Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Fake Review Detection: Taxonomies, Benchmarks, and Intent Modeling Frameworks

Shukla Banik[1*], Ritam Rajak[2]
[1*,2]Department of Computer Science and Engineering- (Artificial Intelligence), Brainware University, Barasat, India

ARTICLE INFO

Received: 12 Dec 2024

Revised: 18 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

Fake reviews on digital platforms are proliferating to a great extent, which is a challenge to the integrity of digital consumer ecosystems. This has put the development of reliable detection mechanisms as a critical research priority in influencing purchasing behavior and business reputation. In this paper, the rule-based feature-engineered machine learning systems, deep neural networks, transformer-based architectures, and emerging intent-aware frameworks are comprehended concerning fake review detection. These approaches are then classified with respect to model structure, feature dependence, and domain adaptability introduced with the help of a structured taxonomy. This study compares existing methods and uses comparative analysis to identify key limitations including domain sensitivity, overreliance on textual content, lack of interpretability, and low adversarial robustness. It particularly focuses on detection strategies in the light of modeler intent while respecting reviewer behavior using persona-based architectures and contrastive embedding alignment. These approaches allow for zero-shot detection as a path to more generalizable and semantically grounded systems. The paper also brings out the need for standardized benchmarking practices, ethically sourced datasets, and interdisciplinary methods incorporating natural language processing, behavioral analytics, and explainable AI. This work aids in bringing these scalability, transparency, and ethical alignment solutions a step closer, by critically evaluating the current landscape and proposing new research trajectories for fake review detection.

Keywords: Fake reviews, taxonomy, zero-shot detection, explainable AI, fake review detection.

1. INTRODUCTION

Consumer engagement has been profoundly influenced by the digitalization of the marketplaces in the modern world: it affects decision-makers behavior in marketplaces. Online reviews are a powerful form of social proof: among many other factors that influence transactional outcomes, these online reviews are especially influential factors [1]. Reviews by the users posted on platforms like Amazon, Yelp, TripAdvisor and Google reviews are not only peer reviews, but also include the ranking signals that affect algorithmic content placement [2]. Here, the authenticity of perceived reviews is directly related to the discoverability of the product, the credibility of the merchant, and the trust of the consumer.

Faced with positive commercial implications of user-generated feedback, both genuine participation as well as malicious manipulation have been incentivized [4]. As the practice of fabricated reviews, made to vanity inflate or belittle a product or service, becomes a systemic threat to digital trust, increasing prevalence of such reviews emerges. Fake or deceptive reviews, which are usually referred to as reviews by these sources, tend to be created by sources of unknown purchaser experience [5]. Most of them are created by third-party agents, paid influencers, and even automated bots representing vested interests. Another estimate has shown that as much as 15% of the published reviews in some sectors such as hospitality or electronics may be deceptive or unauthentic content [6].

The essence of the fake review detection problem is a subjective and subtle issue of natural language [7]. Often, deceptive reviews bear a resemblance to the style, tone, and structure of legitimate ones, rendering it very difficult for even human evaluators to identify these [8]. Moreover, it is not the case that fake reviews are only composed of explicit falsehood, as they may convey opinions that may well be semantically valid, but which are deceptive in intent. Traditionally, attempts to detect such content are complicated by the combination of surface-level plausibility and underlying manipulation. Traditional lines of defense, such as manual moderation, rule-based filtering keyword spotting, etc., which were standard in the past, have not been effective for these systems when generating texts from increasingly sophisticated techniques.

Initial solutions Attempts to address this problem were made by the feature engineering methods – in which we constructed classification models that rely on human-predefined indicators, like review length, polarity of the sentiment, and frequency of reviews per user [10]. Features of these reviews were taken to train supervised learning algorithms such as Naïve Bayes, Support Vector Machines, and Decision Trees to distinguish (separate) fake from genuine reviews [11]. Though these models achieved only small accuracy on domain-specific datasets, they were not very good. Their limited capacity to generalize over other product categories, languages, and review platforms was chief among them. The training data was also of variable quality and representativeness across studies, and these models were also heavily dependent upon it [12].

Later on, deep learning architectures were adopted for review classification when the field of natural language processing grew. From the text, RNNs, Long short-term memory (LSTM) models, and later on using Transformer based architectures like Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized Bidirectional Encoder Representations from Transformers (RoBERTa) have been able to extract semantic vectors of higher dimensions [13]. They showed superior performance on several benchmark datasets and enabled deeper contextual understanding, reduced reliance on handcrafted features, and the ability to capture complex linguistic dependencies. Nonetheless, however, there are still limitations. In supervised settings, deep learning models trained on a domain depend on the pretraining, as they tend to be sensitive to domain shifts and often need retraining or fine-tuning when applied to a new context or a new category of reviews [14]. In addition, these models tend to be black boxes that provide little to no interpretability in the deployment of the real world.

However, the situation has gotten even more confusing with the escalation of generative artificial intelligence. Due to the availability of highly capable language generation systems, such as GPT-based architectures as well as domain-specific synthetic review generators, the distinction between the reviews being human-written and machine-generated reviews is becoming more challenging [15]. These systems can give fluent output, appropriately contextualized, and emotionally nuanced output that is stylistically acceptable for a given review platform. Recent rapidly evolving adversarial tactics [16] hence make current detection models often obsolete.

One of the most persistent problems in this domain is the lack of high-quality labeled datasets [17] and beyond methodological limitations. Cohort of large-scale model training datasets for deception were usually subjective, the law and ethics associated with user privacy have impeded to construct large-scale ground truth datasets for models. However, several benchmark corpora have been developed such as the Ott Deceptive Opinion Spam Corpus and the Yelp Chicago dataset (YelpCHI), but these are not diverse with respect to the domain coverage, linguistic variations, and the complexity of the review intent. In many cases, they are derived from processes or environments that are artificially synthetic and their ecological validity comes up.

Therefore, there has been an emerging research direction that has begun to look at other paradigms that do not simply allow static feature extraction or purely text classification [19]. An interesting and potentially fruitful course of action turns out to be simply modeling the underlying intent behind the review, as opposed to focusing on its linguistic and structural characteristics. Thus, we can solve fake review detection as an intent inference task, to properly understand how close the expressed opinion is to a review and how much the reviewer's inferred motive or behavioral profile matches. This paper addresses a method to overcome mere surface-level measurements by incorporating reviews into the context of user activity, historical consistency, and emotional variance.

This emerging paradigm involves the latent representation of reviewer identity as one of reviewer persona modeling that constructs latent harbinger of reviewer identity over multiple reviews. Temporal frequency, emotional distribution, stylistic entropy, and contextual repetitiveness are all included as features for these personas [18]. Without relying solely on individual review content, inconsistencies can be identified that indicate deception by comparing current reviews to inferred persona. Such models also allow the detection of coordinated fake review campaigns by detecting clusters of similar but coordinated personas that exhibit congruent behavior [7].

A second major direction is developing systems that are capable of zero-shot detection – that is, they do not have to be retrained nor retask adapted when deployed in new domains. This is unlike traditional models where their performance degrades when the models are tested on different categories for which the models are never trained [21]. Specifically, in these systems, fake reviews are discovered not through learned patterns, but using semantic dissimilarity to legitimate reviewer intent spaces over reviews from outside the original training context.

The shift in its conceptual view that aims for intent-driven and domain-independent detection strategies further corresponds with the intended building of a scaleable, resilient, and ethically defensible information system for managing information quality. Such an integration of multi-view learning, that is, the integration of both textual content and behavioral metadata, with interpretability frameworks, provides a strong foundation for the next generation of fake review detection technology. In addition, these methods help with regulatory compliance by facilitating the explanation of model decisions, which has become increasingly important in applications, where the law requires transparency and fairness.

The need for dynamic and flexible detection systems will continually be magnified in the digital review ecosystem and against both human and artificial contributors who are shaping it. This field needs a change of paradigm from retroactively classifying research done in this field on the basis of a static set of features to prospectively modeling research from a behavioral theory, linguistic intention, and cognitive alignment. Such a trajectory represents a more profound understanding of deception, not just as a lapsed phenomenon constructed once and for all, but rather through a communicative gesture given within a relational and socially oriented behavior.

