ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 6, June 2026

Fake Review Detection in E-Commerce: A Critical Review of Robustness, Generalization, and Deployment Challenges

[1] Satwik Pratap Singh*, [2] Smriti Asthana, [3] Vartika Gautam, [4] Pratham Tyagi, [5] Shiv Shankar Yadav

[1] [2] [3] [4] [5] Department of Computer Science and Engineering, Galgotias College of Engineering and Technology, India
*Corresponding Author Email: [1] satwiksingh200@gmail.com

Abstract*— Online reviews play a crucial role as they influence the buying choices of customers in digital platforms. However, there has been a steep rise in fake and misleading reviews which reduces the reliability of such systems, leading to work on developing automated methods to detect them. Over the last ten years, this field of fake review detection has progressively evolved from simple linguistic and psycholinguistic features to behavioral analysis, network-based models, deep learning and the recent transformer-based techniques. Despite these advancements have perceived high accuracy on benchmark datasets, there are still limitations that holds back from real-world deployment. In particular, the existing models shows difficulty in adapting to changes over time (limited robustness), poor cross-domain generalization and in integrating semantic (text-based) and relational (network-based) detection paradigms. Moreover, with large language models in usage, the ai-generated fake reviews have been introduced which propose challenges by making linguistic deception cues unreliable. This paper provides a comprehensive review of fake detection methods, along with their evolution and critically analyzing unresolved challenges that exists across different approaches. By analyzing insights from prior foundational researches and studies, deep learning models, network-based methods, and recent surveys, we focus on outlining existing research gaps and suggest future directions to develop more robust, reliable and adaptable fake review detection systems.

Index Terms*— Fake review detection, opinion spam, deception detection, concept drift, transformer models, e-commerce security.

I. INTRODUCTION

Online consumer reviews play a major role in modern e-commerce platforms by influencing consumer decisions, product visibility, and seller reputation. Aggregated ratings and written feedback are considered proxies of product quality and service reliability, especially when consumers do not have prior experience. Although user-generated content has increased transparency and market efficiency, it has still created vulnerabilities that can be exploited through deceptive practices. Fake reviews are a serious form of online manipulation. In contrast to factual misinformation, deceptive reviews are written with subjective personal experience rather than using objective evidence to verify it; this makes verification inherently difficult. Empirical studies consistently show human evaluators do not reliably distinguish between fake or genuine reviews, and their accuracy is slightly above random guessing. As a result, fake review detection has emerged as a problem that exceeds human judgment and requires automated, data-driven solutions.

Early research was based on the assumption that deceptive intent produces measurable irregularities in language and focused on linguistic and psycholinguistic cues. Controlled experimental settings showed that automated models per-formed better than humans, which decided the feasibility of computational deception detection, but later analysis showed that these models are sensitive to domain context, cultural norms, and language conventions, which in turn decreased their generalization outside narrowly defined datasets.

Addressing the disadvantage of text-only methods, re-searchers gradually shifted toward behavioral and temporal modeling. Observing reviewer activity patterns, rating behavior, and temporal dynamics, behavioral approaches showed greater resistance to superficial linguistic manipulation. These methods proved beneficial in detecting anomalous review behavior at scale, but their reliance on historical data introduced weakness in cold-start scenarios and against advanced adversaries capable of imitating normal behavioral patterns.

As spamming strategies evolved from deceptive acts to coordinated campaigns, network-based and collective inference models started to emerge. These approaches captured collective behavior by modeling relational dependencies between users, reviews, and products. These methods reveal structural patterns that are difficult to hide, but the network-centric model still faces challenges in scalability, data accessibility, and real-time deployment, which limit their applicability to controlled analytical environments.

During this period, advances in machine learning and deep learning reshaped the methodological landscape of fake review detection. Feature-engineered Classifiers were replaced with deep neural architectures, which were capable of learning hierarchical representation from raw data. Attention mechanisms and representation learning made it possible to capture contextual semantics and emotional

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 6, June 2026

signals. More recently, transformer-based language models have achieved state-of-the-art performance on benchmark datasets. However, these advances have also raised concerns about interpretability, computational cost, and vulnerability to highly fluent machine-generated fake reviews.

