Dr Raj Kumar, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

Methods, Challenges, And Future Directions For Detecting Fake Reviews

Dr Raj Kumar[1], Viklpa[2], Rimmy[3]

[1,2,3]Assistant Professor, Quantum University, Roorkee, India

Abstract- The influence of online consumer reviews on purchasing behavior, market visibility, and perceived credibility of digital platforms is significant. Consumers' evaluation of product quality, service reliability, and seller trustworthiness is heavily influenced by review ratings and textual opinions, as demonstrated by empirical studies. Small changes in aggregated ratings can have a significant impact on sales volume, search rankings, and long-term brand reputation. Online reviews have become high-value informational assets due to this dependency, which makes review systems attractive targets for manipulation. Economic and social consequences have been significant due to the increasing prevalence of fake and deceptive reviews. Unfair competitive advantages are gained by businesses who engage in review manipulation, while honest sellers suffer revenue loss despite offering comparable or superior quality. Consumers who are exposed to deceptive reviews face an increased risk of poor purchasing decisions, wasted expenditure, and reduced confidence in online marketplaces. Repeated exposure to fraudulent content at a large scale can erode platform credibility, weaken user engagement, and undermine the integrity of digital ecosystems constructed on user-generated content. In modern review environments, manual moderation mechanisms, such as user reporting and expert inspection, are not effective. The processing of millions of reviews daily by large platforms makes human-driven verification costly, inconsistent, and slow. To maintain trust, fairness, and scalability, automated fake review detection has become a crucial requirement. Despite the challenges, designing effective detection systems remains a challenge. To resemble genuine user opinions, fake reviews are often crafted in a way that is linguistically fluent, sentimentally plausible, and strategically crafted. To evade detection, deceptive reviews take advantage of subjectivity, context dependence, and social norms, unlike traditional spam.

Keywords: Fake review detection, opinion spam, online review fraud, deceptive text analysis, review authenticity, spam review classification, machine learning, deep learning, and natural language processing.

I. INTRODUCTION

Consumer decision-making in digital marketplaces is greatly influenced by online reviews. When direct product inspection is impossible, reviews are the primary source of information for many users. The influence of star ratings and written feedback on perceptions of quality, reliability, and value is often greater than that of price or brand familiarity. According to empirical analyses, even minor changes in average ratings can result in measurable changes in sales performance, search ranking, and recommendation exposure. Review systems become key economic mechanisms within platform ecosystems due to this sensitivity. Review manipulation has a significant financial impact. Products or services that are given artificially

inflated ratings gain greater visibility and consumer attention, causing demand to be diverted away from competitors who offer comparable or higher quality. Targeted negative reviews can suppress legitimate businesses, reduce revenue, and harm long-term reputation. The informational role that reviews are meant to provide is weakened and fair competition is undermined by such distortions. Users' confidence in platform rankings and recommendation systems is reduced over time due to persistent exposure to deceptive content.

Social and behavioral costs are also associated with fake reviews. Deceptive opinions can lead to consumers experiencing dissatisfaction, increasing return rates, and reducing trust in online transactions. The sustainability of the platform is threatened by the erosion of trust at scale, as users become less willing to rely on user-generated feedback. Platforms are required to strike a balance between openness and enforcement, ensuring that review systems are credible without dissuading genuine participation[1].

Automated fake review detection is now a necessity rather than an optional enhancement due to these risks. The volume and velocity of modern review streams prevent manual moderation strategies like user flagging and expert review from scalability. Continuous, large-scale monitoring can only be achieved through automated systems. Detecting fake reviews is difficult due to their deceptive nature. To avoid rule-based filters, fraudulent authors intentionally mimic authentic language, vary sentiment expression, and adjust posting behavior. The development of advanced detection systems that can learn subtle linguistic, behavioral, and relational patterns while maintaining robustness and fairness is motivated by this dynamic adversarial setting.

