Available at https://www.iasj.net/iasj
Iraqi Academic Scientific Journals
Journal homepage:
https://journals.uokerbala.edu.iq/index.php/UOKJ

Research Article

Exploring Approaches for Anomaly Detection in Textual Feedback: A Survey

[1,]Ali Hasan Jasim [2,] Mohsin Hasan Hussain
[1,]Computer Science Department, College of Computer Science and Information Technology, University of Kerbala, Karbala, Iraq.
[2,] College of Computer Science and Information Technology, University of Kerbala, Karbala, Iraq.

Article Info
Article history:
Received 7 -2-2026
Received in revised form10-5-2026
Accepted18-6-2026
Available online 30 -6 -2026
Keywords: Untruthful review, Fake review, Spam review, Opinion Spam, Machine learning, Deep learning. Textual anomaly detection, Large Language Models (LLMs).

Abstract

The present-day uses of the Internet as a platform for e-commerce have broadened the avenues for e-business. With e-commerce platforms generating myriad online reviews, end users, when making online purchases, rely on these reviews, and for businesses, the integrity of these reviews is of utmost importance as it can impact their bottom line. For such systems, reliable methods for identifying genuine reviews from false ones pose an important problem. With this in mind, the present study delves into a number of systems that have been developed to detect fraudulent reviews. The problem of review fraud is compounded by the increasing sophistication of fraudulent writing, the scarcity of reliable datasets with verified labelling, and the linguistic diversity between genuine and fraudulent reviews. This study endeavours to build a survey to cover the existing, to the best of our knowledge, anomaly detection systems aimed at online user reviews, from statistical, machine learning (ML), deep learning (DL), and transformer-based systems, along with their strengths and weaknesses.

Corresponding Author E-mail: ali.hassan.jasim@s.uokerbala.edu.iq , mohsin.h@uokerbala.edu.iq

Peer review under responsibility of Iraqi Academic Scientific Journal and University of Kerbala.

1 Introduction

The rapid evolution of e-commerce and Web 2.0 has transformed the digital landscape, enabling the seamless purchasing of products and services and empowering customers to share their personal experiences, feelings, and emotions through online reviews. Reviews play a crucial role in discovering product flaws and obtaining market intelligence information about their competition, and play an important role in improving the customer experience for new shoppers. Largely positive opinions are proven to entice more customers and generate significant financial gains, while negative reviews often lead to sales losses or a drop in conversions of up to 67%. Consequently, Many retailers rely on this public opinion to reshape their business plans and improve the quality of their products [1], [2].

Unfortunately, the substantial economic benefits derived from positive feedback have incentivised malicious activities ,leading to the wide diffusion of opinion spam or fake reviews. These fraudulent activities include promoting a target product with undeserving positive evaluations or defaming competitors with unreasonable, harmful reviews [2].

Accordingly, statistics indicate that the proportion of spam reviews can reach 14–20% on major online review platforms like Yelp [3]. The consequences of this deception are severe, leading to the loss of consumer trust and brand credibility, which makes customers unlikely to use the e-commerce site again [2]. Recognising the magnitude of this problem ,Anomaly review detection has been a critical subfield of natural language processing and a primary concern for platform owners and researchers [4].

The importance of this problem is clearly evident in multiple application areas. On e-commerce platforms, anomalous reviews directly impact consumer trust, product rankings, and marketing strategies, potentially leading to significant economic losses. On social media, unnatural comments or automated campaigns can be used to manipulate public opinion, spread misinformation, or inflate engagement, making the detection of this type of content both a technological and societal imperative [5], [6].

The subsequent sections are arranged as follows: Section 2 explains the research procedure. Section 3 gives a concise overview of the main foundational concepts. Then, Section 4 addresses the related work and the application of machine learning and deep learning techniques. Following that, Section 5 provides an overview of the currently used evaluation metrics. Section 6 provides an overview of the techniques used in the previous works, and Section 7 describes the obstacles that researchers usually face. Section 8 presents the conclusion. Finally, Section 9 declares the conflict of interest.

2 Survey Methodology

This survey aims to identify and classify research focusing on detecting anomalous reviews. To achieve this, a rigorous process was used, applying the defined inclusion and exclusion criteria, and focusing on research that proposed or assessed automated techniques to identify inappropriate or anomalous reviews in user-generated content. **Figure 1** shows that research on anomaly and fake review detection remained active between 2018 and 2025, despite fluctuations in publication counts. The highest number of publications was recorded in 2020, largely driven by advances in deep learning and BERT-based models, while another notable increase appeared in 2024 with the growing impact of large language models and AI-generated reviews. Overall, the field continues to evolve, with research trends shifting from traditional behaviour-based methods toward transformer-based and deep learning approaches.

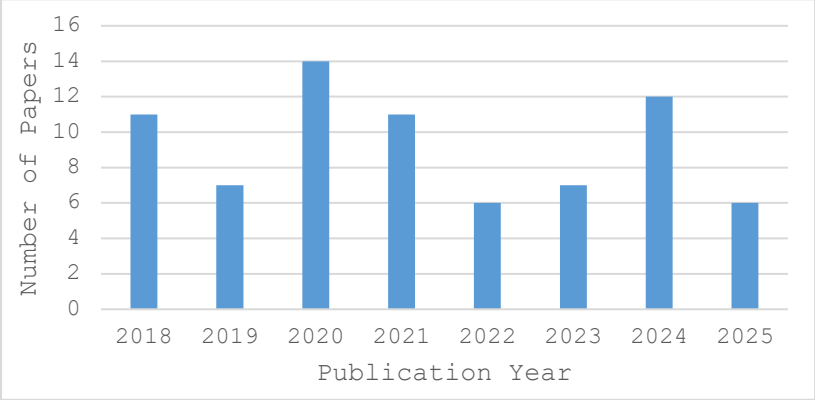

Figure 1: Annual distribution of reviewed publications on anomaly review detection (2018–2025).

2.1 Database selection

The review was performed on five important research databases. The first was the IEEE Xplore Digital Library, which contains research related to electronics and computer engineering. The second was Scopus, which encompasses every discipline. The third was the ACM Digital Library, which contains computer science research. The fourth was Google Scholar, which serves as a search engine to access articles across many databases, and the last was the Web of Science, which contains articles from influential journals. These databases were used to include conference papers and journals related to natural language processing, machine learning, and information systems.

2.2Search Keyword

The search was conducted using a combination of primary and secondary keyword groups. Primary terms included: "fake review detection", "opinion spam detection", "deceptive review", and "anomaly detection in text". Secondary terms were combined using Boolean operators (AND/OR): "machine learning", "deep learning", "natural language processing", "BERT", "transformer", "sentiment analysis", "text classification", "Yelp", "Amazon", and "e-commerce reviews". Searches were carried out in the title, abstract, and keyword fields of each database.

2.3Inclusion and Exclusion criteria

The studies included in this survey were published between 2018-2025, directly addressing fake review detection, opinion spam or anomaly detection in user-generated textual content, were published in peer-reviewed venues or reputable repositories, and were written in English. On the other hand, studies were excluded if they focused on non-textual anomaly detection domains (e.g., image, audio, or sensor data anomaly detection), represented duplicate or extended publications, or consisted of non-peer-reviewed grey literature. In addition, studies published before 2018 were excluded.

3 Foundations of Text-Based Anomaly Detection

This section focuses on the theoretical foundations of anomaly detection in textual data, clarifying the key concepts and scientific principles that distinguish anomalous texts from normal texts. It also aims to provide a conceptual framework that helps in understanding the nature of anomalous revisions and how to analyse them based on their various characteristics and patterns.

3.1Definition of Text Anomalies

Anomaly Detection (AD) is a traditional problem in computer science that revolves around separating normal observations from abnormal ones. Data patterns that do not fit well with a well-defined concept of normal

behaviour are called anomalies [7]. To explain and understand the nature of anomaly reviews, two examples of reviews taken from a Yelp real-life public dataset. The first review is normal, while the second is an anomaly.

• Review 1: "I like this hotel. The staff is very friendly, and you will feel at home. Great location, great hotel to spend the night:)"

• Review 2: "What an awesome place to stay. The staff is amazing and so friendly. The perks, such as free bike rental, are nice. The history (and restoration) of the building is really cool. Thanks for making my stay so memorable."

