Academic Journal of Management and Social Sciences
ISSN: 2958-4396 | Vol. 15, No. 2, 2026

Fake Online Reviews on Digital Platforms: Conceptual Boundaries, Detection Methods and Trustworthy Governance

Xingying Feng [1, a], Rui Wang [1, b], and Jiaming Lu [2, c]

[1] School of Management, Chongqing University of Technology, Chongqing 401311, China
[2] School of Big Data, Fuzhou University of International Studies and Trade, Fuzhou, Fujian, 350202, China
[a] 1217584160@qq.com, [b] wangr@cqut.edu.cn, [c] lujiaming@fzfu.edu.cn

Abstract: Digital evaluation has become one of the necessary contents of the market information system, but the same way that it addresses information asymmetry may also give people reasons to distort their reputations. Gather the results of the 64 investigations on forged web feedback from the survey, and then, based on conceptual boundaries, obvious signs, identification systems and reliable applications, classify them. According to the assessment, there are now disputes over what should be regarded as fake reviews; at the same time, data collection, attribute accuracy and algorithm benchmarks have also been subject to debate. Linguistic patterns are still used in the first stage of screening at the review stage, but now reliable detection is more often based on reviewer patterns, chronological spikes and collaborative dynamics due to the systematic nature of deception in platform engagement architecture. Supervised, unsupervised and semi-supervised methods should be considered different ways to apply various works, not a hierarchy of efficiency. Interpretability is also being used in the regulatory process to connect the results of algorithms with human review, appeals and changes in policy and corporate social responsibility, etc. Based on the above, this paper will introduce a general architecture for object demarcation, deception patterns, visible traits, identification structure and explanatory oversight. Next, we will carry out boundary-aware collection, construct an adaptable cross-website model, perform implementation-oriented evaluation and human-centric management.

Keywords: Fake Review; Online Review; Deceptive Opinion Spam; Feature Engineering; Detection Method; Explainable Artificial Intelligence.

1. Introduction and Design of Review

1.1. Research Background and Problem Statement

Electronic feedback is no longer a small comment added to virtual discussions; it is now an advertisement for distributing attention, status and confidence. Contrasting with vendor-initiated promotion, the endorsement is shown as the experience of customers, and it is integrated into the browsing, suggestion and purchase path. Earlier studies have found that web reviews serve both to provide facts and to mediate information, and the number of reviews, their tone and consistency, etc., can motivate consumers to purchase in different ways [1, 2].

Therefore, there is the same signal system to motivate exploitation. As the rating indicators and the level of placement are exchangeable for visitors' revenue, etc., businesses, shoppers and external vendors may intentionally falsify the evaluations. Yelp data shows that the reasons for a dubious listing are often related to a person's sense of self-worth, local competition, etc. [3] Research in China has also shown that affective catharsis, financial incentives, superficial responses and contracted review agents can be the reasons for the generation of fake reviews [4]. Fabricated listings are thus not merely false words; they are attempts by the platform to manipulate the public's perception.

The Regulations have many problems. Customers will be more hesitant to pay for the appraisal and experience reports. Good businesses will have to compete for prestige in a different way. Hosts are also in the problem of a feedback mechanism; therefore, the accuracy of discovery, recommendation and business cooperation is low. Deceptive testimonials are gradually taking the form of believable experience stories and are no longer obvious spam. Word production, rephrasing and multi-user synchronisation have changed from being a single review phrase to a pattern of behaviour, time and connection [5]. Therefore, the whole literature review should link terminology, detectable indicators, identification methods and management applications.

Although there are some current collections, they are often separate due to the narrowness of focus. An assessment of the digital feedback study has been carried out to clarify the report production, interaction, influence and site architecture of these studies, and generally, attributes, collections and artificial intelligence algorithms are employed to detect fake comments [6, 7]. The reason for this difference is not that more articles have been published; rather, it is that the theoretical limits, attribute engineering, selection of techniques and explanations of their applications have rarely been examined together in a single investigation.

1.2. Literature Collection and Review Design

A system of reviews was applied to classify Chinese and English studies on online reviews and the identification of false reviews. Inquiry keywords include online review, digital word-of-mouth, fake review, misleading opinion spam, opinion spam, review spam identification and interpretable artificial intelligence, etc., along with their corresponding Chinese translations. In addition to the above, backward and forward citation mapping of frequently cited papers and

recent research studies were conducted. Only if it can offer some support to at least one of the five areas in conceptual framing, manipulation incentives, feature indicators, detection techniques, dataset/evaluation and interpretative deployment will research be carried out.

Remove redundancy and filter by topic to get 64 individual articles at the end. This group was not arranged in order. Therefore, the five types of evidence are as follows: conceptual boundary research, incentive and market-effect studies, feature analysis of text, actions, duration and connections, methodological research on supervised, unsupervised and semi-supervised identification, and interpretability research related to human oversight and platform regulation. Therefore, it is easier to organise the materials by the investigative question.

Figure 1 is the Evaluation Method. First, academic references will be collected and qualified; then, thematic evidence will be gathered; Finally, this data will be organised in an organised way and connected with object limits, manipulation techniques, signal architecture, detection models, interpretive governance, etc.