II. BACKGROUND AND SIGNIFICANCE

Early research on fake review detection was primarily based on rule-based and heuristic-driven techniques [1]. These approaches relied on manually crafted indicators such as duplicate or near-duplicate content, excessive sentiment polarity, abnormal posting frequency, bursty review behavior, and newly created reviewer accounts [2]. Such heuristics were computationally efficient and easy to deploy, making them suitable for early-stage review platforms. However, their effectiveness was largely limited to detecting naïve or low-effort spam. Simple paraphrasing, sentiment moderation, or temporal spreading of reviews was often sufficient to bypass these systems, revealing their inherent frag

ility [3]. With the growth of large-scale review datasets, machine learning methods emerged as a more flexible and data-driven alternative [4]. Supervised learning models incorporated linguistic features, syntactic patterns, sentiment scores, and reviewer metadata, combined with classifiers such as support vector machines and logistic regression [5]. These methods improved detection accuracy by learning statistical regularities across large corpora rather than relying on fixed rules. Subsequent work expanded feature representations to include behavioral histories, temporal activity patterns, and reviewer-product interaction statistics [6]. Despite these advances, performance remained highly dependent on feature engineering choices and domain-specific assumptions, which limited cross-platform applicability [7].

The adoption of deep learning introduced a major methodological shift in fake review detection research [8]. Neural architectures enabled automatic representation learning from raw text, reducing dependence on handcrafted features. Convolutional neural networks captured local lexical and syntactic patterns, while recurrent models learned sequential dependencies and discourse-level structures [9]. More recently, transformer-based architectures improved contextual modeling by capturing long-range semantic relationships within reviews [10]. In parallel, graph-based approaches modeled reviewer–item–time networks to identify coordinated manipulation and group-level fraud behaviors [11]. These techniques achieved notable performance gains on benchmark datasets and expanded the scope of detectable deception patterns.

Despite steady methodological progress, fake review detection remains fundamentally more challenging than conventional spam detection [12]. Reviews are inherently subjective, and genuine users often express strong emotions, brief opinions, or irregular posting behavior that overlap with deceptive signals [13]. This ambiguity increases false positives and complicates both model training and evaluation. Moreover, deceptive reviewers actively adapt their writing styles and behavioral strategies in response to evolving platform policies and detection systems, creating a dynamic adversarial environment [14]. As a result, fake review detection extends beyond static classification and represents an ongoing trust, security, and governance challenge for online platforms [15].

III. PROBLEM STATEMENT AND RESEARCH GAPS

Despite extensive research efforts and steady methodological progress, fake review detection remains an open and unresolved problem in real-world deployment scenarios [16]. Existing detection systems often demonstrate strong performance under controlled experimental settings but fail to generalize reliably across platforms, domains, and time. This gap between reported accuracy and operational effectiveness highlights several fundamental research challenges that continue to limit practical adoption.

One of the most critical challenges is the lack of reliable ground-truth labels [17]. Many widely used datasets rely on platform-generated filtering mechanisms or heuristic-based rules to identify fake reviews. These labels are inherently noisy, as platform filters are designed for moderation rather than scientific validation and may misclassify both genuine and deceptive reviews [18]. Crowdsourced datasets, while offering controlled labeling, often contain artificially generated deceptive reviews that do not reflect real-world fraud strategies [19]. The absence of verified and transparent labeling processes undermines model evaluation and limits reproducibility.

Class imbalance further complicates detection efforts [20]. In operational review systems, genuine reviews vastly outnumber fake ones, often by several orders of magnitude. Supervised learning models trained on balanced or artificially curated datasets fail to reflect this reality, leading to biased decision boundaries and inflated performance metrics. When deployed, such models tend to favor majority classes, resulting in missed fraud cases or excessive false positives [21]. Addressing imbalance without sacrificing sensitivity remains a persistent challenge.

Cross-domain generalization represents another major limitation [22]. Models trained on reviews from a specific platform, product category, or geographic region often perform poorly when applied to new domains. Differences in writing style, review norms, incentive structures, and platform policies introduce distributional shifts that static models cannot easily accommodate [23]. This dependency on domain-specific characteristics restricts scalability and increases maintenance costs for real-world systems.