Review 1 is labelled as genuine in the dataset, as it provides specific, grounded feedback with clear personal reference (e.g., 'The staff is very friendly'). In contrast, Review 2 is labelled as anomalous: despite appearing positive, it lacks specific experiential details and relies on generic superlatives ('amazing', 'really cool', 'so memorable') that are common patterns in fraudulent promotional reviews as identified by Ott et al. [8]. The distinction illustrates how surface-level positivity alone does not indicate authenticity."

3.2 Types of Anomalies in Text Review/Feedback

An anomaly in review data is defined in the context of digital commerce as a notable departure from natural or authentic customer feedback patterns. According to recent studies, these anomalies can be divided into a number of different categories according to their structural behaviour, intent, and content:

- **Deceptive and Fake Review :**Fake reviews are authored to deliberately mislead readers. These are categorised based on their polarity: Positive Fake Reviews (Promotion) and Negative Fake Reviews (Demotion). These are deceptive reviews created to artificially inflate a product's reputation or damage the reputation of others [9].
- **Spam Review:** Spam in reviews refers to irrelevant or repetitive content that is considered non-review material, mainly consisting of advertisements or other irrelevant texts that do not contain genuine opinions [2].
- **Fraud Reviews**: Identity Fraud involves a single entity using multiple accounts to create a false sense of consensus. Noekhah et al. [10] identified this as a structural anomaly where multiple "users" share identical behavioural fingerprints or metadata.
- **Statistical Outliers :** outliers represent data points that are mathematically distant from the rest of the observations. In review systems, these are not always fraudulent but are always anomalous. Hussain et al. [11] identified burstiness as behavioural outliers are one of the most reliable indicators of organised spam campaigns. Rating Outliers that provide a rating significantly different from the product's historical mean, often used as a feature in anomaly detection models to flag potential manipulation [12].

With the rapid advancement of Large Language Models (LLMs), a new generation of AI-generated fake reviews has emerged. Recent studies by [12], [13] highlighted that transformer-based language models such as GPT-2 and GPT-4 are capable of generating highly coherent and contextually realistic reviews, making detection increasingly challenging for traditional machine learning approaches. Consequently, advanced deep learning and contextual language models have become essential for identifying sophisticated anomalous reviews and distinguishing them from genuine human-written content.

3.3 Features of Anomaly Reviews

Feature extraction plays a crucial role in fake review detection, as the quality and relevance of the selected features directly influence the performance of detection models. In the literature, these features are generally categorised into two main groups: behavioural features and textual features.

3.3.1 Behavioural Features

Indicators used to measure the spam score of suspected reviewers based on their past and

present reviews. These features can be summarised as follows.

- **Review count:** Compared to regular users, suspected reviewers frequently write multiple reviews with the aim of quickly improving or damaging a product's reputation. Therefore, the total number of user-written reviews is employed to effectively identify this type of abnormal behaviour [14].
- **Maximum amount of Reviews:** The maximum amount of reviews a reviewer posts per day. Often, deceptive reviewers don't spend much time to gain a little benefit. Therefore, they write many deceptive reviews in a single day.[15] .
- **Burstiness or Activity Window:** Deceptive reviewers often attempt to artificially boost ratings to achieve rapid impact. Posting a large number of reviews within a short time frame is therefore considered an anomalous behaviour and may indicate potential spamming activity.[10].
- **Content similarity:** Fraudulent reviewers typically rely on similar words to write a new review each time, without wasting their time writing reviews that contain new words [16].
- **Rating deviation:** How far the reviewer's rating is from the average product rating (indicates an attempt to raise or lower the rating)[17].
- **Review length:** Fraudulent reviewers write reviews with no more than 135 words, while about 92% of genuine reviewers write reviews with more than 200 words [18].
- **Percentage of positive reviews:** A significant percentage of positive reviews written by fraudulent reviewers associated with products may indicate the presence of deceptive reviews. [19].
- **Review gap:** Abnormal reviews are usually written at specific, regular intervals to significantly increase or decrease a product's rating, whereas genuine reviewers post their reviews randomly [14].
- **Time difference:** Fraudulent reviewers often post deceptive reviews at specific times. A common scenario is that a merchant will, in most cases, hire fraudulent reviewers at the early stage of a product launch, who tend to write consistently positive reviews [14].

3.3.2 Textual Features

This approach identifies anomalous reviews by analysing the review text. Review-centric features are features that are constructed from the content of an individual review. Review features can be categorised into many methods, such as POS tags, linguistic features, N-grams, sentiment, textual features and classification-related features [2].

Beyond surface-level textual features, a second family of text representation methods relies on semantic and dimensionality-reduction techniques. Latent Semantic Analysis (LSA) applies Singular Value Decomposition (SVD) to the term-document matrix to obtain a lower-dimensional representation that captures the latent thematic associations between terms and documents; by retaining only the most informative dimensions, this transformation reduces noise while preserving the dominant semantic structure of the corpus. Similarly, Non-negative Matrix Factorisation (NMF) decomposes the term-document matrix into interpretable, non-negative latent components, making it well-suited for topic extraction in review corpora. Principal Component Analysis (PCA) has been employed to reduce the dimensionality of high-dimensional review representations, thereby decreasing redundancy and improving classifier efficiency. Discriminant Latent Analysis (DLA) extends this idea to a supervised setting, seeking projections that maximise class separability between genuine and fraudulent reviews. Collectively, these semantic techniques serve as unsupervised or lightly supervised alternatives to bag-of-words models, enabling detection frameworks to capture broader topical and contextual patterns that surface-level features may miss. More recently, neural embedding techniques, including static embeddings such as Word2Vec, GloVe, and FastText, as well as contextualised embeddings produced by transformer-based models, have become central to modern fake

review detection pipelines, as they encode rich semantic and syntactic relationships directly into dense vector representations used as input to downstream classifiers.

4 Techniques of anomaly review detection

Over the past decade, many researchers have worked to develop sophisticated strategies for detecting anomalous reviews. Some researchers have attempted to apply traditional statistical methods to study different aspects of text characteristics from large-scale datasets, while continuing to refine and evaluate results using different feature extraction tools. Other researchers have begun to use machine learning methods to improve their detection frameworks. In particular, with the increasing use of natural language processing (NLP) methods, researchers are also proposing some neural models to solve the problems of detecting anomaly reviews. **Figure 2** illustrates the taxonomy of anomaly detection techniques, while the corresponding detection approaches are discussed in detail in the following sections.

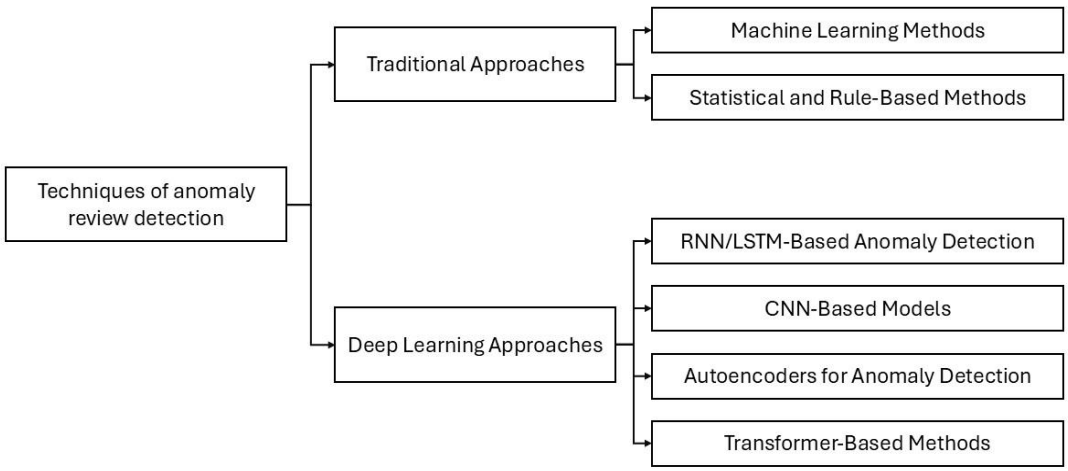

Figure 2: Taxonomy of Anomaly Review Detection Techniques, including traditional statistical and semantic methods, machine learning approaches, deep learning architectures, and transformer-based models with their associated text representation and embedding techniques

4.1 Traditional Approaches

Traditional approaches play an important role in detecting deceptive or anomalous reviews. These methods rely on several statistical and machine-learning techniques and are commonly categorised into supervised, unsupervised, and semi-supervised. Supervised learning relies on labelled training data, whereas unsupervised learning operates on entirely unlabeled datasets. In contrast, semi-supervised learning combines a limited amount of labelled data with a substantially larger pool of unlabeled data to guide the learning process [16].