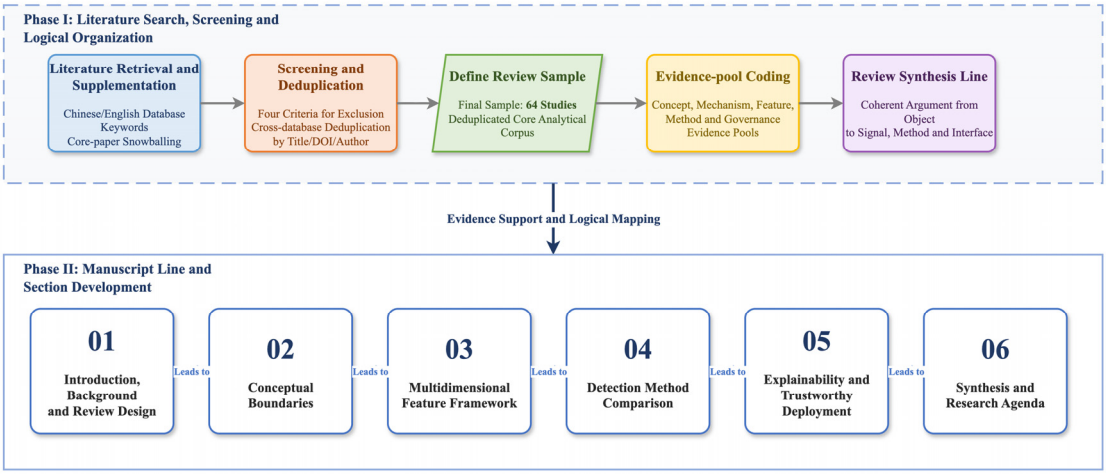


Fig 1. Logic of Literature Screening and Review Organisation.

Table 1. Positioning of This Review in Relation to Existing Review Strands.

Review strand	Representative evidence	Main focus	Contribution	Remaining gap
General online review literature	Online review and eWOM studies [2]	Creation, engagement, helpfulness, market effects and platform design	Explains why reviews affect consumer decisions and platform value	Fake review boundaries and detection logic remain dispersed
Technical fake-review surveys	Detection-oriented surveys [6, 7]	Features, datasets, algorithms and reported performance	Summarizes computational approaches and benchmark limitations	Concept definition, governance interpretation and feature-method fit are often underdeveloped
Current synthesis	Concept, feature, method and governance evidence	Boundary-sensitive review of fake online reviews	Connects detection research with platform trust and accountable deployment	Emphasizes cumulative comparison rather than isolated model ranking

1.3. Review Positioning and Article Structure

The three additions to the prior research are as follows: First, it believes that synthetic evaluations are a problem of platform fairness rather than a problem of limited text classification. Secondly, it can be seen that the linguistic, behavioural, chronological and interactive attributes of a certain type are restrictions on selecting supervised, unsupervised and semi-supervised methods. Finally, to achieve the goal of a system for identifying fraud, it should be possible to explain why a certain entity has been identified, for instance, through an audit trail. Section 2 is an introduction. Section 3 is the feature indicator collection. Section 4 Contrasts Identification Frameworks. Section 5 is about the interpretability and reliable application. Section 6 introduces the overall framework and the research direction of the following section.

2. Conceptual Boundaries and Research Foundation

The first is to determine the scope of counterfeit evaluation. If digital feedback, virtual recommendations, false information, paid endorsements, irrelevant junk, etc., are all given the same importance, then the structure of reliability and analysis of these indicators will be unbalanced. Therefore, we need to refine the concept further before conducting the research.

2.1. Online Reviews and Electronic Word-of-Mouth

Digital feedback is a person's evaluation story and the corresponding metadata, such as scores, time stamps, identity information, engagement records, etc. Electronic word-of-mouth is more spread out; it refers to the spread and influence

of buyers' views in interconnected areas. Based on Chinese studies of consumer feedback, at present, the main reasons that make people willing to accept online complaints are practicality and reliability [8]. The identification system mainly includes the review case itself; generally speaking, many talks about the business applications of eWOM are spread throughout various stages of life's events.

Table 2. Core Concepts and Analytical Boundaries.

Concept	Analytical meaning	Typical evidence	Relation to fake reviews	Boundary rule
Online review	Platform-based evaluation text with metadata	Text, rating, timestamp, reviewer and feedback records	Direct object of detection	Use for observable review units
Electronic word-of-mouth	Networked opinion diffusion and influence process	Adoption, persuasion, helpfulness and market response	Explains why fake reviews matter	Use for impact mechanisms, not for detection labels
Opinion spam	Broad early term for misleading or irrelevant opinion content	Untruthful opinions, brand-only comments and advertisements	Historical umbrella concept	Avoid equating all spam with fake reviews
Deceptive review	A review that appears authentic but is fictitious or intentionally misleading	Linguistic and psycholinguistic deception signals	Important subtype of fake reviews	Best suited to text-centered tasks
Incentivized review	Review produced under rewards, discounts or compensation	Disclosure, motivation and consumer-brand relationship	Boundary case	Only fake when undisclosed distortion misleads users

2.2. Definition, Types and Boundary Cases of Fake Reviews

In the beginning, "review spam" and "opinion spam" were used to refer to this problem. Jindal and Liu have divided them into three types: false opinions, assessments of brands and vendors instead of goods, and promotions or unrelated posts [9, 10]. The category is relatively wide-ranging; therefore, it has gained popularity but also some deliberate deception. Fake reviews are defined as false opinions or organised systems designed to deceive consumers and the platform, and they are the subject of this survey.

A different kind is the category of deceptive opinion spam. Ott and others have studied fabricated critiques that are designed to look genuine and have found that n-gram, psycholinguistic and genre-specific indicators can distinguish crowdsourced fraudulent hotel reviews from sincere ones [11]. The above framework is suitable for text classification, but it cannot cover all kinds of platform alterations. Vindictive negative feedback, organised rating assaults and vendor-funded marketing will have different motives and display differently.

Compensated Feedback is a Boundary Problem. Rewards, discounts or incentives can alter the review-production process, but they are not deception itself. Empirical data show that bribed fake and authentic favorable evaluations can make people feel guilty and reduce their satisfaction with the brand, and honest incentivised comments work differently [12]. Therefore, compensated reviews should be considered a risk situation; only if there is non-disclosure and the encounter is falsely reported or significantly misleading to various consumers does it fall under the category of false reviews.

2.3. Consequences of Review Synthesis

This division will be divided for the next step. Misleading Opinion Spam is based on language form, style and meaning. Organiserkan kan avdekke mønstre i anmeldernes arbeid, tidsforståelser, etc., samt andre delte anmelder-parametere. It has not been disclosed that the reasons for the review process are not transparent, and thus the platform is unable to provide supporting data. As a result, this division is more than just the first division; it explains why different studies have drawn separate conclusions and collected information.

