have incorporated various machine learning and deep learning techniques. However, they are old according to the current context and hence these papers do not encompass novel techniques such as swarms or transformers. Vidanagama *et al*. (2020), Mewada and Dewang (2022) have given limited emphasis on identifying all the features that are used for labelling a review as fake or real. Kaddoura *et al*. (2022) has

***Table 1.*** *Summary of existing literature surveys in fake review detection*

| Reference | Strength | Limitation |
|---|---|---|
| Ren and Ji (2019) | In-depth analysis of features, discussion of various datasets, and machine learning (ML) methods used for FRD | Novel features and techniques have emerged since then, which need to be incorporated into a review |
| Rodrigues *et al*. (2020) | Reviewed various machine and deep learning techniques | Summarized limited existing techniques and no future direction given |
| Tang and Cao (2020) | Gives an overview of various FRD techniques | Not Comprehensive |
| Vidanagama *et al*. (2020) | Various approaches including ML, network-based, and pattern-mining discussed | No mention of identifying feature extraction<br>Summarized limited existing datasets and have not provided future directions |
| Yuanyuan *et al*. (2020) | Have provided future directions in detail and summarized the existing datasets | Summarized limited existing techniques |
| Mohawesh *et al*. (2021) | Feature engineering techniques and datasets are summarized in detail. Given main challenges | Novel features and methodologies/ techniques have emerged since then, which have not been mentioned in these surveys |
| Mewada and Dewang (2022) | Classifies and summarizes the key techniques and features including future directions | Focused on textual features. Reviewer-centric and reviewer-group features have not been considered |
| Kaddoura *et al*. (2022) | Summarized ML and Deep Learning (DL) techniques along with datasets and feature extraction | This survey is not specifically focused on fake reviews but on spam text in social media including fake accounts, fake news, reviews, and rumours |
| Maurya *et al*. (2023) | In-depth review of machine and deep learning techniques till the year 2021 | They have not included any paper after the year 2021 and hence, recent advancements in deep learning have not been incorporated |

encompassed a broader view of spam text with fake reviews being only a minor component. Hence it is not comprehensive. All of them have analysed the existing datasets. Maurya *et al*. (2023) is relatively new and has provided an in-depth analysis of nearly all the techniques including the BERT transformer. However, it still overlooks several recent advancements in the field. These advancements include techniques such as generative pre-trained transformers GPT-3 (Gambetti & Han, 2023; Shukla *et al*., 2023), GPT-4 (Shukla *et al*., 2023), opinion mining (Chopra & Dixit, 2023), graph neural networks to find associations between reviews, users, and products (Ren *et al*., 2022), fitness-based grey wolf optimization (Shringi *et al*., 2022), artificial bee colony-based techniques (Jacob & Rajendran, 2022; Saini *et al*., 2021) and ensemble based on probability (Wu *et al*., 2022). In addition, features such as discourse analysis (Alawadh *et al*., 2023a, b), and the degree of suspicion (Wang *et al*., 2022) were introduced after the publication of the papers listed in Table 1. Furthermore, Shukla *et al*. (2023) have created a novel labelled physician dataset for FRD. Consequently, our research work has more value when compared to its counterparts in presenting an extensive survey encompassing all current techniques for identifying fraudulent reviews.

The rest of the paper has been organized as follows—Section 2 is the overview of the problem that this paper addresses; including what are fake reviews, what is the need for their detection and how can this problem be solved. Section 3 is the literature review based on the various techniques, identifying features, and the datasets used by the current studies along with their shortcomings. Section 4 concludes this paper followed by the limitations of the existing work and identifying the future challenges in section 5.

## 2. Overview of the problem

This section focuses on what fake reviews are, the importance of fake review detection, and showcases how machine learning and deep learning techniques are employed to identify fraudulent reviews. The following subsections discuss the above-mentioned issues in order.

### *2.1. What are fake reviews?*

Fake reviews can manifest in both positive and negative forms, with distinct motivations driving their creation. Numerous factors prompt individuals as well as businesses to participate in fake review creation.

Some of the motives for posting fake reviews are:

- Financial incentives: Positive fake reviews may be driven by monetary rewards or incentives (Zaman *et al*., 2023). Reviewers may receive direct payments, gift cards, free products, discounts, or other forms of compensation. Businesses may also employ individuals to post-fake positive reviews for their brand to boost their sales.
- Endorsement help: Positive fake reviews may also be posted for personal relationships, such as helping a friend or boosting self-esteem, without buying or using a product (Zaman *et al*., 2023; Thakur *et al*., 2018).
- Reputation management: Certain businesses may opt to create an abundance of positive reviews to counteract negative feedback and effectively oversee their online reputation (Barbado *et al*., 2019).
- Monetary compensation: On the other hand, posting negative fake reviews may be motivated by monetary compensation (Zaman *et al*., 2023) or for getting the product free. It might be possible that the brand may have the precedence of offering customers some discount for taking down the negative review.
- Means of seeking revenge: Upset customers or individuals who were once associated with a business may resort to posting negative fake reviews as a method of seeking vengeance (Thakur *et al*., 2018).

***Table 2.*** *Sample fake reviews from an annotated fake review dataset (weblink: https://osf.io/tyue9/) created by Salminen et al. (2022). It contains 20K fake and 20K real product reviews. 'OR' stands for original reviews and 'CG' stands for computer-generated reviews*

| Category | Rating | Label | Review-text |
|---|---|---|---|
| Home and Kitchen | 5 | OR | Excellent product and a much better quality than the one you get at Walmart for $50 |
| Home and Kitchen | 5 | CG | These are just perfect, exactly what I was looking for |
| Kindle_Store | 5 | CG | What a wonderful way to get familiar with this Author!! There were a few typo/errors in the books, but I loved the characters so much!! |
| Kindle_Store | 5 | OR | I was captivated with this book. The characters were well developed and kept my interest. I recommend this book |

- Competitor sabotage: Competing businesses or individuals might write adverse reviews to tarnish the reputation of their rivals and secure a competitive edge (Barbado *et al*., 2019).
- Cold start of products or services: Fake reviews are created for a new product, service, or business, where there is little or no genuine user feedback available. The goal is to give the appearance of popularity and positive reception (Tang *et al*., 2020).
- Competition: Even those with ethical standards in the industry might be enticed to buy fabricated reviews as a last-ditch effort to maintain competitiveness in the market.

The influx of negative reviews targeting a brand can result in unfair competition and adversely impact its ranking on prominent online platforms like Google, Amazon, and TripAdvisor (Salminen *et al*., 2022). However, the impact of fake reviews varies. Negative fake reviews targeting high-quality products can be particularly detrimental to businesses. Similarly, positive fake reviews associated with low-quality products are harmful to consumers. Moreover, competitors offering average or good quality products may suffer from the impact of fake positive reviews on poor quality products, especially if they lack a substantial number of reviews themselves.

In addition to fake positive and fake negative reviews, there are non-reviews known as disruptive spam reviews that offer no relevant opinions or insights regarding the product at hand (Mohawesh *et al*., 2021). Jindal and Liu (2007) were the first ones to identify fake reviews on Amazon and use a supervised learning technique. They categorized opinions into 3 types—type I as deceptive reviews, type II as brand-specific reviews, and type III as disruptive reviews or non-reviews which can be identified by humans easily. These non-reviews have no opinion. For example—'Can anyone confirm this?', 'The other review is too funny', 'Go Eagles Go.' etc. Hence, FRD systems need not bother with them, as they don't misguide the consumer.

Yet another classification was done by Salminen *et al*. (2022). According to them, reviews can be categorized into two types: original reviews (OR) and computer-generated (CG) reviews. This nomenclature is a part of the dataset created by Salminen and is available at https://osf.io/tyue9/. While original reviews have been the primary focus of discussion, computer-generated reviews created automatically by machines are consistently treated as fake. Table 2 shows some examples of OR and CG reviews on a product review dataset.

### *2.2. Why do we need fake review detection?*

The problem of fake review detection involves classifying reviews as either genuine or fake. This problem can be addressed using natural language processing (NLP) techniques, employing machine learning (ML) algorithms or graph networks to ascertain the authenticity or falsehood of a given review. Numerous supervised, semi-supervised, and unsupervised algorithms have been utilized to assess the

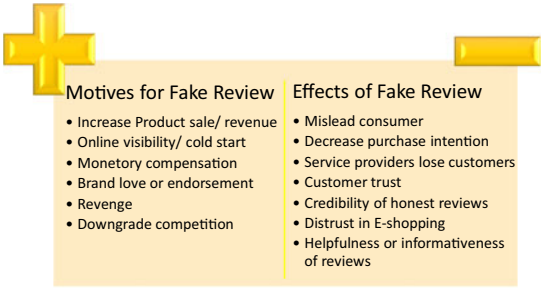


***Figure 2.*** *Fake review—motives and effects.*

effectiveness of different models using datasets from platforms like Yelp, TripAdvisor, Amazon, and YouTube. However, fraudsters continuously adapt by incorporating new features into their reviews to evade detection by existing approaches (Wu *et al*., 2020; Wang & Wu, 2020; Mattson *et al.*, 2021).

Online reviews have become a vital resource for consumers, with a significant level of trust placed in them. According to a survey by Brightlocal in 2019, 76% of consumers trusted online reviews as much as personal recommendations from friends (Kumar & Saroj, 2020). This percentage fell to a mere 46% in 2022 (Paget, 2023). This shows how buyer confidence in the review system has eroded over the years. The reason is the exponential growth of fake reviews on online platforms.

Fake reviews hamper society in three ways. First, they mislead the consumer by providing inaccurate information about products or services. They undermine the credibility of genuine reviews, and erode their trustworthiness, meaning if fake reviews are in a large percentage, they would undermine the trustworthiness of genuine reviews by mitigating the impact of the genuine reviews on the decision-making process of the buyer. Secondly, they create unfair competition for businesses and decrease business reputation. Businesses that engage in writing or purchasing fake reviews gain an unfair advantage over competitors who rely on genuine feedback. They can tarnish a business's reputation by providing false information about its products, services, or customer experiences. This can lead potential customers to form inaccurate perceptions, affecting the company's image. Third, the presence of fake reviews also diminishes trust in e-commerce platforms that struggle to effectively detect and address false reviews. Users may be disappointed if they discover that the reviews on their favourite e-commerce platform are not genuine. The platform's integrity and credibility suffer leading to a decline in user engagement and loyalty toward that platform.

Despite occasional warnings from government bodies and websites, it remains challenging for ordinary consumers to accurately identify fake reviews. Real-life examples from datasets, such as in Table 2, demonstrate this difficulty. In essence, humans struggle to find the authenticity of a review solely through reading it, necessitating the availability of additional features that can aid in making informed decisions (Filho *et al*., 2023). While studies, such as the analysis of Yelp reviews conducted by Kostromitina *et al*. (2021), shed light on the reasons behind extreme star ratings and customer preferences, they often overlook the presence of fake reviews. The emergence of computer-generated fake reviews further exemplifies the increasing sophistication of technology, making it increasingly difficult for humans to distinguish them from genuine reviews (Floridi & Chiriatti, 2020). Companies like Amazon have witnessed sudden increases in unverified reviewers with 5-star ratings, which serves as an indicator of fake reviews (Abdulqader *et al*., 2022). Consequently, online platforms such as Amazon, Yelp, and Google must continually update their strategies for detecting and combating fake reviews (Salminen *et al*., 2022). The abundance of fake reviews significantly undermines the credibility of brands or products in the eyes of consumers (Ismagilova *et al*., 2020). Figure 2 depicts the major contributors and effects of fake reviews.

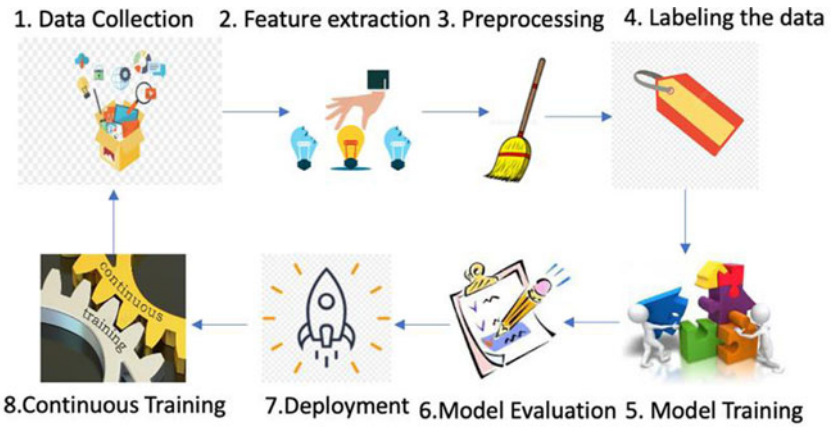


***Figure 3.*** *The process of Fake Review Detection (FRD).*

### *2.3. How is a fake review detected?*

Detecting fake reviews using Artificial Intelligence (AI) involves training a model to analyse various features and patterns within reviews to distinguish between genuine and fake ones. Figure 3 depicts the steps of the fake review detection process which are enumerated below:

1. Data Collection: A dataset is compiled containing a large number of reviews, including both genuine and fake ones. These reviews may be sourced from various platforms, such as e-commerce websites, social media, or review aggregation sites.

2. Feature Extraction: Relevant features are extracted from the reviews, which may include textual information like the review text, user profile details, timestamps, ratings, and other metadata associated with the review. Additional features can be derived, such as sentiment analysis, length of the review, or language patterns.

3. Data Preprocessing: The collected data is preprocessed to ensure consistency and improve the quality of input. This step involves removing noise, normalizing text (e.g. lowercasing, removing punctuation), handling missing data, and transforming features into a suitable format for AI algorithms.

4. Labelling the Data: Each review in the dataset needs to be labelled as either genuine or fake. This can be done manually by human reviewers who are familiar with fake review patterns or by using existing labelling techniques such as rule-based, algorithm-based filtering, and crowdsourced labelling (Mewada & Dewang, 2022). It's crucial to have a balanced dataset with representative samples of both genuine and fake reviews.