2. RELATED WORK

Over the last two decades, studies in deception opinion detection have seen significant improvement. Surface-level heuristics and rule-based methods were used early on; however, it then went towards supervised learning, behavior modeling, and deep contextual language understanding. This trajectory of development is a clear transition from manually engineered indicators to data-driven representations with the help of large-scale pre-trained models [22]. In addition to the content-based analysis, some recent research has also presented multi or modal and intent-based detection techniques to enhance the robustness of domains as well as decrease the dependency on the training data.

2.1 Rule-Based and Statistical Methods

The initial approaches of fake review detection were based on rules and statistical anomalies in textual data. Conditioned on the process of predefined lexical or behavioral rules, these methods were constructed with overly positive adjectives, extremely first-person pronouns, excessive punctuation, and so on [23]. In Jindal et al, they proposed one of the basic models for spam detection using text similarity and duplicate detection methods on Amazon datasets. Specifically, their work pointed out that unnatural patterns of deception can be identified by their observed textual repetition, lexical frequency, and similarity between reviews via the cosine metric [24].

Although these methods worked well on such datasets in the small (and somewhat homogeneous) case, they did not scale well when presented with stylistically varied or adversarially generated reviews. Finally, rule-based systems could not adapt, and subtle or well-crafted deception could not be handled. Consequently, these weaknesses led to the adoption of learning-based models that could learn complex relationships in text + metadata.

2.2 Supervised Machine Learning Approaches

From the early 2010s to this, fake review detection has been dominated by the methodology of supervised learning. Labeled datasets that included fake and truthful reviews were used for training these models to learn to discriminate based on engineered features. In these models, the features used were generally in four domains: the review text's lexical and syntactic features, behavioral features of the reviewer and posting frequency, and product-specific contextual information [11].

Model	Textual Features	Reviewer Behavior	Temporal Patterns	Product Context
Naïve Bayes	Yes	No	No	No
Support Vector Machine (SVM)	Yes	Yes	No	Yes
Random Forest	Yes	Yes	Yes	Yes
Decision Tree	Yes	Yes	Yes	No

Table 1. Representative supervised machine learning methods and their typical feature domains

The use of supervised learning significantly improved performance over rule-based approaches. Supervised learning performed much better than any rule-based approach. Although, these models were still very sensitive to the particular characteristics used at training. For this reason, when models trained on Yelp reviews were applied to Amazon and TripAdvisor datasets, they were not able to generalize because of differences in writing style, length of the content, and vocabulary [25]. Apart from this, the manual extraction of features introduced researcher bias and needed domain expertise.

2.3 Deep Learning and Contextual Representations

This motivated a shift from using feature engineering for dealing with text data, whose limitations prompted the use of deep learning models able of learning hierarchical representations of text data without human intervention [26]. The first deep architectures applied to fake review detection are Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) [27]. RNNs tried to learn a sequential dependency in the sentences while CNNs were trying to learn it on a local n-gram basis.

However, more recent approaches have exploited transformer-based language models (such as BERT, RoBERTa [28], or DistilBERT [11]) for deception detection since they are outperforming in various text classification tasks. They were trained by pretraining with large corpora and then fine-tuned on task-specific ones.

The condition had been eased by deep learning models; it no longer required manual feature extraction, but it brought other problems. Reduced interpretability, dependency on large-scale labeled data for fine-tuning, and poor performance under domain shift were the most troublesome issues they faced. In addition, heavy hardware resources and computation time were usually needed for pre-trained models, making the adoption of pre-trained models impractical for real-time systems or on-device applications.

2.4 Reviewer Behavior and Network-Based Models

Similar to the exploration of text classification model, some authors have investigated the modeling of review behavior for the purposes of spotting suspicious accounts or leafing out coordinated spam campaigns [29]. They leveraged metadata ranging from reviewer rating distribution, temporal burstiness, review duplication rate, and more, even social tie counts among reviewers and products.

The SpEagle model was one of the notable contributions in this domain, by which a reviewer-product-review graph was built and solved using belief propagation under the combination of textual and behavioral indicators

[30]. Reputation score computation, trust propagation, and outlier detection were other methods being focused in temporal series of reviews.

In other cases where text alone is not sufficient to detect the patterns, behavioral models were advantageous in that they could detect imperceptible patterns, such as in certain scenarios where bot-generated content or adversarial text obfuscation is present. However, these models required a large amount of access to reviewer metadata, which is often restricted for these sorts of reasons.

2.5 Domain Adaptation and Cross-Domain Detection

The problem of domain dependence [32] is the one of the most pressing challenges in fake review detection. One type of data, such as restaurant reviews models, tend to perform poorly while tested on a different domain (electronic, apparel [33]). Based on this observation, methods of domain adaptation have been developed with the goal of generalizing detection capabilities to contexts.

Some of the authors used adversarial training frameworks with feature representations that minimized domain-specific divergence. There have been others who applied transfer learning using shared encoder decoder structures, using which the information from different domains was jointly learned. Among such methods proposed by Ren et al. are multi-level feature fusion using GloVe embeddings, syntactic annotation, and document-level emotion classification using DistilBERT [34]. Without the need for retraining, the model achieved competitive accuracy in the restaurant and healthcare domains.

Although domain adaptation has made progress thus far, it is still a partially solved problem. Even many of the state-of-the-art models still rely on a lot of overlap between source and target domains or need partial supervision for the threshold calibration.

2.6 Limitations in the Literature

The key limitation for all existing approaches is content over reviewer intent. Semantically valid deceptive reviews, or any motivated by promotional contracts, incentivization schemes — or adversarial strategies, for example, [35]. Thus far few models have tried to infer why the review was written rather than how they were written.

Last but not least, reviewer persona modeling is another underdeveloped area where the realistic long term behavioral patterns are used to build latent personality profile. The existence of these profiles can also tell when there is deviation indicative of deception, such as a single user writing multiple reviews over time [36].

Finally, existing solutions are still not applicable since no cross-lingual models and multimodal datasets exist. Nearly all models are monolingual and trained on data from English languages, which means they are not very relevant for multilingual or cross-cultural situations.

Time Period	Approach	Models	Focus Area	Limitations / Notes
2007–2011	Rule-based	Pattern matching, filters	Repetition, duplicates	High false positives, no adaptability
2012–2015	Feature-based ML	SVM, RF, Naïve Bayes	Handcrafted features	Domain-specific, poor generalization
2016–2019	Deep learning	CNN, LSTM, BiLSTM	Semantic feature extraction	Opaque, data-intensive
2020–2023	Transformer models	BERT, RoBERTa, DistilBERT	Contextual embeddings	Costly, domain-limited
2024–2025	Intent & persona	Zero-shot, contrastive nets	Reviewer modeling, intent	Underexplored, lacks diverse datasets

Table 3. Timeline of methodological evolution in fake review detection (2007–2025)

3. TAXONOMY OF FAKE REVIEW DETECTION TECHNIQUES

Over several paradigms, the field of fake review detection has incrementally enriched its capabilities and offered unique and new problems. This section presents three principal axes along which the detection techniques can be structured into a taxonomy, namely: the type of computational model, the type of features used, and the degree of generalization between domains. There are again five main categories of methodological taxonomy, including rule-based systems, feature-engineered machine learning models, neural network-based deep learning models, transformer-based contextual models, and finally emerging intent-aware approaches [37]. The difference between each class is not only how it computes but also how it is interpretable, what types of domains it can be flexible, and whether it can be scaled.

3.1 Rule-Based Systems

Rule-based detection systems were the first class of fake review detection techniques. These systems have relied on a predefined set of heuristic rules that are developed based on observations of deceptive language patterns or behavioral anomalies. For instance, rules may be put in place, defining, for example, a maximum number of characters, the frequency of punctuation, the presence of aforementioned adjectives, as well as the frequency of self-referencing pronouns [38]. In other cases, passive voice or other non-standard grammatical structures can also be encoded in the detection criteria.

On the other hand, rule-based systems are easy to implement and interpretable, however with poor adaptability. Rule design is also selectively handcrafted, and as spam evolves, making rules brittle. Moreover, such systems are also incapable of discovering latent semantic evolutions or strategic manipulations not following known heuristics. For such reasons, they are usually not used as standalone solutions in modern systems.