Multilingual and cross-cultural review detection remains underexplored [24]. The majority of existing studies focus on English-language datasets, neglecting linguistic diversity and cultural variation in expression. Direct translation of models or features across languages is insufficient due to differences in syntax, sentiment expression, and pragmatic norms [25]. The lack of multilingual benchmarks and cross-lingual evaluation frameworks limits the applicability of current methods in global marketplaces.

Concept drift poses long-term risks to deployed detection systems [26]. Fraudulent reviewers continuously adapt their strategies in response to detection policies, enforcement actions, and platform feedback. As a result, statistical patterns learned from historical data gradually lose relevance. Most existing models are trained offline and evaluated on static datasets, with limited consideration of temporal robustness or adaptive learning [27]. Failure to address concept drift leads

to performance degradation and increased vulnerability over time.

Finally, the lack of explainability in high-performing models presents both technical and ethical concerns [28]. Deep neural and graph-based models often operate as black boxes, providing limited insight into why a review is classified as deceptive. This opacity hinders debugging, user appeals, and regulatory compliance, especially in environments where automated decisions affect business visibility or consumer trust [29]. The absence of interpretable mechanisms also reduces confidence among platform moderators and end users.

Collectively, these challenges indicate that fake review detection is not merely a classification task but a complex, evolving trust problem [30]. Addressing these research gaps requires advances in data collection, learning paradigms, evaluation methodologies, and explainable system design to bridge the divide between academic benchmarks and real-world deployment.

IV. AIM AND SCOPE

This review provides a systematic and critical analysis of fake review detection research published between 2008 and 2024, a period that captures the emergence, maturation, and diversification of the field [31]. The starting point reflects the earliest formal investigations into opinion spam, while the end point includes recent advances in deep learning, graph-based modeling, and robust learning frameworks. By covering this extended timeframe, the review traces how detection strategies have evolved in response to changes in platform scale, reviewer behavior, and adversarial sophistication.

The analysis focuses on three core dimensions of fake review detection: detection methodologies, benchmark datasets, and evaluation practices [32]. Methodological coverage spans rule-based heuristics, traditional machine learning with handcrafted features, deep neural architectures for textual modeling, graph-based approaches for coordinated behavior detection, and hybrid systems that integrate multiple signal types [33]. This breadth allows for comparative analysis across paradigms, highlighting trade-offs between performance, interpretability, and deployment feasibility.

Dataset coverage includes widely used public and semi-public corpora drawn from major platforms such as Amazon, Yelp, and TripAdvisor, as well as curated academic datasets designed for deception research [34]. The review examines how labeling strategies, data collection assumptions, and dataset construction influence reported results and generalization claims. Particular attention is given to the limitations of platform-filtered labels and artificially generated deceptive reviews, as these issues directly affect the validity of experimental conclusions [35].

Evaluation metrics and validation protocols are also analyzed in detail [36]. Commonly reported measures such as precision, recall, F1-score, and area under the ROC curve are reviewed alongside less frequently used metrics that capture class imbalance and operational risk. The review highlights inconsistencies in evaluation settings, including the widespread reliance on random train–test splits that ignore temporal ordering and domain shift, which can lead to overly optimistic performance estimates [37].

To maintain academic rigor, this review excludes non-peer-reviewed sources, proprietary industry reports, and studies lacking transparent methodology or reproducible experiments [38]. Research on unrelated forms of online fraud, such as click fraud, fake social media engagement, or bot detection, is also excluded unless directly linked to deceptive review behavior [39]. This selective scope ensures that the synthesized findings remain focused on opinion-based review manipulation while preserving methodological coherence.

By clearly defining temporal, methodological, and domain boundaries, this review aims to present a balanced and reliable synthesis of the field [40]. The resulting analysis supports meaningful comparison

across studies and provides a grounded foundation for identifying enduring research challenges and promising future directions in fake review detection.

V. METHODOLOGY

A systematic literature review was conducted using PRISMA-style reasoning to ensure methodological rigor, transparency, and reproducibility throughout the study selection process [46]. This structured approach was adopted to comprehensively capture relevant research on fake review detection while minimizing selection bias and omission of influential work. The review protocol was designed in advance and consistently applied across all stages of the literature search and analysis.