5. Model Training: A model is trained on the labelled dataset using various algorithms like logistic regression, decision trees, random forests, or more advanced techniques like support vector machines (SVMs) or deep learning models such as recurrent neural networks (RNNs) or convolutional neural networks (CNNs). The model learns to recognize patterns and relationships between the features and the review's authenticity.

6. Model Evaluation: The trained model is evaluated using different evaluation metrics such as accuracy, precision, recall, and F1-score metrics. This step helps assess the model's performance and determine if further adjustments or improvements are needed.

7. Model Deployment: Once the model demonstrates satisfactory performance, it can be deployed to detect fake reviews in real time. New reviews can be fed into the model, and it will predict authenticity based on the learned patterns from the training phase.

8. Continuous Learning: Fake review patterns can evolve, so it's important to continuously monitor and update the model to adapt to emerging trends. Feedback from users and human reviewers can be incorporated to refine the model and improve its detection capabilities.

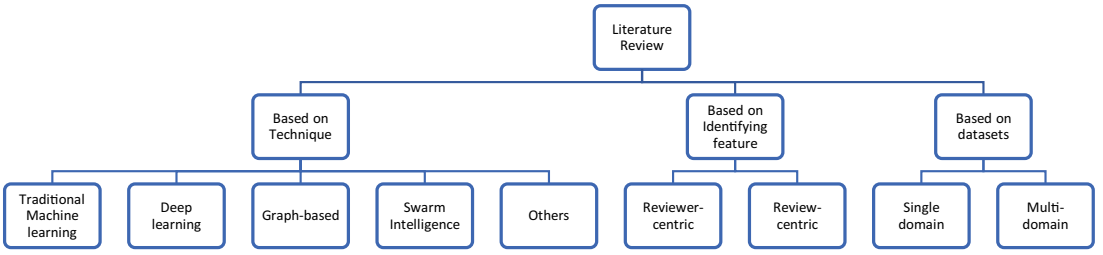


***Figure 4.*** *Structure of this literature review.*

## 3. Literature review

Many researchers have tried to solve this problem of fake review detection using various artificial intelligence techniques. This section presents a comprehensive literature review of their work. We conducted three distinct types of reviews using the same set of research papers. The first one focused on the techniques utilized, the second examined the features employed, and the third analysed the datasets utilized. These three categorizations are illustrated in Figure 4 and discussed in Sections 3.1, 3.2, and 3.3 respectively. Section 3.1 classifies the studies based on the algorithms or techniques utilized, including machine learning, deep learning, transformers, and swarm intelligence. Section 3.2 analyses the linguistic and behavioural features used to distinguish between fake and genuine reviews. Further, Section 3.3 provides an overview of the datasets used or created in previous research endeavours. Figure 4 presents a diagrammatic representation of the three distinct types of reviews performed in this paper.

### *3.1. Review based on techniques*

Several algorithms such as rule-based (using a set of rules to classify), graph-based (data representation as nodes and edges), machine learning (ML—learning from data and improving the performance) and deep learning (DL) have been employed to detect fake reviews. This section highlights the most recent algorithms used in this field of research. These algorithms are continually evolving as researchers strive to improve the accuracy and effectiveness of fake review detection techniques.

#### *3.1.1. Machine learning techniques*

Machine learning algorithms are computational models that learn patterns and relationships from data without being explicitly programmed. They use statistical techniques to automatically detect patterns, make predictions, or make decisions based on the input data. Some commonly used machine learning algorithms include Decision Trees, Random Forest (RF), Support Vector Machines (SVM), Naïve Bayes (NB), K-Nearest Neighbours (KNN), Logistic Regression (LR), etc. Researchers frequently employ these supervised, semi-supervised, and unsupervised machine learning techniques for fake review detection.

**Supervised ML Algorithms**

Supervised ML algorithms like SVM, KNN, and LR are among the most commonly used methods in the field. These ML algorithms have been used by Tufail *et al*. (2022) to detect fake reviews on the Yelp hotel review dataset. Their model proved to be a robust one, but it only focused on supervised models. Similarly, Kumaran *et al*. (2021) have used naïve Bayes, Eristic regression, and SVM for detecting fake reviews in a dictionary based on social media keywords and online reviews. Their model uses a very limited feature set of uni-, bigrams, and length of review. Poonguzhali *et al*. (2022) have implemented SVM for fake review detection by creating an online e-commerce interface. Fake reviews are predicted in this system and are not used in the database while recommending a product on this interface. But their method can't upload more than one review by one user. So, their prevention method is not feasible in large databases.

Along similar lines, Hussain *et al*. (2020) have given some important behavioural and linguistic pointers to identify fake reviews. For example, the absence of profile data of the reviewer, posting duplicate reviews, short and often grammatically erroneous reviews, groups of reviews with the same timestamp, and excessive use of positive or negative words. However, they have studied the behavioural model and linguistic model separately. The chances of reviews being classified accurately decrease when only one model is used. Hence, Alsubari *et al*. (2022) have extracted features such as sentiment score, four grams, number of verbs, nouns, and strong positive or negative words to identify fake reviews. Then they applied four different supervised classifiers—SVM, Naïve Bayes, Random Forest, and Adaptive Boost and compared their accuracy for fraudulent reviews detection. The limitation of this study was fewer extracted features and, their dataset was limited to the hotel domain. According to the findings of Abdulqader *et al*. (2022), non-verbal features carry greater significance than verbal features, and their combination can enhance the accuracy of detection. However, they have only given a pure theory-based model, which may or may not apply to datasets other than the one used here. Alawadh *et al*. (2023a) experimented with a benchmark, balanced hotel review dataset and proved that real-time application of deep learning-based, semantically aware text features on web portals can effectively detect fraudulent reviews. However, they have only used a small dataset which doesn't fully utilize the neural network advantage. A novel ML framework based on M-SMOTE has been created by Kumar *et al*. (2022) to address the class imbalance problem. The study's results confirm that combining reviewer-centric features with review-centric features significantly improves the performance of FRD models. A major limitation of this study is that the features extracted from the datasets used in this study, may not be present in other domains. Theuerkauf and Peters (2023) employ labelled reviews sourced from the iOS App Store and combine them with two statistical approaches. The simultaneous utilization of multiple feature sets is demonstrated to enhance the detection of fraudulent reviews, but undetected false positives might affect the evaluation metrics.

**Semi-supervised ML Algorithms**

The supervised machine algorithms use extensive labelling, which is laborious as well as subjective. Hence, the semi-supervised approach uses a pre-defined set of features to train classifiers. In their research, Jing-Yu *et al*. (2022) employed a semi-supervised approach called AspamGAN (Generative Adversarial Network), which utilizes an attention mechanism in the classifier to detect fake reviews. They have used the data from the TripAdvisor dataset. But, if the data generated by their model is insufficient, it may result in poor accuracy. The earlier version of this method spamGAN given by Stanton and Irissappane (2020) was a powerful tool that used a simple classifier. But it has the disadvantage that it may lose important information such as context, the focus of the sentence, etc., and thus won't be able to detect fake information with high probability. AspamGAN has the advantage that it has better performance with limited label data, than SpamGAN.

The semi-supervised approach relies on a pre-defined set of features to train classifiers. Due to the laborious labelling task of the fake review datasets, researchers such as Ligthart *et al*. (2021) used four semi-supervised techniques in their study. This consisted of self-training (training on the labelled portion of the data), co-training (utilizing additional perspectives of data), Transductive SVM (variation of traditional SVM used in a semi-supervised setting), and label propagation plus spreading (again used for semi-supervised training). They train on one dataset and perform experiments on two more datasets from Yelp. Although their effort alleviates the labelling task, they have not considered metadata or reviewer-based features. In contrast, Wu *et al*. (2022) implement a semi-supervised probabilistic ensemble that collectively captures the individual behavioural characteristics of reviewers as well as the reviewer network. They use ten behavioural features such as the number of reviews in a day, burstiness, popularity of product that the user has reviewed, average distance between a user and other users, etc. They assume that the reviewer network presents homophily.

**Unsupervised ML Algorithms**

As mentioned earlier, detecting fake reviews accurately is highly improbable by humans. Hence, the availability of labelled data is less. Unsupervised algorithms have a good scope here. The work done by Mothukuri *et al*. (2022) creates clusters using the extracted features. They perform K-means clustering, GMM Full covariance clustering, and GMM Diagonal covariance unsupervised techniques to detect fake reviews taken from the café dataset of Yelp. They found that K-means shows the highest accuracy

among the three. Many other unsupervised algorithms could have been explored in the aforementioned work, and different domains.

There is a major problem of concept drift (Mohawesh *et al*., 2021; Tommasel & Godoy, 2019) while using machine learning techniques. Concept drift refers to the adaptation of fake reviewers over some time. They start writing in such a way that their writing skills are similar to real reviews and hence, can't be detected by the detection algorithms in place. For this, Wang *et al*. (2022) have incorporated the temporal patterns of reviews. They added a degree of suspicion for FRD by analysing 3D time series, the number of reviews, and the information entropy. Their framework seems comprehensive, but they have only used supervised algorithms. One major drawback of machine learning algorithms is that they require the extraction of features manually and a huge utilization of computational resources (Kaddoura *et al*., 2022).

Many supervised, semi-supervised and unsupervised algorithms have been used to detect fake reviews, to date. These traditional machine learning algorithms perform adequately on small datasets and are highly valued by researchers. In addition to this, they are simpler to implement and computationally cheap but, their performance deteriorates on larger datasets as compared to deep learning models. Also, their solution doesn't sustain over time and the problem of concept drift arises. In addition to the above shortcomings, they also require manual feature extraction and huge computational resources. Hence, deep learning algorithms were explored by researchers and their studies are discussed next. Table 3 is a summarized representation of the traditional machine learning algorithms used in FRD research work.

#### *3.1.2. Deep learning techniques*

Several deep learning algorithms are suitable for fraudulent review identification. These algorithms leverage the power of deep neural networks to automatically learn features and structural patterns from review data. Some commonly used deep learning algorithms for FRD are Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Transformer-based Models, Deep Belief Networks (DBN) and Siamese Neural Networks. The researchers Alsubari *et al*. (2021) have used four datasets including hotels, restaurants, Yelp, and Amazon to perform convolutional and max-pooling layers of the CNN to reduce dimensions and extract features. They have done cross-domain analysis, but the performance of this model decreases on a single-domain dataset. Jing-Yu and Ya-Jun (2022) utilized a semi-supervised method called AspamGAN (Generative Adversarial Network) that incorporates an attention mechanism to detect fake reviews in their assembled, partially labelled dataset. Stanton and Irissappane (2020) previously introduced the SpamGAN model, but AspamGAN addresses some of its limitations and performs better with limited labelled data.

The work done by Basyar *et al*. (2020) built a Long-short-term memory (LSTM) model as well as a Gated Recurrent Unit (GRU) model to detect e-mail spam. The former outperformed the latter, but the result was not significant, and they only used a training dataset due to time constraints. The authors Liu *et al*. (2022) give a layered attention network employing two stratums to capture semantic information. The authors then integrate a convolutional structure and Bi-LSTM to extract crucial semantics resulting in superior performance compared to other algorithms. Their limitation is the use of a supervised algorithm and a smaller number of extracted features. The work done by Crawford *et al*. (2021) uses inductive transfer learning to detect hotel review spam. However, they have limited their work to a single domain. As a different approach, Sadiq *et al*. (2021) train deep learning algorithms to predict the star ratings that will match the review. But they don't predict whether the review is fake or real. Salminen *et al*. (2022) fine-tuned the RoBERTa model, which is an optimized model of the BERT transformer, and called it fakeRoBERTa. They also create a fake review dataset with the help language generation model GPT-2. They experiment with fakeRoBERTa on their created dataset and conclude that AI algorithms can outperform humans in detecting fake reviews. However, datasets and techniques need to keep evolving to outmanoeuvre the concept drift problem.

Recently, language generation models such as BERT, and GPT-3 have also been used for the classification of fake reviews. Gambetti and Qiwei (Gambetti & Han, 2023) used the OpenAI GPT-3 model to generate a fake review dataset and then fine-tuned a pre-trained GPT-Neo model for FRD. They have

***Table 3.*** *Machine learning techniques used for FRD*

| Reference | Dataset used | Techniques used/Methodology | Features | Result | Strength/Limitation |
|---|---|---|---|---|---|
| Abdulqader *et al*. (2022) | YelpChi, YelpNYC, YelpZip | LR, NB, DT, RF<br>Provide evidence supporting the notion that non-verbal features carry greater significance than verbal features and that a combination of both can enhance the accuracy of detection | Deceptive constructs are gathered from theoretical theories of deception, encompassing both verbal and non-verbal elements. | Accuracy<br>LR = 86.75<br>NB = 76.72<br>DT = 79.84<br>RF = 86.08 | Theory-based deception model.<br>The proposed model may not apply to all domains |
| Alawadh *et al*. (2023) | Benchmark dataset of 1600 hotel reviews (Chicago) | NB, SVM radial basis, LR, DT, and RF combined to form deceptive review detection using semantic-aware deep features | Frequency-based features | Accuracy = 87% | The proposed technique uses a small dataset |
| Alsubari *et al*. (2020) | 30 476 reviews of electronic products in the USA collected from Yelp | DT, RF, AdaBoost | Review-centric | Accuracy<br>RF = 94%<br>DT = 96%<br>AdaBoost = 97% | Behaviour features not included in the technique |
| Alsubari *et al*. (2022) | TripAdvisor | NB, SVM, AdaBoost, RF | Review-centric such as sentiment score, four grams, no. of verbs, nouns, etc. | LR = 86%<br>NB = 88%<br>SVM = 93%<br>Adaboost = 94%<br>RF = 95% | Fewer number of extracted features in the model<br>Also, the dataset is limited to the hotel domain |

