
ARTIFICIAL INTELLIGENCE FOR DETECTING FAKE ONLINE REVIEWS

Dr. Rohan Mehta

Assistant Professor

Department of Computer Science and Engineering

School of Engineering and Technology

Vellore Institute of Technology

Vellore, Tamil Nadu, India

Abstract

Online reviews have become an important source of product information, service evaluation, and reputation signals in e-commerce, travel, health, education, hospitality, and platform-based work. Their influence also creates incentives for deceptive activity. Fake, incentivised, coordinated, or automatically generated reviews can distort ratings, mislead consumers, harm legitimate businesses, and weaken trust in digital marketplaces. Detecting such reviews is difficult because deceptive content may resemble genuine opinion, and abusive actors can alter wording, timing, account behaviour, and interaction patterns to avoid simple rules.

This review-based study examines how artificial intelligence can support the detection of fake online reviews through the combined analysis of review text, rating behaviour, reviewer activity, temporal patterns, product relationships, and network structure. It synthesises established approaches based on linguistic indicators, supervised machine learning, anomaly detection, deep neural models, transformer-based language models, graph learning, and explainable AI. The study also considers current risks created by coordinated campaigns and synthetic review generation, where content may be grammatically fluent and semantically plausible.

The analysis finds that reliable detection depends on multi-signal assessment rather than reliance on a single textual cue or user attribute. Stronger systems integrate content features with behavioural and relational evidence, evaluate performance under class imbalance and domain shift, and preserve human review for consequential decisions. The paper concludes that AI should be deployed as a decision-support mechanism within transparent platform governance, combining calibrated risk scores, meaningful explanations, appeal procedures, privacy safeguards, and continuous monitoring for new manipulation tactics.

Keywords: Fake review detection; Opinion spam; Artificial intelligence; Machine learning; Natural language processing; Transformer models; Graph neural networks; Online reputation; Explainable AI; Platform trust

1. INTRODUCTION

Digital platforms increasingly organise consumer discovery and commercial trust through ratings and written reviews. A prospective buyer may compare hotels, restaurants, medicines, electronics, online courses, delivery services, or professional providers by reading the experiences of earlier users. This arrangement gives reviews a public role that extends beyond casual commentary: they can influence rankings, visibility, revenue, and reputation. When the review environment contains fabricated or strategically manipulated content, the resulting harm reaches consumers, sellers, and the credibility of the platform itself.

Fake online reviews, often described as opinion spam or review fraud, include several

forms of manipulation. A review may be written by a person who has not used the product or service, commissioned by a seller or competitor, copied from another source, produced by a coordinated group of accounts, or generated with automated tools. Some reviews are overtly promotional or malicious, while others imitate ordinary customer language and discuss plausible features. The practical challenge is not simply to identify positive or negative sentiment. It is to assess whether a review represents an authentic, independent experience.

Earlier research examined deceptive reviews through lexical patterns, sentiment intensity, unusual wording, duplicate text, rating deviations, and account-level behaviour. Studies by Jindal and Liu, Ott and colleagues,

Lim and colleagues, and Mukherjee and colleagues demonstrated that text and behaviour can provide useful signals, but also showed that no single feature is definitive. A short enthusiastic review can be genuine, and a highly active reviewer can be legitimate. Detection must therefore use evidence in context and avoid treating normal variation as proof of deception.

Artificial intelligence offers a scalable way to process large review streams and identify suspicious patterns for further assessment. Classical machine learning can combine manually designed text and metadata features; deep learning can represent semantic and discourse information; transformer models can capture contextual language signals; and graph-based systems can model relationships among reviewers, products, sellers, and review events. These methods are promising, but they introduce operational concerns involving data quality, biased labels, model drift, privacy, explanation, and the consequences of false positive decisions.

This paper presents a structured review of artificial intelligence for detecting fake online reviews. It focuses on the analytical signals used by detection systems, the methodological choices that shape validity, and the governance safeguards required for responsible use. Rather than treating AI as an automatic truth engine, the study frames it as a risk-assessment tool that should help platforms prioritise review, investigate coordinated manipulation, and protect trustworthy participation.

2. LITERATURE REVIEW

The literature on review fraud developed from early work on duplicate, misleading, and deliberately biased reviews. Jindal and Liu (2008) classified opinion spam into deceptive reviews, reviews that target brands or products without discussing the actual offering, and non-review content. This framing was important because it separated authenticity from sentiment. A review can be negative yet genuine, or positive yet deceptive. Subsequent studies expanded the analysis to author behaviour, product targeting, review timing, group coordination, and the structure of reviewer-product networks.