A major outcome of existing research is that fake review detection is still insufficient for real-world deployment. Many models hold assumptions that are not operational in dynamic environments, like static data distributions, stable domains, and isolated deceptive behavior. In addition, reported performance gains don't hold adaptability against long term robustness, cross-domain reliability, and evolving adversarial strategies.

Motivated by these challenges, this paper presents a com-prehensive thematic review of fake review detection research, tracing its evolution from early linguistic methods to mod-ern transformer-based approaches. Rather than providing a descriptive survey, it critically examines fragmentation be-tween reviews of persistent field-level weaknesses like temporal concept drift, cross-domain generalization, and semantic and network-based paradigms. By synthesizing insights from multiple methodologies, this work helps in clarifying the fundamental challenges that still limit progress and outlines principled directions for the development of robust, adaptive, and trustworthy fake review detection systems.

II. LITERATURE REVIEW

A. Foundations of Linguistic and Psycholinguistic Deception Detection

Early fake review detection research observation was based on the fact that deceptive reviews try to look genuine, but still their language contains systematic differences. Studies show that humans are also weak in finding fake reviews; their accuracy is around 55–60% [1], [2]. Researchers thought that automated linguistic analysis is more reliable than manual judging because language can be an early signal of deception. Initial approaches used shallow linguistic features like n-grams, parts-of speech patterns, lexical diversity, and sentiment polarity. Psycholinguistic theory states that when a human writes a lie, he carries a burden of cognitive and emotional load with him that reflects in his choice of words and exaggeration. These ideas were supported in experiments, where automated classifiers wrote better than many humans; in many cases accuracy was above 80% [1], [3].

Moreover, the problem was that the results came in con-trolled settings. Most of the time datasets were domain-specific or artificially generated. When a model that was trained on one domain (like hotel reviews) was used on another domain, then accuracy fell 15–25% [1], [4]. In addition, linguistic cues depend mostly on language and cultural context; that is why English-based features don't fit well with multilingual or informal reviews. this approach was conceptually strong, but limited for real-world scalability and generalization, which motivated upcoming research.

B. Behavioral and Temporal Footprints of Review Spammers

The main idea of this section is that when detecting fake reviews only from text became difficult, researchers began to explore behavioral modeling. The basic assumption is that a spammer's intent is more clearly reflected through their actions, such as when a review is posted, how often it is posted, and how the activity is distributed over time. Analysis of long-term data collected from large-scale review platforms has observed that spammers' activity patterns are generally abnormal, and these patterns tend to remain mostly unchanged even after the review text is modified.

Certain behavioral features have been repeatedly found to be effective in the literature, such as review burstiness, abnormal review frequency, early-review bias, deviation of ratings from the consensus, and high activity within a short time window [5]–[8]. Mukherjee et al. formalized these signals and proposed a Bayesian Author Spamicity Model, where a reviewer's spam behavior are modeled as a latent variable [5]. This model is unsupervised in nature and is especially useful in scenarios where ground-truth labels are not fully reliable, and it has shown better performance as compared to heuristic-based methods.

Temporal modeling also revealed that a large portion of opinion spam comes from singleton reviewers, that is, accounts that post only a single review. Xie et al. 0 showed that in some datasets, more than 90% of reviewers fall into the singleton category, making traditional reviewer centric features ineffective [6]. Coordinated attacks were detected by analyzing the temporal correlation between review volume and rating changes.

Overall, evaluations indicate that behavioral models perform better than linguistic-only models in adversarial settings, but since they depend on historical data, they cannot serve as a standalone solution in cases of cold-start users, new products, and behavior-mimicking smart spammers [6].

C. Network and Collective Models for Opinion Spam Detection

When spamming strategies began to shift from individual reviews toward coordinated campaigns, individual-level detection approaches became insufficient. As a result, network-based and collective inference models were introduced, in which fake reviews are treated as group-level behavior rather than isolated cases. Under this approach, relational structures between users, products, and reviews are analyzed to uncover hidden collusion patterns.

Rayana and Akoglu's SpEagle framework is considered an important contribution in this direction [9]. This model combines textual, temporal, and rating metadata in a unified graph-based inference framework. By jointly estimating spamicity scores for users, reviews, and products, SpEagle

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 6, June 2026

shows better scalability and robustness in coordinated spam detection compared to single-entity classifiers.