***Table 3.*** *Continued*

| Reference | Dataset used | Techniques used/Methodology | Features | Result | Strength/Limitation |
|---|---|---|---|---|---|
| Budhi *et al*. (2021) | YelpChi Hotel (5854), YelpChi restaurant (61 541), YelpNYC (359 052), YelpZip (608 598) | Random sampling techniques for class imbalance problem. 3 single—LR, SVM, Multilayer perceptron (ANN); 3 ensembles—bagging predictor, RF, Adaboost ensemble | Textual features | Accuracy LR = 86.78% SVM = 85.80% MLP = 84.13% (for YelpChi Hotel) | The proposed model shows that Under sampling (for solving imbalance) works well with ensembles but not on single classifiers And only textual features used |
| Chuttur and Bissonath (2022) | TripAdvisor | Adaboost compared with baseline as RF, DT, and SVM | Linguistic features | F1 score of SVM = 97% | Only review-centric features used |
| Elmogy *et al*. (2021) | Yelp (5853) | Compared KNN, NB, SVM LR, RF | Stylometric and behavioural | Accuracy SVM = 86.9% KNN = 86.23% NB = 86.08% LR = 86.89% RF = 86.82% | Doesn't work well on large datasets |
| Jain *et al*. (2021) | Yelp hotel and restaurant review | DT, SVM, KNN, LR, AdaBoost, NB | Max no. of reviews, review length, etc. | Accuracy of Hotel dataset SVM = 74.55% KNN = 82.51% NB = 80.71% LR = 88.11% DT = 84.57% AdaBoost = 85.68% | Only textual features have been incorporated which is a major drawback |

***Table 3.*** *Continued*

| Reference | Dataset used | Techniques used/Methodology | Features | Result | Strength/Limitation |
|---|---|---|---|---|---|
| Khan *et al*. (2021) | Spam reviews 5573 of which 747 were spam and 4825 were HAM | Support Vector Machine supervised machine learning | Review-centric | Accuracy = 98.92% | The limited-sized and imbalanced dataset used in the proposed technique |
| Kumar *et al*. (2022) | Yelp restaurant dataset (5044r restaurant reviews) Amazon dataset | M-SMOTE algo Modified SMOTE to solve class imbalance also. Then applied XGBoost, LSTM, GBM, Ann, RNN, SVM, LR, KNN, NB, and RF with and without preprocessing | Both review and reviewer-centric features | >80% AUC score | Features extracted from these datasets may not be present in other domains |
| Kumaran *et al*. (2021) | 1000 reviews collected from Twitter, 800 user reviews from IMDB | NB multinomial | Review-centric linguistic features | Accuracy on tweets >82% Accuracy on online movie review db >94% | Behavioural features need to be incorporated for better accuracy |
| Mohawesh *et al*. (2021) | Four real-world ds—YelpCHI, Yelp NYC, YelpZIP, and Yelp consumer electronics ds | SVM, LR, and Perceptron neural network | Review-centric | Assessed the correlation between concept drift and FRD problem and found they are negatively correlated | An efficient method is needed to handle the problem of concept drift |

***Table 3.*** *Continued*

| Reference | Dataset used | Techniques used/Methodology | Features | Result | Strength/Limitation |
|---|---|---|---|---|---|
| Oh and Park (2021) | First Korean dataset. (1735 comments on social issues) | SVM, Deep neural network | linguistic | Accuracy<br>SVM = 80.8%<br>DNN = 89.4% | The proposed technique is based on opinion spam rather than reviews, and metadata needs to be used |
| Poonguzhali *et al*. (2022) | – | Support Vector Machine | Textual (Sentiment analysis) | Product recommendation based on the removal of fake reviews | One user can only upload one review and not more than that |
| Salminen *et al*. (2022) | Used GPT-2 to create fake reviews to corresponding real reviews from the Amazon dataset | Baseline models—NBSVM and fine-tuning openAI fake detection model RoBERTa called fakeRoBERTa | Numerical vectors created by RoBERTa itself | Accuracy<br>NBSVM = 95.82<br>fakeRoBERTa = 96.64<br>OpenAI = 83 | Amazon dataset might also contain fake reviews which have been taken as authentic in this study. Hence, biases may creep in |
| Shan *et al*. (2021) | Yelp.com (24 539 reviews) | RF, CART, SVM, NB, MLPNN | Content, language, and non-verbal | Accuracy<br>RF = 92.9%<br>CART = 90.7%<br>SVM = 84.9%<br>NB = 73.5%<br>MLPNN = 83.6% | The feature set used may not apply to other domains as Yelp is a localized dataset |

***Table 3.*** *Continued*

| Reference | Dataset used | Techniques used/Methodology | Features | Result | Strength/Limitation |
|---|---|---|---|---|---|
| Theuerkauf and Peters (2023) | A dataset previously compiled by Martens & Maalej, comprising balanced reviews (16,000 balanced reviews) of apps in the iOS Apple App Store | Random Forest | Reviewer-based, product-based, review-based—different combinations of these features | 4 combinations of feature sets Accuracy = 79.17%, 80.92%, 92.16%, 94.38% respectively | False-negative predictions in the dataset may go undetected |
| Tufail *et al*. (2022) | Yelp dataset (1900 reviews) | SKL model (consisting of SVM, KNN, and LR) | Textual features such as Length count, bigram type, relationship words, etc. | Accuracy = 95% | Not much work has been done on unlabelled datasets in this model<br>It is only applicable to a limited domain |
| Wang and Kuan (2022) | Labelled Yelp review dataset | Differences in psycholinguistic features by using 3 different LR models | Message variables, formulation variables, and control variables | Found that deceptive reviews show fake writing styles in content but not in their expressions. | Focus mainly on linguistic summary variables rather than composite |
| Ligthart *et al*. (2021) | benchmark dataset of 1600 hotel reviews (Chicago) + Yelp dataset of 1560 reviews + Yelp restaurant dataset of 1600 reviews | Semi-supervised learning techniques, including self-training, co-training, Transductive SVM, label propagation, and spreading | Review-centric | Self-training came out best with accuracy = 93% with NB as the base classifier | Classified reviews on content only. No metadata or behavioural features were analysed |
| Tian *et al*. (2020) | Yelp (500 reviews) and benchmark dataset of 1600 hotel reviews (Chicago) | Semi-supervised PU learning called Ramp One class SVM | Textual | Accuracy<br>Ott ds: 92.13%<br>Yelp ds: 74.37% | Sentiment-based analysis makes it a domain and geographically dependent study |

***Table 3.*** *Continued*

| Reference | Dataset used | Techniques used/Methodology | Features | Result | Strength/Limitation |
|---|---|---|---|---|---|
| Wu *et al*. (2022) | Amazon dataset | semi-supervised probabilistic ensemble model including the individual behavioural features and the collusion-based behaviours via propagating the partial labels in the reviewer network | Time-related and rating related behavioural features | F-score on 50%labelled = 86.6% F-score on 5% labelled = 83.7% | The feature set used is small |
| Mothukuri *et al*. (2022) | Cafes from yelp | Unsupervised –K-means clustering, GMM Full covariance clustering, and GMM Diagonal covariance | Quantitative score, sentiment score, date of review, Property-ID, etc. | K-means shows the highest accuracy among the three algorithms | Does mainly clustering and not classification |
| Neisari *et al*. (2021) | Benchmark dataset of 1600 hotel reviews (Chicago) Second benchmark multidomain dataset. | Unsupervised learning utilizing the combination of self-organizing maps (SOM) and a CNN | Review-centric | Accuracy Hotel: 0.86 Doctor: 0.94 Restaurant: 0.88 Multidomain: 0.82 | Limited feature set and limited performance |
| Wang *et al*. (2020) | YelpNYC, YelpZIP | Markov random field-based method ColluEagle | Metadata features | Higher precision | Only considers review rating and date |
| Wang *et al*. (2022) | AmazonCn and Yelp | Unsupervised graph-based | Pairwise features from user reviews | AP on AmazonCn>0.85 AP on YelpHotel > 0.60 | More feature sets can be incorporated to improve the algorithm |

also correlated fake reviews of a restaurant with customer visits. Shukla *et al*. (2023) have used the latest language transformer model GPT-3, on a novel, annotated dataset of physician reviews. Then they compare the results with LR, XG, RF, and SVM and find that GPT-3 is superior to all in terms of accuracy. However, the major disadvantage of using GPT-3 is that it can't be fine-tuned to train datasets larger than 10 million characters, and the whole dataset can't be tested at a go.

It is observed that deep learning algorithms outperform traditional machine learning models when applied to large-scale datasets. However deep learning models are computationally expensive. In addition to the above, deep learning models are susceptible to overfitting and thus, they can't be used for smaller data. Some transformer-based models such as GPT-3 are uninterpretable. They operate as a black box and have higher computational requirements. Table 4 shows a summary of deep learning models used in FRD so far.

*3.1.3. Graph-based techniques*

Graph-based techniques in machine learning refer to approaches that leverage graph structures or networks to represent and analyse data. In these techniques, data elements are represented as nodes, while relationships or connections between the elements are represented as edges or links in the graph. Graph Neural Networks (GNNs), Graph Convolutional Networks (GCNs), and Graph Embeddings such as node2vec, GraphSAGE, and DeepWalk are some commonly used graph-based techniques in machine learning. They can capture complex dependencies and make informed predictions.

The authors Ren *et al*. (2022) use graph neural networks to find associations between reviews, users, and products. This is based on the premise that there is a dependency between the user and product, the reviewer and the time, ratings, etc. Then they introduce the idea of suspicious values based on the TrustRank (Gyongyi *et al*., 2004) method. Their model finds more fake reviews, but the accuracy has not increased significantly. Manaskasemsak *et al*. (2023) have used two novel graph-partitioning algorithms BeGP and BeGPX for FRD. They use the snowball effect to capture all fraudulent users. In the extended version of BeGPX, they also capture the semantic and emotional content of text.

Although graph-based techniques have several advantages, they suffer from scalability and data sparsity issues. As the datasets grow, the size of the graph and in turn, computational complexity grows exponentially. If the labelled data is sparse, as is in the case of most fake review datasets, it can result in weaker signals in graphs. Further, interpreting the results of graph-based techniques is a challenge in itself. Table 5, given below, summarizes the graph-based models used for fake review detection in the recent past.

*3.1.4. Swarm intelligence techniques*

Swarm techniques for fake review detection draw inspiration from the behaviour of swarms in nature, where collective intelligence emerges from the interactions of simple individuals. These techniques leverage swarm intelligence principles to detect fake reviews by considering the collective behaviour of reviewers or reviews within a dataset. Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), Bee Algorithm (BA), and Firefly Algorithm (FA) are a few examples of swarm techniques used in fake review detection. The work done by Shringi *et al*. (2022) uses the Fitness-based Grey Wolf Optimization (FGWOK) technique and k-means clustering to classify fake and authentic reviews. The datasets used are synthetic spam review (from the Database and Information System Lab, University of Illinois, TripAdvisor dataset), movie review dataset from IMDB, and Yelp review dataset. After comparing their results with various metaheuristic clustering methods, such as genetic algorithm (GA), particle swarm optimization (PSO), cuckoo search (CS), and artificial bee colony clustering, they found that their algorithm performs better than the current ways. However, they have not considered the feature interactions and could have used better optimizers. Similarly, Jacob and Rajendran (2022) give a Fuzzy Artificial Bee colony-based CNN-LSTM approach for fake review classification. After data preprocessing, they use the chi-squared technique for feature extraction and selection and CNN-LSTM-FABC is applied, which is found to outperform the earlier approaches. Here also, contextual features are not considered, which may give higher accuracy when combined with the extracted features. Previously, Saini

***Table 4.*** *Summary of Deep Learning models used for FRD*

| Reference | Dataset | Technique used | Features | Result | Strength/Limitation |
|---|---|---|---|---|---|
| Budhi *et al*. (2021) | YelpChi Hotel (5854), YelpChi restaurant (61 541), YelpNYC (359 052), YelpZip (608 598) | Random sampling techniques for class imbalance problem. 3 single—LR, SVM, Multilayer perceptron (ANN). 3 ensembles—bagging predictor, RF, Adaboost ensemble | Textual features | Accuracy LR = 86.78% SVM = 85.80% MLP = 84.13% (for YelpChi Hotel) | The proposed model shows that Under sampling (for solving imbalance) works well with ensembles but not on single classifiers. • Only textual features used |
| Gambetti and Han (2023) | Yelp + GPT3 created fake reviews | Human survey + GPT-Neo benchmarked with Bi-LSTM, LR, NB, RF, XGBoost, GPT2 and RoBERTa | Review-based, user-based, service provider base, writing based | Accuracy of human survey = 57.13% Accuracy of GPT-Neo = 95.21% | The analysis is restricted to New York City and cannot be generalized |
| Jing-Yu *et al*. (2022) | 1596 tagged real hotel reviews + 32 297 unlabelled reviews from TripAdvisor | AspamGAN | Review-centric | Accuracy on 10%labelled = 71.87% 50%labelled = 86.56% 90%labelled = 87.18% 100%labelled = 84.50% | Better recognition than baselines on the limited size of the dataset |
| Kumar *et al*. (2022) | The real-world dataset collected from Yelp.com 5044 restaurants in four US states + Amazon dataset | M-SMOTE algorithm Modified SMOTE to solve class imbalance also. Then applied XGBoost, LSTM, GBM, Ann, RNN, SVM, LR, KNN, NB, and RF with and without preprocessing | Features focusing on both review and reviewer | >80% AUC score | Features extracted from these datasets may not be present in other domains |