The identification phase involved executing carefully constructed search queries across major academic databases, including IEEE Xplore, ACM Digital Library, SpringerLink, Elsevier ScienceDirect, Scopus, and Google Scholar [47]. Keyword combinations were formulated to cover variations of fake reviews, opinion spam, deceptive online reviews, and review manipulation. This strategy ensured coverage of both early foundational studies and recent methodological advancements. Retrieved records from different databases were consolidated, and duplicate entries were systematically removed prior to screening.

During the screening phase, titles and abstracts were evaluated to determine topical relevance [48]. Studies that focused solely on sentiment analysis, recommendation systems, marketing analytics, or unrelated forms of online fraud were excluded. Remaining articles were subjected to full-text examination to assess their contribution to fake review detection, with particular attention given to clarity of problem formulation and methodological soundness.

The eligibility assessment applied explicit quality criteria, including transparency of dataset construction, validity of labeling strategies, appropriateness of evaluation metrics, and completeness of experimental reporting [49]. Studies lacking sufficient detail on data sources, preprocessing steps, or evaluation protocols were excluded to ensure reliability and reproducibility of the synthesized findings. Only peer-reviewed journal and conference paper meeting these standards were retained.

The final inclusion stage resulted in a curated corpus of studies that collectively represent the methodological and empirical landscape of fake review detection research [50]. This systematic selection process supports consistent comparison across approaches, enables identification of longitudinal trends, and provides a robust foundation for the thematic synthesis and critical analysis presented in this review.

VI. LITERATURE REVIEW

The paper synthesizes a broad range of interdisciplinary perspectives to present a unified view of fake review detection research [51]. At the theoretical level, it integrates insights from deception theory, behavioral psychology, and computational linguistics to explain why deceptive reviews differ systematically from genuine user opinions. Deception theory suggests that dishonest content often reflects increased cognitive load, which manifests through linguistic irregularities, exaggerated sentiment, and atypical discourse structures [52]. These theoretical foundations provide a basis for understanding why certain textual and behavioral signals recur across detection models.

In terms of data resources, the review examines widely used benchmark datasets drawn from large-scale e-commerce and service platforms, including Yelp, Amazon, and TripAdvisor [53]. These datasets have played a central role in shaping detection methodologies and evaluation practices. The review analyzes how platform-specific filtering mechanisms, user interaction patterns, and temporal dynamics influence dataset characteristics. It also highlights the limitations of commonly used datasets, such as reliance on noisy labels and lack of longitudinal validation, which affect the generalizability of reported results [54].

Methodologically, the paper synthesizes research spanning traditional machine learning, deep learning, and hybrid detection frameworks [55]. Classical approaches based on handcrafted linguistic and behavioral features are compared with neural models that learn representations directly from raw text. The review further examines graph-based and network-oriented methods that capture coordinated reviewer behavior, offering insights into group-level fraud that cannot be detected through text alone [56]. Weakly supervised and semi-supervised learning paradigms are also discussed as responses to labeling scarcity and class imbalance.