A key advantage of the network-centric model is that it exploits dependencies, which is difficult for the spammers to hide. At the individual level norms can be mimicked, but to maintain consistency in an interconnected review network is challenging. Studies show that collective inference performs 8–15% better in comparison to independent classifiers [5], [9]. Graph construction, complete relational data requirements, and increasing computational complexity make real-time deployment difficult. TABLE I summarizes commonly used benchmark datasets in network based fake review detection. [9].

Table I. Dataset Statistics Used in Fake Review Detection Studies

Dataset	#Reviews (Filtered %)	#Users (Spammer %)	Product
YelpChi	67,395 (13.23%)	38,063 (20.33%)	201
YelpNYC	359,052 (10.27%)	160,225 (17.79%)	923
YelpZip	608,598 (13.22%)	260,277 (23.91%)	5,044

D. Feature-Engineered Machine Learning Paradigms and Their Structural Limitations

After identifying linguistic and behavioral deception cues, fake review detection was formulated as a supervised classification problem. In this phase, traditional machine learning algorithms such as Support Vector Machines, Naive Bayes, Logistic Regression, and Random Forests were commonly used. These models were trained using carefully engineered feature sets, which usually combine textual, behavioral, and sometimes relational attributes [1], [5], [7], [10].

There are two main reasons that this paradigm became famous. Firstly, ML classifiers integrated heterogeneous features in one framework, which were analyzed separately. Secondly, it was experimentally observed that combining linguistic and behavioral features improves performance. Multiple studies showed a gain in F1-score of about 6–12% on using hybrid feature vectors rather than text-only models [3], [7]. An important outcome of this phase was also the establishment of feature importance hierarchies, where features like rating deviation, review burstiness, and early-review bias emerged as strong spam indicators [5], [7].

Despite the advantages, feature-engineered ML approaches had structural limitations. The feature engineering process is labor-intensive and domain dependent— for example, features optimized for hotel reviews do not perform well on electronics or service reviews [3], [10]. Along with this, these models assume spammer behavior is stationary, which is the vice versa of the real world, where they adapt their strategies continuously. A major critical issue is that these depend on pipeline-labeled data where ground truth labels are the result of heuristics and crowd sourcing, due to which reported results overestimate real-world effectiveness. TABLE II summarizes the comparative strengths and limitations of different signal types used in feature-engineered detection paradigms.

E. Deep Learning and Representation Learning for Fake Review Detection

Deep learning approaches emerged when it became clear that handcrafted features were not sufficient to fully capture the complexity and evolving nature of deceptive behavior. Deep neural networks are able to learn hierarchical representations directly from data, which in turn reduce human bias and makes the models more adaptable over time. This shift was mainly influenced by recent advancements in natural language processing, where representation learning has shown strong results across different text analysis tasks.

According to literature, CNN, RNN, LSTMs, and attention-based architectures are widely used in this category [11], [12], [14], [15]. Better results in capturing contextual semantics and long-range dependencies acted as a base motivation for these models. Empirical studies show that when sufficiently large datasets are available, then deep learning models perform better than traditional baseline machine learning models.

Table II. Utility of Different Signal Types in Fake Review Detection

Signal Type	Observed Utility	Limitations Identified in Literature
Linguistic (Text)	Effective for semantic deception detection	Poor cross-domain and multilingual generalization
Behavioral	Stable against textual camouflage	Ineffective for new users and sparse histories
Temporal	Identifies bursty and singleton spam	Requires precise times tamp data
Network / Relational	Detects collusion and group attacks	High computational and data access cost
Emotional / Psycholinguistic	Improves recall in deceptive reviews	Sensitive to writing style and culture
Hybrid Signals	Consistently strongest performance	Higher system complexity

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 6, June 2026

Attention-argumented CNN-BiLSTM showed 5–10% bet-ter results on benchmark datasets in comparison to SVM-based classifiers [12]. Hybrid models showed better robustness when word embeddings and reviewer behavior or emotional indicators were combined [11], [14]. But still, deep learning approaches are data hungry and less interpretable; if the dataset is imbalanced, then their performance sharply degrades.