***Table 4.*** *Continued*

| Reference | Dataset | Technique used | Features | Result | Strength/Limitation |
|---|---|---|---|---|---|
| Mir *et al*. (2023) | Benchmark multidomain dataset | BERT is used for word embeddings, and classifiers used are SVM, RF, Bagging, KNN, AB, Gaussian NB | Review-centric | Accuracy<br>SVM = 87.81%<br>RF = 83.43%<br>Bagging = 79.06%<br>KNN = 77.18%<br>AB = 78.43%<br>GNB = 78.43% | Behavioural features may be included for better result |
| Mohawesh *et al*. (2021) | Four real-world datasets—YelpCHI, Yelp NYC, YelpZIP, and Yelp consumer electronics ds | SVM, LR, and Perceptron neural network | Review-centric | Assessed the correlation between concept drift and FRD problem and found they are negatively correlated | An efficient method is needed to handle the problem of concept drift |
| Salminen *et al*. (2022) | Used GPT-2 to create fake reviews to corresponding real reviews from the Amazon dataset | Baseline models—NBSVM and fine-tuning openAI fake detection model RoBERTa called fakeRoBERTa | Numerical vectors created by RoBERTa itself | Accuracy<br>NBSVM = 95.82<br>fakeRoBERTa = 96.64<br>OpenAI = 83 | Amazon dataset might also contain fake reviews which have been taken as authentic in this study. Hence, biases may creep in |
| Shan *et al*. (2021) | Yelp.com (24 539 reviews) | RF, CART, SVM, NB, MLPNN | Content, language, and non-verbal | Accuracy<br>RF = 92.9%<br>CART = 90.7%<br>SVM = 84.9%<br>NB = 73.5%<br>MLPNN = 83.6% | The feature set used may not apply to other domains as Yelp is a localized dataset |

***Table 4.*** *Continued*

| Reference | Dataset | Technique used | Features | Result | Strength/Limitation |
|---|---|---|---|---|---|
| Shukla *et al*. (2023) | A healthcare dataset generated by them (novel pre-labelled dataset of 38048 physician reviews) 38,048 reviews, with 29,630 (77.88%) labelled as authentic and 8,418 (22.12%) as fake | GPT-3 and compares with the previous ML Then GPT-4 is also employed to look for content similarity in reviews | Linguistic features | F1 score of GPT-3 = 0.713 (for a dataset of 10 000) F1 score of GPT-3 for cold start dataset = 0.345 (for dataset of 10 000) | Tries to solve cold start problems along with the creation of a new database. However, the problem of concept drift remains The training size constraint of GPT-3 limits the testing on the entire dataset |
| Tang *et al*. (2020) | Subsets of Yelp Hotel and Restaurant datasets | bfGAN (behaviour feature generating model) | Behavioural feature – real and synthetic | Accuracy of Hotel dataset = 83% Accuracy on restaurant = 75.7% | Handled cold start problem but GAN cannot be effectively trained in all cases |
| Alawadh *et al*. (2023) | Benchmark dataset of 1600 hotel reviews (Chicago) | Used multi-channel convolutional neural network with discourse markers as various n-grams, on different proportions of data split | Discourse markers as n-grams | Accuracy = 87.5% | The dataset's size could be increased Number of channels used in the CNN is less Imbalance of the dataset is not considered |

***Table 4.*** *Continued*

| Reference | Dataset | Technique used | Features | Result | Strength/Limitation |
|---|---|---|---|---|---|
| Alsubari *et al*. (2022) | A. 110 reviews from 3 Indian restaurants<br>B. benchmark dataset of 1600 hotel reviews (Chicago)<br>C. Yelp electronic product reviews (9461 reviews)<br>D. Amazon dataset of 30 products (21 000 reviews) | CNN-LSTM<br>CNN for dimensionality reduction and LSTM after that | n-grams of review content | In-domain accuracy<br>A-77%<br>B-85%<br>C-86%<br>D-87%<br>Cross-domain accuracy<br>= 89% | Cross-domain results are better than in-domain because of the larger size of the dataset |
| Bhuvaneshwari *et al*. (2021) | YelpZip dataset (1 935 038 reviews) | Novel framework Self Attention-based CNN Bi-LSTM (ACB) | Review-centric | Accuracy = 0.86 | Performs better than comparable models |
| Cao *et al*. (2022) | Multidomain benchmark dataset, Constructed dataset (9256), Yelp Restaurant (16 606) | ST-MFLC (using multi-feature fusion) | Local, temporal, and weighted semantic features | Higher results and good stability | The relationship and weights of distinguishing features need to be considered<br>Various modes of reviews may be incorporated |
| Hmoud and Waselallah (2022) | Yelp Hotel reviews | Evaluation of Bi-LSTM and CNN | Hybrid (Textual and behavioural) | Accuracy<br>Bi-LSTM = 85%<br>CNN = 84% | Focused on a single domain |
| Javed *et al*. (2021) | Yelp filtered dataset | Bag of n-grams and CNN ensemble | Textual and non-textual | F1 score = 92% | Focused on a single domain |

***Table 4.*** *Continued*

| Reference | Dataset | Technique used | Features | Result | Strength/Limitation |
|---|---|---|---|---|---|
| Neisari *et al*. (2021) | Benchmark dataset of 1600 hotel reviews (Chicago)<br>The second benchmark multidomain dataset | Unsupervised learning using self-organizing maps (SOM) combined with a convolutional neural network (CNN) | Review-centric | Accuracy = 87.63% | Performance is directly proportionate to the size of the SOM map and the neighbouring radii<br>Metadata could be included |
| Oh and Park (2021) | First Korean dataset. (1735 comments on social issues) | SVM, Deep neural network | linguistic | Accuracy<br>SVM = 80.8%<br>DNN = 89.4% | Based on opinion spam rather than reviews, and metadata needs to be used |
| Sadiq *et al*. (2021) | App review dataset from Google Play Store – 502 658 records | Deep Learning framework for predicting contradictions between numeric reviews and ratings in Google Apps.—CNN, LSTM, Bi-LSTM, and GRU<br>classify the star ratings to match the review | Reviewer-centric | Accuracy<br>CNN = 89%<br>LSTM = 87%<br>RNN = 83%<br>Bi-LSTM = 86%<br>GRU = 77% | Not much work done in this area (combination of review and corresponding rating) |
| Liu *et al*. (2022) | Benchmark multidomain dataset | Hierarchical attention network through N-gram CNN at word-to-sentence level and Bi-LSTM at sent-to-doc level | Review-centric | In-domain<br>Hotel = 83%<br>Restaurant = 77.5<br>Doctor = 91%<br>Cross-domain<br>Restaurant = 77.5%<br>Doctor = 67.3%<br>Mix domain = 86.5% | Uses all labelled data |

***Table 4.*** *Continued*

| Reference | Dataset | Technique used | Features | Result | Strength/Limitation |
|---|---|---|---|---|---|
| Kumar *et al*. (2022) | The real-world dataset that 0f 5044 restaurants in the USA + Amazon dataset | M-SMOTE algo Modified SMOTE to solve class imbalance also. Then applied XGBoost, LSTM, GBM, Ann, RNN, SVM, LR, KNN, NB, and RF with and without preprocessing | Both review and reviewer-centric features | >80% AUC score | Features extracted from these datasets may not be present in other domains |
| Mohawesh *et al*. (2021) | Yelp consumer electronic dataset (9653) + Deception dataset (2082 reviews) + Benchmark multidomain dataset | convolutional—LSTM (C-LSTM), character-level convolutional—LSTM, Hierarchical attention network (HAN), convolutional HAN, BERT, DistilBERT, and RoBERTa | Hybrid | Accuracy on deception dataset C-LSTM = 58% HAN = 75.1% Conv. HAN = 68.1% Char level C_LSTM = 79.4% BERT = 86.2% DistilBERT = 83.2% RoBERTa = 91.02% | This is a survey paper and also compares the DL methods on datasets |
| Saumya and Singh (2022) | Reviews (or comments) on 5 popular YouTube videos | LSTM and LSTM-autoencoder | Textual | F1 score of OneHot embedding = 0.99 | Very small dataset |
| Zhang *et al*. (2023) | Datasets from YelpZIP | Used layers—Bi-LSTM and attention mechanism on CNN layer | Behavioural and textual | | A novel approach that tries to use hybrid features for deep learning |

***Table 5.*** *Summary of graph-based techniques used for FR*

| Reference | Dataset | Technique used | Features | Result | Strength/ Limitation |
|---|---|---|---|---|---|
| Manaskasemsak *et al*. (2023) | YelpNYC, YelpZIP | Behavioural graph-partitioning approach BeGP and BeGPX | Hybrid—Similar behaviour of reviewers, emotions expressed in reviews. | precision@100 BeGPX on YelpNYC = 0.96 BeGP on YelpNYC = 0.85 (both for reviewer ranking) | Graph based on similar characteristics of reviewers is explored but is limited to a single domain |
| Rathore *et al*. (2021) | Real review DS from Google Play Store | The DeepWalk method is applied to reviewers' graph data, along with a (modified) semi-supervised clustering technique that allows for the integration of partial background knowledge | Reviewer-centric | Comparable accuracy score for detecting fraud reviewers | This needs to be explored more |
| Ren *et al*. (2022) | Yelp Restaurant dataset | Improved Graph neural networks to find associations | Textual | F1 score of Tr-GSAGE = 0.76 D_Tr-GSAGE = 0.77 | The accuracy of the model is compromised to find more undetected targets |
| Tang *et al*. (2021) | Amazon reviews dataset on 3 categories – baby, music instruments, and automotive | Fraud Aware Heterogeneous Graph Transformer (FAHGT) | User, rating, time, and text | F1 score Baby = 65.58 Musical instruments = 76.14 Automotive = 58.85 | Claims to discover camouflage in reviews |

***Table 5.*** *Continued*

| Reference | Dataset | Technique used | Features | Result | Strength/ Limitation |
|---|---|---|---|---|---|
| Fang *et al*. (2020) | Data from Amazon, Netflix, and Movielens | Dynamic knowledge graph-based method after adding time series feature and features extracted via ST-Bi-LSTM | Textual | Accuracy<br>Netflix dataset = 92.65%<br>Movielens dataset = 94.38%<br>Amazon dataset = 93.41% | Metadata and contextual features need to be added for a complete picture |
| Wang and Wu (2020) | The review dataset provided by Tencent consists of 85 025 users, 302 097 reviews, and 7584 apps | Detect, Defense, and forecast (DDF). Finds fraud reviewers with many iterations of Graph convolutional network, compared with baseline methods LR, RF, DeepWalk, LINE | Textual, behavioural, and temporal | The precision of DDF is approx. 0.95 across all thresholds | Domain-specific results |
| Wang *et al*. (2022) | AmazonCn and Yelp | Unsupervised graph-based | Pairwise features from user reviews | AP on AmazonCn > 0.85<br>AP on YelpHotel > 0.60 | More feature sets can be incorporated to improve the algorithm |

*et al*. (2021) used k-means artificial bee colony for feature selection after data preprocessing. Then they optimized clusters using ABC with k-means on three datasets namely Synthetic Spam (containing 478 total reviews), Yelp (containing 4952 reviews), and Movie (a subset of IMDB containing 8544 reviews). Earlier Pandey and Rajpoot (2019) used a combination of cuckoo search and Fermat spiral to identify spam reviews. It is compared to six metaheuristics clustering methods and was found to be better than these six. But still, the accuracy could have been improved further.

Thus, swarm intelligence approaches give an optimized result for the fake review detection problem, but this area has not been explored fully to date. Many bio-inspired techniques can be utilized and investigated. Table 6 summarizes the few swarm techniques employed for FRD so far.

#### *3.1.5. Other techniques*

In addition to machine learning techniques, several other approaches and techniques can be used for fake review detection. Some commonly used methods are linguistic analysis, metadata analysis, reviewer behaviour analysis, etc. These identifying features will be discussed in detail in Section 3.2. Hussain *et al*. (2020) focused on identifying behavioural and linguistic indicators of fake reviews. Filho *et al*. (2023) experimented on five theoretical studies.

At first, they implemented persuasion knowledge acquisition in which potential Customers can gain insights into the distinguishing features that set apart counterfeit reviews from authentic ones. Alternatively, in the second study, they are just told about which reviews are fake. Their research shows, that persons who know about the linguistic features of fake reviews are better able to detect them. But, due to the structure and vocabulary of fake and real reviews being very similar, their premise seems to fail in real life. The dichotomy of fake and real fails in real-life settings. Hlee *et al*. (2021) collected 4450 reviews from Yelp and showed that online reviews of new restaurants are manipulated. They elucidated the correlation between extreme ratings and counterfeit reviews. But again, they have just given a theoretical model. Li *et al*. (2021) use a reviewer grouping method. This is in the context of data mining; the objective is to categorize reviews from reviewers into distinct groups. These groups then contribute to the creation of innovative grouping models that can effectively identify both positive and negative deceptive reviews. In their study, Wang *et al*. (2022) investigate the significance of emotional and cognitive cues in detecting Fake Review Deception. Upon experimentation, they found that fake reviews require deliberation at the writer's end. Hence, emotional, and cognitive cues both play a significant role together. But writing real reviews is stress-free and hence cognitive cues are mostly absent there. However, their experiment is limited to the hospitality domain only and it is yet to be seen whether the results apply to other domains also. Table 7 shows theoretical models and other metadata-based techniques used for deceptive review identification.

It is important to highlight that these techniques have the potential to be employed either individually or in conjunction with machine learning approaches, thereby enhancing the precision and dependability of fake review detection systems. Each technique has its strengths and limitations, and a holistic approach combining multiple methods often yields more robust results. Figure 5 depicts the number of publications year-wise, under various techniques. The analysis of this graph indicates that the trend in the number of publications employing purely traditional machine learning has seen a steep decline as compared to publications employing deep learning. The reason may be the advent of language generator transformers like BERT and GPT-3.