Text-based research has shown that deception can sometimes be associated with broad

praise, limited verifiable detail, repeated templates, excessive emotional language, unusual part-of-speech patterns, and inconsistencies between rating and text. However, linguistic signals are domain dependent and can be affected by culture, language proficiency, review length, and platform norms. The work of Ott et al. (2011), Perez-Rosas and Mihalcea (2015), and Ren and Zhang (2016) highlights both the potential of language analysis and the difficulty of transferring a model trained in one setting to another.

Behavioural studies shift attention from an isolated review to the activity pattern of the account that produced it. Relevant indicators include review frequency, rating extremity, product concentration, temporal bursts, account age, verified transactions, repeated review structures, and deviations from typical community behaviour. Lim et al. (2010) and Mukherjee et al. (2013) showed that rating behaviour and behavioural footprints can expose suspicious activity even when review text is difficult to classify.

Network-based approaches consider the relationships among users, reviews, products, sellers, and time. A coordinated campaign may involve many accounts reviewing a small group of products within a narrow period, while sharing templates, ratings, or review targets. Akoglu, Chandy, and Faloutsos (2013) and Rayana and Akoglu (2015) demonstrated that relational patterns can reveal collective fraud that is not apparent from a single review. Recent work also explores graph neural networks and multi-relational models that can learn from these connections at scale.

Modern language models have made review content analysis more flexible. Transformer architectures such as BERT can learn contextual representations without relying entirely on manually engineered features. Yet linguistic fluency is not equivalent to authenticity. This limitation is especially relevant when reviews are written or refined by generative systems. An effective contemporary approach should therefore combine semantic representation with transaction, timing, account, and network evidence, while also retaining transparent procedures for human validation.

2.1 FORMS OF FAKE REVIEWS AND OBSERVABLE SIGNALS

Fake reviews can be categorised by motive, production method, and coordination structure. Promotional reviews aim to raise the perceived quality of a target product or seller. Defamatory reviews attempt to damage a competitor or service provider. Incentivised reviews may result from gifts, discounts, or other benefits that are not disclosed. Organised campaigns can use groups of accounts, review exchanges, paid intermediaries, or automated generation. These categories can overlap, so a practical detection system should measure risk rather than assume a fixed fraud type.

Content-level indicators concern the wording and internal structure of a review. A model may evaluate semantic specificity, product-feature references, polarity, emotion, readability, repetition, semantic similarity to other reviews, mismatch between the rating and written opinion, and the presence of templated phrases. Such signals should be interpreted cautiously. Reviews written by new users, non-native speakers, or users with limited time may be brief or formulaic for legitimate reasons. A fair system must not rely on linguistic style alone.

Behavioural indicators describe how an account participates over time. Examples include a sudden burst of reviews, repeated maximum or minimum ratings, reviews posted immediately after account creation, unusually narrow targeting of related products, or a review pattern inconsistent with the user's earlier activity. These signals are often more useful when compared with relevant peer groups rather than global averages. A frequent reviewer in a local guide community, for example, may be normal even though the same volume could be suspicious for an otherwise inactive account.

Relational indicators capture coordination and influence. A group of accounts may review the same products in a similar order, use related wording, appear within the same time window, or interact with the same sellers. Graph representations can model reviewer-product, reviewer-review, seller-product, and time-based connections. They can also reveal dense subgroups, repeated campaigns, and abnormal propagation patterns. Their usefulness depends

on access to reliable relational data and careful safeguards against inferring misconduct merely from social or commercial proximity.

2.2 AI APPROACHES FOR REVIEW-FRAUD DETECTION

Rule-based systems provide clear and auditable checks, such as flagging duplicate text, excessive posting frequency, impossible transaction timelines, or violations of disclosure requirements. They are useful for high-confidence cases and rapid enforcement, but they are vulnerable to adaptation. Fraudsters can rewrite templates, space out activity, use aged accounts, and vary ratings. Rules are therefore most effective when combined with statistical and machine-learning models.

Traditional supervised learning represents reviews through engineered features and predicts a class such as genuine or suspicious. Common algorithms include logistic regression, support vector machines, random forests, gradient boosting, and ensemble methods. Their advantages include manageable training requirements and relatively clear feature contribution analysis. Their limitations include dependence on labels that may be incomplete, inconsistent, or based on platform moderation decisions rather than verified ground truth.