F. Transformer-Based Models and the Emergence of Machine-Generated Opinion Spam

The recent phase of fake review detection research starts with transformer based language models. These models work on self-attention mechanisms and large-scale pretraining, which understand text deeply along with the context. Trans-formers provide better contextual understanding and transfer learning capabilities, which perform better on varied datasets than traditional models.

Literature sections consistently report that pretrained trans-former models like BERT and RoBERT provide better results than earlier CNN- and LSTM based architectures. These models showed 4–9% improvement in F1-score in domain evaluation settings [16], [17]. An important aspect of this phase is the evaluation of detection tasks against machine-generated fake reviews, creating a “machines detecting ma-chines” scenario.

Moreover, transformer-based detectors still face challenges. These models have a very high computational cost and inherit biases from pretraining corpora. In addition, their interpretability is limited, so real-world deployment still holds a concern. As generative models increasingly optimize for stylistic real-ism and semantic coherence, the assumption that deception al-ways leaves detectable linguistic artifacts is becoming weaker. TABLE III summarizes the evolution of fake review detection paradigms.

Table III. Evolution of Fake Review Detection Paradigms

Paradigm	Primary Signal	Typical Models	Key Strength	Fundamental Limitation
Linguistic Psycholinguistic	Textual cues (n-grams, sentiment, LIWC)	SVM, Na¨ıve Bayes	Captures deception-related language patterns	Strong domain and language dependence
Behavioral / Temporal	Review timing, rating deviation, burstiness	Bayesian models, regression	Harder to manipulate behavior than text	Cold-start users and reliance on history
Network / Collective	User–review–product relations	Graph-based inference	Detects coordinated spam campaigns	Scalability and privacy constraints
Feature-Engineered ML	Hybrid text + behavior	SVM, Random Forest	Integrates heterogeneous signals	Manual feature engineering and feature drift
Deep Learning	Learned representations	CNN, LSTM, Attention	Reduces handcrafted feature bias	Data-hungry and limited interpretability
Transformer-Based	Contextual semantics	BERT, RoBERTa	Strong semantic generalization	Vulnerable to LLM-generated re-views

G. Systematic Reviews, Meta-Analytical Insights, and Persis-tent Field-Level Failures

Systematic reviews analyze the fake review detection literature from a meta-analytical perspective, where the results of multiple paradigms are synthesized to highlight structural weaknesses at the field level [15], [20]. Studies concluded a recurring observation: absence of reliable ground truth, as no objective verification mechanism exists for fake reviews. Datasets generally depend on heuristic or indirect labeling strategies, which in turn impact evaluation validity. Another persistent issue is related to the evaluation of realism. Most of the studies use static datasets where spammer activities are fixed and the stationarity assumption is made. Moreover, longitudinal analyses show that concept drift occurs rapidly and spammers adapt their tactics faster than detection systems.

Due to this mismatch, experimental performance cannot be sustained in real-world settings.

A central conclusion of systematic reviews is that no single modality or model can fully capture the multidimensional nature of opinion spam. It is insufficient to treat fake review detection only like a classification task because it is a dynamic adversarial system that operates under economic incentives, social dynamics, and platform-specific constraints. Table IV summarizes the main strengths and structural limitations of major fake review detection paradigms based on meta-analytical studies.

III. RESEARCH GAPS AND CHALLENGES

A. Robustness as the Central Unresolved Challenge in Fake Review Detection

Researchers have been trying to detect fake reviews from both online and offline platforms for many years. Over time, their methods of review have improved from basic techniques that examined writing styles and behavioral patterns to more advanced AI models, like deep learning and

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 6, June 2026

transformers.

While these models get high accuracy in tests, but they don't perform very well in real-life situations. This is mainly because of three main reasons: A model trained on a specific domain cannot transfer its learning to other domains; fake reviews evolve constantly; therefore, most review methods don't adapt over time—as a result, they become outdated and less effective. Additionally, semantic (word-based) and relational (graph-based) review methods operated in separate worlds, which limited the strength of the overall detection systems.