### *3.2. Review based on features*

Up until now, the majority of the research has concentrated on either the textual attributes of the reviews or the behavioural characteristics of the reviewers. Hence, the former is known as the review-centric model and the latter is known as the reviewer-centric model. Jindal and Liu (2007) categorized features based on information associated with a review: reviewer-centric, and review-centric. Wang *et al*. (2022) have given a comprehensive fake review detection framework that combines both these models. They use

***Table 6.*** *Summary of swarm techniques for FRD*

| Reference | Dataset | Technique used | Result | Strength/ Limitation |
|---|---|---|---|---|
| Jacob and Rajendran (2022) | 1500 reviews from various sources like Expedia, Yelp.com, Hotels.com, Amazon, Trip Advisor as well as Priceline. Finally, 80 reviews formed the dataset | CNN-LSTM FABC (Fuzzy Artificial Bee Colony) | Accuracy approx. 98% | Optimization of the classification process but Contextual features are not taken into account. And the dataset is very small |
| Pandey and Rajpoot (2019) | Spam reviews (1600), synthetic spam reviews (479), yelp hotel (5678), yelp restaurant (58 517), and Twitter spam (10 000). | Uses the strength of cuckoo search and Fermat spiral | Accuracy with an optimal set of features<br>Spam review – 63.78%<br>Synthetic spam review – 70.48%<br>Yelp hotel = 69.64%<br>Yelp restaurant = 72.56%<br>Twitter = 97.82% | Finds optimal solution in lesser no. of iterations. However, accuracy could be improved |
| Saini *et al*. (2021) | Synthetic Spam (478) Yelp (4952), and Movie (a subset of IMDB 8544 reviews). | Clusters using K-means ABC | Comparable performance over all the experimented datasets in this work | The 'bees' have been used for feature selection, which is a step towards optimization, but a combination of features would have given better performance |
| Shringi *et al*. (2022) | Synthetic spam reviews, movie reviews from IMDB, Yelp hotel and restaurant review dataset | FGWOK (Fitness-based Grey Wolf Optimizer Clustering Method) | Accuracy on<br>Synthetic spam dataset = 82.68%<br>IMDB = 65.91%<br>Yelp hotel and restaurant = 78.51% | Better cluster optimization results than other swarm techniques. However, have not considered the feature interactions<br>Better optimizers can be proposed |

***Table 7.*** *Summary of techniques used for FRD, other than ML, DL, Graph-based or bio-inspired*

| Reference | Dataset | Technique used | Features | Result | Strength/ Limitation |
|---|---|---|---|---|---|
| Ansari and Gupta (2021) | Corpus of mobile phone reviews from Flipkart (120) | The method used is how customers perceive the reviews available on the e-commerce platform | Textual and metadata | Various combinations of variables give better results on models with control and without control variables as compared to baseline models | Identify a very important behavioural feature, that is, linguistic style for finding the reviewer's intentions but it can't be generalized to other domains.<br>Further, a very small dataset and a reviewer's emotions need to be considered |
| Chopra and Dixit (2023) | Amazon dataset from Kaggle (5 68 454 reviews)<br>Yelp dataset with 10 000 reviews | Opinion-mining approach | Review-centric | Amazon accuracy – 89.29%<br>Yelp accuracy – 70.79%<br>(90-10 split) | Removes PRs and NRs to avoid false recommendations. However, PRs and NRs affect large datasets more than smaller datasets. Also, the use of BoW creates sparse matrices |
| Crawford *et al*. (2021) | A model trained on Wikipedia and used after fine-tuning on hotel review DS | Inductive transfer training | Textual | Inductive transfer can perform better than the traditional BoW approach | The use of transfer learning has not been explored in FRD earlier. But using this, source dataset bias might hinder the performance<br>Cannot be generalized to all domains |
| Filho *et al*. (2023) | Questionnaires were given for this theoretical model | Persuasion knowledge acquisition study | Textual | Deduced that fake will go unnoticed by naïve users | Their theoretical model has been supported by 5 psychological studies, but the concept drift problem is not considered<br>Fake and real reviews are so similar that knowledge won't be able to help users much |

***Table 7.*** *Continued*

| Reference | Dataset | Technique used | Features | Result | Strength/ Limitation |
|---|---|---|---|---|---|
| Hlee *et al*. (2021) | 4450 from yelp.com | Inductive method for pattern search in Online reviews of newly opened restaurants vs the long-running ones | Text content, review extremity, reviewer network | Deduced that time trend contributes as a major factor in finding fake reviews | Time-trend is a new feature explored here but it is only for restaurant reviews. Also, their study relies on people to find fake and real and thus may contain bias |
| Hussain *et al*. (2020) | Real amazon dataset | Calculating drop score using behavioural and linguistic features separately (SRD-BM and SRD-LM) | Behavioural, linguistic | SRD-LM (Unigram) with<br>NB = 84.0<br>LR = 84.2<br>SVM = 86.5<br>RF = 73.3 | In-depth analysis of behavioural features but decreased accuracy due to two separate models on behavioural and linguistic features<br>Features may be combined to improve accuracy |
| Li *et al*. (2021) | Reviews from dianping.com<br>Hotel 89 741<br>Personal care and services 67 953<br>Movie 75 453 | Review group method | Features that are not dependent on language are created | Improves the precision by up to 7% in comparison to baseline techniques | Novel language-independent features are created. The collision of groups and their relationships is a novel feature used here |

***Table 7.*** *Continued*

| Reference | Dataset | Technique used | Features | Result | Strength/ Limitation |
|---|---|---|---|---|---|
| Moon *et al*. (2021) | 250 000 real-world hotel reviews | All Terms procedure | Word patterns | Also finds the features that determine fake reviews such as emotional exaggeration | Study using linguistic features as well as user's motives. Reviews dataset is created using crowdsourcing and human bias may creep in. In addition, it is domain-specific |
| Plotkina *et al*. (2020) | Created review dataset via 1041 respondents | Micro-linguistic automatic detection, binomial regression | Linguistic | High accuracy of 81% with a better deceit rate | Findings revealed that the addition of quality labels to reviews adds truth bias. Failed to investigate linguistic cues |
| Wang and Kuan (2022) | 66 940 reviews from Yelp restaurant | Computational linguistic analysis. And differences in psycholinguistic features by using 3 different LR models | Psycholinguistic features | A framework is provided to find the psychology behind fake reviews and thus identify them | The psychology behind fake reviews has been identified as just a theoretical model. Might not be successful in practice |
| Wang *et al*. (2022) | Opinion spam corpus (688 323 instances) Amazon's cell phones dataset (1 039 833 instances) | A fake review identification framework with a degree of suspicion and temporal patterns | Suspicion degree, review-centric and reviewer-centric features | Precision = 95.3% | Uses suspicion degree, reviewer, and review-centric features together but does not apply to large datasets. Geographical factors were not taken into account during the feature extraction process |

***Table 7.*** *Continued*

| Reference | Dataset | Technique used | Features | Result | Strength/ Limitation |
|---|---|---|---|---|---|
| Wang *et al*. (2022) | Restaurant n Yelp Hotel SF (60 464) | Emotional cues as features to detect | Emotional cues, cognitive cues, and their combination | Find that cognitive cues come into play along with emotional cues, mostly in writing fake reviews. Not in real ones | Reveals the competition of mental resources between emotional and cognitive cues but is a single-domain study only |
| Zhang *et al*. (2022) | Yelp.com | ImDetector—a system designed to identify fraudulent reviewers, addressing data imbalance through the utilization of weighted latent Dirichlet allocation (LDA) and Kullback–Leibler (KL) divergence | Latent topics | The data imbalance problem addressed | Addresses data imbalance in fake reviews using topic modelling but doesn't consider the features of text or the reviewers |

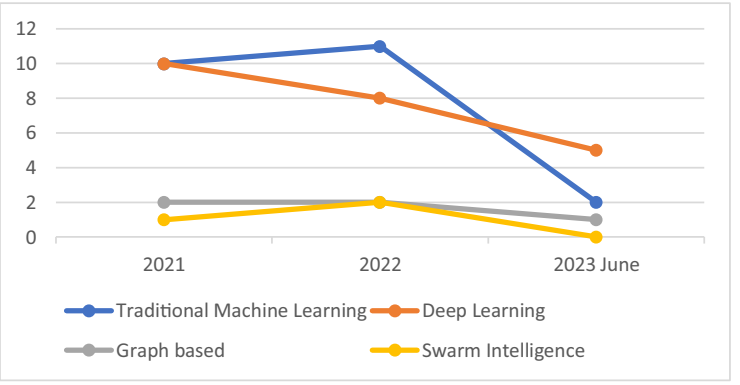


***Figure 5.*** *Distribution of various AI techniques for FRD. (The decline in the number for 2023 is because publications are taken only up to May 2023).*

Yelp and Amazon public datasets to find the effectiveness of their model. However, they have only used supervised learning algorithms. Earlier, Budhi *et al*. (2021) found 133 unique features that encompass a combination of textual and behavioural elements for detecting fake reviews through the utilization of ML techniques. Their approach was limited to balanced datasets only while in real life most of the datasets are imbalanced and the number of fake and real reviews is different. Furthermore, they employed two sampling methods to enhance the accuracy of balanced datasets.

Mattson *et al*. (2021) introduce a feature engineering approach known as the M-SMOTE (modified-synthetic minority over-sampling technique) model, which combines review- and reviewer-centric features for their analysis. In addition to incorporating the M-SMOTE model, they also aimed to address the issue of class imbalance in their study. They extract six reviewer-centric features namely rating entropy, review gap, review count, rating deviation, time of review, and user tenure. They also identified six review-centric features namely review length, word density, part-of-speech ratio, sentiment polarity ratio, SpamHitScore, and sentiment probability. A combination of both these models resulted in higher accuracy. However, their work is limited to the above-mentioned features only, which may not apply to other datasets. Alawadh *et al*. (2023a) perform discourse analysis on the reviews, based on the premise that factual reviews have strong coherence, whereas fraudulent reviews lack structure and semantics. However, their dataset is limited to a smaller review base as compared to the deep learning ones. In their other work, Alawadh *et al*. (2023b) have given a discourse analysis-based credibility check scheme, which gives higher performance. They have used the convolutional neural network with discourse markers as various n-grams, on different proportions of data split. However, they have used a small, balanced dataset, whose size could be increased to produce a scalable solution. Similarly, Chopra and Dixit (2023) investigate Push ratings (PRs) and nuke ratings (NRs) in fake reviews. Push ratings are higher ratings and nuke are lower ratings. They utilize opinion-mining techniques to determine the authenticity of the reviews. But they have used bag-of-words for preprocessing which results in a sparse matrix, which may lead to computational inefficiency. Tables 3, 4, 5 and 7 show the reviews used by the researchers in the past.

*3.2.1. Reviewer-centric models*

This model focuses on the atypical and suspicious behaviour of the reviewer as well as identifying the connections between a group of reviewers. These are non-linguistic characteristics of the reviews.

Tufail *et al*. (2022) have extracted the behavioural features of the user. For example, review time, writing style, relationship words, grammatical errors, punctuation, etc. These features may contribute to the classification of the review. This robust model only uses supervised techniques which may not give apt results for big datasets. Alsubari *et al*. (2022) ponder upon the fact that 90% of genuine reviewers usually write one review a day, based on the product they bought or the service they used. But 70% of

fraudsters may write up to 5 reviews a day. However, their dataset is limited to the hotel domain. Hence, the number of reviews per day is an important characteristic of detecting fake reviews (Heydari *et al*., 2015). However, the number of review features may not apply to other domains. According to Sadiq *et al*. (2021), the reviewers tend to give a higher proportion of positive fake reviews compared to the proportion of negative fake reviews. But they also have limited their work to a single domain. The rating of the fake reviewer tends to be different from the ratings of the genuine reviewer. This can also help in identifying fraudsters (Ott *et al*., 2013). The research conducted by Wang *et al*. (2022) encompasses the analysis of the non-verbal behaviour of the reviewer. They say that usually, the fake reviewer attempts to mislead customers by posting reviews as early as possible. They either give maximum rating stars or minimum, depending upon the fake positive or negative review. This may be attributed to the algorithms of e-commerce engines that detect fake reviews after a certain period of days and then only remove them. But the damage has already been done till then. The reviewer's credibility can also be assessed by determining the proportion of the reviewer's friends and followers. In addition to these characteristics, users with longer profile timelines would be more authentic than the users who have recently created their profiles. Even with a limited number of non-verbal features employed, the computational cost is less as compared to review-centric feature selection. But theirs is a purely theoretical model and its practical implementation is yet to be seen.

Another view of the authors Sadiq *et al*. (2021) says that the difference between the ratings given in a review and the emotional intensity of the review is also a giveaway. But they don't predict whether the review is fake or real. Hlee *et al*. (2021) collected 4450 reviews from Yelp and showed that online reviews of new restaurants are manipulated. They elucidated the correlation between extreme ratings and fake reviews. But again, they have just given a theoretical model. Tang *et al*. (2020) used a generative adversarial network (GAN) trained model to identify six behaviour features including text, rating, and attribute features.

Thus, some of the important reviewer-centric features can be quantity, user profile, timespan, source credibility, non-immediacy, etc. The fake review detection algorithms show good performance with the incorporation of these non-textual features. This means analysing reviews using the behaviour of the review's author or creator. But which behavioural features are to be selected for the detection problem, is a big task.

*3.2.2. Review-centric model*

This kind of fake review detection primarily centres on the textual content of the reviews. Research has indicated that there are substantial linguistic distinctions between authentic and fake reviews, which prove instrumental in their identification. These features may include micro-linguistic content or semantic content such as product characteristics.