3.2 Feature-Based Supervised Learning

Feature-based supervised learning is a critical milestone in fake review discovery. In contrast to rule-based systems which are based solely on manually created heuristics, these models are trained on labeled corpora and learn to classify reviews by statistical pattern discovery over the feature vectors. Typically, these vectors are constructed from attributes taken from four key domains: Textual content, reviewer behavior, transient exercise, and product area-related metrics [40].

Typically, textual features have lexical richness, sentiment polarity scores, part of speech frequency distributions, syntactic complexity measures, and vocabulary diversity. The reviewer-centric attributes are based on the number of reviews posted by a user, average review length, deviation of rating from the product average, and temporal distribution of how they are posted [41]. The temporal features capture burst activity or frequency change over time as well as intervals between consecutive reviews. Product-centric features are either the representation of an overall reputation of the item being reviewed, average sentiment divergence, or anomaly in aggregated ratings [42].

In this category simple classifiers including SVM, Naïve Bayes, Decision Trees, or Random Forests are often used. The training data is used to learn a decision boundary or ensemble decision space for these models, and this is then used to make decisions on the reviews that are new, or unseen. This methodology has its main strength in the possibility of being computationally cheap, whereby domain knowledge in the form of an interpreter can be incorporated into the detection process.

Still, to a large extent, the big limitation here is that feature extraction is handcrafted. The selection and design of the features are normally subjective and require domain knowledge as well as iterative optimization. For example, since reviews from one domain (e.g., restaurants) rarely generalize well to other domains (electronics, apparel, etc.), due to differences concerning language usage, customer expectations, and customer demographics, models trained on reviews from one domain will tend to perform badly on other domains [43]. Furthermore, the models' vulnerability to adversarial manipulation where the deceptive content is placed such that it evades all known feature thresholds but mimics real reviews in appearance is highly sensitive to minor stylistic changes.

3.3 Deep Learning Architectures

Although deep learning brought in a new paradigm towards fake review detection [44]. Whereas, in neural models, the raw review text is used directly to construct hierarchical representations without the need to manually create features. Since there is little or no long-range dependency in text, the most commonly used architectures are Convolutional Neural Networks (CNNs) to capture local n-gram dependencies and Recurrent Neural Networks (RNNs) e.g. Long Short Term Memory (LSTM) networks to capture long-range dependencies [45].

We made great progress using deep learning methods in terms of accuracy and robustness. LSTMs were good at capturing stylistic patterns over whole reviews whilst CNNs did effectively model short phrases and sentiment markers [46]. For these more complex architectures, we also saw performance increase with the use of Bidirectional Long Short-Term Memory (BiLSTMs) as well as attention-based mechanisms to give higher weights to semantically rich tokens [47].

However, deep models suffer from lack of interpretable and they need a large block of labeled data to train. In addition, their domain specific training is portable only if very well-tuned for a different dataset.

3.4 Transformer-Based Contextual Models

Transformer-based architectures like BERT (Bidirectional Encoder Representations from Transformers) [28], Robustly Optimized BERT Approach (RoBERTa) [28], DistilBERT[28] have redefined text classification tasks by robustly developed contextual embedding. These models are pre trained on large corpora using language modeling objectives, then they are fine tuned on downstream classification such as outputting fake review detection.

The key advantage of transformer models is their capacity to learn bidirectional context, polysemy, and to pay attention to relevant parts of the input using self attention mechanisms. On standard benchmarks like YelpCHI, Ott and Amazon [48] these models are able to achieve superior performance.

Nevertheless, transformer models also have high computational overhead and often are very hard to first train, involving the fine-tuning in the specific domain. This is because of the black box nature of their behavior in explainability, presenting a problem when they are used in regulatory or customer facing applications.

3.5 Intent-Aware and Persona-Based Modeling

We are at the start of the emergence of a new class of models that rethink fake review detection as an intent classification problem instead of a binary classification problem on text. These models reason about the intender behind the review, rather than the content of the review, such as if the intent is malicious, promotional, malign, or neutral feedback [7].

The reviewers' persona is formed by integrating reviewer signals into the history, sentiment trajectory, review style, and emotional variance. These can be thought of as latent profiles constituting reviewer consistency, credibility, and engagement style over time and products [49]. Such models allow for zero-shot detection where a reviewer's behavior in a new domain can be detected by highlighting misalignments between the current behavior of the reviewer and his/her historical intent profile [50].

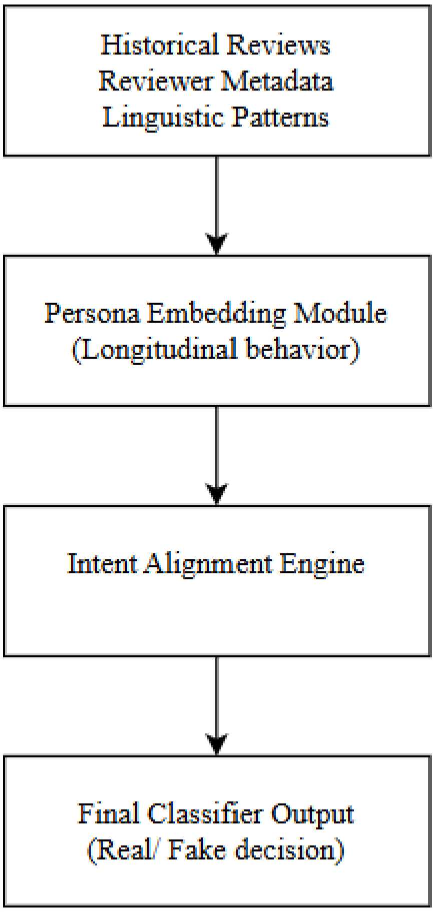

Figure 1. Conceptual architecture of persona-based fake review detection using intent alignment

The workflow of a reviewer persona based fake review detection model as illustrated in this architecture. Through its usage of historical behavioral and linguistic patterns, it profiles the reviewers, aligns the current review's latent intent with the persona model, and classifies the review as genuine or unauthentic based on this alignment.

Category	Interpretability	Cross-Domain Performance	Labeled Data Requirement	Current Maturity
Rule-Based	High	Low	Low	Legacy
Feature-Based ML	Medium	Low	Medium	Common
Deep Learning (CNN/LSTM)	Low	Medium	High	Mature
Transformer Models	Low	Medium to High	High	Cutting-Edge
Intent/Persona-Based Models	Medium to High	High	Low to Medium	Emerging

Table 3. Comparative overview of fake review detection methodologies

A taxonomy of the techniques is presented both as a view of the progression of techniques and the current gaps in the field. The traditional models are not adaptive, whereas the current models provide interpretability.

Performance gains come with opacity and limited transfferability from the deep models [51]. By choosing intent-driven paradigm, we demonstrate a conceptual shift towards modeling of deception as a cognitive process in contrast to a surface level artifact that would afford new breakthroughs in robustness and generality.

4. COMPARATIVE ANALYSIS OF APPROACHES

Since there exist many techniques in fake review detection, a comparative framework beyond accuracy metrics is required [7]. Building model effectiveness on benchmark datasets is an important indicator of the model's effectiveness but other factors like generalizability, computational complexity, interpretability, data dependence, and adversarial manipulation robustness are also important measures if this model will be deployed in the real world. This section provides quantitative benchmarks and qualitative characteristics of the major methodological classes introduced in the taxonomy.

4.1 Benchmark Accuracy Across Datasets

In all of the studies metrics used for evaluation include accuracy, precision, recall, F1-score, and AUC-ROC. Nevertheless, these metrics are often reported inconsistently about unnecessarily different datasets, data splits, and preprocessing methods. Results of the most commonly used benchmark datasets (Ott, YelpCHI, Amazon Reviews) are summarized in Table 4 and these scores are compared to common representative models across each methodological class.

Model Type	Model Name	Dataset	F1-Score (%)	Reference
Rule-Based	Heuristic Engine	YelpCHI	63.5	[52]
Feature-Based ML	SVM	Ott	82.1	[53]
Deep Learning	LSTM	YelpCHI	86.7	[54]
Transformer-Based	RoBERTa	Ott	90.8	[55]
Intent/Persona Modeling	ContrastiveNet	Cross-domain	88.2	[56]

Table 4. Reported F1-scores of representative models on standard fake review detection datasets

Results indicate that transformer-based models are generally more superior than earlier ones at F1 score. Nevertheless, our findings show that intent modeling and persona synthesis of the reviewer increase perfromance at the cross-domain generalization without great sacrifice in performance while relying less on superficial lexical patterns and more on the context of behavior.