Deep neural methods can learn semantic and sequential patterns directly from text. Convolutional networks, recurrent networks, attention mechanisms, and hierarchical encoders have been used to capture document-level signals. Transformer-based models offer a strong foundation for contextual text representation and can be fine-tuned for review classification. However, models trained only on review text may learn platform-specific shortcuts, sentiment distribution, or genre cues instead of genuine deception. Evaluation across domains and time periods is necessary to determine whether a model has learned a reliable signal.

Graph-based learning extends the analysis beyond individual content. A graph can connect reviewers to products, reviews to ratings, users to accounts, and events to time windows. Graph neural networks can aggregate local structure and feature information, while anomaly-detection methods

can identify unexpected clusters or rare behavioural combinations. These systems are especially relevant for coordinated campaigns, but they require robust data governance because relationship data may be sensitive and can encode unequal exposure, popularity, or marketplace structure.

A practical AI system is often best designed as a layered pipeline. Rules can identify clear policy breaches; text models can estimate linguistic inconsistency; behavioural models can measure account-level anomalies; graph methods can detect collusion; and a final calibrated score can prioritise cases for review. Human moderators should be able to inspect the evidence behind a flag, request additional verification, correct model errors, and feed confirmed outcomes back into model monitoring.

3. OBJECT OF STUDY

The primary object of this study is to examine how artificial intelligence can be used to identify fake online reviews while preserving reliability, fairness, and procedural accountability. The paper considers individual reviews, reviewer accounts, products or services, seller relationships, and temporal review activity as interconnected sources of evidence. It recognises that authenticity is not directly observable from wording alone and that detection must combine heterogeneous signals.

The study focuses on review environments in which user-generated ratings and narratives influence consumer choice. It examines the analytical contribution of textual, behavioural, metadata, and network information; the role of supervised, unsupervised, and semi-supervised learning; and the use of explainable outputs to support platform decisions. The study also considers the limitation that many available labels are proxies, such as filtered reviews or policy removals, rather than independently verified fraud cases.

A further objective is to identify conditions under which detection systems are more likely to be valid and useful. These conditions include clear platform definitions, representative training data, attention to class imbalance, temporal validation, domain transfer testing, separation of detection from automatic punishment, and meaningful human

oversight. It also includes safeguards for reviewers who may be incorrectly flagged because of writing style, accessibility needs, cultural expression, or atypical but legitimate purchasing behaviour.

Finally, the study treats review-fraud detection as part of a broader trust-and-safety process. The purpose is not only to remove suspicious content but to improve the integrity of online information environments. This requires transparent platform policies, disclosure mechanisms, evidence preservation, user notification where appropriate, reliable appeals, and ongoing evaluation as adversarial strategies evolve.

3.1 TEXTUAL AND SEMANTIC ASSESSMENT

- Assessment of semantic specificity, including references to product features, service encounters, use context, and verifiable experience.
- Evaluation of duplicated, near-duplicated, translated, paraphrased, or templated review content across products, accounts, and time periods.
- Analysis of rating-text consistency, sentiment intensity, exaggerated certainty, unsupported claims, and abrupt changes in writing style.
- Use of transformer embeddings and similarity measures to identify unusual content clusters without treating a specific language style as fraudulent by itself.
- Testing across product categories, languages, review lengths, and time periods to reduce reliance on platform-specific shortcuts.
- Human review of high-impact flags, especially where a model's conclusion would remove content, restrict an account, or affect a seller's ranking.

3.2 BEHAVIOURAL, TEMPORAL, AND NETWORK ASSESSMENT

- Measurement of posting frequency, account age, verified interaction indicators, rating distribution, review length variation, and product targeting patterns.
- Detection of temporal bursts, synchronised activity, campaign-like rating shifts, and sudden reviewer-product concentration.

- Construction of reviewer-product-review graphs that represent shared targets, similar timing, repeated content patterns, and suspected coordinated groups.

- Comparison of accounts with contextually appropriate peer groups instead of using a single global threshold for all users.

- Use of anomaly scores and calibrated confidence values so that uncertainty can be represented rather than hidden.

- Monitoring for adaptive behaviour, including spaced-out campaigns, account farming, review exchanges, content rewriting, and synthetic review generation.

3.3 TRUST, FAIRNESS, AND GOVERNANCE

- Clear definitions of prohibited review manipulation, including undisclosed incentives, fabricated experience, coordinated fraud, and deceptive automation.

- Privacy-conscious collection and retention of account, transaction, device, and relationship data used for risk assessment.