Text features such as the count of nouns, verbs, and adjectives, along with the usage of strong positive and negative words in the review, have been recognized as potential indicators of a fake review (Jindal & Liu, 2007). The writing style of the reviewer can have lexical characters such as the number of characters and their proportion to uppercase letters, numeric characters, to the rate of spaces or tabs, called stylometric features can help in detecting fake reviews. However, they used the common features, and no new features were introduced for fake review detection. Jain *et al*. (Shan *et al*., 2021), employed Linguistic Inquiry and Word Count (LIWC), which is considered a deep linguistic feature. Its output such as emotions, self-reference, social words, overall cognitive words, etc. can be incorporated to find the fake review. Moon *et al*. (2021) obtain fake and authentic reviews of hotels via a survey and determine that features such as lack of details, temporal bias, and hyperbole are all part of fraudulent reviews. However, their work is based on surveys and is limited to just the hotel domain. Kumaran *et al*. (2021) have used language features like unigrams and their frequency, and bigrams and their frequency and length of reviews but conclude that behavioural features need to be added to accurately identify the reviews. However, their work focuses more on sentiment classification as positive or negative. The work done by Abdulqader *et al*. (2022) finds that the short length of the online review, review replication,

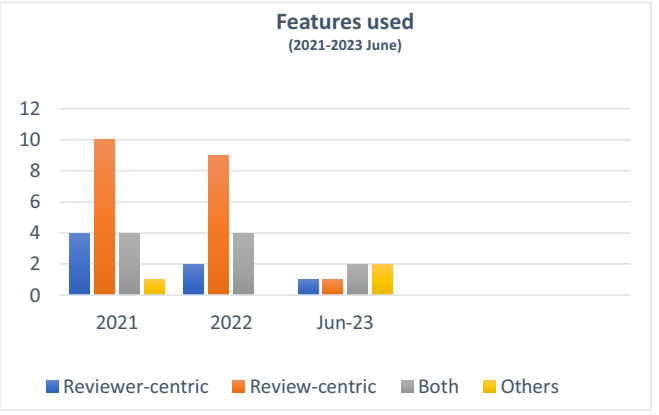


***Figure 6.*** *Number of publications in FRD according to features.*

Term frequency-inverse document frequency (TF-IDF), cohesion and coherence measures, and stylometric features are some of the telltale signs of a spam reviewer. Along with these, LIWC such as less usage of personal pronouns (less use of 'I',, 'we' etc.), less information about time and location, and strong use of positive as well as negative words, all point to fake reviewers. However, their work is based on theory, which may or may not apply to other datasets. Text similarity is widely used as an indication of fake reviews because spammers tend to copy the reviews to save effort (Hussain *et al*., 2020). But their work has studied textual and behavioural models separately which decreases the accuracy of fake review detection. The authors Wang and Kuan (2022) went a step further and divided the features into message level (review length and psychological cues), formulation level (readability and linguistic variables), and control level (rating extremity, review age, etc.). Although the dataset they used was Yelp (till 2012), their idea was to understand the process human brains undergo to formulate a message. Their limitation was that they focused on single domain and linguistic summary variables rather than composite variables.

Although review-centric features have been used extensively by researchers, it was found that they alone, or reviewer-centric features alone, are not sufficient to determine fake reviews accurately. Thus, a combination of important features of both should be incorporated to make any fake review detection algorithm outperform others. Table 8 shows some examples of the two types of identifying features used in the existing research. Figure 6 depicts the distribution of publications according to features. From the graph in this figure, it is observed that researchers have predominantly investigated the review-centric features over the reviewer-centric ones. This preference may stem from the fact that textual features are easier to identify as compared to behavioural or network-related ones. Additionally, hybrid (combination of review and reviewer-centric) and other features such as discourse and metadata (depicted in Table 8), are gaining popularity due to their enhanced performance in detecting fake reviews.

### *3.3. Review based on datasets*

Fake review detection involves the use of various datasets to train and evaluate models. Here are some commonly used datasets in fake review detection research:

1. Yelp Dataset: The Yelp dataset contains a large collection of reviews from the Yelp platform, including both genuine and fake reviews. It is widely used for training and evaluating fake review detection models.
2. Amazon Product Reviews Dataset: This dataset comprises product reviews from the Amazon platform and is frequently utilized for fake review detection tasks. It covers diverse product categories and contains both authentic and deceptive reviews.

***Table 8.*** *Summary of identifying features in each publication*

| Identifying features | Domain/Dataset | Examples of features used | References |
|---|---|---|---|
| Review-centric (textual) | Hospitality (Hotels and Restaurants)<br>Yelp—YelpChi, YelpNYC, YelpZip<br>TripAdvisor<br>Hotel reviews | Lack of details, emotional exaggeration, etc. No. of punctuations and emojis; linguistic inquiry, word count, etc. Weight of each word in sentence Max reviews, review length, rating deviation; emotional and cognitive cues | Ligthart *et al*. (2021), Mothukuri *et al*. (2022), Budhi *et al*. (2021), Chuttur and Bissonath (2022), Elmogy *et al*. (2021), Jain *et al*. (2021), Wang and Kuan (2022), Bhuvaneshwari *et al*. (2021), Hmoud and Waselallah (2022), Wang *et al*. (2022), Moon *et al*. (2021) |
| | Twitter dataset (1000 instances) | Unigrams, bigrams and their frequency, length of reviews | Kumaran *et al*. (2021) |
| | Subset of IMDB (800 instances)<br>5 popular YouTube videos | Unigrams, bigrams, and their frequency, length of reviews, and other textual features | Kumaran *et al*. (2021), Saumya and Singh (2022), Fang *et al*. (2020) |
| | Online survey on Qualtrics.com about 120 online products<br>Amazon products | Argument structure, flattering, etc. | Wang *et al*. (2022), Ansari and Gupta (2021) |
| | Korean dataset | | Oh and Park (2021) |

***Table 8.*** *Continued*

| Identifying features | Domain/Dataset | Examples of features used | References |
|---|---|---|---|
| | Multidomain | Semantic similarity of words, review length, frequency of words, etc. | Liu *et al*. (2022), Neisari *et al*. (2021), Mir *et al*. (2023), Cao *et al*. (2022) |
| Reviewer-centric (contextual or behavioural) | Hospitality (Hotels and Restaurants)<br>Yelp—YelpChi, YelpNYC, YelpZip<br>TripAdvisor<br>Hotel reviews | Review time, relationship words, sentiment word count, etc.; Relationship between extreme ratings and fake reviews; rating deviation, burstiness, entropy of temporal gaps, etc. | Tang and Cao (2020), Tufail *et al*. (2022), Manaskasemsak *et al*. (2023), Hlee *et al*. (2021) |
| | Apps reviews from Google Play store (502,648 instances); reviewer ID from Google Play store (38,123 instances) | Difference between ratings and emotional intensity of review; reviewer-ids to detect groups | Sadiq *et al*. (2021), Rathore *et al*. (2021) |
| | Crawled dianping.com (opinion sharing website) (4189 instances) | Reviewer group characteristics; inner group content similarity | Li *et al*. (2021) |
| | Amazon product reviews | Collusive behaviour between reviewers and individual behaviour | Wu *et al*. (2022) |

***Table 8.*** *Continued*

| Identifying features | Domain/Dataset | Examples of features used | References |
|---|---|---|---|
| Hybrid (Review and reviewer-centric) | Hospitality (Hotels and Restaurants)<br>Yelp—YelpChi, YelpNYC, YelpZip<br>TripAdvisor<br>Hotel reviews | Length of review, similar reviews for different products, cohesion, coherence, etc.; No. of nouns, verbs, adjectives, no. of reviews per day, strong emotional words, etc.; Emotional intensity, the proportion of nouns, suspicion degree, etc.; lexical diversity, hyperlink count, photo count, self-reference diversity, friend/follower count | Gambetti and Han (2023), Chopra and Dixit (2023), Wang *et al*. (2022), Abdulqader *et al*. (2022a), Alsubari *et al*. (2022), Kumar *et al*. (2022), Shan *et al*. (2021), Hmoud and Waselallah (2022), Javed *et al*. (2021), Javed *et al*. (2021), Zhang *et al*. (2023), Budhi *et al*. (2021) |
| | Product Reviews on Amazon<br>Apple IOS app reviews | Rating entropy, review gap, rating deviation, review length, word density, part-of-speech ratio, etc.; Ratings of review and opinion; length of username, time entropy, review text sentiment, etc. | Chopra and Dixit (2023), Wang *et al*. (2022), Kumar *et al*. (2022), Theuerkauf and Peters (2023), Tang *et al*. (2021) |
| | Tencent | Review quantity, length, repeat times, similar review number, time-based quantity distribution, etc. | Wang and Wu (2020) |
| Other features | Hotel reviews (1600 instances) | Discourse: Coherence, structure, and semantics; Discourse markers as various n-grams | Alawadh *et al*. (2023b), Alawadh *et al*. (2023a) |
| | YelpNYC, YelpZip | Metadata—textual plus product-related features like names or IDs | (Wang *et al*., 2020) |
| | Apple IOS app reviews | | Theuerkauf and Peters (2023) |

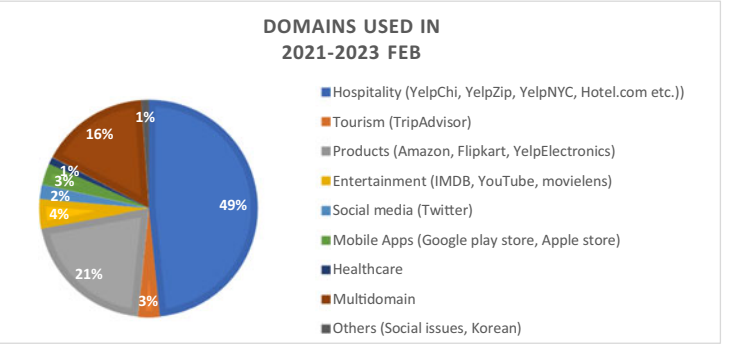


***Figure 7.*** *Domain distribution in FRD.*

3. TripAdvisor Dataset: The TripAdvisor dataset consists of reviews from popular travel websites, encompassing different destinations, hotels, and attractions. It serves as a valuable resource for fake review detection research in the travel domain.
4. IMDB Movie Reviews Dataset: The IMDB dataset includes a vast collection of movie reviews. It has been used in fake review detection studies to identify deceptive or fraudulent reviews among the genuine ones.
5. Deceptive Opinion Spam Dataset: This dataset specifically focuses on deceptive opinion spam, which involves generating fake reviews to manipulate public perception. It contains hotel reviews labelled as either truthful or deceptive, making it suitable for studying fake review detection.
6. Yelp Challenge Dataset: The Yelp Challenge dataset is a subset of the Yelp dataset and was released as part of a research competition. It consists of reviews and associated metadata, providing researchers with a resource to explore fake review detection techniques.

These datasets serve as valuable resources for studying and developing fake review detection models. This facilitates researchers in training and evaluating their algorithms using a wide range of diverse and realistic data. A descriptive summary of the major datasets used in existing research is given in Table 9.

This section examined multiple review datasets encompassing various domains such as products, stores, hotels, restaurants, and movies. This analysis is based on a thorough review of more than 80 research papers over a period spanning two and a half years. Figure 7 visualizes the distribution of datasets used across publications since 2021. These datasets vary in terms of size and composition. Figure 7 clearly shows the imbalance in the domains used for FRD. While the hospitality domain has been extensively utilized due to the availability of benchmark and labelled datasets, other domains such as healthcare and education have been largely overlooked. Annotated datasets in these domains are scarce, if not non-existent. For instance, the Gold standard dataset by Ott *et al*. (2013) consists of a collection of 400 genuine five-star reviews sourced from 20 hotels located in the Chicago area on TripAdvisor. Furthermore, the authors acquired 400 fabricated positive (deceptive) reviews for the identical set of 20 hotels from Amazon Mechanical Turk (AMT) to meet the requirements of their study.

By employing a word bag of features approach, they reported achieving an accuracy of 89.6% for their classification task. Mukherjee *et al*. (2013) conducted an analysis stating that reviews obtained from Amazon Mechanical Turk (AMT) are not truly fake as they lack adequate domain knowledge and do not exhibit a similar psychological mindset as expert authors who write genuine deceptive reviews. To address this issue, they utilized deceptive as well as truthful reviews from Yelp's real-life data, specifically the YelpChi dataset, which consisted of reviews for well-known restaurants and hotels in the Chicago area. By employing n-gram features, they achieved an accuracy of 67.8%. To further enhance accuracy, they put forward a collection of behavioural features related to the users and their reviews. Li *et al*. (2014) developed a benchmark dataset that spans multiple domains, including Restaurants, Hotel, and Doctors. This dataset encompasses three distinct types of opinions: authentic customer

***Table 9.*** *Summary of existing datasets used by various research*

| Dataset name | Description | Limitation | Reference |
|---|---|---|---|
| Single-domain hotel review dataset | This benchmark dataset comprises 1600 hotel reviews extracted from the TripAdvisor website, with an equal distribution of 800 spam and 800 ham (legitimate) reviews. These reviews are from 20 popular hotels in Chicago | Limited size, Single Annotation, crowdsourced spam, dataset specificity | Alawadh *et al*. (2023), Ligthart *et al*. (2021), Alsubari *et al*. (2021), Tian *et al*. (2020), Neisari *et al*. (2021), Pandey and Rajpoot (2019), Ott *et al*. (2013) |
| Multidomain deception dataset | Another benchmark dataset of Hotels, Restaurants, and doctors reviews (2836 reviews). Hotel reviews comprise 1880 reviews, Restaurants 400, and Doctor 556 reviews | Lack of metadata, labelling inaccuracy, limited size | Mohawesh *et al*. (2021), Liu *et al*. (2022), Neisari *et al*. (2021), Mir *et al*. (2023), Cao *et al*. (2022), Li *et al*. (2014) |
| Yelp Hotel and Restaurant datasets<br>– YelpChi<br>– YelpNYC<br>– YelpZip | The YelpChi dataset includes a subset of reviews and associated data from Yelp's platform, specifically about businesses in the Chicago area<br>The YelpNYC dataset comprises a subset of reviews and related information from businesses located in the New York City area<br>The YelpZip dataset encompasses reviews and associated data from various zip codes, providing a more diverse sample of businesses and user reviews<br>In FRD, they have been used mostly for hotel and restaurant reviews | Imbalance, dataset specificity | Ren *et al*. (2022), Shringi *et al*. (2022), Abdulqader *et al*. (2022), Mohawesh *et al*. (2021), Budhi *et al*. (2021), Jain *et al*. (2021), Wang *et al*. (2020), Manaskasemsak *et al*. (2023), Pandey and Rajpoot (2019), Bhuvaneshwari *et al*. (2021), Javed *et al*. (2021), Zhang *et al*. (2022), Rayana and Akoglu (2015) |