4.2 Cross-Domain Generalizability

It is well known that many models, especially supervised learning-based models, heavily rely on coming from a domain with specific training data. Classifier trained on restaurant reviews often underperform when used on electronics or fashion reviews because style deviates, consumer expectation varies and the reviewer demographics change. Finally, Table 5 depicts the cross-domain performance of transfers among studies that report the performance on a specific domain explicitly.

Source Domain	Target Domain	Model Type	Performance Drop (%)	Reference
Yelp (food)	Amazon (tech)	SVM	18.3	[53]
Amazon (books)	TripAdvisor (hotels)	CNN	12.6	[54]
Yelp (hotels)	Doctor reviews	RoBERTa	9.1	[55]

Table 5. Cross-domain performance drop in fake review detection

With traditional models, when using new domains, performance severely degrades, there is no retraining required. On the other hand, intent aware models and contrastive alignment techniques have shown a much better transfer performance of abstraction over domain specific vocabulary and using intent distribution as opposed to text patterns.

4.3 Interpretability and Transparency

The factor of interpretability continues to be an important one for applications that require that fake review detection outputs be auditable, explainable, or legally justifiable. Rule-based systems are the one of most interpretable systems whose decisions are based on explicit heuristics. Second, feature-based models can also be transparent, if the features are understood and visualizing decision boundaries is possible [57]. Despite this, deep learning and transformer models are typically "black box" opaque, leaving it to post hoc explainability techniques like SHAP or LIME to discover which tokens or segments influence the model the most [58].

Introducing an intermediate level of interpretability is done using persona-based models. Although these internal embeddings are abstract, their alignment of current and historical intent is visualized, which is an intuitive justification for classification outcomes. For instance, if an author who has always been writing detailed, neutral reviews suddenly submits an unbalanced highly polarized piece, that too can be flagged for personal deviation [59].

4.4 Robustness to Adversarial Manipulation

With the coming of large language models that can spew out coherent, human-like text, fake review generators have gotten much more sophisticated. Paraphrasing or token substitution is trivial for bypassing rule-based and feature-based classifiers. Despite being very semantic sensitive, even transformer models can be fooled with style-preserving adversarial attacks [38]. Few studies have explicitly tested robustness concerning adversarial settings and even less with active defenses.

In this area, intent-based models promise a piece of good news. These can, however, detect deception even where the linguistic plausibility of the review text is taken care of because they are made up of behavioral patterns and historical context [7]. For example, suppose a reviewer began writing with a new style, in tonality that is different from a review-persona's past writing, or posted reviews with a much higher or lower review density compared to their past writing, the model can signal the review without referring to the superficial text content.

4.5 Data Efficiency and Annotation Constraints

The creation of large and error-free labeled datasets is highly expensive, and error-prone, and also requires the use of expensive human labelers. The problem of labeling deception is inherently subjective and all publicly available datasets that utilize deception, are either synthetically generated or derived from opaque platform-level filters [11]. Deep learning models require even bigger datasets for training, and this issue is amplified by it.

In particular, unsupervised or semi-supervised contrastive learning based data efficiency are achieved using persona based systems. Since they can be trained using fewer labeled examples and can still separate meaningful deceptive versus genuine profile distributions (by modeling distributions rather than absolute classes), they are more robust to small amounts of outliers and are more efficient in requiring data labeled only for intent. They are also useful in settings that lack labeled data or such data is unreliable.

Methodology	Accuracy	Generalization	Interpretability	Robustness	Data Efficiency
Rule-Based	Low	Poor	High	Very Low	High
Feature-Based ML	Medium	Poor	Medium	Low	Medium
Deep Learning (CNN/LSTM)	High	Medium	Low	Medium	Low
Transformer Models	Very High	Medium-High	Low	Medium	Low
Intent-Persona Modeling	High	High	Medium-High	High	Medium-High

Table 6. comparison of fake review detection approaches

The results of this assessment show the tradeoffs inherent in existing approaches. However, transformer models are more resource-intensive and harder to explain. Intent-aware frameworks have a balanced profile across all categories, which may provide the most balanced profile and future systems should incorporate all of the three methods (intent modeling, longitudinal reviewer analysis, and contrastive alignment) to attain robust, interpretable, and scalable detection.

5. DATASETS AND BENCHMARKING PRACTICES

The availability and quality of benchmark datasets play a critical role in the development of fake review detection systems. The datasets used for model training, evaluation, and comparison are used in these. However, building deception-oriented corpora is difficult for reasons of the lack of clear ground truth, the subjectivity of deception, and ethical constraints on user data. As a result, researchers have used a combination of controlled, synthetic, and weakly labeled datasets that, have their respective advantages and disadvantages [62]. This section provides a list of the most commonly used datasets in the literature and explains their architectural properties, compares such properties to the commonly used datasets in the literature, and analyzes the benchmarking common practices such as evaluation metrics and methodological gaps.

5.1 Publicly Available Datasets

Ott Deceptive Opinion Spam Corpus was one of the first and most structured datasets of fake review detection. Specifically, it contains 800 deceptive and 800 truthful hotel reviews, which used deceptive content populated through Amazon Mechanical Turk, and truthful reviews collected on TripAdvisor [1]. Being able to control it, and also balance it out to have classes this way, this dataset has been used so widely for foundational model training and testing. Nevertheless, the synthetic generation process of these deceptive reviews precludes ecological validity, since crowd-generated reviews may be devoid of the subtlety, and the real-world motivation observed in naturally occurring opinion spam.

YelpCHI is a dataset that is based on Yelp's internal filtering system, whereby reviews labeled as 'filtered' are considered to be deceptive, and those labeled 'recommended' are considered truthful. The variety of businesses included in this dataset includes restaurants, service providers, and retail facilities [53]. First, due to its size and diversity, it's superior in scale and diversity than smaller datasets; and second, it lives about reality through its foundation in actual platform behavior. However, though the labeling criteria are proprietary and opaque, the validity of ground truth is still an issue of concern. Furthermore, the dataset does not include any detailed behavioral metadata of reviewers which reduces its application to modeling by behavior and persona.

The second very large resource is the Amazon Review Corpus, which covers hundreds of millions of product reviews across various categories, e.g. electronics, clothes, books, and house things. However, Amazon does not itself annotate fake reviews—researchers instead take heuristic labeling strategies like detecting nonverified purchase reviews, extreme rating deviations, or too high review frequency as proxies of deception [54]. The large-scale experiments are realized with this weak supervision and it is applicable for domain adaptation studies. In addition, the metadata richness of the dataset with timestamps, reviewer IDs, and product hierarchies is perfect for time, behavior, and multi-modal research. However, the use of heuristics leads to label noise and the scale of the dataset necessitates a large amount of preprocessing and resource requirements.

Widely used corpora include YelpZip and YelpNYC which are region-specific reviews from Yelp. These datasets have an identical labeling methodology to YelpCHI but make possible spatial and regional analysis [54]. These corpora were used to study geo-linguistic variation in spam patterns and to explore the effect of location on review trustworthiness. Furthermore, several less formally standardized datasets are extracted from TripAdvisor, Expedia, and medical review plate forms [55]. Rich domain-specific content like doctor or hospital reviews are offered by these sources and they are widely used in transfer learning and cross-domain generalization setups. However, they are rarely publicly accessible and are annotated in terms of internal or researcher-defined criteria.

5.2 Dataset Properties and Comparison

Table 7 presents a comparative summary of the most influential datasets. This table compares datasets based on their size, domain, labeling type, metadata availability, and public accessibility. It points out the inherent tradeoff between realism and control. Ott (and related datasets like FB15k) are datasets with high control but low realism, Amazon and/or Yelp are datasets with more general domain coverage but have ambiguity in label fidelity that comes from heuristic or proprietary label mechanisms.