Table IV. Strengths and Limitations of Major Fake Review Detection Paradigms

Paradigm	Primary Strength	Limitation
Linguistic Models	Semantic deception analysis	Easily adapted by skilled spammers
Behavioral Models	Resistant to text manipulation	Cold-start and mimicry issues
Network Models	Captures coordinated attacks	Limited scalability in real-time systems
Feature Engineered ML	Interpretability	Feature drift and domain dependence
Deep Learning	High in domain accuracy	Black box nature and data dependence
Transformer Models	Contextual understanding	Weak robustness against LLM generated reviews

B. Underlying Causes of the Research Gap

Most researchers use outdated and static datasets, such as Yelp, Amazon, and Trip Advisor. These kinds of datasets don't evolve; they don't incorporate changes, and researchers typically train and test their models using random splits of the same data. The problem arises here when real fake reviews change over time, but the datasets don't. Static datasets create a false sense of stability, so models trained on outdated and fixed datasets look good in tests but fail in real use.

Deep learning doesn't solve the problem, even though the models are more advanced and trained, as the evaluation methods used are outdated and need new datasets to be trained on. Most models are trained and tested within a single platform, such as either on Trip Advisor or Amazon, and since models only work well in one domain, this creates a generalization gap: good accuracy in one domain, bad accuracy in another.

To resolve these problems, there exist two approaches, but they don't work together: text-centric (semantic) methods and network-centric (graph) methods. Each approach is valid and developed separately because hybrid (combined) models are rare, and when they exist, they are either too shallow, limited by datasets, or too computationally heavy. Since the models never fully capture the whole scenario of fraud and fake reviews, a structural disconnection occurs.

C. Evidence from the Literature

Most researchers assume that reviews stay the same; only a few studies consider that reviews change over time. Studies that analyze burst-based patterns in review posting demonstrate that detection accuracy degrades as data distributions evolve. However, most deep learning models do not include any method to handle this drift and still rely on static, mixed-time datasets. Feature-based or emotion-aware models achieve strong in-domain results but are not validated across heterogeneous product categories.

Transformer-based detection further reveals that models trained on one review distribution struggle when confronted with stylistically distinct or machine-generated fake reviews. So, no strong reviewed paper offers a strong and principled solution for domain-invariant deception representation. Network-based studies convincingly show that coordinated spammer behavior is detectable at the relational level; however, these methods largely ignore textual semantics.

Conversely, deep and transformer-based models focus on the meaning, style, and linguistic cues and detect subtle deception in writing but miss coordinated fraud where individual reviews appear authentic. Lack of a unified framework that brings together semantic intent and relational behavior is still a key limitation in this field. Models either excel in understanding what a review says or how reviewers behave, but not both.

D. Consequences of the Gap

From a scientific perspective, the lack of robustness weakens the reliability of fake-review detection. Reported accuracy improvements may reflect dataset familiarity rather than actually learning how deception works, and this leads to overfitting at the research-field level. In real-world situations, models quietly degrade because of concept drift, inconsistency of performance on product categories, and the inability to find coordinated fraud beyond the capabilities of text filters. All these directly affect consumer trust, the credibility of online platforms, and the ability of regulators to ensure fairness.

E. Unresolved Challenges Preventing Solutions

The continuing existence of these gaps can be attributed to many factors. For one, the lack of longitudinal multi-domain datasets representing how reviews change over time and between different platforms creates obstacles to under-standing this area; likewise, the complexity of performing text–network analysis together means it is

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 6, June 2026

difficult to implement these techniques in real-world situations. In addition, there are several additional issues that have yet to be fully addressed in a systematic way within the body of literature. These include the lack of standardization in protocols to test temporal robustness, as well as the inherent trade-offs between scalability, interpretability and model performance.

IV. FUTURE ENHANCEMENTS AND RESEARCH DIRECTIONS

A. Drift-Aware Transformer–GNN Hybrid Architectures with Temporal Attention

In the present scenario, a transformer works best when we have to understand the review's context and emotional tone; models are trained generally on fixed datasets, but with time, when the review's style, language, and behavior change, the model finds it difficult to handle, and the concept drift problem arises. Only using a transformer is not enough because fake reviews are not only dependent on text but also on who the reviewer is. What is the product? and also the time period in which activity increases, using a transformer with a graph neural network provides both understanding of semantic content, like wording and tone, and the relationship between reviewer, product, and time; this provides the model an overview of the surrounding interaction pattern.