3. Multivariate Feature Model

The chosen attributes will have a certain modifying effect. The section on language indicators in reviews has been moved to a secondary position, and now a multi-level structure that includes text, reviewer behaviour, time patterns and social dynamics is used [11, 6]. Fig. 2 is the structure of this many attributes, and Table 3 shows the first signal types.

3.1. Textual Features

The characteristics of text will continue to be used for a beginning point because fraudulent criticism is regarded by users as spoken words. Ott and his colleagues have found that local lexical motifs are better than general deceptive-opinion lexicons and syntactic genre indicators in a regulated deceptive-opinion corpus [11]. It was not found that there are general dishonest words, but rather that fabricated testimonials often contain redundant expressions, exaggerated emotions and insufficient context-specific information.

The flaws of pure text identification were soon evident. Li and others have carried out cross-domain research to find that a model trained in one domain may learn the style of deception for products in that domain rather than general deception methods [13]. There is also a problem of large language model rephrasing; although the superficial templates look more elegant, there may still be a consistency in the creation patterns of AI-generated text [5]. Linguistic attributes can be used to narrow down the initial scope, but they are too weak to be relied upon as sole evidence of platform identification.

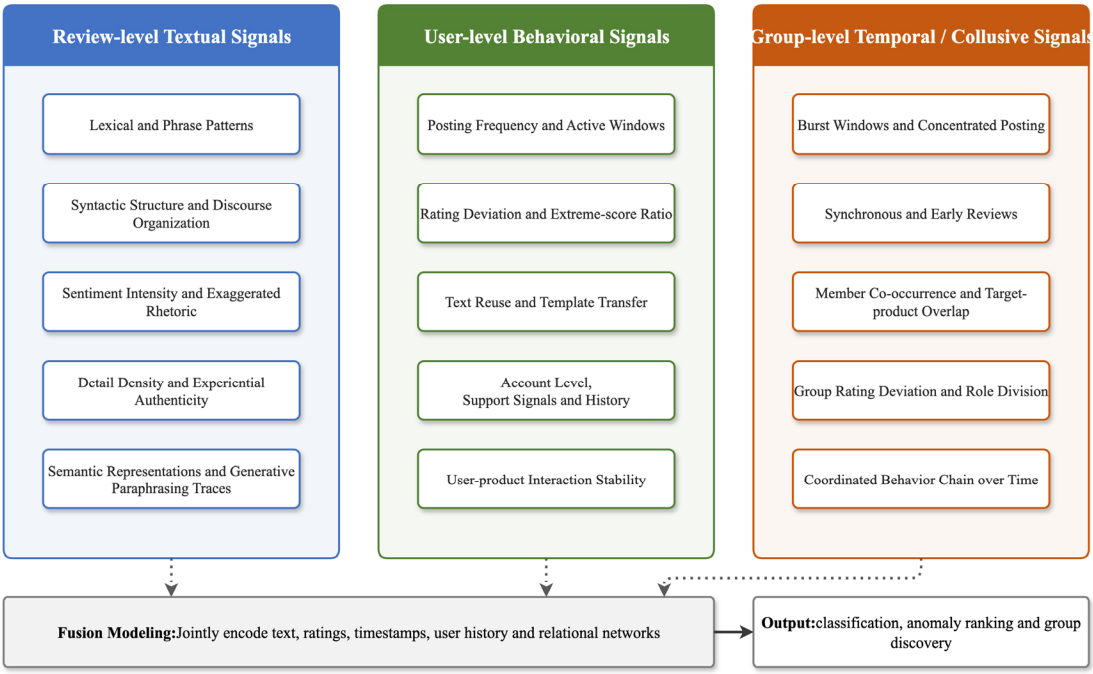


Fig 2. Multidimensional Feature Framework for Fake Online Reviews.

3.2. Reviewer Behaviour Features

Behavioral indicators have altered the direction of research, and now it is not only about the authenticity of the document but also about whether the profile behaves like a typical consumer. The frequency of posting, variation in ratings, proportion of high/low scores, activity periods, contribution records and peer feedback can all reveal the cost-effective patterns of review fabrication. Studies on trustworthiness in China's online commerce have shown that to assess the reliability of reviews, one can consider combining reviewer status, endorsement cues and viewpoint coherence [14].

The Behavioural Data will be analyzed. Further research has shown that real users will also modify their ratings after a long time of use, service failure, etc. Research on supplementary feedback indicates that coherence, discrepancy, quality and timing are all related to the sense of usefulness and the intention to purchase [15-16]. Therefore, short-term intensity or score fluctuation does not have to be fraud. Behaviors will be more easily considered unreliable content in the chronology of an account by the platform.

3.3. Fusion of Review Text and Reviewer Behaviour

Univariate Attributes are weak in the Environment of actual Platforms. Skilled deceivers can produce reasonable materials; genuine people will have some abnormal behaviour temporarily due to the campaign or technical problems. An integrative way to deal with the problem is to show how the content is distributed, who the author is, when it was released, what kind of relationship the profile has with items and other assessors, etc.

SpEagle is a typical integration method that can consider the text, timestamps, scores and interactional information of participants to spread suspicion among them, evaluations and items [17]. Research on the development of features has also been divided into two directions: review-oriented and reviewer-focused indicators, and affective expressions, direct attributes and activity patterns have been added [18]. At the same time, domestic research has also combined review text and reviewer behaviour in the prioritized multi-dimensional attributes and dual convolutional pipelines [19]. Synthesis is to build the platform structure of the interference itself, not to add more variables.

3.4. Time and Group Manipulation Signals

A regular assessment will be held in the group. Reviewers who wish to increase or reduce the importance of a particular item must do so over a short period, adjust their scores, and periodically check the same set of goods. Therefore, Mukherjee and others did not use the single-reviewer method but started to employ frequent itemset mining, behavioural pattern analysis, etc., to discover clusters of fraudulent reviewers [20]. Burstiness analysis is also a type of network indicator for the authors who appear together in clustered phases [21].