- Bias testing across languages, regions, product categories, seller size, review length, and customer participation patterns.

- Decision records that distinguish a model flag from a confirmed policy violation and preserve the evidence used in moderation.

- Explanation tools that identify the meaningful evidence behind a score without exposing detection thresholds that can be easily gamed.

- Accessible notice and appeal procedures for users, with human reconsideration for contested, high-consequence actions.

4. RESEARCH METHODOLOGY

The present study uses a review-based, descriptive, analytical, and comparative methodology. It synthesises research on opinion spam, deceptive language, review behaviour, fraud networks, machine learning, deep learning, transformer models, graph-based detection, and responsible AI. The aim is not to claim a new benchmark result or to present an operational classifier. Instead, the study compares the evidence types and design choices that shape the reliability of fake-review detection systems.

Secondary material is interpreted through a multi-signal framework. The framework organises evidence into four linked dimensions: review content, reviewer behaviour, product or service context, and relational structure. The analysis also evaluates model development concerns such as label quality, dataset imbalance, contamination between training and test samples, temporal drift, cross-domain performance, explanation, fairness, and human review.

The methodology treats platform enforcement data cautiously. A removed review is not always confirmed fraud, and an unremoved review is not necessarily authentic. Where possible, studies should distinguish between directly verified cases, policy-based labels, synthetic datasets, and proxy labels. This distinction is essential because a model can appear accurate against a platform filter while failing to identify actual deceptive behaviour in a different environment.

4.1 RESEARCH DESIGN

Component	Description
Research Type	Review-based, descriptive, analytical, and comparative study
Approach	Multi-signal artificial intelligence analysis for review-fraud detection
Data Source	Peer-reviewed research, conference proceedings, policy materials, and responsible-AI guidance
Analysis Method	Thematic synthesis, comparative model assessment, and trust-and-safety evaluation
Focus Area	Text, metadata, reviewer behaviour, time patterns, network structure, explainability, and governance

4.2 SOURCES AND ANALYTICAL CATEGORIES

The study considers literature that addresses artificial intelligence and online review

manipulation through multiple methodological traditions. This includes experiments in deceptive text classification, studies of reviewer rating behaviour, graph-based fraud

analysis, neural and transformer models, systematic reviews of review-spam detection, and policy discussions concerning disclosure, transparency, and consumer protection.

Each source is examined according to the following analytical categories:

- Definition of fake or deceptive reviews and the type of ground-truth evidence used.
- Features extracted from text, ratings, reviewer history, transaction context, timestamps, and relational graphs.
- Learning paradigm, including rule-based, supervised, unsupervised, semi-supervised, deep-learning, and graph-learning approaches.
- Evaluation design, including class distribution, train-test separation, cross-domain testing, temporal robustness, and error analysis.
- Operational implications for explanation, human moderation, privacy, fairness, user notice, appeals, and platform governance.

5. DATA ANALYSIS

The comparative analysis shows that fake-review detection is strongest when it is formulated as an evidence-fusion problem. A suspicious review rarely provides a single conclusive signal. A text-only model may detect repeated promotional language but miss a carefully written paid review. A behavioural model may detect a burst of ratings but misclassify a legitimate user who reviews many products after returning from travel. A network model may identify dense clusters but incorrectly interpret a genuine enthusiast community as coordinated fraud. Combining evidence reduces dependence on any one imperfect indicator.

Content analysis is valuable for identifying patterns of duplication, semantic vagueness, incompatible product claims, unusual sentiment alignment, and similarity across many reviews. Modern language representations can help detect nuanced paraphrases and semantic overlap. Yet review text should not be treated as a lie detector. The same writing feature can be influenced by marketing vocabulary, translation tools, literacy differences, accessibility constraints, or ordinary consumer preferences. Text

features are best used as one component of a broader risk profile.

Behavioural analysis provides a different view by examining account trajectories. Sudden account activation, repeated extreme ratings, abnormal product concentration, timing bursts, and atypical rating deviations may signal manipulation. The value of these indicators grows when compared against a relevant context, such as the expected pattern in a product category, locality, or reviewer cohort. Time-aware models are particularly important because adversaries can adapt by slowing campaigns, varying content, or using older accounts.

Network analysis can reveal forms of collective manipulation that do not appear in isolated records. Reviewer-product graphs can expose repeated co-targeting, shared campaign timing, and suspicious communities. The relationship structure becomes especially informative when it is combined with text similarity and metadata. However, graph predictions can be difficult to explain and may amplify historical enforcement patterns. Systems should therefore document what relationships contribute to a flag and test whether model performance remains stable across seller size, product popularity, and user communities.