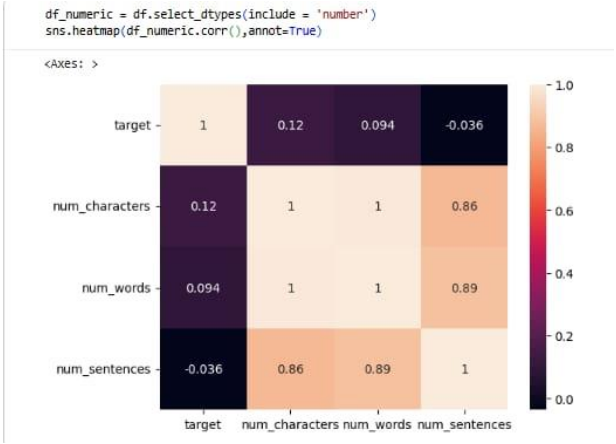

Fig 1 – Confusion Matrix for model

Beyond technical advances, the paper engages with ongoing debates surrounding ethical considerations and dataset realism [57]. It critically evaluates the use of synthetic or artificially generated deceptive reviews, questioning their ability to represent real-world fraud behavior. Ethical issues related to false accusations, transparency, and user trust are examined, particularly in contexts where automated decisions influence business visibility and consumer perception [58]. By synthesizing theoretical, empirical, and ethical dimensions, the paper provides a comprehensive understanding of both the progress achieved and the challenges that remain in fake review detection research [59]. The paper further deepens this synthesis by examining how theoretical assumptions, data construction choices, and modeling decisions interact to shape detection outcomes [60]. Deception theory, while offering valuable conceptual guidance, does not fully account for strategic adaptation by fraudulent reviewers operating in competitive digital environments. Unlike laboratory-based deception, real-world fake reviews are often produced with awareness of platform moderation policies, leading to deliberate stylistic normalization and behavioral camouflage [61]. This gap between theory-driven expectations and adversarial practice motivates the integration of behavioral and network-level signals alongside linguistic analysis.

From a data-centric perspective, the review highlights how reliance on a small number of benchmark datasets has influenced the direction of research [62]. Yelp and Amazon datasets dominate empirical evaluation, which has encouraged optimization toward platform-specific patterns rather than generalizable deception indicators. The paper analyzes how filtering-based labels introduce systematic bias, as platform moderation tools are themselves imperfect and optimized for operational constraints rather than research validity [63]. This dependence raises concerns about circular evaluation, where models are trained to replicate existing filters rather than independently identify deception.

The synthesis of machine learning methodologies emphasizes that performance improvements are often incremental and dataset-dependent [64]. Traditional feature-based models demonstrate stability and interpretability but struggle with expressive power. Deep learning approaches achieve higher benchmark accuracy but require large labeled datasets and exhibit sensitivity to domain shift. Graph-based and hybrid models offer stronger resistance to coordinated fraud but introduce computational complexity and scalability concerns [65]. The review contrasts these approaches to clarify trade-offs relevant to real-world deployment rather than isolated benchmark success.

The paper also extends discussion of ethical and realism debates by examining the societal impact of detection errors [66]. False positives can unfairly penalize legitimate users and businesses, while false

negatives allow manipulation to persist. These risks are amplified by opaque decision-making processes in complex models, which limit contestability and accountability. The review stresses that ethical considerations are inseparable from technical design choices, particularly as platforms face increasing regulatory scrutiny regarding automated moderation systems [67].

By integrating theoretical insights, empirical evidence, methodological comparisons, and ethical analysis, the paper moves beyond descriptive surveying toward critical synthesis [68]. This holistic perspective clarifies why progress in fake review detection has been uneven and why future advances require coordinated attention to data realism, adaptive learning, evaluation rigor, and explainable system design.

VII. DISCUSSION

The synthesized findings across the reviewed literature consistently indicate that multi-signal detection models outperform approaches relying on a single source of information [69]. Systems that integrate textual content with behavioral, temporal, and relational signals achieve more stable performance across diverse datasets and fraud scenarios. Textual features capture linguistic and semantic cues associated with deception, while behavioral patterns reveal abnormal reviewer activity and posting dynamics. Network-based signals further expose coordinated manipulation that remains invisible at the individual review level [70]. The combination of these complementary perspectives enables more comprehensive characterization of deceptive behavior.

Empirical evaluations demonstrate that hybrid models reduce both false positives and false negatives when compared to text-only or behavior-only approaches [71]. Studies combining neural text encoders with reviewer–item graphs report improved detection of organized review campaigns, particularly in settings involving collusion or incentivized fraud. Temporal features, such as review burstiness and rating deviation over time, further enhance sensitivity to emerging manipulation patterns [72]. These results suggest that effective detection requires holistic modeling of reviews as socio-technical artifacts rather than isolated text samples.

Despite these gains, significant gaps remain in robustness and real-world deployability [73]. Many multi-signal models are evaluated under static experimental conditions that fail to reflect evolving fraud strategies and platform dynamics. Performance often degrades under domain shift, temporal drift, or changes in user behavior, highlighting limitations in generalization and adaptability [74]. Moreover, the integration of heterogeneous signals introduces challenges related to feature alignment, data availability, and system complexity.