Time is to be expected, and clever spammers may try to hide a regular pattern but still show promotion or deletion. Frameworks for the identification of trustworthy fake reviews have added explicit chronological and behavioural structure to detect temporary manipulation [22]. To solve the problem of rating anomalies, a time-series method can be used to handle the textual data and user behaviour at the time of the anomaly [23]. Nevertheless, there may be spikes in festivals, discounts and popular culture. A chronological proof will be more convincing if it can be based on the foundation of objective overlap, textual similarity and co-review architecture.

4. Detection Method System

Due to the difference in platforms, there have been many ways of identification. Few platforms have many good labels, detailed metadata and uniform review rules at the same time. Supervised methods are employed to classify data that has been annotated; Unsupervised methods are to find outliers in the absence of labels, and semi-supervised algorithms use a small number of labelled examples to guide the learning of a

large amount of unlabeled data [24, 7].

Table 3. Main Feature Signals and Their Detection Implications.

Feature type	Observation unit	Typical signals	Main value	Main limitation
Textual features	Review	n-grams, LIWC, syntax, semantics, writing style	Directly screens review-level plausibility	Weak cross-domain transfer and vulnerable to paraphrasing
Reviewer behavior	Account	Frequency, rating deviation, history, helpfulness and feedback	Reveals persistent manipulation routines	Requires metadata and may confuse genuine activity changes
Text-behavior fusion	Review-account-product	Text, rating, timestamp, account history and relation links	Connects content plausibility with user suspiciousness	Sensitive to missing platform logs and cross-view noise
Temporal and group signals	Group or event window	Bursts, synchronization, target overlap and co-review networks	Detects organized manipulation	May resemble promotions or organic popularity shocks

4.1. Supervised Detection

The old way established a first good experimental system: classify evaluations, derive attributes and have a model predict whether a critique is fake. Li and the others expanded on this model to add review, reviewer and item data, and they pointed out that deceivers often produce spam uniformly in the comments [24]. Patterns have been applied to many dimensions and are therefore very precise; tagging is still an issue affecting reliability.

The first problem is that the annotated fraudulent reviews are diverse objects. Crowdsourced deceptive-commentary repositories may have high precision in their own domains, but cross-domain tests have shown that many of the attributes are related to specific genre styles rather than deception itself [11, 13]. Therefore, later supervised studies have developed more complex models that combine text and behaviour, layered attention and class dataset imbalance handling [19, 25, 26]. Supervised identification is still feasible in the event that the tags are uniform, but it will not work well when those tags are constrained by the platform's algorithm or a specific data collection method.

Figure 3a is supervised detection, and it is a sequence of consistent labels, features, categorisation and hazard prediction. The Structure will be functional if the labelling Strategy is reliable; when there are changes in the protocol, the model's results will adjust.

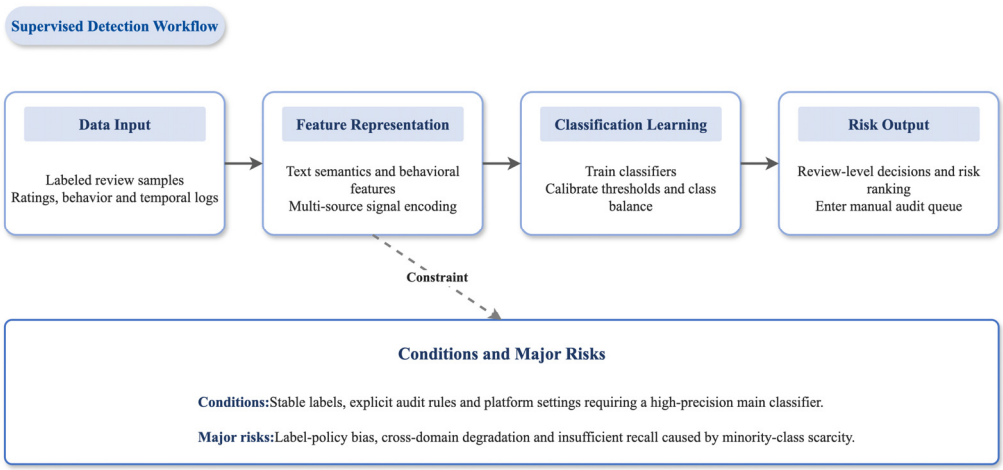


Fig 3a. Typical Workflow of Supervised Fake-Review Detection.

4.2. Unsupervised Detection

Unsupervised learning is suitable when there are no good labels. At this time, these methods do not know that the comment is fake; rather, they will identify that the user, person or group has left the normal platform environment. Dynamic clustering is used to find the abnormal clusters in the attributes of reviews and inter-comment relationships that do not have a fixed correct-or-incorrect threshold [27]. Heterogeneous graph frameworks evaluate the credibility of reviewers, the authenticity of reviews and the store, etc., through connections among reviewers, evaluations and vendors [28].

Unsupervised Detection will be applied to solve the above problem. Group-discovery methods are synchronous review clusters that do not have individual profiles [20], and at the same time, other strategies have also studied web feedback fraud from the perspectives of behaviour and structure [29]. Unsupervised research has also found that transferable trust indicators can be obtained from Yelp-like datasets in the present day [30]. The restriction is both theoretical and practical; it is not a lack of construction. Marketing, time pressure and the true spread of information may have uneven developments. Therefore, the results of the unsupervised method should be used as hazard ranking tools and not as final

answers.