***Table 9.*** *Continued*

| Dataset name | Description | Limitation | Reference |
|---|---|---|---|
| Opinion spam corpus | Reviews collected from hotel and restaurant websites and has 688 323 instances | Imbalance | Wang *et al*. (2022), Jindal and Liu (2008) |
| Electronic products of USA | Again, a subset of the Amazon dataset (30 476 reviews) | | Alsubari *et al*. (2020) |
| TripAdvisor | Travel and tourism spam detection dataset | Limited size, Single Annotation | Alsubari *et al*. (2022), Jing-Yu *et al*. (2022), Chuttur and Bissonath (2022), Ott *et al*. (2013) |
| Amazon dataset | Product review subset taken from Kaggle containing 5 68 454 reviews<br>The subset containing cell phone reviews contains 1 039 833 reviews<br>Subset containing reviews on 3 categories—baby, musical instruments, and automotives | Imbalance, Lack of temporal or metadata, age of the dataset | Chopra and Dixit (2023), Kumar *et al*. (2022), Alsubari *et al*. (2020)<br>Wang *et al*. (2022)<br>Tang *et al*. (2021) |
| Twitter | 1000 tweets | Limited size | Kumaran *et al*. (2021), Pandey and Rajpoot (2019) |
| Korean dataset | First review dataset in Korean language with 1735 comments on social issues | Limited size, imbalance | Oh and Park (2021) |
| Apple app store | App review datasets containing 16 000 reviews | Dataset specificity | Theuerkauf and Peters (2023), Martens and Maalej (2019) |
| Healthcare | A novel dataset with 38 048 reviews with 29 630 as authentic and 8418 as fake | Imbalance | Shukla *et al*. (2023) |

***Table 9.*** *Continued*

| Dataset name | Description | Limitation | Reference |
|---|---|---|---|
| Google Play Store dataset | Contains 14 different categories of mobile app reviews comprising 502 658 records | Lack of metadata, imbalance | Sadiq *et al*. (2021), Rathore *et al*. (2021) |
| YouTube videos | A small dataset of reviews on 5 YouTube videos | Very small dataset | Saumya and Singh (2022) |
| Tencent | A dataset containing 85 025 users with 302 097 reviews on 7584 apps | Labelling accuracy | Wang and Wu (2020) |
| Multidomain Expedia, Yelp, Hotels.com, Amazon, TripAdvisor, Priceline | A total of 80 reviews | Very small size | Jacob and Rajendran (2022) |
| Synthetic spam | Containing 478 reviews | Limited size | Shringi *et al*. (2022), Saini *et al*. (2021); Kumaran *et al*. (2021), Li *et al*. (2021) |
| IMDB subset | 8544 reviews | Limited size, imbalanced dataset | |
| Multidomain from Dianping.com | A multidomain dataset containing 89 741 hotel reviews, 67 953 reviews on personal care items, and 75 453 movie reviews | Labelling inaccuracy | Li *et al*. (2021) |
| Flipkart | A subset of 120 reviews of mobile phones | Very small dataset | Ansari and Gupta (2021) |

reviews (submitted by actual customers), domain-expert-generated fake opinion spam (fabricated by employees or experts), and crowdsourced fake reviews (produced by Turkers, i.e. workers on Amazon Mechanical Turk). Rayana and Akoglu (2015) utilized three datasets, namely YelpChi, YelpNYC, and YelpZip, which were collected from Yelp.com. The YelpChi dataset consists of 67,395 reviews for 201 restaurants and hotels located in the Chicago area. The YelpNYC dataset contains 359 052 reviews for 923 restaurants situated in New York City. The YelpZip dataset comprises 608 598 reviews for 5044 restaurants located in various zip codes within the states of NY, NJ, VT, CT, and PA. To identify deceptive opinion spam and fraudulent reviewers, the researchers introduced a user-product bipartite graph model, specifically FraudEagle (unsupervised) and SpEagle (semi-supervised) approaches. These models leverage the graph structure to analyse patterns and detect potential instances of deceptive behaviour.

There are some limitations of these datasets. For instance, most of the datasets have been curated via crawling. Thus, they contain only limited features. Sometimes only review-centric features are there. This creates a big challenge for a benchmark dataset having multiple features. Secondly, Concept drift over time also poses a problem as the spam features also tend to drift with changes in fraud writings. Thirdly, there is a need for new datasets in languages other than English too. Korean, Arabic, and Chinese datasets have been created but they are small, and their labelling may be biased. Lastly, the problem of imbalance is a common issue with FRD. Class distribution is not uniform between fake and real. Thus, all researchers must put extra time into first balancing the dataset and then experimenting with it. Some work has been done by sampling techniques or ensembles in this area, but their performance is still not up to the mark.

## 4. Conclusion

The identification of deceptive reviews has emerged as a crucial concern for both researchers and unsuspecting consumers. To bridge the gap between earlier surveys and current research, this paper focuses on the research done between 2019 to 2023 and presents an updated overview of the various techniques, identifying features, and datasets employed in fake review detection. This paper conducted a comprehensive literature review in the field of Fake Review Detection (FRD) by examining three key aspects: the employed techniques, identifying features, and the datasets involved in the existing body of work.

In the first aspect, involving the FRD techniques, it was seen that the traditional machine learning methods depend on pre-defined features and require significant effort in feature engineering. Although these techniques are easy to implement and often perform well with small datasets, they face challenges when dealing with large datasets. Their effectiveness is limited due to a shortage of sufficient labelled data for input, which restricts the scope of their application. In contrast, deep learning models have the advantage of unsupervised learning of features from data, thus alleviating the need for manual feature engineering. While these models tend to perform better with large datasets, they may require more computational resources for training and can lead to overfitting on smaller datasets. Along with this, some language generation models used for classification, such as GPT-3, lack interpretability making it complex and difficult to comprehend how they arrive at their predictions. Further, it was found that swarm-based intelligence techniques yield optimized results, although these approaches have not been fully explored in the current research. In addition to the above, graph-based techniques in FRD leverage the relationships and structure within a graph representation to identify fake reviews as well as reviewers by detecting anomalies, identifying communities, or incorporating graph-based features into machine learning models.

The second categorization of this literature review focused on examining the identifying features. Analysis of these features reveals a decline in the popularity of review-centric features over time and hybrid features involving both textual and behavioural characteristics have emerged as efficient for FRD. However, the detection techniques need to be one step ahead of the fake reviews, necessitating the exploration of novel features. For instance, businesses may strategically promote positive fake reviews for their newly launched brand to give it a cold start. If these features can be scraped from their website, they can be used as a novel business-centric feature set, that can contribute to more effective FRD.

The final classification in this paper pertains to the use of various datasets in the current research. It has been noted that the datasets commonly employed in current studies are constructed through crowdsourced labelling and are prone to human perspective errors. Additionally, the findings reveal a predominant focus on the hospitality domain in the existing research, primarily due to the availability of labelled datasets. However, there is a need to gather and investigate data from domains such as education and the healthcare sector. This paper undertakes a comprehensive comparison and analysis of existing techniques to underscore the challenges within this field. This not only aids in identifying the most effective method but also facilitates further research on the intricate issue of detecting fake reviews.

## 5. Future directions

Most of the existing research primarily concentrates on training and testing models within a single domain. However, if a model can be trained in one domain and effectively applied to another, it can help address the scarcity of labelled datasets. This cross-domain classification will help a lot in areas where the availability of annotated datasets is much less. Another significant challenge in fake review detection is the problem of class imbalance, as most datasets have a substantial majority of reviews annotated as authentic rather than fake. Although some research has been done in this direction using sampling methods, the results have not been satisfactory. Hence, machine learning techniques like Siamese neural networks may be used for handling imbalances in datasets.

Another critical challenge is multilingual fake review detection, where research and data in languages other than English are severely lacking. Algorithms need to be trained to reflect global languages. Another demanding issue of fake review detection algorithms is the need to adapt and combat the problem of concept drift, ensuring they can identify fake reviews even as they evolve to resemble genuine ones. Classifiers need to be updated frequently to remain on top of the fake review detection. Moreover, existing feature extraction and selection methods exhibit various limitations, necessitating the exploration of novel feature sets. For instance, incorporating business-centric features such as fake reviews generated to boost sales of newly launched products could contribute to the development of a unique feature set.

Furthermore, the existing research has predominantly focused on employing machine learning, deep learning, and swarm intelligence techniques. The datasets used in these studies have primarily been obtained from platforms like Yelp and Amazon, but they are often outdated Many of these datasets are a result of web crawlers which causes many features to be left out or included unnecessarily. Consequently, the solutions developed on these datasets may or may not generalize well to newer datasets or different domains. Hence, there is a pressing need for a scalable, new benchmark dataset. In addition, previous research primarily focused on utilizing linguistic features for fraudulent review detection, neglecting the challenge of handling multimodal reviews that may include images, audio, video, metadata, etc. This is an emerging challenge as more and more reviews are described with the help of images and metadata.

Lastly, the FRD techniques employed by major companies like Google and Amazon are not publicly available. This highlights the need for a robust tool accessible to the public, accurately identifying fraudulent reviews and benefiting both consumers and e-commerce platforms.

**Competing interests.** On behalf of all authors, the corresponding author states that there is no conflict of interest.

## References

Abdulqader, M., Namoun, A. & Alsaawy, Y. 2022. Fake online reviews: A unified detection model using deception theories. *IEEE Access* **10**, 128622–128655.

Alawadh, H. M., Alabarah, A., Meraj, T. & Rauf, H. T. 2023b. Discourse analysis based credibility checks to online reviews using deep learning based discourse markers. *Computer Speech and Language* **78**, 101450.

Alawadh, H. M., Alabrah, A., Meraj, T. & Rauf, H. T. 2023a. Semantic features-based discourse analysis using deceptiveand real text reviews. *Information* **14**(1), 34.

Alsubari, N. S., Deshmukh, S. N., Al-Adhaileh, M. H., Alsaade, F. W. & Aldhyani, T. H. 2021. Development of integrated neural network model for identification of fake reviews in e-commerce using multidomain datasets. *Applied Bionics and Biomechanics* **1**, 5522574.

Alsubari, S. N., Deshmukh, S. N., Alqarni, A. A., Alsharif, N., Aldhyani, T. H., Alsaade, F. W. & Khalaf, O. I. 2022. Data analytics for the identification of fake reviews using supervised learning. *CMC-Computers, Materials & Continua* **70**, 3189–3204.

Alsubari, S. N., Shelke, M. B. & Deshmukh, S. N. 2020. Fake reviews identification based on deep computational linguistic. *International Journal of Advanced Science and Technology* **29**, 3846–3856.

Amos, R. 2022. Consumer Protection on the Web with Longitudinal Web Crawls and Analysis. Princeton University ProQuest Dissertations & Theses, 29061151.

Ansari, S. & Gupta, S. 2021. Customer perception of the deceptiveness of online product reviews: A speech act theory perspective. *International Journal of Information Management* **57**, 102286.

Barbado, R., Araque, O. & Iglesias, C. A. 2019. A framework for fake review detection in online consumer electronics retailers. *Information Processing & Management*, 1234–1244.

Basyar, I., Adiwijaya, M. D. & Murdiyansah, D. 2020. Email spam classification using gated recurrent unit and long short-term memory. *Journal of Computer Science* **16**, 559–567.

Bhuvaneshwari, P., Rao, A. N. & Robinson, Y. H. 2021. Spam review detection using self attention based CNN and bi-directional LSTM. *Multimedia Tools and Applications* **80**, 18107–18124.

Budhi, G. S., Chiong, R. & Wang, Z. 2021. Resampling imbalanced data to detect fake reviews using machine learning classifiers and textual-based features. *Multimed Tools Applications* **80**(9), 13079–13097.

Budhi, G. S., Chiong, R., Wang, Z. & Dhakal, S. 2021. Using a hybrid content-based and behaviour-based featuring approach in a parallel environment to detect fake reviews. *Electronic Commerce Research and Applications* **47**, 101048.

Cao, N., Ji, S., Chiu, D. K. & Gong, M. 2022. A deceptive reviews detection model: Separated training of multi-feature learning and classification. *Expert Systems with Applications* **187**, 115977.

Chopra, A. B. & Dixit, V. S. 2023. Detecting biased user-product ratings foronline products using opinion mining. *Journal of Intelligent Systems* **32**(1), 20229030.

Chuttur, M. Y. & Bissonath, R. 2022. A comparison of AdaBoost and SVC for fake hotel reviews detection. In 3rd International Conference on Computation, Automation and Knowledge Management (ICCAKM), Dubai.

Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N. & Al Najada, H. 2021. Using inductive transfer learning to improve hotel review spam detection. In 2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI).

Elmogy, A. M., Tariq, U., Ammar, M. & Ibrahim, A. 2021. Fake reviews detection using supervised machine learning. *International Journal of Advanced Computer Science and Applications* **12**(1).

Fang, Y., Wang, H., Zhao, L., Yu, F. & Wang, C. 2020. Dynamic knowledge graph based fake-review detection. *Applied Intelligence*, 4281–4295.

Filho, C., Rafael, M. D. N., Barros, L. S. G. & Mesquita, E. 2023. Mind the fake reviews! Protecting consumers from deception through persuasion knowledge acquisition. *Journal of Business Research* **156**, 113538.

Floridi, L. & Chiriatti, M. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines* **30**, 681–694.

Gambetti, A. & Han, Q. 2023. Combat AI With AI: Counteract Machine-Generated Fake Restaurant Reviews on Social Media, arXiv preprint arXiv:2302.07731.

Gyongyi, Z., Garcia-Molina, H. & Pederson, J. 2004. Combating web spam with trustrank. In Proceedings of the 30th International Conference on Very Large Data Bases (VLDB).

He, S., Hollenbeck, B. & Prosperpio, D. 2022. The market for fake reviews. *Marketing Science* **41**, 896–921.

Heydari, A., Tavakoli, M. A., Salim, M. N. & Heydari, Z. 2015. Detection of review spam: A survey. *Expert Systems with Applications* **42**(7), 3634–3642.

Hlee, S., Lee, H., Koo, C. & Chung, N., Fake Reviews or Not: Exploring the relationship between time trend and online restaurant reviews, *Telematics and Informatics*, **59**, 101560.