4.1.1 Statistical and Rule-Based Methods

Traditional methods require extracting various features from reviews and are usually presented as a linguistic model. Due to anomalies, reviews contain many specific features, and feature engineering is essential for statistical models. The following is a summary of previous researchers who have used various feature engineering techniques.

Dong et al. (2018) [20] focuses on unsupervised topic – sentiment joint probabilistic modelling (UTSJ), which enhances the basic LDA model by adding a sentiment level. It uti-

lised Gibbs sampling to extract topic-sentiment distribution vectors as features, and fed them to a random forest classifier, tested on Yelp multi-domain balanced and unbalanced datasets. The model gained better semantics and sentiments of text reviews and became more accurate in filtering deceptive reviews.

On the other side, Asghar et al. (2020) [17] developed a Supervised Hybrid Classification framework using a Rule-based Feature Weighting Scheme and Spamicity Score Computation .The model used a comprehensive Hybrid Feature Set (17 Features) covering review feature, reviewer feature, and product features. Tested on an Amazon Dataset (cell phone accessories). This indicates that using the aggregation feature can enhance the efficiency of the classification model. Although this is considered a limitation because the set of spam features used by the researcher is limited and can be increased to obtain more robust results.

In the study by Noekhah et al. (2020) [10] proposed an unsupervised method using a graph-based model (MGSD). The primary objective of the proposed model is to uncover both internal and external relationships among entities. It assesses the spamicity of these entities through a multi-iterative algorithm that updates spamicity scores by incorporating various influencing factors. The model was evaluated using real-world Amazon review data in conjunction with a crowdsourced dataset. They calculated spamicity by applying feature combination methods to find the most effective combination of features.

More recently, Pan and Xu (2024) [12] introduced an unsupervised rule-based method for detecting deceptive reviews by analysing the features of ratings, text similarity, pictures(identical pictures), reviewer level, and the model is evaluated to verify the performance, since the accuracy of the recommendation increases when removing the fake reviews.

4.1.2 Machine Learning Methods

Machine learning approaches have been widely adopted for addressing anomaly review detection tasks due to their ability to automatically learn complex patterns from large-scale textual data. Researchers have increasingly applied machine learning–based techniques to tackle such classification problems. Researchers widely use these approaches to detect anomaly reviews, as shown below.

In the study, Elmurngi et al. (2018) [21] focused on the detection of Unfair Reviews on Amazon. The model involved a comparative evaluation of four Supervised ML Classifiers (Logistic Regression, SVM, Naïve Bayes, and Decision Tree). It exclusively used Linguistic Features derived from Sentiment Analysis of the review text. Tested on three Amazon product review datasets (Clothing, Baby, Pet Supplies), the models achieved varying accuracies. Logistic Regression achieved the highest accuracy of 81.61% on the Clothing dataset.

Yilmaz and Durahim (2018) [22] propose a semi-supervised framework (SPR2EP) to detect spam reviews that extract features from the textual content of the reviews and features obtained by taking advantage of the underlying reviewer product network structure. They use two learning algorithms (Doc2vec and node2vec). Embeddings generated using Doc2Vec for textual context and Node2Vec for network nodes, respectively. They evaluated the proposed model on three real-life Yelp datasets. The model got results on three sets of data, which were 80.71% AUC, 81.29% AUC, and 83.18% AUC. However, the proposed method was not compared with the neural network method to demonstrate its effectiveness.

Pandey et al. (2019) [23] Focus on spam review detection using a metaheuristic approach is achieved through the Spiral Cuckoo Search Clustering. The model employs LIWC features, refined using WOASA. They evaluated the proposed model on four spam datasets and one Twitter dataset. Furthermore, the proposed approach has been compared with several algorithms, including K-means, PSO, DE, and others. It turns out that the proposed model is more efficient than the compared methods.

Navastara et al. (2019) [24] introduced a self-training approach based on two types of features, Review Centric features and bi-gram features and used an SVM as a classification model. First, the initial training set was prepared from 90 labelled and 100 unlabeled

reviews, and then the SVM was used to predict the unlabelled data of 100 reviews and in the next iteration added to the current training set. The proposed model demonstrated good performance in spam review classification, achieving an average accuracy of 86.33%.

Hussain et al. (2020) [11] proposed a Behavioural Method (SRD-BM) to label data and a Linguistic Method (SRD-LM) for classification (using LR, SVM, RF, NB) .SRD-BM utilised 13 Rich Behavioural Features (e.g., Content Similarity, Review Burstiness). Tested on an Amazon Dataset, SRD-BM achieved 93.1% Accuracy, significantly outperforming the Linguistic Method SRD-LM best accuracy: 86.525%. A comparison of the two proposed models showed that the SRD-BM model obtained higher accuracy than the SRD-LM model because the SRD-BM model uses behavioural features of the dataset, which provide more support for identifying spammers and spam reviews.

Wang et al. (2020) [25] proposed a model for detecting fake reviews by combining multiple features and rolling collaborative training. They evaluated the proposed model on the YelpCHI dataset demonstrate that this method is more powerful than traditional algorithms. The author improves the effectiveness of the classification model by using unlabeled data.

Tian et al. (2020) [26] attempt to mitigate the scarcity of labelled data by adopting a semi-supervised one-class SVM approach. To enhance robustness against noise and outliers, they incorporated a Ramp loss function, giving rise to what they termed the Ramp one-class SVM. Their experimental framework consisted of several stages: text preprocessing (such as removing stop words and stemming), they used TF-IDF for feature extraction, and model validation via k-fold cross-validation to avoid overfitting. Finally, the parameters of the Ramp loss function were optimised using a grid search algorithm.

On their part, Yao et al. (2021) [27] proposed a model for fake review detection using an ensemble (Stacking/Voting) based on hybrid features (Reviewer-centric and Review-centric, TF-IDF Bigram). The author uses a grid search algorithm to handle the unbalanced data and resampling by finding the best sampling ratio for each classifier. Then, each classifier separately gets extracted features. The experimental result on two imbalanced Yelp datasets showed that the proposed model did not perform better than the state-of-the-art methods. Further, the proposed model exhibits temporal complexity.

Ligthart et al. (2021) [28] aims to evaluate the effectiveness of semi-supervised approaches for opinion spam classification. Four semi-supervised methods were implemented and compared against traditional supervised classifiers. Experimental results indicate that the self-training approach can outperform supervised methods when labelled data are scarce. The evaluation was conducted on the TripAdvisor dataset, where the Naïve Bayes classifier obtained the highest performance, reaching an accuracy of 93%. Nevertheless, these findings have limited generalizability to real-world scenarios.

Shan et al. (2021) [9] investigated the importance of Review Inconsistency in detection .RI Features (22 F) included Content, Sentiment, and Language inconsistency, tested with multiple ML models (RF, SVM, NB, etc.). Applied to Yelp Restaurant data, the Random Forest model incorporating all RI features achieved the highest F1-score. Motivated by this, Li et al. (2021) [29] proposed a new approach rely on aspect-oriented sentiment mining that can detect spam groups supported by nominated topics. Experimental results demonstrate that their approach is competitive and outperforms many modern solutions based on content duplication and time burstiness metrics.

Krishnan et al. (2022) [30] introduced comparing traditional supervised ML models (LR, SVM, KNN, DT) using Reviewer Actions derived from semantic analysis. Tested on a Mixed Product Dataset, the results show that the logistic algorithm has a greater accuracy. On the other side, Rizali et al. (2024) [31] compared Traditional ML models for spam review detection on the Shopee E-commerce platform .Features were combined (N-grams, Sentiment Analysis, Heuristic features). Tested on Shopee Reviews. Here, the SVM classifier

gives the best performance compared to all the different algorithms.