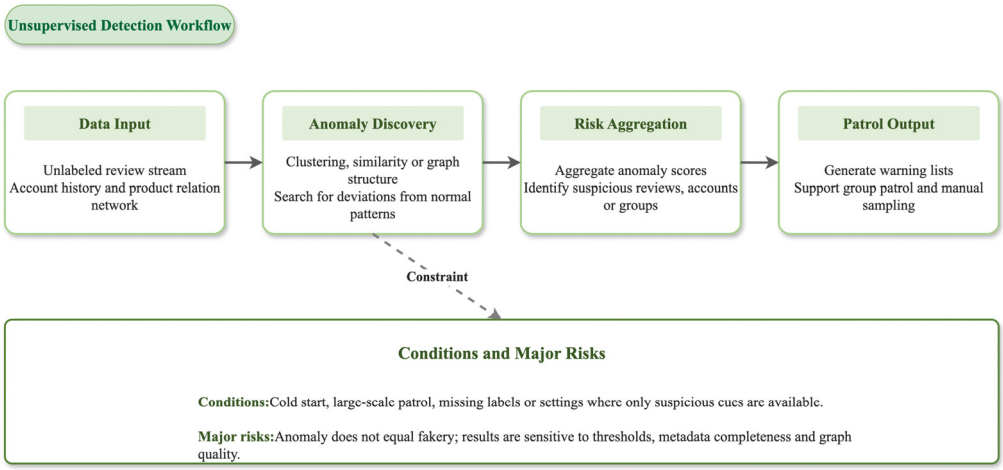


Fig 3b. Typical Workflow of Unsupervised Fake-Review Detection.

4.3. Semi-supervised Detection

Semi-supervised methods are in line with the actual situation of a platform; there is little trustworthy annotation data, but we cannot ignore all annotations. Co-training is based on the joint information of review text and reviewer behaviour to add more training data from unsupervised sources [24]. At the same time, domestic research has also been conducted combining computational linguistics with psycholinguistic features in semi-supervised schemes [31] and built co-training frameworks based on term characteristics and deep syntactic rules [32].

PU learning is very effective when the system can recognise some positive and dubious examples but does not have the security to classify the rest of the data as negative. Ren and others considered the misclassified boundary examples to be part of the problem and used reliable negative filtering before preparing the classifier [33]. Later, in order to reduce the error of pseudo-labels, researchers have employed vertical ensemble Tri-training, generative adversarial networks and reviewer-network propagation [34-36]. Semi-supervised methods still have a labelling problem; only now do we have weak-label management.

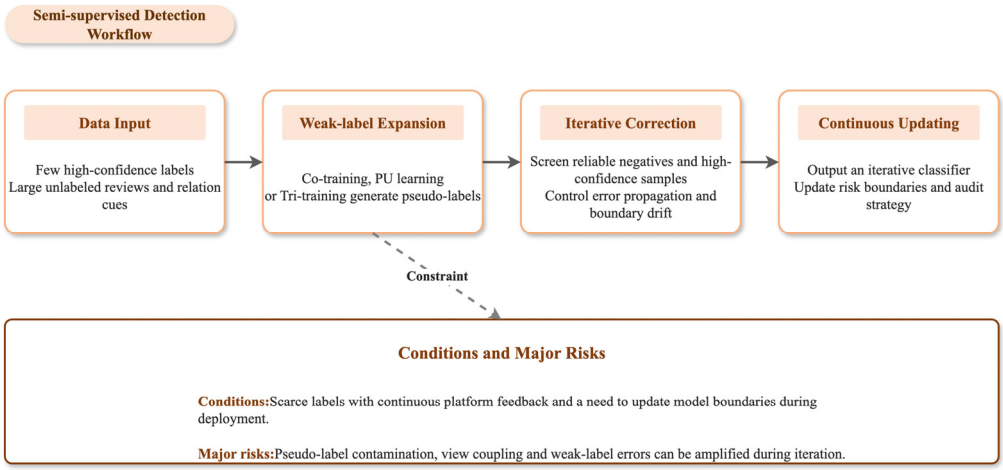


Fig 3c. Typical Workflow of Semi-supervised Fake-review Detection.

4.4. Datasets, Metrics and Method Comparison

The data sets are not the same, so the methods of study are different. Structured crowdsourced data have clear labels, but they may be the result of composition instructions rather than commercial interests. Yelp-approved reviews are still somewhat closer to reality and more in line with the policies of the platform [37], so they are considered platform-moderated information. Custom-built collections are suitable for some enterprises but cannot be applied to all institutions. Accuracy is misleading when there are few counterfeit assessments. Generally speaking, Precision, Recall, F1 score, AUC, top-k accuracy and human-validation consistency are all applicable to the problem described above.

5. Explainability and Trustworthy Deployment

Given that the goal of identifying fake reviews is to build a regulatory system, it should not only predict them. A system often provides the reasons for the penalty of a comment, why a profile was flagged, and why a set of reviews entered manual review. The evidence is based on a few studies of fake reviews and related research on spam accounts, fake identities and forensic artificial intelligence [14, 36, 38, 39, 40]. Figure 4 is the cycle of explainability, and it includes model results, explanations, human supervision, platform intervention and rule modification.

Table 4. Comparison of the Main Detection Paradigms.

Paradigm	Label condition	Main strength	Main constraint	Suitable deployment role
Supervised	Requires stable labeled data	Strong review-level classification under known policy	Sensitive to label bias, domain shift and class imbalance	Main detector when audit labels are mature
Unsupervised	Does not require true-false labels	Discovers anomalies, groups and emerging patterns	Anomaly does not equal fakery; threshold selection is unstable	Patrol, cold start and candidate recall
Semi-supervised	Uses limited reliable labels with large unlabeled data	Balances guidance and scalability	Pseudo-label errors and view dependence can accumulate	Continuous learning under limited audit feedback