Temporal attention layers use is important because old patterns are only relevant for the same level. Some deception strategies weaken after some time; vice versa, new patterns emerge. Temporal attention lets the model learn which time period is meaningful, which prevents it from retraining and adjusts its prediction gradually. The model adapts with time, which is more practical for real-world platforms.

B. Cross-Platform, Multilingual Datasets with Aligned Deception Labels

The dataset used for fake review detection is mostly focused on the English language and limited platforms, due to which the model shows narrow behavior, and when applied to different platforms or language reviews, performance often drops. It is still difficult to properly capture real-world diversity, and generalization is still a major challenge. In addition to tackling this problem, future datasets must be multilingual and cross-platform compatible. Deception labels' meaning should be consistent for each platform or language in which text, along with behavioral signals and timing patterns, should be considered. If the annotation process is aligned, systematic testing of models would be possible, and it would be easier to understand its performance in different domains and languages.

C. LLM-Resilient Detection via Adversarial Training Against Generated Reviews

Old fake reviews were mostly human-written; therefore, they had obvious spam patterns, but the introduction of LLMs has changed the scenario. Generated reviews are very fluent and also emotionally consistent, due to which traditional detectors, which depend upon spelling errors, repetition, or simple cues, fail to detect them. In the future, generated reviews should be considered as an adversarial class; if they are considered as traditional spam, then the model cannot learn their unique pattern properly. A practical solution can be adversarial training. In this approach, the detection model can be trained against those reviews directly, which are generated by different language models. These reviews are created using different prompts, styles, and constraints so that the model can be prepared for real-world attacks. Through this model, one can learn how the behavior of machine-generated content is different from human deception.

D. Explainability-Guided Architectures Linking Semantic and Behavioral Signals

Deep learning and transformer-based models perform better in fake review detection, but to understand their decision process is difficult. It is important that the model should clearly state why a review is marked fake. Future systems should focus on architectures that directly connect review text signals with user behavior and network-level patterns. It means the model shows which words, tones, or phrases look suspicious and which behavioral signals, like repeated reviews or coordinated activity, contribute to decision making. In this model, attention visualization, attribution methods, or simple rule based layers can be used so that internal decisions can be converted into human-readable form. Its goal is that moderators and auditors can easily understand on which basis fraud is detected; this type of explainability increases model trust and makes deployment on platforms more practical.

E. Online Evaluation Benchmarks Simulating Real Platform Deployment

To date, fake review detection uses a random train-test split, which mixes past and future data, which is the vice versa of real platforms. Data comes in order of date and time on real platforms, labels get delayed, and attack strategies change. That is why offline evaluation often doesn't reflect real-world performance accurately. To make future benchmarks more realistic, chronologically ordered data streams, delayed labeling, and evolving attack patterns should be included in the evaluation. This would test if the model will work only for a fixed dataset or also be stable for real deployment conditions. Online evaluation protocol, where the model is tested on streaming data and limited retraining is allowed, provides a more practical picture.

V. CONCLUSION

This paper presents a thematic review of fake review detection research, analyzing the field's progression from early linguistic and psycholinguistic cue-based approaches to

behavioral modeling, network-based inference, deep learning architectures, and recent transformer-driven methods. The paradigms comparison showed improvement in detection ac-curacy, but many practical challenges are yet to be solved in real-world deployment.

One of the key conclusions of fake review detection is that it is still not robust enough for real-world use. In different approaches, the model still struggles with temporal concept drift, shows weak cross-domain generalization, and treats text-based analysis differently from behavioral or network-level information. The main reason due to which problems arise is broader assumptions in the field, such as static data distributions and stable spammer behavior. Generally, e-commerce platforms are dynamic in nature, which is the opposite of these assumptions. The reviews also show the gains captured in the performance of deep learning and transformer-based models are only limited to benchmark settings. Lack of interpretability, high computational overhead, and vulnerability against machine-generated opinion spam put constraints on the performance of these models in the real world. As generative models are becoming more fluent and diverse, relying on textual deception cues is becoming increasingly fragile.

Overall, fake review detection should be viewed as a continuously evolving adversarial problem, requiring system-level robustness rather than small model improvements. To build more trustworthy detection systems, future research should focus on drift-aware learning, domain-invariant representations, better integration of semantic and relational signals, and more realistic evaluation settings.

REFERENCES

[1] M. Ott, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," in *Proc. ACL*, 2011, pp. 309–319.