Dataset Name	Size	Domain	Labeling Type	Metadata Included	Publicly Available
Ott [62]	1600	Hotels	Crowdsourced	No	Yes
YelpCHI [63]	>60,000	Mixed (Yelp)	Platform-filtered	Limited	Yes
Amazon Reviews [64]	>1M	Multi-product	Heuristic	Yes	Partially
YelpZip/YelpNYC [65]	~50,000	Food/Retail	Filtered	Yes	Yes
TripAdvisor/Expedia [66]	Varies	Hotels	Mixed	Partial	Partially

Table 7. Summary of key datasets used in fake review detection

5.3 Evaluation Metrics

Fake review detection models are usually tested using a set of binary classification metrics including accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC-ROC). Accuracy is a measure of the percentage of correctly classified instances, and this can mislead in imbalance problems where the fraction of one class is much larger than another. Recall measures how well the model can predict all the fake reviews, and precision indicates the fraction of fake reviews correctly predicted from all the fake reviews regarded as fake by the model [11]. Especially in the case of skewed class distributions, the F1 score as the harmonic mean of precision and recall is a balanced indicator. AUC-ROC is a threshold-independent measure of a model's ability to separate classes and is useful for comparing probabilistic models.

New metrics, emerging research, have been proposed based on intent based and persona aware systems. For example, intent divergence scores represent the deviation of a review's latent semantics from the behavior of a reviewer. However, these metrics are not a standardized as yet and are largely experimental. On the other hand, their integration into mainstream evaluation frameworks may provide further insights on, for example, the model's robustness and deception nuances.

5.4 Benchmarking Challenges

The problem of benchmarking in fake review detection has substantial inconsistencies. First, due to many often markedly different preprocessing steps, dataset splits and filtering techniques used by researchers, reported results cannot be compared [32]. This is because many studies don't share specific data partitions nor publish any code for replication. Furthermore, almost every evaluation is conducted within the same domain as their training data and loses the inquiry of the real world, i.e., cross-domain applicability. To our knowledge, few of the studies assess model stability in adversarial attacks, low resource settings, and multilingual review corpora. Omissions of this point mask the generalizability and robustness of proposed models [67].

Furthermore, some studies employ derived or rebalanced versions of existing corpora, which may produce artificially high results on performance metrics by removing edge cases or simplifying the classification boundaries [11]. As methodological innovations and dataset artifacts are hard to distinguish without constantly shared experimental setups and consistent baselines, it is difficult to discern what causes improvements to be observed.

5.5 Toward Benchmarking Standards

The need for benchmarking in the domain of fake review detection is becoming one of the standards. In future research, optimized unified benchmark corpora should include several domains, languages and labeling strategies validated by humans or semi-automated methods [48]. The preprocessing routines, the cross-validation method, and the report formats about the performance should be clearly defined.

In addition, it would also be beneficial for the community to have domain generalizable, zero-shot, and robust to adversarial tasks in dedicated challenge datasets. By making leaderboards publicly available and baselines shared, reproducibility and comparative fairness would be incentivized. There is also a need to report auxiliary metrics — training time, memory consumption, and inference latency— in particular for models that are going to be deployed in real-time environments [51]. The field can move towards more ferrite, generalizable, and practically deployable solutions to fake review detection through the alignment of evaluation practices to real-world conditions.

6. CHALLENGES AND OPEN ISSUES

Although there have been numerous advances in fake review detection, several persistent challenges still prevent the development of generalizable, interpretable, and generalizable systems [32]. Conceptual ambiguity, data limitation, architectural problem, and some deployment related problems are these challenges. Specifically, in this section these issues are explored in depth and with particular focus given to the structural tensions that frame ongoing research efforts.

6.1 Lack of Ground Truth and Annotation Ambiguity

In the same spirit, there is no question that the task of labeling deceptive content is inherently subjective and epistemic uncertain. Fake reviews are unlike factual misinformation, where claims can readily be tested (verified or nullified), as they are based on opinions that may not be at all easily confirmed or falsified [68]. Malicious intent is review writing—that is, a review can be written such that, so long as it appears linguistically truthful, review writing may be written with the intent of writing a review and with the intent of retaining a negative opinion of the entity under review [7]. Thus, human annotators disagree on labels as often as not and platforms compensate for that with automated heuristics or crowd-sourced proxies to infer deception. The disadvantage of these methods is that they provide different noises and biases in benchmark datasets, which in turn violate the training and evaluation of the model.

Involving indirect labeling mechanisms of many widely used datasets, such as YelpCHI and Amazon, the labeling can be either malicious or incomplete in many cases [63], [64]. Now this hides the line that separates right from wrong behavior and asks whether the models we have today are really learning intent and are not just memorizing platform-specific filtering logic. A foundational impediment in the field has been the absence of authoritative, large-scale, ethically sourced ground truth datasets.

6.2 Domain Sensitivity and Generalization Gaps

Poor generalization of models across domains is a recurring limitation of studies. Many systems trained on reviews of restaurants or hotels fare poorly when deployed in domains that are different from electronics, healthcare or digital services [69]. The reason for this combines with the lexical divergence and change of user behavior, sentiment expression and review intent.

Adversarial learning and fine-tuning have been used to attempt domain sensitivity, however, they have achieved only partial success and come along with overfitting to an intermediate representation and dependency on auxiliary domain labels [70]. Such a lack of domain agnostic modeling frameworks makes fake review detection systems hard to scale especially in multiple environment, for example, in multilingual, and across platforms [71].

However, zero-shot detection frameworks, although conceptually promising, are in the early stages of implementation and evaluation. First, the potential of their ability to recognize deception without domain-specific supervision has been under-explored, especially in low-resource language and heterogeneous content [21].

6.3 Overreliance on Textual Features

To date, most existing models heavily depend on the linguistic content from the review itself. In that case, model robustness suffers from the textual dependence. LLMs that generate sophisticated fake reviews can mimic the natural syntax, the coherent structure, and the plausible sentiment flows with great fidelity. In such cases, shallow or deep based content based classifiers are becoming increasingly ineffective [72].

Also, models which overlook context relating to temporal and behavioral events and relational dependencies favour missed cues such as reviewer rating burstiness, recurring submission pattern or odd product category interactions [73]. However, only a few studies leverage behavioral metadata but there is no unified framework for the multi-modal feature fusion that leverages text, time and user history in a meaningful way in a single pipeline.

To address this, the emerging field of persona based modeling compares reviews to latent representations of a person's historical behavior. Nevertheless, we lack the ability to capture reviewer consistency, scale to users, and maintain privacy preserving computation [74]. Without these components, deception detection systems will always be re active not a nticipatory.

6.4 Lack of Interpretability in Deep Models

With the state of the art on most of the benchmark tasks, deep learning + transformer architectures have sacrificed transparency. These are black-box systems and have limited insight into how predictions are generated for them [75]. For applications of high stakes (e.g. healthcare services, financial platforms, or consumer protection agencies), opacity of the model is not leveragable unless is complemented by post-hoc explainability techniques.

Further methods like attention visualization, feature importance maps, and SHAP (Shapley Additive exPlanations) values have been used, but the interpretability is usually superficially high [76]. These explanations are also lacking with theoretical groundings thereby making it difficult to audit or regulate them. Furthermore, explainable AI is becoming a necessity for the algorithms seen in public space due to regulatory frameworks, yet at the same time usually interpretability is sacrificed for predictive performance.

However, intent aware systems, by virtue of behavioral modeling build interesting avenues to interpretability. Such systems make intuition a justification by allowing us to justify flagging a review as suspicious by discovering semantic deviation of a review from prior reviewer intent. However, there is a tremendous opportunity to develop and standardize interpretability metrics and visualization frameworks for these approaches.

6.5 Inadequate Adversarial Robustness

Adversarial techniques, such as automated review generators, paraphrasing engines, and style transfer models, are growing and constitute a new threat to fake review detection. However, most of the current literature evaluates most systems on static datasets that presume stationary linguistic behaviors and the same systems in their undetermined states [77]. Nevertheless, spam agents can produce adaptive content that takes advantage of model blind spots.