Evaluation must reflect the practical distribution of fraud. Fake reviews are often a minority class, which means a high overall accuracy may hide poor detection of the cases that matter. Precision, recall, F1 score, area under the precision-recall curve, calibration, and error analysis should be reported alongside accuracy. A platform should also measure downstream outcomes: the proportion of flags confirmed by reviewers, appeal reversal rates, decision time, repeat-offender discovery, and the effect of interventions on review quality.

The analysis also identifies a growing problem of synthetic and assisted review writing. Generative systems can create fluent and varied text, reducing the usefulness of simple grammatical or sentiment cues. This does not make detection impossible; it changes the balance of evidence. Transaction verification, reviewer history, timing patterns, device and account integrity signals where lawfully

available, and graph relationships become more important. The most resilient defence is to evaluate whether the review fits a credible pattern of authentic participation, not merely whether its wording appears machine-generated.

6. DISCUSSION

The findings indicate that artificial intelligence can improve the scalability and consistency of fake-review detection, but only when platforms avoid overreliance on automated classification. Review integrity is an adversarial problem. Detection models are observed, tested, and eventually challenged by actors who can vary language, change account behaviour, distribute activity across time, and purchase access to apparently legitimate profiles. A static model with fixed thresholds will gradually lose effectiveness unless it is monitored and updated.

A central lesson is that authenticity and sentiment should not be conflated. Negative reviews can be valuable signals of genuine customer experience, and positive reviews can be authentic even when strongly expressed. Models that penalise emotional wording or unusual grammar may disadvantage particular groups without improving fraud detection. The objective should be to identify a credible combination of anomalous evidence, with uncertainty clearly represented and with a path for correction.

Explainability has an important operational role. A moderator needs more than a binary label; they need to understand whether a case was flagged because of duplicated text, impossible timing, coordinated product targeting, unresolved account integrity concerns, or a combination of factors. Explanations should support investigation rather than expose detailed thresholds to abuse. They should also help platforms identify systematic error, such as a model that over-flags certain product categories, languages, or new sellers.

Governance is therefore inseparable from model design. Platforms should define prohibited conduct, communicate disclosure requirements, protect user data, preserve audit records, and establish proportionate enforcement. High-confidence cases may warrant immediate action when there is clear

policy evidence, while lower-confidence flags should receive human review or additional verification. Users and businesses should have meaningful avenues to challenge incorrect decisions, particularly where rankings, account access, or commercial participation are affected.

The most sustainable approach is a socio-technical one. AI can triage volume, surface patterns, and prioritise investigations; trained reviewers can assess context; policy teams can refine definitions and remedies; and researchers can test models for transferability and bias. This approach treats fake-review detection as an evolving trust-and-safety capability rather than a one-time classification task.

7. CONCLUSION

This study concludes that artificial intelligence can make a meaningful contribution to detecting fake online reviews when it integrates textual, behavioural, temporal, and network evidence. Classical machine-learning methods, neural models, transformers, anomaly detection, and graph-based techniques each provide useful capabilities, but none is sufficient in isolation. Reliable systems should combine complementary signals, measure uncertainty, and use calibrated risk scores to direct attention toward cases that require investigation.

The analysis shows that model quality depends heavily on the quality of labels, the relevance of training data, the handling of class imbalance, the testing of domain and temporal transfer, and the availability of accountable human oversight. A model that performs well on a narrow benchmark may not remain effective when language, product categories, platforms, or adversarial strategies change. Continuous validation, post-deployment monitoring, and careful error analysis are therefore essential.

Effective platform policy should pair AI detection with transparent review rules, meaningful explanations, fair process, privacy safeguards, and accessible appeals. The goal is not to automate the removal of every unusual review; it is to reduce manipulation while protecting legitimate voices. By treating fake-review detection as a multi-signal, human-governed, and continuously evaluated process,

online platforms can strengthen the reliability of ratings and preserve trust in digital marketplaces.