Purohit et al. (2025) [32] proposed a novel Parallel Hybrid k-means and Henry Gas Solubility Optimisation (H-KHGSO) algorithm for detecting fake reviews on social media platforms. The approach employs the Linguistic Inquiry Word Count (LIWC) tool to extract semantic, emotional, and linguistic features from textual data. Comprehensive experiments were carried out on three benchmark spam review datasets—SSR, MR, and YHRR—and the proposed method was compared with several state-of-the-art clustering algorithms, including PSO, DE, GA, CS, GWOK, and conventional k-means. The results confirm the superior performance of the H-KHGSO algorithm.

Summary: Traditional machine learning methods rely on learning from data using predefined features to predict outcomes. These methods are easy to implement and require fewer computational resources compared to deep learning models. They also often perform well when dealing with small datasets. However, feature engineering is a major challenge, as it requires prior domain knowledge to extract the appropriate features from the raw data. Additionally, traditional machine learning methods perform less efficiently on large datasets compared to deep learning models. **Table 1** shows the significant results in the traditional approaches in the literature.

Table 1: Summary of Studies Using Traditional Machine Learning Methods

No.	Author	Year	Dataset	Methods	Key Results
1	[20]	2018	• Yelp Chi dataset	• Unsupervised Topic-Sentiment Joint Probabilistic Model + Random Forest	F1-scores up to **85.41%** (unbalanced hotel).
2	[21]	2018	• Amazon product reviews	• Logistic Regression (LR) • SVM • Naïve Bayes (NB) • Decision Tree	Logistic Regression accuracy on clothing, shoes and jewellery reviews dataset: 81.61%
3	[22]	2018	• YelpChi • YelpNYC • YelpZip	• Semi-supervised SPR2EP framework (Doc2vec, node2vec)	Achieved 80.71%, 81.29%, and 83.18% AUC.
4	[24]	2019	• 90 labelled • 100 unlabeled reviews	• Self-training SVM approach	Achieved an average accuracy of 86.33%.
5	[17]	2020	• Amazon Dataset	• Supervised Hybrid Classification • Feature Weighting Scheme	Best Accuracy: 98%.
6	[10]	2020	• Gold Standard Dataset: 1,600 Hotel Reviews[8] • Amazon dataset	• Unsupervised Multi-iterative Graph-based opinion Spam Detection (MGSD)	Achieved 91.2% Accuracy on the Amazon dataset and 95.3% Accuracy on Ott's dataset.
7	[11]	2020	• Amazon Dataset	• SRD-BM (Behavioural Method - for labelling data	SRD-BM Accuracy: 93.1% SRD-LM accuracy with Naive

				• SRD-LM (Linguistic Method - for	Bayes classifier: 86.525% SRD-LM accuracy
8	[26]	2020	• Gold Standard Dataset: 1,600 Hotel Reviews[8] • Yelp	• Semi-supervised Ramp one-class • SVM + TF-IDF	Best Accuracy: 92.13% on gold standard dataset
9	[27]	2021	• Yelp Chi	• Ensemble Model: (RF, Xgboost, Lightgbm, Catboost) • Data Resampling (RUS) and Feature Pruning.	Best F1-score: 79.65% (Stacking, Restaurant Dataset).
10	[28]	2021	• Gold Standard Dataset: 1,600 Hotel Reviews[8] • Yelp Dataset.	• Self-Training (Base: Naïve Bayes).	Best Accuracy: 93% (Self-Training + NB)
11	[9]	2021	• Yelp	• ML models • Review Inconsistency (RI) Features	Best F1-score: RF accuracy:92.9%
12	[29]	2021	• Dataset collected from JD.com and TMALL.com	• Sentiment Mining K-means Clustering)	Competitive performance; outperforms solutions based on content duplication and burstiness metrics.
13	[30]	2022	• Mixed Product Dataset (retrieved from the web).	• Logistic Regression (LR), • Support Vector Machine(SVM), • K-Nearest Neighbour (KNN), • Decision Tree (DT).	Best Accuracy: 92% (Logistic Regression).
14	[12]	2024	• Dianping restaurant reviews, Yelp	• Unsupervised Fake-Review Detection (FRD)	The accuracy of recommendations increased significantly when fake reviews were removed.
15	[31]	2024	• Shopee Reviews (E-commerce - Storage Devices).	• SVM • Random Forest (RF) • Naïve Bayes (NB). • Feature Extraction: TF-IDF, Sentiment Analysis.	Best F1-score: 0.96 (SVM).
16	[32]	2025	• SSR, MR, YHRR (Synthetic, Movie, Yelp Hotel/Restaurant Reviews)	• Parallel Hybrid H-KHGSO algorithm • LIWC features	H-KHGSO outperformed all comparisons (k-means, PSO, DE, GA, CS, GWOK).

4.2 Deep Learning Approaches

Deep learning approaches achieve strong results in classification problems within natural language processing tasks. Deep learning can quickly extract valuable features from data, compared to traditional machine learning. Deep learning approaches also learn semantic representations of text through word embeddings. Recently, information science researchers have been using deep learning techniques in various NLP tasks. This section provides a brief review of some of the work that has used deep learning techniques for anomaly detection.

4.2.1 RNN/LSTM-Based Anomaly Detection

Recurrent Neural Networks (RNNs) are designed to process sequential data by maintaining internal memory that captures dependencies within input sequences. Although RNNs are theoretically capable of retaining information over long sequences, they suffer in practice from vanishing and exploding gradient problems, which limit their ability to model long-term dependencies. To overcome these limitations, several advanced architectures have been proposed, including Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Bidirectional LSTM, Stacked LSTM, and an attention-based LSTM model.

Wang et al. (2018) [33] employed a Long Short-Term Memory (LSTM) recurrent neural network to identify spammers using dictionary-based features. The proposed architecture comprised a multilayer perceptron with three main components: an input layer that accepts the textual features, an LSTM layer followed by a hidden layer for dimensionality reduction, and a single-neuron output layer that classifies reviewers as either genuine (0) or spam (1). The dataset was collected from product review webpages and the Taiwanese platform moblil01.com, with annotations derived from internal confidential documents. Experimental results demonstrated that the LSTM-based model outperformed the SVM baseline, achieving an accuracy of 89.4%, and highlighted the advantage of LSTM over traditional RNNs due to its ability to capture long-term dependencies. However, the study didn't compare the proposed model with other neural networks. In addition, the proposed model relied solely on textual features, neglecting behavioural and metadata information that could further enhance detection effectiveness.

To address the limitations of RNN, Liu et al. (2020) [34] proposed a bidirectional LSTM model to learn document-level representations of reviews for detecting fake reviews by combining heterogeneous features. The proposed framework integrates part-of-speech (POS) features, first-person pronoun indicators, and document representations derived from GloVe word embeddings. The experimental result on the AMT dataset across three domains (restaurant, hotel, and doctor). Results show the proposed approach consistently outperforms state-of-the-art baselines, including paragraph averaging, SWNN, SWNN+POS+I, BiLSTM, and basic CNN+POS+I. In mixed-domain settings, the model achieved superior performance compared to methods such as SWNN, Deep CNN, CNN–LSTM, and CLSTM, attaining an accuracy of 83.9%. Additionally, domain-specific evaluations reported accuracies of 83.9% for hotels, 85.8% for restaurants, and 83.8% for doctors. The findings further highlight the significant contribution of first-person pronoun features in detecting deceptive reviews. Nevertheless, the proposed model demands substantial computational cost, which may limit its practical deployment.

Shinde et al. (2024) [35] proposed a Deep Hybrid Ensemble Model combining Bi-LSTM with a Capsule Neural Network ,optimised with Borderline-SMOTE for imbalance .Features were Hybrid (Text, Behaviour, Deceptive Scoring). Mewada et al. (2025) [36] focused on deceptive opinion detection using Stacking-Based Deep Ensemble Learning .The ensemble consisted of four base classifiers (MLP, CNN, LSTM, BiLSTM) with a Logistic Regression Meta-Learner .Features were Deep Contextual Features via GloVe embeddings. Tested on the Amazon Fake Review Dataset.