In the literature, adversarial testing is still quite rare. There are a few models that are evaluated under threat models that simulate real attack vectors like character substitution, sentiment flipping, or deliberate ambiguity injection [78]. In addition, while transformer models are more adversarial robust than traditional classifier models, the adversarial robustness of transformer models persists in corrupted attack scenarios such as paraphrased or templated attacks. There remains considerable unexplored work, including developing adversarially robust architectures, training strategies in the presence of synthetic noise injection, and counterfactual data augmentation. However, models need to be tested not only on how well they classify but how robust they are to deception designed with a full understanding of the model's decision boundaries [39].

6.6 Ethical and Privacy Considerations

Fake review detection usually involves a behavioral metadata analysis (IP addresses, device fingerprints, and review history among other things). However, this increased data can increase detection accuracy but with a high ethical and legal safety issue regarding user privacy [7]. For example, techniques like Persona modeling or behavioral clustering can violate the dignity of a user by profiling of user based on protected characteristics or subject the individual to undue scrutiny.

Automated deception detection is ignored by much current research on deception. It discusses little, for instance, of informed consent, data retention policies or there might be exploitation of algorithmic discrimination. In addition, platform-provided labels as ground truth without the use of transparency to their users may not align with the principles of algorithmic fairness [79]. The future systems must incorporate privacy-preserving techniques like federated learning, differential privacy, or secure multiparty computation to ensure ethical integrity. It is also essential to subject detection outputs to human review especially in high-stakes settings to avoid making incorrect decisions based on incomplete data [56].

7. FUTURE RESEARCH DIRECTIONS

The advanced sophistication of deceptive opinion generation techniques leads the future fake review detection to the proactive, context-aware, behaviorally grounded models. This section identifies some interesting research directions that might help overcome its limitations and proposition new foundations for the robust and robustly adaptable detection systems.

One way of exploring is the formal integration of intent modeling into the architecture of detection. Instead of just figuring out what are the linguistic anomalies or outlier tokens, the future should judge the reviewed behavior of a reviewer concerning their current expression in terms of semantic and emotional alignment. Capturing subtle tests of tone, sentiment, or rhetorical framing that might signal strategic deception is possible when reviews are embedded in the context of a reviewer's behavior. One way to operationalize this concept is to develop intent divergence scores calibrated using contrastive learning or probabilistic embedding techniques.

More related is the advancement of the synthesis of the reviewer persona by constructing latent representations of reviewer behavior over time. The personas can then be used as reference points to check new reviews to see if they deviate from an established communication pattern. This can support both individualized and group-level analysis to enable systems to detect coordinated fake review campaigns. Future datasets have to capture longitudinal user histories, temporal posting patterns, and cross-product review behavior, masquerading it in anonymized or federated storage to enable such methods.

The development of zero-shot and few-shot learning frameworks is another important area of research. As there are a variety of platforms, product categories, and targeted demographics, it is neither practicable nor scalable to retrain the models on such domains. The models must be able to detect deception in completely new environments without fine-tuning to be deployed in dynamic ecosystems. This direction can potentially be achieved through domain invariant embedding spaces, adversarial alignment, and pre-trained behavioral encoders which need to be empirically evaluated and benchmarked.

It is also important to achieve improved adversarial robustness. With such large language models becoming increasingly accessible, it will become trivial to produce highly plausible fake reviews. This should be anticipated by future detection systems that include adversarial training of their architectures, generative, model of noise, and style transfer resistance. It should become a standard part of the evaluation protocol to simulate attacks while developing the model.

There should finally be more research on ethical and regulatory dimensions – such as explainability, fairness, and data governance – as this is the direction, I think, of what will happen shortly. Detection decisions may entail reputational or legal consequences for users and businesses alike. Hence, models require the transparency of their justifications to adhere to frameworks such as respecting data privacy, avoiding discrimination, and having human oversight. To have trust and accountability, explainable AI principles will need to be integrated with regulatory alignment and auditability.

Taken together, these future directions constitute a move into detection systems that are more accurate, but much more in sync with the realities of practical, ethical, and technical constraints of the environments in which they appear.

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

8. CONCLUSION

While digital platforms and online consumer decision-making will continue to see the persistence and evolution of fake reviews as a significant challenge to their credibility, it is also significant in that it can no longer be limited to a small number of sites. Despite the significant gains in deploying the linguistic, statistical and learning based models for the task of detecting deceptive reviews over the years, existing approaches have not successfully generalized across domains, are prone to adversarial manipulation, and are not interpretable in high stakes applications.

The presented paper provides extensive analysis and background to the journey of fake review detection techniques from basic rules to feature-engineered classifiers followed by supervised learning processes and finally to the modern transformer-based language method and more recently the Intent Aware methods that are trending. Existing approaches were organized according to assumptions about methodological assumptions, data requirements, and operational characteristics by introducing a structured taxonomy. Comparative analysis suggested trade-offs between accuracy, generalizability, transparency, and robustness and highlighted the need for more holistic models suited for user-generated content that can adapt to the dynamism of deception.

In particular, a particular focus was to learn from the conceptual innovation in composing review persona modeling and intent-based detection architectures. They seek to go beyond the surface textual analysis and move to include longitudinal behavioral cues, emotional consistency, and patterns of the reviewer's identity. In conjunction with techniques such as zero-shot learning and contrastive embedding alignment such models may be well suited to perform well across domains without much retraining.

Benchmark datasets and evaluation protocols were discussed by research practices existing in current research; they pointed out the fragmentation in the research and the need for standardization. A large number of models are tested in domain-specific, narrow settings that fail to reproduce real-world variability in concepts mentioned in reviews, platform structure, or culture. There remains a lack of publicly available diverse, and ethically curated datasets that inhibit the development of scalable and inclusive solutions.

The field has to look forward, where thinking on an interdisciplinary research line that includes natural language processing, user analysis, adversarial learning, and ethical AI design must be done. Future models shouldn’t just be accurate (in terms of detecting deception) but must also explain their reasoning and be private for the user. Also, they need to be vaguely legal as society evolves with evolving regulatory frameworks. Future systems can help restore trust to such ecosystems that depend on online reviews by deliberately focusing on generalizability, interpretability, and robustness.

REFERENCES

[1] Pramiarsih, E. E. (2024). Consumer behavior in the digital era. *INTERNATIONAL JOURNAL OF FINANCIAL ECONOMICS*, *1*(3), 662-674.

[2] George, R., Stainton, H., & Adu-Ampong, E. (2021). Word-of-mouth redefined: A profile of influencers in the travel and tourism industry. *Journal of Smart Tourism*, *1*(3), 31-44. DOI: https://doi.org/10.52255/smarttourism.2021.1.3.6

[3] Kim, M., & Kim, J. (2020). The influence of authenticity of online reviews on trust formation among travelers. *Journal of Travel Research*, *59*(5), 763-776. https://doi.org/10.1177/0047287519868307

[4] Huang, N., Burtch, G., Gu, B., Hong, Y., Liang, C., Wang, K., ... & Yang, B. (2019). Motivating user-generated content with performance feedback: Evidence from randomized field experiments. *Management Science*, *65*(1), 327-345. https://doi.org/10.1287/mnsc.2017.2944

[5] Petrescu, M., Kitchen, P., Dobre, C., Ben Mrad, S., Milovan-Ciuta, A., Goldring, D., & Fiedler, A. (2022). Innocent until proven guilty: suspicion of deception in online reviews. *European Journal of Marketing*, *56*(4), 1184-1209. https://doi.org/10.1108/EJM-10-2019-0776

[6] Vasist, P. N., & Krishnan, S. (2022). Demystifying fake news in the hospitality industry: A systematic literature review, framework, and an agenda for future research. *International Journal of Hospitality Management*, *106*, 103277. https://doi.org/10.1016/j.ijhm.2022.103277

[7] Poojary, K. K. (2024). *Deciphering Deception-Detecting Fake Review using NLP by analysis of stylistic, sentiment-based, and semantic features* (Doctoral dissertation, Dublin Business School).

[8] Abdulqader, M., Namoun, A., & Alsaawy, Y. (2022). Fake online reviews: A unified detection model using deception theories. *IEEE Access*, *10*, 128622-128655. DOI 10.1109/ACCESS.2022.3227631

[9] Crothers, E. N., Japkowicz, N., & Viktor, H. L. (2023). Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*, *11*, 70977-71002.

[10] Li, D., Li, M., Han, G., & Li, T. (2021). A combined deep learning method for internet car evaluation. *Neural Computing and Applications*, *33*, 4623-4637.