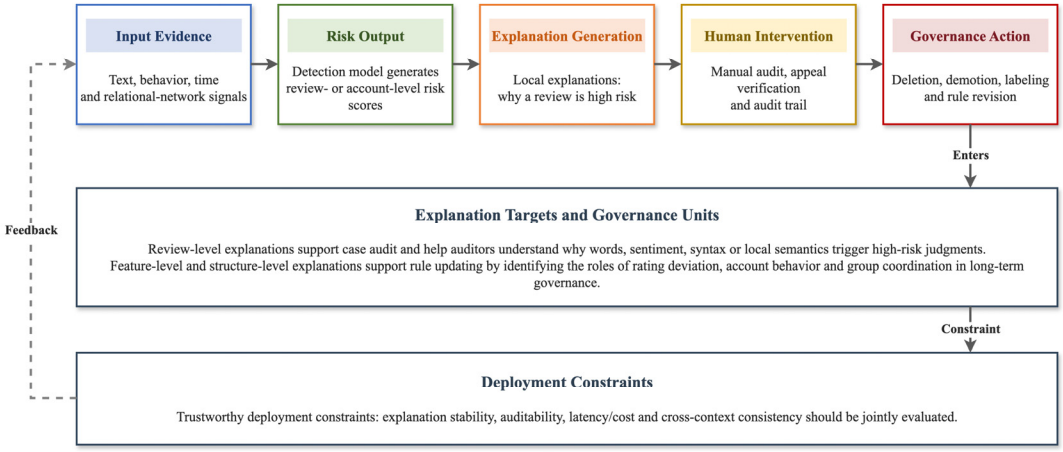


Fig 4. Explanation of the Governance Loop for Fake-Review Detection.

5.1. Explanation Targets and Granularity

Clarifications can be made on different parts. Review-based clarifications identify terms, expressions of emotion and depictions of aspects that may be the objects of criticism. Feature-oriented clarification is the quantification of the effect of score deviation, behaviour windows, user information and other cues. Framework-based clarification refers to the situation where a person or group of people are under suspicion due to a common review link, network proximity or label diffusion [36]. Clarification serves the purpose of modification and evaluation, such as the assessment of people and models, policy adjustment, etc.

5.2. Local and Global Explanations

Local explanations are the reasons for the occurrence of a particular case. They are useful for human audits, the collection of evidence and appeals, etc., and can be connected to particular results. Global explanations know which patterns the framework usually employs and are more in line with policy understanding. Research on SHAP and LIME in the field of forensics and synthetic profiles has shown that local and global explanations can be used to determine the consistency of explanations, the weight of features and practical application value [38, 40]. Efficient explainable spam filtering also needs to be improved in terms of feature refinement so as not to have too high a functional cost [39].

5.3. Deployment Constraints

Explainability is not a guarantee of the reliability of application. Post-hoc justifications may not be in line with changes in the data profile, and generally, they are too simplistic to show the small coordinated attacks. There will be no problems when the rules of the platform are too explicit, either. To enhance the fairness of the management of fake reviews, one can specify the reasons for disabling a particular review, provide an overview of the regulatory changes, and submit organised deception reports. The aim of reliability is to make decisions better and responsibly; it is not about having a graph.

6. Synthesis, Research Agenda and Conclusion

6.1. Convergences and Disagreements

Many parts have been agreed upon in academic research. Theories of Bounds are Sets of Data and Performance Evaluations. Only text identification is not standardised. As deception is organised in the form of profiles and engagement, action-based, chronological and associative indicators have gradually been given more weight. Now we will use supervised, unsupervised and semi-supervised methods instead of the previous three.

Still different, however. Fake reviews may contain false experiences, biased opinions, malicious comments, financial fraud, etc. Databases based on collective sourcing, system screening and academically developed subsets are not the same in terms of the subjects of observation. Therefore, the framework in the same environment will be able to learn the particular classification protocol and avoid general interference. The whole thing needs to be specified for the sides.

6.2. Integratve Framework

Figure 5 is composed of the five strata in the domain. The object boundary is the scope of the evaluation, profile or group being studied. Manipulation indicates the root reason for the deviation. Observable signals are transformed by the above mechanism into text, behaviour, time and other attributes. According to the amount of data available, one of the three modes is selected. Explanatory governance will decide whether the model's output can be used in the audit, appeal, policy adjustment, etc. Therefore, all studies that aim to find the cause of fake reviews may have different results.

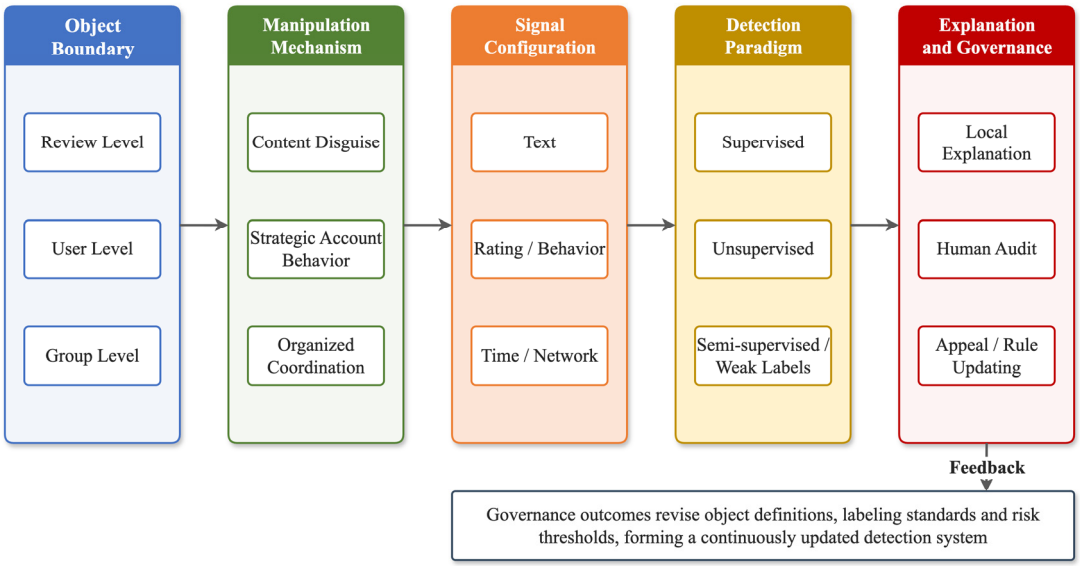


Fig 5. Integrative Framework for Fake Online Review Research.