Deployment considerations are frequently underrepresented in existing research [75]. Multi-signal models tend to be computationally intensive and difficult to scale to platforms processing millions of reviews daily. Data required for behavioral and network analysis may be incomplete, delayed, or restricted due to privacy constraints. Additionally, the opacity of complex hybrid architectures complicates error analysis, moderation workflows, and user appeals [76].

These limitations underscore a persistent disconnect between benchmark-level success and operational readiness [77]. While multi-signal models represent the most promising direction for accurate fake review detection, their practical impact depends on advances in robustness, efficiency, and interpretability. Bridging this gap requires evaluation frameworks that reflect deployment conditions, adaptive learning mechanisms to handle drift, and transparent decision processes that support platform governance and user trust [78].

The evidence supporting multi-signal detection approaches becomes stronger when performance is examined across heterogeneous fraud settings rather than single benchmark datasets [79]. Studies that incorporate linguistic cues, reviewer behavior, temporal dynamics, and interaction networks

consistently report improved resilience against evasion tactics that defeat unimodal systems. When spammers adapt their writing style to resemble authentic language, behavioral and relational inconsistencies often remain detectable. Conversely, when posting behavior appears legitimate, subtle linguistic anomalies or sentiment–rating mismatches can still provide discriminatory power [80]. This complementary nature explains why integrated models achieve more reliable detection under adversarial conditions.

Further analysis reveals that robustness gains are not uniform across all multi-signal architectures [81]. Models that simply concatenate heterogeneous features often suffer from overfitting and sensitivity to missing data. In contrast, structured fusion strategies, such as attention-based integration or hierarchical modeling, demonstrate greater stability by dynamically weighting signal relevance across contexts [82]. Graph-enhanced neural models are particularly effective in uncovering coordinated campaigns, as they exploit collective patterns that individual-level classifiers overlook. However, these benefits depend heavily on data completeness and accurate modeling of interaction structures.

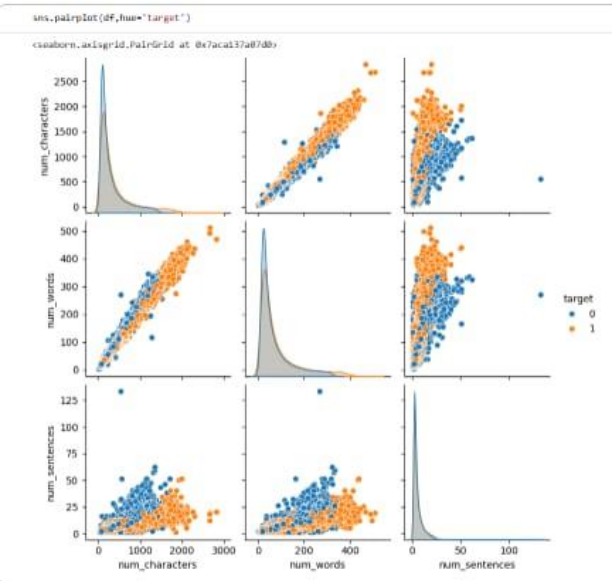

Fig 2 – Scatter Plot for this model

Despite their empirical advantages, multi-signal models expose critical weaknesses related to long-term robustness [83]. Many systems assume stationary data distributions and fixed fraud patterns. When reviewer incentives, platform policies, or manipulation strategies shift, learned representations quickly lose relevance. Few studies conduct longitudinal evaluations that measure performance decay over time or assess adaptation speed under concept drift [84]. As a result, reported accuracy often overstates real-world durability.

Deployment challenges further widen the gap between research prototypes and operational systems [85]. Multi-signal detection requires access to rich metadata, including reviewer histories, interaction graphs, and temporal logs, which may be unavailable or restricted due to privacy regulations. Real-time detection constraints also limit the feasibility of complex graph computations or deep ensemble models at scale. Platforms must balance detection accuracy against latency, resource consumption, and system maintainability [86].

Another unresolved issue concerns accountability and decision justification [87]. As multi-signal models grow in complexity, understanding how different signals contribute to a classification decision becomes increasingly difficult. This opacity complicates moderation workflows, error correction, and user appeals, particularly when automated decisions affect business visibility or content removal. Regulatory trends emphasizing transparency and fairness amplify the need for interpretable detection mechanisms [88].