8. REFERENCES

1. Aghakhani, H., Machiry, A., Nilizadeh, S., Kruegel, C., and Vigna, G. (2018). Detecting deceptive reviews using generative adversarial networks. *Proceedings of the IEEE Security and Privacy Workshops*, 89-95.
2. Akoglu, L., Chandy, R., and Faloutsos, C. (2013). Opinion fraud detection in online reviews by network effects. *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, 2-11.
3. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171-4186. <https://doi.org/10.18653/v1/N19-1423>
4. Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv:1702.08608*.
5. Feng, S., Banerjee, R., and Choi, Y. (2012). Syntactic stylometry for deception detection. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 171-175.
6. Feng, S., Xing, L., Gogar, A., and Choi, Y. (2012). Distributional footprints of deceptive product reviews. *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, 98-105.
7. He, S., Hollenbeck, B., Overgoor, G., Proserpio, D., and Tosyali, A. (2024). Detecting fake review buyers using network structure: Direct evidence from Amazon. *arXiv:2410.17507*.
8. Hussain, N., Mirza, H. T., Hussain, I., Iqbal, F., and Memon, I. (2019). Spam review detection techniques: A systematic literature review. *Applied Sciences*, 9(5), 987. <https://doi.org/10.3390/app9050987>
9. Jindal, N., and Liu, B. (2008). Opinion spam and analysis. *Proceedings of the 2008 ACM International Conference on Web Search and Data Mining*, 219-230. <https://doi.org/10.1145/1341531.1341560>
10. Li, J., Ott, M., Cardie, C., and Hovy, E. (2014). Towards a general rule for identifying deceptive opinion spam. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 1566-1576. <https://doi.org/10.3115/v1/P14-1147>
11. Lim, E.-P., Nguyen, V.-A., Jindal, N., Liu, B., and Lauw, H. W. (2010). Detecting product review spammers using rating behaviors. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 939-948. <https://doi.org/10.1145/1871437.1871557>
12. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*.
13. Maurya, S. K., Kumar, A., and Biswas, B. (2023). Deceptive opinion spam detection approaches: A literature review. *Applied Intelligence*, 53, 12276-12317. <https://doi.org/10.1007/s10489-022-03427-1>
14. Mukherjee, A., Kumar, A., Liu, B., Wang, J., Hsu, M., Castellanos, M., and Ghosh, R. (2013). Spotting opinion spammers using behavioral footprints. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 632-640. <https://doi.org/10.1145/2487575.2487580>
15. National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. <https://doi.org/10.6028/NIST.AI.100-1>
16. Ott, M., Choi, Y., Cardie, C., and Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 309-319.
17. Pan, L., Pan, W., Chen, W., Nakov, P., Kan, M.-Y., and Wang, W. Y. (2023). On the risk of misinformation pollution with large language models. *Findings of the Association for Computational Linguistics: ACL 2023*, 1389-1403.

18. Paul, H., and Nikolaev, A. (2021). Fake review detection on online e-commerce platforms. *Data Mining and Knowledge Discovery*, 35, 1830-1881. <https://doi.org/10.1007/s10618-021-00772-6>
19. Perez-Rosas, V., and Mihalcea, R. (2015). Experiments in open domain deception detection. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1120-1125. <https://doi.org/10.18653/v1/D15-1133>
20. Rayana, S., and Akoglu, L. (2015). Collective opinion spam detection: Bridging review networks and metadata. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 985-994. <https://doi.org/10.1145/2783258.2783370>
21. Ren, Y., and Ji, D. (2017). Neural networks for deceptive opinion spam detection: An empirical study. *Information Sciences*, 385-386, 213-224. <https://doi.org/10.1016/j.ins.2017.01.015>
22. Ren, Y., and Zhang, Y. (2016). Deceptive opinion spam detection using neural network. *Proceedings of COLING 2016*, 140-150.
23. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206-215. <https://doi.org/10.1038/s42256-019-0048-x>
24. Shajalal, M., Atabuzzaman, M., Boden, A., Stevens, G., and Du, D. (2024). What matters in explanations: Towards explainable fake review detection focusing on transformers. *arXiv:2407.21056*.
25. Stanton, G., and Irissappane, A. (2019). GANs for semi-supervised opinion spam detection. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 5204-5210. <https://doi.org/10.24963/ijcai.2019/723>
26. U.S. Federal Trade Commission. (2024). Trade regulation rule on the use of consumer reviews and testimonials. *Federal Register*, 89(161), 68035-68166.
27. Wang, G., Xie, S., Liu, B., and Yu, P. S. (2011). Review graph based online store review spammer detection. *Proceedings of the 11th IEEE International Conference on Data Mining*, 1242-1247.
28. Ye, J., Kumar, S., and Akoglu, L. (2016). Temporal opinion spam detection by multivariate indicative signals. *Proceedings of the 10th International AAAI Conference on Web and Social Media*, 743-746. <https://doi.org/10.1609/icwsm.v10i1.14801>