6.3. Future Research Agenda

Next, we will address the problem of dynamic precision benchmarks. A boundary-aware dataset needs to be built, and the annotations for the different types of reviews, users and groups must be saved independently. With the modification of ranking algorithms and generation software, many platforms will have to be adjusted repeatedly to meet the new ban on false promotion. Calibration, recall for the minority class, robustness, interpretability stability and human-auditor agreement are also to be considered in the evaluation. In the end, interpretability should be assessed within the context of governance and not merely be a secondary attribute of classification.

Table 5. Future Research Agenda and Practical Implications.

Agenda	Research need	Operational implication	Expected contribution
Boundary-sensitive data	Separate review, account and group labels	Clearer audit policy and dataset documentation	More comparable evidence across studies
Dynamic modeling	Track manipulation across time, platforms and model updates	Early warning for strategy drift and coordinated bursts	Higher robustness under changing platform ecology
Deployment-oriented evaluation	Report recall, calibration, top-k utility and audit agreement	Align metrics with operational risk control	Less reliance on misleading aggregate accuracy
Human-in-the-loop explanation	Test whether explanations improve audit and appeal quality	Create traceable evidence for moderation actions	Accountable and contestable platform governance

6.4. Conclusion

Phony web evaluations have gone from being about the dubious content to a problem of platform trust. The primary aim of the region is not to find an ideal general algorithm, but rather to coordinate the purpose of objective presentation, indications for evidence, models of methodology and regulations. Linguistic characteristics are still necessary, but more reliable identification can now be based on behaviour, time and network structure. Supervised algorithms are good at addressing normal tags, unsupervised methods can be employed to detect anomalies, and semi-supervised methods are often very convenient for dealing with rare but stable audit data. The next stage of research should be to investigate the stability of the screening framework in keeping the review ecosystem stable, and not just how much it can raise the benchmark index.

References

[1] Park, D.-H., & Lee, J. (2008). eWOM overload and its effect on consumer behavioral intention depending on consumer involvement. *Electronic Commerce Research and Applications*, *7*(4), 386–398. https://doi.org/10.1016/j.elerap.2007.11.004.

[2] Pocchiari, M., Proserpio, D., & Dover, Y. (2023). Online reviews: A literature review and roadmap for future research. *SSRN Electronic Journal*. https://doi.org/10. 2139/ssrn. 456 5563.

[3] Luca, M., & Zervas, G. (2013). Fake it till you make it: Reputation, competition, and Yelp review fraud. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2293164.

[4] Zhang, W., Wang, Q., Ma, Z., Li, J., & Xie, R. (2022). Research on the publishing motivation and formation mechanism of online commodity fake reviews. *Chinese Journal of Management Science*, *30*(7), 176–188. (In Chinese). https://doi.org/10.16381/j.cnki.issn1003-207x.2019.2022.

[5] Xylogiannopoulos, K. F., Xanthopoulos, P., Karampelas, P., & Bakamitsos, G. A. (2024). ChatGPT paraphrased product reviews can confuse consumers and undermine their trust in genuine reviews. Can you tell the difference? *Information Processing & Management*, *61*(6), 103842. https://doi.org/10. 1016/j. ipm.2024.103842.

[6] Mohawesh, R., Xu, S., Tran, S. N., Ollington, R., Springer, M., Jararweh, Y., et al. (2021). Fake reviews detection: A survey. *IEEE Access*, *9*, 65771–65802. https://doi.org/10. 1109/access. 2021. 3075573.

[7] Gupta, R., Jindal, V., & Kashyap, I. (2024). Recent state-of-the-art of fake review detection: A comprehensive review. *The Knowledge Engineering Review*, *39*. https://doi.org/ 10.1017/ s0269888924000067.

[8] Song, X., & Sun, X. (2015). A review of consumer online review adoption research. *Journal of Modern Information*, *35*(1), 164–169. (In Chinese). https://doi.org/10. 3969/j.issn. 1008-0821.2015.01.032.

[9] Jindal, N., & Liu, B. (2007). Review spam detection. In Proceedings *of the 16th International Conference on World Wide Web* (pp. 1189–1190). https://doi.org/10.1145/ 1242572. 1242759.

[10] Jindal, N., & Liu, B. (2008). Opinion spam and analysis. In Proceedings *of the International Conference on Web Search and Web Data Mining - WSDM '08* (p. 219). https://doi. org/10. 1145/ 1341531.1341560.

[11] Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 309–319).

[12] Bozkurt, S., Welch, E., Gligor, D., Gligor, N., Garg, V., & Pillai, K. G. (2023). Unpacking the experience of individuals engaging in incentivized false (and genuine) positive reviews: The impact on brand satisfaction. *Journal of Business Research*, *165*, 114077. https://doi.org/10.1016/j.jbusres.2023.114077.

[13] Li, J., Ott, M., Cardie, C., & Hovy, E. (2014). Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1566–1576). https://doi.org/10.3115/v1/P14-1147.

[14] Li, C., Xiang, J., & Xiang, J. (2019). A credibility evaluation method for online product reviews. *Journal of Computer Applications*, *39*(1), 181–185. (In Chinese). https://doi.org/10. 11772/ j.issn. 1001-9081.2018051154.

[15] Wang, C., He, C., & Wang, K. (2015). Perceived usefulness of follow-up reviews in online word-of-mouth. *Management Science*, *28*(3), 102–114. (In Chinese). https://doi. org/10. 3969/j. issn.1672-0334.2015.03.009.

[16] Zheng, J., & Duan, Y. (2022). The influence of follow-up reviews on consumers' purchase intention on e-commerce platforms: Evidence from Taobao. *Business Economics Research*, (13), 80–84. (In Chinese).