[11] 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. *Computers, Materials & Continua*, *70*(2), 3189-3204. DOI:10.32604/cmc.2022.019625

[12] Singhal, P., Walambe, R., Ramanna, S., & Kotecha, K. (2023). Domain adaptation: challenges, methods, datasets, and applications. *IEEE access*, *11*, 6973-7020. doi: 10.1109/ACCESS.2023.3237025

[13] Eang, C., & Lee, S. (2024). Improving the Accuracy and Effectiveness of Text Classification Based on the Integration of the Bert Model and a Recurrent Neural Network (RNN_Bert_Based). *Applied Sciences*, *14*(18), 8388. https://doi.org/10.3390/app14188388

[14] Talaei Khoei, T., Ould Slimane, H., & Kaabouch, N. (2023). Deep learning: systematic review, models, challenges, and research directions. *Neural Computing and Applications*, *35*(31), 23103-23124. https://doi.org/10.1007/s00521-023-08957-4

[15] Goyal, M., & Mahmoud, Q. H. (2024). A systematic review of synthetic data generation techniques using generative AI. *Electronics*, *13*(17), 3509. https://doi.org/10.3390/electronics13173509

[16] Truong, V. (2024). Textual emotion detection–A systematic literature review. https://doi.org/10.21203/rs.3.rs-4673385/v1

[17] Guerra, J. L., Catania, C., & Veas, E. (2022). Datasets are not enough: Challenges in labeling network traffic. *Computers & Security*, *120*, 102810. https://doi.org/10.1016/j.cose.2022.102810

[18] Harris, C. G. (2024, October). Exploring Transformer Models and Domain Adaptation for Detecting Opinion Spam in Reviews. In *2024 36th Conference of Open Innovations Association (FRUCT)* (pp. 249-255). IEEE. doi: 10.23919/FRUCT64283.2024.10749897.

[19] Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., ... & He, L. (2022). A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *13*(2), 1-41. https://doi.org/10.1145/3495162

[20] Hou, J., Tan, Z., Zhang, S., Hu, Q., & Wang, P. (2025). Detecting fake review intentions in the review context: A multimodal deep learning approach. *Electronic Commerce Research and Applications*, 101485. https://doi.org/10.1016/j.elerap.2025.101485

[21] Baashirah, R. (2024). Zero-Shot Automated Detection of Fake News: An Innovative Approach (ZS-FND). *IEEE Access*. doi: 10.1109/ACCESS.2024.3462151

[22] Chen, W., Liang, Y., Zhu, Y., Chang, Y., Luo, K., Wen, H., ... & Zheng, Y. (2024). Deep learning for trajectory data management and mining: A survey and beyond. *arXiv preprint arXiv:2403.14151*.

[23] Vidanagama, D. U., Silva, A. T. P., & Karunananda, A. S. (2022). Ontology based sentiment analysis for fake review detection. *Expert Systems with Applications*, *206*, 117869. https://doi.org/10.1016/j.eswa.2022.117869

[24] Gupta, R., Jindal, V., & Kashyap, I. (2024). Recent state-of-the-art of fake review detection: a comprehensive review. *The Knowledge Engineering Review*, *39*, e8. DOI: https://doi.org/10.1017/S0269888924000067

[25] Garner, B., & Kim, D. (2022). Analyzing user-generated content to improve customer satisfaction at local wine tourism destinations: an analysis of Yelp and TripAdvisor reviews. *Consumer behavior in tourism and hospitality*, *17*(4), 413-435.

[26] Liang, H., Sun, X., Sun, Y., & Gao, Y. (2017). Text feature extraction based on deep learning: a review. *EURASIP journal on wireless communications and networking*, *2017*, 1-12. https://doi.org/10.1186/s13638-017-0993-1

[27] Bahad, P., Saxena, P., & Kamal, R. (2019). Fake news detection using bi-directional LSTM-recurrent neural network. *Procedia Computer Science*, *165*, 74-82. https://doi.org/10.1016/j.procs.2020.01.072

[28] Krishnan, A. (2023). Exploring Machine Learning and Transformer-based Approaches for Deceptive Text Classification: A Comparative Analysis. *arXiv preprint arXiv:2308.05476*.

[29] Jiang, M., Cui, P., & Faloutsos, C. (2016). Suspicious behavior detection: Current trends and future directions. *IEEE intelligent systems*, *31*(1), 31-39. doi: 10.1109/MIS.2016.5

[30] Manaskasemsak, B., Tantisuwankul, J., & Rungsawang, A. (2023). Fake review and reviewer detection through behavioral graph partitioning integrating deep neural network. *Neural Computing and Applications*, 1-14.

[31] Ferrara, E. (2023). Social bot detection in the age of ChatGPT: Challenges and opportunities. *First Monday*. DOI: https://doi.org/10.5210/fm.v28i6.13185

[32] Duma, R. A., Niu, Z., Nyamawe, A. S., Tchaye-Kondi, J., Jingili, N., Yusuf, A. A., & Deve, A. F. (2024). Fake review detection techniques, issues, and future research directions: a literature review. *Knowledge and Information Systems*, *66*(9), 5071-5112. DOI: https://doi.org/10.1007/s10115-024-02118-2

[33] Khan, M. U., Javed, A. R., Ihsan, M., & Tariq, U. (2023). A novel category detection of social media reviews in the restaurant industry. *Multimedia Systems*, *29*(3), 1825-1838. DOI: https://doi.org/10.1007/s00530-020-00704-2

[34] Ren, G., Wang, H., & Yang, Y. (2025). Cross-domain Fake Review Detection Based on Deep Learning MultiLevel Generic Features Extraction Fusion. *Informatica*, *49*(18). https://doi.org/10.1007/s00530-020-00704-2

[35] Lim, W. M., Agarwal, R., Mishra, A., & Mehrotra, A. (2024). The Rise of Fake Reviews: Toward a Marketing-Oriented Framework for Understanding Fake Reviews. *Australasian Marketing Journal*, 14413582241283505. https://doi.org/10.1177/14413582241283505

[36] Abdulqader, M., Namoun, A., & Alsaawy, Y. (2022). Fake online reviews: A unified detection model using deception theories. *IEEE Access*, *10*, 128622-128655. doi: 10.1109/ACCESS.2022.3227631.

[37] Sarker, I. H. (2021). Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN computer science*, *2*(6), 1-20. DOI: https://doi.org/10.1007/s42979-021-00815-1

[38] Alsubari, S. N., Deshmukh, S. N., Aldhyani, T. H., Al Nefaie, A. H., & Alrasheedi, M. (2023). Rule-based classifiers for identifying fake reviews in e-commerce: a deep learning system. In *Fuzzy, Rough and Intuitionistic Fuzzy Set Approaches for Data Handling: Theory and Applications* (pp. 257-276). Singapore: Springer Nature Singapore. DOI: https://doi.org/10.1007/978-981-19-8566-9_14

[39] Islam, E., Moon, M. R., Vasha, T. K., & Mahdi, M. T. (2023). *Unmasking Deception: Analyzing Fake Product Reviews through Machine and Deep Learning* (Doctoral dissertation, East West University).

[40] Fontanarava, J., Pasi, G., & Viviani, M. (2017, October). Feature analysis for fake review detection through supervised classification. In *2017 IEEE international conference on data science and advanced Analytics (DSAA)* (pp. 658-666). IEEE. doi: 10.1109/DSAA.2017.51.

[41] 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. https://doi.org/10.1016/j.dss.2021.113728

[42] Paul, H., & Nikolaev, A. (2021). Fake review detection on online E-commerce platforms: a systematic literature review. *Data Mining and Knowledge Discovery*, *35*(5), 1830-1881. DOI: https://doi.org/10.1007/s10618-021-00772-6

[43] Ho, B., Mayberry, T. R., Nguyen, K. L., Dhulipala, M., & Pallipuram, V. K. (2024). ChatReview: A ChatGPT-enabled natural language processing framework to study domain-specific user reviews. *Machine Learning with Applications*, *15*, 100522. https://doi.org/10.1016/j.mlwa.2023.100522

[44] Islam, M. R., Liu, S., Wang, X., & Xu, G. (2020). Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, *10*(1), 82. DOI: https://doi.org/10.1007/s13278-020-00696-x

[45] Chandra, N., Ahuja, L., Khatri, S. K., & Monga, H. (2021). Utilizing gated recurrent units to retain long term dependencies with recurrent neural network in text classification. *J. Inf. Syst. Telecommun*, *2*, 89.