4.2.2 CNN-Based Models

Convolutional Neural Networks (CNNs) play a crucial role in extracting salient features for anomaly text classification. Initially, review texts are converted into word vector

representations and used as input to the convolutional layer, which applies multiple filters of varying dimensions to capture local textual patterns. Activation functions such as ReLU or tanh are then employed to generate convolutional feature maps. Subsequently, max-pooling is applied to these feature maps to retain the most informative features while reducing dimensionality. The resulting representations are fed into fully connected dense layers, where activation functions such as softmax or sigmoid are used for classification, and the output layer produces the final prediction indicating whether a review is anomalous or normal.

Zhang et al. (2018) [37] presented a method for detecting misleading reviews using a recurrent neural network (DRI-RCNN) and the mathematical formula for the same method to identify misleading reviews using deep learning. They have tested the effectiveness of their model on two datasets: the spam dataset of Ott et al. [8] and the deception dataset of Li et al. [38] and observed that their proposed skip-gram based word embedding has outperformed other state-of-the-art methods in the detection of deceptive reviews.

Also, Archchitha and Charles (2019) [39] presented a convolutional neural network (CNN) model for opinion spam detection that leverages features derived from pretrained Global Vectors (GloVe) for word representation. The proposed architecture integrates behavioural features and conventional textual features through three parallel convolutional layers with varying filter sizes. To further improve performance, additional word-level and character-level features identified in prior studies were extracted and fused with the feature representations learned from the CNN layers.

4.2.3 Autoencoders for Anomaly Detection

An autoencoder is an unsupervised neural network that learns to associate inputs with itself. By adjusting the output values to be equal to the inputs. Thus, the dimensions of the inputs and outputs are equal. Due to the problem of labelled data, autoencoders can be applied as anomaly detection systems. The hidden layer representations of genuine reviews and abnormal reviews vary greatly, which is used to separate them into different subspaces.

Dong et al. (2020) [40] developed an end-to-end trainable unified model to use the power of an auto-encoder along with the random forest. This study has implemented a stochastic decision tree model to guide the global parameter learning process. Authors argue that an auto-encoder can preserve the representation of feature patterns in an unsupervised way. In the second layer of their model, they combined this feature representation with an ensemble-driven random forest to generate the final output. They considered various features of spam, such as historical records, rating signals, feedback signals, time signals, product comment information, and user review information for spam detection. The author has considered the Amazon review dataset, and their proposed method has outperformed other baseline methods.

Similarly, Saumya and Singh (2022) [41] present an unsupervised learning model that combines LSTM with an Autoencoder (LSTM-Autoencoder) for learning patterns of real reviews and overcoming the labelled dataset problem. It uses the reconstruction loss (the error between input and output) as the primary feature to cluster reviews into "real" or "spam". The study used multiple word embedding techniques, including One-Hot encoding, GloVe, and Word2Vec, to represent text, and the results of the experiment showed that One-Hot encoding performed better than other embedding methods.

4.2.4 Transformer-Based Methods (BERT, GPT, Large Language Model Embeddings)

Earlier representation models, such as Word2Vec and GloVe, assign a single, static vector to each word, regardless of its contextual usage. More recent approaches extend representation learning to sentences and documents, enabling models to generate context-aware word embeddings. In this setting, models trained on similar tasks can transfer learned knowledge rather than being trained from scratch, giving rise to the concept of transfer learning, where a pre-trained model is adapted to a new NLP task. Transformer-based architectures further advance this paradigm by employing encoder–decoder structures and multi-

head self-attention mechanisms, which effectively capture contextual dependencies and support sequence-to-sequence learning.

Refaeli & Hajek (2021) [42] focused on intrinsic fake review detection using Fine-tuned BERT .The model employed Context-Aware Embeddings from BERT. Tested on the Crowdsourced reviews dataset (balanced) and Yelp (imbalanced) datasets, the model achieved %91 on the Crowdsourced data and %73Accuracy on the Yelp dataset.

Salminen et al. (2022) [43] explored the Creation (GPT-2) and Detection (fakeRoBERTa) of machine-generated reviews. The fakeRoBERTa Classifier used Implicit Textual Patterns learned by the Transformer. Tested on an Amazon Dataset, the model achieved nearly perfect metrics.

Mohawesh et al. (2024) [44] introduced a Hybrid Model integrating RoBERTa with an LSTM Layer for fake review detection .Features were Semantic and Linguistic-aware Contextualised Embeddings .Tested on the Op-Spam and Deception benchmark datasets.

Geetha et al. (2025) [13] introduced MBO-DeBERTa ,a Transformer (DeBERTa) model optimised with the Monarch Butterfly Optimiser (MBO) .Features were Deep Contextual Features (DeBERTa's disentangled attention). Tested on three datasets (Amazon, Fake Review, Deceptive Opinion Spam), the model achieved high Accuracy on the Fake Review Dataset.

Summary: Deep learning and neural networks are among the most effective machine learning methods, and deep learning techniques have achieved outstanding results in anomaly review detection. One of the key advantages of deep learning is its ability to automatically extract features from input data without requiring manual feature engineering or prior domain knowledge. However, these models face several challenges when applied to fake review detection. Deep learning models are often considered “black boxes” due to their limited ability to provide a clear explanation of their decision-making mechanism and the generated results. In addition, these models require large amounts of data to achieve effective performance, making them less efficient with small datasets. Deep learning models also require significant computational resources compared to traditional machine learning methods. **Table 2** shows the summary of deep learning techniques with performance results.

Table 2: Summary of Studies Using Deep Machine Learning and Transformers Methods

No.	Author	Year	Dataset	Method	Key Results
1	[33]	2018	• Product webpages. • moblil01.	LSTM recurrent neural network + dictionary-based features	Outperformed SVM baseline, achieving 89.4% accuracy.
2	[37]	2018	• Gold Standard Dataset: 1,600 Hotel Reviews [8]. • deception dataset [38].	DRI-RCNN + skip-gram word embedding	Outperformed state-of-the-art methods in detecting deceptive reviews.
3	[39]	2019	• Gold Standard Dataset: 1,600 Hotel Reviews [8].	CNN + GloVe word representation	Successfully integrated behavioral and textual features via parallel convolutional layers.

4	[34]	2020	• AMT dataset (hotel, restaurant, doctor).	BiLSTM + heterogeneous features (POS, GloVe)	Achieved 83.9% accuracy, consistently outperforming state-of-the-art baselines.
5	[40]	2020	• Amazon review dataset.	Autoencoder + Random Forest	Outperformed baseline methods while preserving feature patterns unsupervised.
6	[42]	2021	• AMT dataset (hotel, restaurant, doctor). • Yelp.	Fine-tuned BERT + Context-Aware Embeddings	Best Accuracy: 91% and 73% Accuracy on Yelp
7	[41]	2022	• YouTube video reviews.	Unsupervised LSTM-Autoencoder	Achieved best performance with OneHot Embedding (F1-score: 0.99 for spam/real, MCC: 0.98).
8	[43]	2022	• Amazon Dataset.	GPT-2 (Creation) + fakeRoBERTa (Detection)	Best F1-score: 0.97 (fakeRoBERTa).
9	[44]	2024	• Deceptive Opinion Spam Corpus. • Deception.	Hybrid Model (RoBERTa + LSTM Layer)	Best Accuracy: 96.03% (OpSpam).
10	[35]	2024	• Yelp.	Bi-LSTM + Capsule Neural Network	Best Accuracy: 99% (Both Hotel and Restaurant datasets).
11	[36]	2025	• Amazon Fake Review Dataset.	Stacking-Based Deep Ensemble (MLP, CNN, LSTM, BiLSTM) + Logistic Regression	Best Accuracy: 90.04% (Stacked Model).
12	[13]	2025	• Amazon. • Fake Review. • Deceptive Opinion Spam Corpus.	MBO-DeBERTa (Transformer optimised with MBO)	Best Accuracy: 98% (Fake Review Dataset).