Hmoud, A.-A. M. & Waselallah, F. 2022. Detecting and analysing fake opinions using artificial intelligence algorithms. *Intelligent Automation & Soft Computing* **32**(1), 643–655.

Hussain, N., Mirza, H. T., Hussain, I., Iqbal, F. & Memon, I., 2020. Spam review detection using the linguistic and spammer behavioural methods. *IEEE Access* **8**, 53801–53816.

Ismagilova, E., Slade, E., Rana, N. P. & Dwivedi, Y. K. 2020. The effect of characteristics of source credibility on consumer behaviour: A meta-analysis. *Journal of Retailing and Consumer Services* **53**, 101736.

Jacob, M. S. & Rajendran, P. S. 2022. Fuzzy artificial bee colony-based CNN-LSTM and semantic feature for fake product review classification. *Concurrency and Computation: Practice and Experience* **34**(1), e6539.

Jain, P. K., Pamula, R. & Ansari, S. 2021. A supervised machine learning approach for the credibility assessment of user-generated content. *Wireless Personal Communications* **118**, 2469–2485.

Javed, M. S., Majeed, H., Mujtaba, H. & Beg, M. O. 2021. Fake reviews classification using deep learning ensemble of shallow convolutions. *Journal of Computational Social Science* **4**, 883–902.

Jindal, N. & Liu, B. 2007. Analyzing and detecting review spam. In Seventh IEEE International Conference on Data Mining (ICDM 2007).

Jindal, N. & Liu, B. 2008. Opinion spam and analysis. In Proceedings of the 2008 International Conference on Web Search and Data Mining.

Jing-Yu, C. & Ya-Jun, W. 2022. Semi-supervised fake reviews detection based on AspamGAN, *Journal of Artificial Intelligence* **4**(1), 17–36.
Kaddoura, S., Chandrasekaran, G., Popescu, D. E. & Duraisamy, J. H. 2022. A systematic literature review on spam content detection and classification. *PeerJ Computer Science* **8**, e830.
Khan, H., Asghar, M. U., Asghar, M. Z., Srivastava, G. P. K., Maddikunta, R. & Gadekallu, T. R. 2021. Fake review classification using supervised machine learning. In Pattern Recognition. ICPR International Workshops and Challenges.
Kostromitina, M., Keller, D., Cavusoglu, M. & Beloin, K. 2021. His lack of a mask ruined everything." Restaurant customer satisfaction during the COVID-19 outbreak: An analysis of Yelp review texts and star-ratings. *International Journal of Hospitality Management* **98**, 103048.
Kumar, A., Gopal, R. D., Shankar, R. & Tan, K. H. 2022. Fraudulent review detection model focusing on emotional expressions and explicit aspects: Investigating the potential of feature engineering. *Decision Support Systems* **155**, 113728.
Kumar, A. & Saroj, K. 2020. Impact of customer review on social media marketing strategies. *International Journal of Research in Business Studies* **5**(2), 105–114.
Kumaran, N., Chowdhary, C. H. & Sreekavya, D. 2021. Detection of fake online reviews using semi supervised and supervised learning. *International Research Journal of Engineering and Technology (IRJET)* **8**, 650–656.
Li, H., Chen, Z., Liu, B., Wei, X. & Shao, J. 2014. Spotting fake reviews via collective positive-unlabelled learning. In 2014 International Conference on Data Mining.
Li, Y., Wang, F., Zhang, S. & Niu, X. 2021. Detection of fake reviews using group model. *Mobile Networks and Applications* **26**(1), 91–103.
Ligthart, A., Catal, C. & Tekinerdogan, B. 2021. Analyzing the effectiveness of semi-supervised learning approaches for opinion spam classification. *Applied Soft Computing* **101**, 107023.
Liu, Y., Wang, L., Shi, T. & Li, J. 2022. Detection of spam reviews through a hierarchical attention architecture with N-gram CNN and Bi-LSTM. *Information Systems* **103**, 101865.
Manaskasemsak, B., Tantisuwankul, J. & Rungsawang, A. 2023. Fake review and reviewer detection through behavioural graph partitioning integrating deep neural network. *Neural Computing and Applications* **35**(2), 1169–1182.
Marco-Franco, J. E., Pita-Barros, P., Vivas-Orts, D., González-de-Julián, S. & Vivas-Consuelo, D. 2021. COVID-19, fake news, and vaccines: should regulation be implemented?. *International Journal of Environmental Research and Public Health* **18**(2), 744.
Martens, D. & Maalej, W. 2019. Towards understanding and detecting fake reviews in app stores. Empirical Software Engineering **6**, 3316–3355.
Mattson, C., Bushardt, R. L. & Artino Jr., A. R. 2021. When a measure becomes a target, it ceases to be a good measure. *Journal of Graduate Medical Education* **13**(1), 2–5.
Maurya, S. K., Singh, D. & Maurya, A. K. 2023. Deceptive opinion spam detection approaches: A literature survey. *Applied intelligence* **53**(2), 2189–2234.
Mewada, A. & Dewang, R. K. 2022. Research on false review detection Methods: A state-of-the-art review. *Journal of King Saud University –Computer and Information Sciences* **34**(9), 7530–7546.
Mir, A. Q., Khan, F. Y. & Chishti, M. A. 2023. Online Fake Review Detection Using Supervised Machine Learning And BERT Model, arXiv preprint arXiv:2301.03225.
Mohawesh, R., Tran, S., Ollington, R. & Xu, S. 2021. Analysis of concept drift in fake reviews detection, *Expert Systems with Applications* **169**, 114318.
Mohawesh, R., Xu, S., Tran, S. N., Ollington, R., Springer, M., Jararweh, Y. & Maqsood, S. 2021. Fake reviews detection: A survey. *IEEE Access* **9**, 65771–65802.
Moon, S., Kim, M.-Y. & Lacobucci, D. 2021. Content analysis of fake consumer reviews by survey-based text categorization. *International Journal of Research in Marketing* **38**(2), 343–364.
Mothukuri, R., Aasritha, A., Maramella, K. C., Pokala, K. N. & Perumalla, G. K. 2022. Fake review detection using unsupervised learning. In 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS).
Mukherjee, A., Venkataraman, V., Liu, B. & Glance, N. 2013. Fake review detection: Classification and analysis of real and pseudo reviews, UIC-CS-03-2013. Technical Report.
Neisari, A., Rueda, L. & Saad, S. 2021. Spam review detection using self-organizing maps and convolutional neural networks. *Computers & Security* **106**, 102274.
Oh, Y. W. & Park, C. H. 2021. Machine cleaning of online opinion spam: Developing a machine-learning algorithm for detecting deceptive comments. *American Behavioural Scientist* **65**(2), 389–403.
Ott, M., Cardie, C. & Hancock, J. T. 2013. Negative deceptive opinion spam. In Proceedings of the 2013Conf. of the North American Chapter of the Association for Computational Linguistics: Human LanguageTechnologies, Atlanta, Georgia.
Paget, S., Brightlocal, 7 February 2023. [Online]. Available: https://www.brightlocal.com/research/local-consumer-review-survey/ [Accessed 26 October 2023].
Pandey, A. C. & Rajpoot, D. S. 2019. Spam review detection using spiral cuckoo search clustering method. *Evolutionary Intelligence* **12**(2), 147–164.
Plotkina, D., Munzel, A. & Pallud, J. 2020. Illusions of truth—experimental insights into human and algorithmic detections of fake online reviews. *Journal of Business Research* **109**, 511–523.
Poonguzhali, R., Sowmiya, S. F., Surendar, P. & Vasikaran, M. 2022. Fake reviews detection using support vector machine. In 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS).
Rathore, P., Soni, J., Prabakar, N., Palaniswami, M. & Santi, P. 2021. Identifying groups of fake reviewers using a semisupervised approach. *IEEE Transactions on Computational Social Systems* **8**(6), 1369–1378.

Rayana, S. & Akoglu, L. 2015. Collective opinion spam detection: Bridging review networks and metadata. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Ren, X., Yuan, Z. & Huang, J. 2022. Research on fake reviews detection based on graph neural network. In International Symposium on Computer Applications and Information Systems (ISCAIS 2022), 290–297.

Ren, Y. & Ji, D. 2019. Learning to detect deceptive opinion spam: A survey. *IEEE Access* **7**, 42934–42945.

Rodrigues, J. C., Rodrigues, J. T., Gonsalves, V. L. K., Naik, A. U., Shetgaonkar, P. & Aswale, S. 2020. Machine & deep learning techniques for detection of fake reviews: A survey. In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE).

Sadiq, S., Umer, M., Ullah, S., Mirjalili, S., Rupapara, V. & Nappi, M. 2021. Discrepancy detection between actual user reviews and numeric ratings of Google App store using deep learning, *Expert Systems with Applications* **181**, 115111.

Saini, P., Shringi, S., Sharma, N. & Sharma, H. 2021. Spam review detection using K-means artificial bee colony. In *Communication and Intelligent Systems*, Proceedings of ICCIS 2020. Springer, Singapore, 731–744.

Salminen, J., Kandpal, C., Kamel, A. M., Jung, S.-g. & Jansen, B. J. 2022. Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services* **64**, 102771.

Saumya, S. & Singh, J. P. 2022. Spam review detection using LSTM autoencoder: An unsupervised approach. *Electronic Commerce Research* **22**, 113–133.

Shan, G., Zhou, L. & Zhang, D. 2021. From conflicts and confusion to doubts: Examining review inconsistency for fake review detection. *Decision Support Systems* **144**, 113513.

Shringi, S., Sharma, H. & Suthar, D. L. 2022. Fitness-based grey wolf optimizer clustering method for spam review detection. *Mathematical Problems in Engineering* **1**, 6499918.

Shukla, A. D., Agarwal, L., Mein, J. & Agarwal, R. 2023. Catch Me If You Can: Identifying Fraudulent Physician Reviews with Large Language Models Using Generative Pre-Trained Transformers, arXiv preprint arXiv:2304.09948.

Stanton, G. & Irissappane, A. A. 2020. GANs for semi-supervised opinion spam detection. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19).

Tang, H. & Cao, H. 2020. A review of research on detection of fake commodity reviews. *Journal of Physics: Conference Series* **1651**, (1), 012055,

Tang, S., Jin, L. & Cheng, F. 2021. Fraud detection in online product review systems via heterogeneous graph transformer. *IEEE Access* **9**, 167364–167373.

Tang, X., Qian, T. & You, Z. 2020. Generating behaviour features for cold-start spam review detection with adversarial learning. *Information Sciences* **526**, 274–288.

Thakur, R., Hale, D. & Summey, J. H. 2018. What motivates consumers to partake in cyber shilling?. *Journal of Marketing Theory and Practice* **26**, 181–195.

Theuerkauf, R. & Peters, R. 2023. Detecting fake reviews: Just a matter of data, In Proceedings of the 56th Hawaii International Conference on System Sciences.

Tian, Y., Mirzabagheri, M., Tirandazi, P. & Bamakan, S. M. H. 2020. A non-convex semi-supervised approach to opinion spam detection by ramp-one class SVM. *Information Processing & Management* **57**(6), 102381.

Tommasel, A. & Godoy, D. 2019. Short-text learning in social media: A review. *The Knowledge Engineering Review* **34**, e7.

Tufail, H., Ashraf, M. U., Alsubhi, K. & Aljahdali, H. M. 2022. The effect of fake reviews on e-commerce during and after Covid-19 pandemic: SKL-based fake reviews detection. *IEEE Access* **10**, 25555–25564.

Vidanagama, D. U., Silva, T. P. & Karunananda, A. S. 2020. Deceptive consumer review detection: A survey. *Artificial Intelligence Review* **53**(2), 1323–1352.

Wang, B. & Kuan, K. K. 2022. Understanding the message and formulation of fake online reviews: A language-production model perspective. *AIS Transactions on Human-Computer Interaction* **14**(2), 207–229.

Wang, E. Y., Fong, L. H. N. & Law, R. 2022. Detecting fake hospitality reviews through the interplay of emotional cues, cognitive cues and review valence. *International Journal of Contemporary Hospitality Management* **34**(1), 184–200.

Wang, J. & Wu, C. 2020. Camouflage is NOT easy: Uncovering adversarial fraudsters in large online app review platform. *Measurement and Control* **53**, 2137–2145.

Wang, N., Yang, J., Kong, X. & Gao, Y. 2022. A fake review identification framework considering the suspicion degree of reviews with time burst characteristics. *Expert Systems with Applications* **190**, 116207.

Wang, Z., Hu, R., Chen, Q., Gao, P. & Xu, X. 2020. ColluEagle: Collusive review spammer detection using Markov random fields. *Data Mining and Knowledge Discovery*, 1621–1641.

Wang, Z., Wei, W., Mao, X.-L., Guo, G., Zhou, P. & Jiang, S. 2022. User-based network embedding for opinion spammer detection. *Pattern Recognition* **125**, 108512.

Wu, Y., Ngai, E. W., Wu, P. & Wu, C. 2020. Fake online reviews: Literature review, synthesis, and directions for future research. Decision Support Systems 132, 113280.

Wu, Z., Liu, G., Wu, J. & Tan, Y. 2022. Are Neighbors Alike? A Semi-supervised Probabilistic Ensemble for Online Review Spammers Detection.

Zaman, M., Vo-Thanh, T., Nguyen, C. T., Hasan, R., Akter, S., Mariani, M. & Hikkerova, L. 2023. Motives for posting fake reviews: Evidence from a cross-cultural comparison. *Journal of Business Research* **154**, 113359.

Zhang, D., Li, W., Niu, B. & Wu, C. 2023. A deep learning approach for detecting fake reviewers: Exploiting reviewing behaviour and textual information. *Decision Support Systems* **166**, 113911.

Zhang, W., Xie, R., Wang, Q., Yang, Y. & Li, J. 2022. A novel approach for fraudulent reviewer detection based on weighted topic modelling and nearest neighbors with asymmetric Kullback–Leibler divergence. *Decision Support Systems* **157**, 113765.