[46] Rehman, A. U., Malik, A. K., Raza, B., & Ali, W. (2019). A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis. *Multimedia Tools and Applications*, *78*, 26597-26613. DOI: https://doi.org/10.1007/s11042-019-07788-7

[47] Chen, J., Zhang, T., Yan, Z., Zheng, Z., Zhang, W., & Zhang, J. (2025). Attention-based BiLSTM with positional embeddings for fake review detection. *Journal of Big Data*, *12*(1), 83. DOI: https://doi.org/10.1186/s40537-025-01130-9

[48] Ennaouri, M., & Zellou, A. (2023). Machine learning approaches for fake reviews detection: A systematic literature review. *Journal of Web Engineering*, *22*(5), 821-848. doi: 10.13052/jwe1540-9589.2254.

[49] Hu, S., Kumar, A., Al-Turjman, F., Gupta, S., & Seth, S. (2020). Reviewer credibility and sentiment analysis based user profile modelling for online product recommendation. *Ieee Access*, *8*, 26172-26189. doi: 10.1109/ACCESS.2020.2971087

[50] Tan, Z., & Jiang, M. (2023). User modeling in the era of large language models: Current research and future directions. *arXiv preprint arXiv:2312.11518*.

[51] Rohera, D., Shethna, H., Patel, K., Thakker, U., Tanwar, S., Gupta, R., ... & Sharma, R. (2022). A taxonomy of fake news classification techniques: Survey and implementation aspects. *IEEE Access*, *10*, 30367-30394. doi: 10.1109/ACCESS.2022.3159651.

[52] Joshi, H. (2024). Transformer-Based Language Deep Learning Detection of Fake Reviews on Online Products. *J. Electrical Systems*, *20*(3), 2368-2378.

[53] Deshai, N., & Rao, B. B. (2022). A detection of unfairness online reviews using deep learning. *J Theor Appl Inf Technol*, *100*(13), 4738-4779.

[54] Alsaad, M. M. B., & Joshi, H. (2024). Combination between Deep Learning and Transformer Models to Detect Fake Yelp Electronic Product Reviews. *Journal of Computational Analysis and Applications*, *33*(7).

[54] Mohawesh, R., Salameh, H. B., Jararweh, Y., Alkhalaileh, M., & Maqsood, S. (2024). Fake review detection using transformer-based enhanced LSTM and RoBERTa. *International Journal of Cognitive Computing in Engineering*, *5*, 250-258 https://doi.org/10.1016/j.ijcce.2024.06.001

[55] Nawara, D., Aly, A., & Kashef, R. (2024). Shilling attacks and fake reviews injection: Principles, models, and datasets. *IEEE Transactions on Computational Social Systems*. doi: 10.1109/TCSS.2024.3465008

[56] 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. doi: 10.1109/ACCESS.2021.3075573

[57] Stoffel, F. (2018). Transparency in Interactive Feature-based Machine Learning: Challenges and Solutions.

[58] Mohammadi, H., Bagheri, A., Giachanou, A., & Oberski, D. L. (2025). Explainability in Practice: A Survey of Explainable NLP Across Various Domains. *arXiv preprint arXiv:2502.00837*.

[58] Das, R., Ahmed, W., Sharma, K., Hardey, M., Dwivedi, Y. K., Zhang, Z., ... & Filieri, R. (2024). Towards the development of an explainable e-commerce fake review index: An attribute analytics approach. *European Journal of Operational Research*, *317*(2), 382-400. https://doi.org/10.1016/j.ejor.2024.03.008

[59] Valiaiev, D. (2024). Detection of machine-generated text: Literature survey. *arXiv preprint arXiv:2402.01642*.

[60] Mckenzie, G. (2024). *Hiding in Plain Site: A Turing Test on Fake Persona Spotting* (Doctoral dissertation, Lancaster University (United Kingdom)).

[61] Rout, J. K., Singh, S., Jena, S. K., & Bakshi, S. (2017). Deceptive review detection using labeled and unlabeled data. *Multimedia Tools and Applications*, *76*, 3187-3211. DOI: https://doi.org/10.1007/s11042-016-3819-y

[62] Ott, M., Cardie, C., & Hancock, J. T. (2013, June). Negative deceptive opinion spam. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: human language technologies* (pp. 497-501).

[63] Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. (2013). What yelp fake review filter might be doing?. In *Proceedings of the international AAAI conference on web and social media* (Vol. 7, No. 1, pp. 409-418).

DOI: https://doi.org/10.1609/icwsm.v7i1.14389

[64] Jindal, N., & Liu, B. (2008, February). Opinion spam and analysis. In *Proceedings of the 2008 international conference on web search and data mining* (pp. 219-230). https://doi.org/10.1145/1341531.1341560

[65] Rayana, S., & Akoglu, L. (2015, August). Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th acm sigkdd international conference on knowledge dis* https://doi.org/10.1145/2783258.2783370

[66] Li, F., Huang, M., Yang, Y., & Zhu, X. (2011, July). Learning to identify review spam. In *IJCAI proceedings-international joint conference on artificial intelligence* (Vol. 22, No. 3, p. 2488).

[67] Rananga, S., Isong, B., Modupe, A., & Marivate, V. (2024). Misinformation Detection: A Review for High and Low-Resource Languages. *Journal of Information Systems and Informatics*, *6*(4), 2892-2922. DOI: 10.51519/journalisi. v6i4.931

[68] Kumar, S., & Shah, N. (2018). False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559*.

[69] Luo, Y., & Xu, X. (2021). Comparative study of deep learning models for analyzing online restaurant reviews in the era of the COVID-19 pandemic. *International Journal of Hospitality Management*, *94*, 102849. https://doi.org/10.1016/j.ijhm.2020.102849

[70] Kininis, P. (2025). Robustness and domain generalization in computer vision by using adversarial data augmentation.

[71] Boumber, D. A., Qachfar, F. Z., & Verma, R. (2024, May). Domain-agnostic adapter architecture for deception detection: Extensive evaluations with the difraud benchmark. In *Proceedings of The 2024 Joint International Conference On Computational Linguistics, Language Resources And Evaluation (LREC-COLING 2024)* (pp. 5260-5274).

[72] Kwon, S., & Jang, B. (2025). A Comprehensive Survey of Fake Text Detection on Misinformation and LM-Generated Texts. *IEEE Access*. doi: 10.1109/ACCESS.2025.3538805.

[73] Santosh, K. C. (2018). *Anomalous Behavior Analysis in Social Networks and Consumer Review Websites* (Doctoral dissertation, University of Houston).

[74] Salminen, J., Guan, K., Jung, S. G., & Jansen, B. J. (2021). A survey of 15 years of data-driven persona development. *International Journal of Human–Computer Interaction*, *37*(18), 1685-1708. https://doi.org/10.1080/10447318.2021.1908670

[75] Zaheer, H., & Bashir, M. (2024). Detecting fake news for COVID-19 using deep learning: a review. *Multimedia Tools and Applications*, *83*(30), 74469-74502. DOI: https://doi.org/10.1007/s11042-024-18564-7

[76] Mohawesh, R., Xu, S., Springer, M., Jararweh, Y., Al-Hawawreh, M., & Maqsood, S. (2023). An explainable ensemble of multi-view deep learning model for fake review detection. *Journal of King Saud University-Computer and Information Sciences*, *35*(8), 101644. https://doi.org/10.1016/j.jksuci.2023.101644

[77] Kassem, A. (2023). *Mitigating the shortcomings of language models: Strategies for handling memorization & adversarial attacks* (Master's thesis, University of Windsor (Canada)).

[78] Chiang, H. Y., Chen, Y. S., Song, Y. Z., Shuai, H. H., & Chang, J. S. (2023, August). Shilling black-box review-based recommender systems through fake review generation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 286-297). https://doi.org/10.1145/3580305.3599502

[79] Koltay, A. (2021). The protection of freedom of expression from social media platforms. *Mercer L. Rev.*, *73*, 523.