[17] Rayana, S., & Akoglu, L. (2015). Collective opinion spam detection. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 985–994). https://doi.org/10.1145/ 2783258. 2783 370.

[18] 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.

[19] Yang, C. (2019). Fake review detection based on review text content and reviewer behavior features [Master's thesis]. Shanghai Normal University. (In Chinese).

[20] Mukherjee, A., Liu, B., & Glance, N. (2012). Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st International Conference on World Wide Web* (pp. 191–200). https://doi.org/10.1145/2187836.2187863.

[21] Fei, G., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., & Ghosh, R. (2013). Exploiting burstiness in reviews for review spammer detection. *Proceedings of the International AAAI Conference on Web and Social Media*, *7*(1), 175–184. https://doi.org/10.1609/icwsm.v7i1.14400.

[22] Wu, X., Dong, Y., Tao, J., Huang, C., & Chawla, N. V. (2017). Reliable fake review detection via modeling temporal and behavioral patterns. In *2017 IEEE International Conference on Big Data (Big Data)* (pp. 494–499). https://doi.org/10.1109/ bigdata. 2017.8257963.

[23] Heydari, A., Tavakoli, M., & Salim, N. (2016). Detection of fake opinions using time series. *Expert Systems with Applications*, *58*, 83–92. https://doi.org/10.1016/j. eswa.2016. 03. 020.

[24] Li, F., Huang, M., Yang, Y., & Zhu, X. (2013). Learning to identify review spam. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence* (pp. 2488–2493). https://doi.org/10.5591/978-1-57735-516-8/IJC AI11-414.

[25] Yan, M., Ji, D., & Ren, Y. (2019). Fake review identification based on a hierarchical attention neural network model. *Journal of Computer Applications*, *39*(7), 1925–1930. (In Chinese). https://doi.org/10.11772/j.issn.1001-9081. 2018 112 340.

[26] Du, S., Yang, M., & Qiu, R. (2023). Online product fake review identification based on SMOTE-RF and multidimensional feature vectors. *Journal of Intelligence*, *42*(4), 156–164. (In Chinese). https://doi.org/10.3969/j.issn.1002-1965.2023. 04. 022.

[27] Song, H., Yan, X., Yu, Z., Shi, L., & Su, F. (2013). Fake review detection based on adaptive clustering. *Journal of Nanjing University (Natural Sciences)*, *49*(4), 433–439. (In Chinese). https://doi.org/10.13232/j.cnki.jnju.2013.04.005.

[28] Wang, G., Xie, S., Liu, B., & Yu, P. S. (2011). Review graph based online store review spammer detection. In *2011 IEEE 11th International Conference on Data Mining* (pp. 1242–1247). https://doi.org/10.1109/icdm.2011.124.

[29] Lin, Y., Zhu, T., Wang, X., Zhang, J., & Zhou, A. (2014). Towards online review spam detection. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 341–342). https://doi.org/10.1145/2567948.2577293.

[30] Mothukuri, R., Aasritha, A., Krishna Chaitanya, M., Kesava Nandini, P., & Gopi Krishna, P. (2022). Fake Review Detection using Unsupervised Learning. In *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)* (pp. 119–125). https://doi. org/ 10. 1109/icscds53736.2022.9760908.

[31] Ren, Y., Ji, D., & Yin, L. (2014). Research on deceptive review identification based on semi-supervised learning algorithms. *Journal of* Sichuan *University (Engineering Science Edition)*, *46*(3), 62–69. (In Chinese). https://doi.org/10. 15961/j. jsuese. 2014.03.022.

[32] Zhang, W., Wang, Q., Bu, C., Li, J., & Zhang, S. (2020). Online fake review identification based on co-training. *Systems*

Engineering - Theory & Practice, *40*(10), 2669–2683. (In Chinese). https://doi.org/10.12011/1000-6788-2019-1089-15.

[33] Ren, Y., Ji, D., Zhang, H., & Yin, L. (2015). Research on deceptive review identification based on PU learning. *Journal of Computer Research and Development*, *52*(3), 639–648. (In Chinese). https://doi.org/10.7544/issn1000-1239.2015. 201 31473.

[34] Yin, C., & Zhu, Y. (2020). A fake review detection model based on vertical ensemble Tri-training. *Journal of Computer Applications*, *40*(8), 2194–2201. (In Chinese). https://doi.org/ 10. 11772/ j.issn.1001-9081.2019112046.

[35] Chen, J.-Y., & Wang, Y.-J. (2022). Semi-Supervised Fake Reviews Detection based on AspamGAN. *Journal of Artificial Intelligence and Capsule Networks*, *4*(1), 17–36. https:// doi. org/ 10.36548/jaicn.2022.1.002.

[36] Wu, Z., Liu, G., Wu, J., & Tan, Y. (2022). Are neighbors alike? A semi-supervised probabilistic ensemble for online review spammers detection. *SSRN Electronic Journal*. https://doi.org/ 10.2139/ ssrn.4013130.

[37] Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. (2013). What Yelp fake review filter might be doing?. *Proceedings of the International AAAI Conference on Web and Social Media*, *7*(1), 409–418. https://doi.org/ 10.1609/ icwsm. v7i1.14389.

[38] Hermosilla, P., Berríos, S., & Allende-Cid, H. (2025). Explainable AI for forensic analysis: A comparative study of SHAP and LIME in intrusion detection models. *Applied Sciences*, *15*(13), 7329. https://doi.org/10.3390/app15137329

[39] AlZeyadi, H., Sert, R., & Duran, F. (2025). A lightweight, explainable spam detection system with Rüppell's Fox Optimizer for the social media network X. *Electronics*, *14*(21), 4153. https://doi.org/10.3390/electronics14214153.

[40] Praveen Goud, V., & Morusupalli, R. (2025). SHAP for explainable feature engineering and LIME-based interpretability to enhance fake profile detection. In *Lecture Notes in Computer Science* (pp. 152–163). https://doi.org/ 10. 1007/ 978-981-95-4963-4_13.