5 . Evaluation Metrics

The evaluation metrics measure the model's performance in terms of its effectiveness in detecting anomalies in texts. Effectively evaluating the performance of a text anomaly detection model is crucial for understanding its strengths and weaknesses. The primary goal is to compare the performance of different anomaly detection models. The metrics used in the literature can be broadly categorised into three groups: (1) standard classification metrics, applicable to supervised models; (2) task-specific textual metrics, designed for text-based evaluation; and (3) unsupervised evaluation metrics, used when ground-truth labels are unavailable.

5.1Standard Classification Metrics

There are several metrics available in supervised methods, but the most popular used by researchers in anomaly review detection are: F1score, recall, precision, accuracy and AUC, which are ideally suited for training data labelled normal and anomalous instances:

• **Accuracy**: This metric comprehensively estimates the number of cases that have been accurately identified, which can be determined as [45]:

$$Accuracy = \frac{number\ of\ correct\ prediction}{total\ number} = \frac{TP + TN}{TP + TN + FP + FN} \quad (5.1)$$

where TP represents the "true positive", TN indicates the "true negative", FP is the "false positive", and FN is the "false negative".

• **Precision**: This metric describes the proportion of successfully predicted reviews to the total number of reviews for a given class, which can be calculated as follows [45]:

$$Precision = \frac{number\ of correct\ predictions\ of\ each\ class}{total\ number\ of\ predictions\ of\ each\ class} = \frac{TP}{TP + FP} \quad (5.2)$$

• **Recall:** This metric shows the proportion of relevant reviews achieved from the total number of reviews and is calculated as follows [45]:

$$Recall = \frac{number\ of\ correct\ predictions}{total\ number\ of\ predictions} = \frac{TP}{TP + FN} \quad (5.3)$$

• **F1 score**: This metric shows the average of precision and recall and can be calculated as follows [45]:

$$F1 = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (5.4)$$

• **Matthews Correlation Coefficient (MCC):** is considered one of the most robust metrics for binary classification on severely imbalanced datasets. Unlike F1, it accounts for all four cells of the confusion matrix — TP, TN, FP, and FN — and returns a value between -1 and +1, where +1 indicates perfect prediction, 0 indicates random performance, and -1 indicates total disagreement as follows [46]:

$$MCC = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TN} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TP} + \text{FP})}} \quad (5.5)$$

• **AUC-ROC (Area Under the Receiver Operating Characteristic Curve):** evaluates a model's discriminative ability across all classification thresholds. An AUC of 1.0 represents a perfect classifier, while 0.5 indicates random performance. AUC is particularly valuable in comparing models on imbalanced datasets, where accuracy alone is insufficient [47].

5.2 Task-Specific and Text-Based Metrics

Beyond standard classification metrics, several metrics are specific to the textual and sentiment analysis aspects of fake review detection. These metrics complement classification metrics by capturing the semantic and opinion-related characteristics of the review content:

• **Text Polarity:** Score measures the overall positive or negative sentiment orientation of a review. Tools such as VADER and TextBlob are commonly used to compute polarity, which is then incorporated as a detection feature.

• **Text Similarity Score (Cosine Similarity):** quantifies the degree of overlap between reviews from the same user or about the same product.

• **Spamicity Score:** is a composite behavioural score used to rank reviewers or reviews by their likelihood of being spam. It is typically computed by aggregating weighted behavioural features such as review frequency and rating deviation.

• **Correlation Score (Pearson/Spearman):** is used in some studies to measure the statistical relationship between review ratings and review text sentiment

5.3 Unsupervised Evaluation Metrics

In unsupervised settings, evaluation becomes more challenging in unsupervised anomaly detection methods where labelled data is unavailable. Identifying true anomalies becomes more difficult. Several alternative metrics have been proposed to evaluate these models:

- **Anomaly Score:** assigns a numerical weight to each review or reviewer based on the degree of deviation from the modelled normal behaviour. Reviews that exceed a threshold anomaly score are classified as abnormal [48].
- **Reconstruction Error (used in autoencoder):** measures how accurately a model can reconstruct an input review after encoding it into a compressed latent representation. Normal reviews are expected to be reconstructed with low error, while anomaly reviews yield high reconstruction error and are classified as anomalous [41].
- **Recommendation-Based Metric**: proposed by [12] It evaluates fake review detection indirectly by measuring the improvement in recommendation system accuracy after removing detected fake reviews.

6 .Discussion

While it is widely accepted that anomaly detection algorithms employed in assessment reviews have improved over the years, significant trade-offs must also be acknowledged when comparing traditional machine learning (ML) and deep learning (DL) methodologies. SVM and Random Forest, for example, fall under the traditional ML bracket, and describe algorithms that are faster and more interpretable when compared to their deep learning counterparts. Traditional ML models perform well on carefully selected behavioural features (e.g. contrast in reviews and identity spoofing of reviewers), but are unable to capture complex transformations in the semantics of texts because they are limited by feature engineering.
On the one hand, and in comparison with traditional ML models, deep learning models and more specifically, Transformer architectures, excel in the domain of text and take an even more superior approach to automatic feature and contextual relationship extraction. Bi-LSTMs and more recently BERT-based models, for example, have been shown to outperform traditional ML models in the task of implicit linguistic anomaly detection because they are able to capture the textual context in both a sequential and bidirectional manner. On the other hand, a fundamental trade-off for deep learning models is that they operate as "black boxes", are less interpretable, and require more resources and larger datasets.
Furthermore, the reported effectiveness of these models is heavily dataset-dependent. High accuracy rates achieved on single-domain, balanced datasets frequently drop when models are tested in cross-domain scenarios or on highly imbalanced datasets, such as real-world datasets like Amazon reviews. Models trained on a small dataset often fail to generalise to larger, more complex review datasets. This dependency highlights that a model's superior performance is often a result of overfitting to a specific dataset's characteristics rather than a universal capability to detect anomaly behaviour.

7 .Challenges

Despite significant progress in the field of detecting anomalies in textual reviews, there remain many critical challenges that have yet to be addressed, paving the way for future research directions:

- **Multilingual and Cross-Lingual Reviews**

The vast majority of current detection models are heavily optimised for English text. As e-commerce platforms expand globally, the volume of reviews in multiple languages is exponentially increasing. such as English, Chinese, Malay or Arabic. Detecting deceptive reviews across different languages is highly challenging due to linguistic nuances; therefore, few studies have used datasets from different languages. [6], [49]. So, there is still a need to address this issue for detecting multilingual anomaly reviews.

- **Scarcity of High-Quality Labelled Datasets**

One of the major challenges in anomaly review detection is the lack of high-quality labelled datasets. Existing datasets often suffer

from limitations related to size, labelling reliability, and representation of real-world deceptive behaviour. For example, Jindal and Liu [16] labelled duplicated reviews from Amazon reviews as fake, although such behaviour may also occur legitimately. Similarly, Ott et al. [8] created a dataset using reviews generated through Amazon Mechanical Turk, but the dataset was relatively small and lacked diversity. Moreover, supervised and deep learning models heavily depend on accurately labelled data, while obtaining reliable ground-truth labels remains difficult due to the complexity of distinguishing sophisticated anomaly reviews from normal ones. Consequently, recent research has increasingly emphasised unsupervised, semi-supervised, and self-supervised approaches to reduce dependency on manual labelling and improve robustness in real-world anomaly detection tasks.

- **The cross-domain challenge**

The scarcity of annotated datasets hampers anomaly review detection. Other studies emphasised single-domain detection. As a result, improvements in model detection across the same domain led to poor performance when examining different domains. For instance, other studies [50] demonstrated that models trained on one domain achieved significantly lower accuracy when examined in different domains. Thus, an important area that needs research is cross-domain anomaly review detection.

8. Conclusion

The rapid growth of online platforms has greatly increased how reviews and ratings affect purchasing behaviour and how companies adapt their products. Consequently, maintaining online reviews' integrity has become a serious issue for both the academic and practical sides of this field. This survey paper describes and categorises research on fake and anomalous review detection and considers the use of machine learning and the somewhat older machine learning approaches, the deep learning frameworks, and some of the more recently proposed transformers.