[2] E. Plotkina, M. Munzel, and R. Pallud, "Fake reviews: The effects of emotionality and deceptive cues on consumer perceptions," *J. Business Research*, vol. 119, pp. 346–355, 2020.

[3] P. Ha´jek, T. Olej, and M. Myskova, "Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining," *Neural Computing and Applications*, vol. 32, no. 23, pp. 17259–17274, 2020.

[4] K. Ji, Y. Wang, and X. Liu, "Hybrid deep learning models for fake review detection," in *Proc. Int. Conf. Artificial Intelligence and Big Data*, 2020, pp. 1–8.

[5] A. Mukherjee, B. Liu, J. Wang, N. Glance, and N. Jindal, "Spotting opinion spammers using behavioral footprints," in *Proc. 19th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, Chicago, IL, USA, 2013, pp. 632–640.

[6] S. Xie, G. Wang, S. Lin, and P. S. Yu, "Review spam detection via time series pattern discovery," in *Proc. 21st Int. World Wide Web Conf. (WWW)*, Lyon, France, 2012, pp. 635–644.

[7] D. Savage, X. Zhang, X. Yu, P. Chou, and Q. Wang, "Detection of opinion spam based on anomalous rating deviation," *Expert Systems with Applications*, vol. 42, no. 22, pp. 8650–8657, 2015.

[8] B. Liu, "Opinion spam and analysis," in *Proc. 1st ACM Int. Conf. Web Search and Data Mining (WSDM)*, Palo Alto, CA, USA, 2008, pp. 3–4.

[9] S. Rayana and L. Akoglu, "Collective opinion spam detection: Bridging review networks and metadata," in *Proc. 21st ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, Sydney, NSW, Australia, 2015, pp. 985–994.

[10] K. Cai, H. Yang, and J. Li, "Feature engineering for spam detection: A survey and empirical evaluation," in *Proc. Int. Conf. Data Mining and Big Data*, 2017, pp. 1–10.

[11] P. Bhuvaneshwari, A. Nagaraja Rao, and Y. H. Robinson, "Spam review detection using self-attention-based CNN–BiLSTM," *Multimedia Tools and Applications*, vol. 80, no. 25, pp. 33013–33035, 2021.

[12] A. Neisari, L. Rueda, and S. Saad, "Spam review detection using self-organizing maps and convolutional neural networks," *Computers & Security*, vol. 106, Art. no. 102274, 2021.

[13] M. Ha´jek, P. Olej, and M. Myskova, "Emotion-aware fake review detection," in *Proc. Int. Conf. Neural Information Processing*, 2020, pp. 1–10.

[14] H. Mohawesh, S. M. Al-Harbi, and A. Al-Zoubi, "Deep learning-based opinion spam detection," *Expert Systems with Applications*, vol. 185, Art. no. 115620, 2021.

[15] J. Paul and A. Nikolaev, "Fake review detection on e-commerce plat-forms: A systematic literature review," *Data Mining and Knowledge Discovery*, vol. 35, no. 5, pp. 1834–1888, 2021.

[16] J. Salminen, C. Kandpal, A. M. Kamel, S.-G. Jung, and B. J. Jansen, "Creating and detecting fake reviews of online products," *J. Retailing and Consumer Services*, vol. 64, Art. no. 102771, 2022.

[17] K. Khoirotulmuadiba, R. Prasetyo, and A. Wibowo, "RoBERTa-based fake review detection," in *Proc. Int. Conf. Intelligent Systems and Data Science*, 2024, pp. 1–8.

[18] B. Shehnepoor, M. Salehi, R. Farahbakhsh, and N. Crespi, "Network-aware fake review detection," *IEEE Access*, vol. 10, pp. 91234–91249, 2022.

[19] H. Paul and A. Nikolaev, "Systematic literature review of fake review detection," *Data Mining and Knowledge Discovery*, vol. 35, no. 6, pp. 2101–2144, 2021.

[20] R. Gupta, V. Jindal, and I. Kashyap, "Recent state-of-the-art of fake review detection: A comprehensive review," *The Knowledge Engineering Review*, vol. 39, Art. no. e8, pp. 1–47, 2024, doi: 10.1017/S0269888924000067.