The performance of anomaly detection techniques is heavily reliant on the training data and the way the features are represented.

Traditional machine learning approaches can be very useful due to their interpretability and the ease of obtaining results when there are sufficient behavioural and metadata cues. On the other hand, CNNs, LSTMs and deep learning approaches tend to be more effective at extracting complicated semantic relationships and contextual information from the text. Recently, transformer-based approaches such as BERT, RoBERTa and DeBERTa, which use deep, context-sensitive representations, have surpassed prior techniques and have become the state-of-the-art approaches to identifying subtle, machine-generated fake reviews. The survey also notes that detection frameworks combining contextual language modelling and a behavioural and metadata approach are among the best in detection frameworks.

Despite the advances made over the previous years, there are still challenges that remain open. A common problem in the existing approaches is the reliance on supervised learning, which lacks scalability due to the need for large datasets with correct labels. Acquiring reliable ground-truth labels is quite difficult. For instance, manual annotations and labels from Amazon Mechanical Turk lead to labelling biases. The resultant datasets have low reliability. To resolve these issues, more advanced, unsupervised, semi-supervised, and self-supervised methods, especially those that are not reliant on manual labelling to a large degree, are required. The field also recognises the challenges of effective cross-domain generalisation, explainability, and scalability when attempting to identify sophisticated AI-driven reviews that are generated by large, advanced AI language models.

Anomaly detection in online reviews continues to be a fast-evolving area of research. Progress in machine learning and natural language processing is expected to lead to more sophisticated detection that is more adept at recognising trickery. Similar to this line of thought, we hope that this survey provides a meaningful reference to researchers by identifying challenges, major research avenues, and emerging research directions in this field with respect to anomalies and fake review detection.

9 .Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this survey. No funding was received from any organisation that could have influenced the outcomes or conclusions presented in this work.

10 .References:

[1] S. N. Alsubari, "Data Analytics for the Identification of Fake Reviews Using Supervised Learning," *Comput. Mater. Contin.*, vol. 70, no. 2, pp. 3189–3204, 2022, doi: 10.32604/cmc.2022.019625.

[2] Dushyanthi Vidanagama *et al.*, "Deceptive consumer review detection: a survey," *Artif. Intell. Rev.*, vol. 53, no. 2, pp. 1323–1352, Feb. 2020, doi: 10.1007/s10462-019-09697-5.

[3] L. He, X. Wang, H. Chen, and G. Xu, "Online Spam Review Detection: A Survey of Literature," *Hum.-Centric Intell. Syst.*, vol. 2, no. 1, pp. 14–30, Jun. 2022, doi: 10.1007/s44230-022-00001-3.

[4] M. Ennaouri and A. Zellou, "Machine Learning Approaches for Fake Reviews Detection: A Systematic Literature Review," *J. Web Eng.*, vol. 22, no. 5, pp. 821–848, Jul. 2023, doi: 10.13052/jwe1540-9589.2254.

[5] L. S. Abdulzahra and A. J. Obaid, "Detection of Fake Reviews in Yelp Dataset Using Machine Learning and Chain Classifier Approach," in *Micro-Electronics and Telecommunication Engineering*, D. K. Sharma, S.-L. Peng, R. Sharma, and G. Jeon, Eds., Singapore: Springer Nature, 2024, pp. 331–346. doi: 10.1007/978-981-99-9562-2_27.

[6] O. Ignat, X. Xu, and R. Mihalcea, "MAiDE-up: Multilingual Deception Detection of GPT-generated Hotel Reviews," Jun. 19, 2024, *arXiv*: arXiv:2404.12938. doi: 10.48550/arXiv.2404.12938.

[7] K. Boutalbi, F. Loukil, H. Verjus, D. Telisson, and K. Salamatian, "Machine Learning for Text Anomaly Detection: A Systematic Review," in *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, Torino, Italy: IEEE, Jun. 2023, pp. 1319–1324. doi: 10.1109/COMPSAC57700.2023.00200.

[8] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding Deceptive Opinion Spam by Any Stretch of the Imagination," 2011, doi: 10.48550/ARXIV.1107.4557.

[9] G. Shan, L. Zhou, and D. Zhang, "From conflicts and confusion to doubts: Examining review inconsistency for fake review detection," *Decis. Support Syst.*, vol. 144, p. 113513, May 2021, doi: 10.1016/j.dss.2021.113513.

[10] S. Noekhah, N. B. Salim, and N. H. Zakaria, "Opinion spam detection: Using multi-iterative graph-based model," *Inf. Process. Manag.*, vol. 57, no. 1, p. 102140, Jan. 2020, doi: 10.1016/j.ipm.2019.102140.

[11] N. Hussain, H. Turab Mirza, I. Hussain, F. Iqbal, and I. Memon, "Spam Review Detection Using the Linguistic and Spammer Behavioral Methods," *IEEE Access*, vol. 8, pp. 53801–53816, 2020, doi: 10.1109/ACCESS.2020.2979226.

[12] Y. Pan and L. Xu, "Detecting Fake Online Reviews: An Unsupervised Detection Method With a Novel Performance Evaluation," *Int. J. Electron. Commer.*, vol. 28, no. 1, pp. 84–107, Jan. 2024, doi: 10.1080/10864415.2023.2295067.

[13] S. Geetha, E. Elakiya, R. S. Kanmani, and M. K. Das, "High performance fake review detection using pretrained DeBERTa optimized with Monarch Butterfly paradigm," *Sci. Rep.*, vol. 15, no. 1, p. 7445, Mar. 2025, doi: 10.1038/s41598-025-89453-8.

[14] J. Luo, J. Luo, G. Nan, and D. Li, "Fake review detection system for online E-commerce platforms: A supervised general mixed probability approach," *Decis. Support Syst.*, vol. 175, p. 114045, Dec. 2023, doi: 10.1016/j.dss.2023.114045.

[15] S. Rayana and L. Akoglu, "Collective Opinion Spam Detection: Bridging Review Networks and Metadata," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in KDD '15. New York, NY, USA: Association for Computing Machinery, Aug. 2015, pp. 985–994. doi: 10.1145/2783258.2783370.

[16] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proceedings of the international conference on Web search and web data mining - WSDM '08*, Palo Alto, California, USA: ACM Press, 2008, p. 219. doi: 10.1145/1341531.1341560.

[17] M. Z. Asghar, A. Ullah, S. Ahmad, and A. Khan, "Opinion spam detection framework using hybrid classification scheme," *Soft Comput.*, vol. 24, no. 5, pp. 3475–3498, Mar. 2020, doi: 10.1007/s00500-019-04107-y.

[18] H. Ahmed, I. Traore, and S. Saad, "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques," in *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, vol. 10618, I. Traore, I. Woungang, and A. Awad, Eds., in Lecture Notes in Computer Science, vol. 10618. , Cham: Springer International Publishing, 2017, pp. 127–138. doi: 10.1007/978-3-319-69155-8_9.

[19] X. Tang, T. Qian, and Z. You, "Generating behavior features for cold-start spam review detection with adversarial learning," *Inf. Sci.*, vol. 526, pp. 274–288, Jul. 2020, doi: 10.1016/j.ins.2020.03.063.

[20] L. Dong *et al.*, "An unsupervised topic-sentiment joint probabilistic model for detecting deceptive reviews," *Expert Syst. Appl.*, vol. 114, pp. 210–223, Dec. 2018, doi: 10.1016/j.eswa.2018.07.005.

[21] Elshrif Ibrahim Elmurngi, Abdelouahed Gherbi, and A. Gherbi, "Unfair Reviews Detection on Amazon Reviews using Sentiment Analysis with Supervised Learning Techniques," *J. Comput. Sci.*, vol. 14, no. 5, pp. 714–726, May 2018, doi: 10.3844/jcssp.2018.714.726.

[22] C. M. Yilmaz and A. O. Durahim, "SPR2EP: A Semi-Supervised Spam Review Detection Framework," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Aug. 2018, pp. 306–313. doi: 10.1109/ASONAM.2018.8508314.

[23] Avinash Chandra Pandey, A. C. Pandey, and D. S. Rajpoot, "Spam review detection using spiral cuckoo search clustering method," *Evol. Intell.*, vol. 12, no. 2, pp. 147–164, Feb. 2019, doi: 10.1007/s12065-019-00204-x.

[24] D. A. Navastara, A. A. Zaqiyah, and C. Fatichah, "Opinion Spam Detection in Product Reviews Using Self-Training Semi-Supervised Learning Approach," in *2019 International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation (ICAMIMIA)*, Batu, Indonesia: IEEE, Oct. 2019, pp. 169–173. doi: 10.1109/ICAMIMIA47173.2019.9223407 .

[25] J. Wang, H. Kan, F. Meng, Q. Mu, G. Shi, and X. Xiao, "Fake Review Detection Based on Multiple Feature Fusion and Rolling Collaborative Training," *IEEE Access*, vol. 8, pp. 182625–182639, 2020, doi: 10.1109/ACCESS.2020.3028588.

[26] Y. Tian, M. Mirzabagheri, P. Tirandazi, and S. M. H. Bamakan, "A non-convex semi-supervised approach to opinion spam detection by ramp-one class SVM," *Inf. Process. Manag.*, vol. 57, no. 6, p. 102381, Nov. 2020, doi: 10.1016/j.ipm.2020.102381.

[27] J. Yao, Y. Zheng, and H. Jiang, "An Ensemble Model for Fake Online Review Detection Based on Data Resampling, Feature Pruning, and Parameter Optimization," *IEEE Access*, vol. 9, pp. 16914–16927, 2021, doi: 10.1109/ACCESS.2021.3051174.

[28] A. Ligthart, C. Catal, and B. Tekinerdogan, "Analyzing the effectiveness of semi-supervised learning approaches for opinion spam classification," *Appl. Soft Comput.*, vol. 101, p. 107023, Mar. 2021, doi: 10.1016/j.asoc.2020.107023.

[29] J. Li, P. Lv, W. Xiao, L. Yang, and P. Zhang, "Exploring groups of opinion spam using sentiment analysis guided by nominated topics," *Expert Syst. Appl.*, vol. 171, p. 114585, Jun. 2021, doi: 10.1016/j.eswa.2021.114585.

[30] H. Muthu Krishnan *et al.*, "Detection of Fake Reviews on Online Products Using Machine Learning Algorithms," *Lect. Notes Netw. Syst.*, pp. 314–319, Jan. 2022, doi: 10.1007/978-3-030-96299-9_31.

[31] M. N. Rizali, M. M. Rosli, and N. A. S. Abdullah, "Spam Review Detection in E-

Commerce Using Machine Learning," in *2024 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, Aug. 2024, pp. 189–193. doi: 10.1109/IICAIET62352.2024.10730100.

[32] R. Purohit, K. R. Chowdhary, S. D. Purohit, R. Pal, and H. Sharma, "A Novel Parallel Hybrid k-Means and HGSO Based Approach for Detecting Fake Reviews in Social Media," *SN Comput. Sci.*, vol. 6, no. 5, pp. 1–16, Jun. 2025, doi: 10.1007/s42979-025-03990-7.

[33] C.-C. Wang, M.-Y. Day, C.-C. Chen, and J.-W. Liou, "Detecting spamming reviews using long short-term memory recurrent neural network framework," in *Proceedings of the 2nd International Conference on E-commerce, E-Business and E-Government*, Hong Kong Hong Kong: ACM, Jun. 2018, pp. 16–20. doi: 10.1145/3234781.3234794.

[34] W. Liu, W. Jing, and Y. Li, "Incorporating feature representation into BiLSTM for deceptive review detection," *Computing*, vol. 102, no. 3, pp. 701–715, Mar. 2020, doi: 10.1007/s00607-019-00763-y.

[35] S. A. Shinde *et al.*, "Deceptive opinion spam detection using bidirectional long short-term memory with capsule neural network," *Multimed. Tools Appl.*, vol. 83, no. 15, pp. 45111–45140, May 2024, doi: 10.1007/s11042-023-17348-9.

[36] A. Mewada, S. K. Maurya, Mohd. A. Ansari, O. P. Sharma, S. Avdhesh Yadav, and S. Ahmad, "Deceptive Opinion Detection Using Stacking-Based Deep Ensemble Learning," in *2025 3rd International Conference on Disruptive Technologies (ICDT)*, Mar. 2025, pp. 1614–1617. doi: 10.1109/ICDT63985.2025.10986298.

[37] W. Zhang, Y. Du, T. Yoshida, and Q. Wang, "DRI-RCNN: An approach to deceptive review identification using recurrent convolutional neural network," *Inf. Process. Manag.*, vol. 54, no. 4, pp. 576–592, Jul. 2018, doi: 10.1016/j.ipm.2018.03.007.

[38] J. Li, M. Ott, C. Cardie, and E. Hovy, "Towards a General Rule for Identifying Deceptive Opinion Spam," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Baltimore, Maryland: Association for Computational Linguistics, 2014, pp. 1566–1576. doi: 10.3115/v1/P14-1147.

[39] K. Archchitha and E. Y. A. Charles, "Opinion Spam Detection in Online Reviews Using Neural Networks," in *2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer)*, Sep. 2019, pp. 1–6. doi: 10.1109/ICTer48817.2019.9023695.

[40] M. Dong, L. Yao, X. Wang, B. Benatallah, C. Huang, and X. Ning, "Opinion fraud detection via neural autoencoder decision forest," *Pattern Recognit. Lett.*, vol. 132, pp. 21–29, Apr. 2020, doi: 10.1016/j.patrec.2018.07.013.

[41] S. Saumya and J. P. Singh, "Spam review detection using LSTM autoencoder: an unsupervised approach," *Electron. Commer. Res.*, vol. 22, no. 1, pp. 113–133, Mar. 2022, doi: 10.1007/s10660-020-09413-4.

[42] D. Refaeli and P. Hajek, "Detecting Fake Online Reviews using Fine-tuned BERT," presented at the ICEBI 2021: 2021 5th International Conference on E-Business and Internet, Singapore Singapore: ACM, Oct. 2021, pp. 76–80. doi: 10.1145/3497701.3497714.

[43] J. Salminen, C. Kandpal, A. M. Kamel, S. Jung, and B. J. Jansen, "Creating and detecting fake reviews of online products," *J. Retail. Consum. Serv.*, vol. 64, p. 102771, Jan. 2022, doi: 10.1016/j.jretconser.2021.102771.

[44] R. Mohawesh, H. B. Salameh, Y. Jararweh, M. Alkhalaileh, and S. Maqsood, "Fake review detection using transformer-based enhanced LSTM and RoBERTa," *Int. J. Cogn. Comput. Eng.*, vol. 5, pp. 250–258, Jan. 2024, doi: 10.1016/j.ijcce.2024.06.001.

[45] D. C. Blair, "Information Retrieval, 2nd ed. C.J. Van Rijsbergen. London: Butterworths; 1979: 208 pp. Price: $32.50," *J. Am. Soc. Inf. Sci.*, vol. 30, no. 6, pp. 374–375, Nov. 1979, doi: 10.1002/asi.4630300621.

[46] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, Dec. 2020, doi: 10.1186/s12864-019-6413-7.

[47] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve.," *Radiology*, vol. 143, no. 1, pp. 29–36, Apr. 1982, doi: 10.1148/radiology.143.1.7063747.

[48] M. Nikolić, M. Stojanović, and M. Marjanović, "Anomaly Detection in Hotel Reviews: Applying Data Science for Enhanced Review Integrity," in *2024 32nd Telecommunications Forum (TELFOR)*, Belgrade, Serbia: IEEE, Nov. 2024, pp. 1–4. doi: 10.1109/TELFOR63250.2024.10819036.

[49] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "What Yelp Fake Review Filter Might Be Doing?," *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 7, no. 1, pp. 409–418, Aug. 2021, doi: 10.1609/icwsm.v7i1.14389.

[50] H. Li, Z. Chen, B. Liu, X. Wei, and J. Shao, "Spotting Fake Reviews via Collective Positive-Unlabeled Learning," in *2014 IEEE International Conference on Data Mining*, Dec. 2014, pp. 899–904. doi: 10.1109/ICDM.2014.47.