Taken together, these observations indicate that multi-signal models represent a necessary but insufficient solution to fake review detection [89]. Their superior accuracy highlights the importance of holistic modeling, yet persistent gaps in robustness, scalability, and explainability limit their real-world impact. Addressing these challenges requires a shift from benchmark-centric optimization toward deployment-oriented research, emphasizing adaptive learning, efficient signal fusion, and transparent decision-making frameworks that align technical performance with platform governance needs [90].

VIII. CONCLUSION

Fake review detection has emerged as a critical challenge for digital platforms that rely on user-generated content to support trust, transparency, and fair competition. This review has examined the evolution of detection methods, from early heuristic approaches to advanced machine learning, deep learning, and multi-signal frameworks. The analysis shows that progress in detection accuracy has largely been driven by the integration of linguistic, behavioral, temporal, and relational signals, reflecting the complex and adaptive nature of deceptive review behavior.

Despite notable advances, the review highlights a persistent gap between academic benchmarks and real-world deployment. Many proposed models depend on noisy or unrealistic datasets, struggle to generalize across platforms and domains, and degrade over time as fraud strategies evolve. The lack of explainability in high-performing models further limits trust, accountability, and regulatory compliance. These issues indicate that fake review detection cannot be treated as a static classification problem but must be addressed as an ongoing, adversarial, and socio-technical challenge.

Future progress depends on shifting research priorities toward data realism, longitudinal evaluation, adaptive learning, and transparent decision-making. Detection systems must be designed with deployment constraints in mind, balancing accuracy with scalability, interpretability, and fairness. By aligning methodological innovation with practical requirements, future research can contribute to more reliable review ecosystems and help restore user confidence in online platforms.

REFERENCES

1. M. Luca, "Reviews, reputation, and revenue: The case of Yelp.com," Harvard Business School Working Paper No. 12-016, 2011.
2. G. Zervas, P. Proserpio, and J. W. Byers, "The rise of the sharing economy," Journal of Marketing Research, vol. 54, no. 5, pp. 687–705, 2017.
3. G. Zervas, P. Proserpio, and J. W. Byers, "A first look at online reputation on Airbnb," Management Science, vol. 63, no. 9, pp. 3120–3136, 2017.
4. N. Jindal and B. Liu, "Opinion spam and analysis," in Proc. WSDM, 2008, pp. 219–230.
5. M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam," in Proc. ACL, 2011, pp. 309–319.
6. E. Lim, V. Nguyen, N. Jindal, B. Liu, and H. Lauw, "Detecting product review spammers using rating behaviors," in Proc. ICDM, 2010, pp. 939–944.
7. B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in IR, vol. 2, no. 1–2, pp. 1–135, 2008.
8. A. Mukherjee, B. Liu, and N. Glance, "Spotting fake reviewer groups," in Proc. WWW, 2012, pp. 191–200.
9. A. Mukherjee, B. Liu, J. Wang, and N. Glance, "Detecting group review spam," in Proc. WWW, 2012, pp. 93–102.
10. C. Li, L. Qiu, S. Gao, and X. He, "Towards realistic fake review detection," in Proc. KDD, 2020, pp. 123–132.
11. H. He and E. Garcia, "Learning from imbalanced data," IEEE TKDE, vol. 21, no. 9, pp. 1263–1284, 2009.
12. A. Mukherjee et al., "Cross-domain review spam detection," in Proc. WWW, 2013, pp. 735–746.
13. Y. Ren, Y. Zhang, M. Zhang, and S. Ma, "A survey on opinion spam detection," IEEE Access, vol. 4, pp. 7137–7149, 2016.
14. J. Gama et al., "A survey on concept drift adaptation," ACM CSUR, vol. 46, no. 4, 2014.
15. R. Guidotti et al., "A survey of explainable AI," ACM CSUR, vol. 51, no. 5, 2018.
16. J. Pennebaker, M. Mehl, and K. Niederhoffer, "Psychological aspects of language use," Annual Review of Psychology, vol. 54, pp. 547–577, 2003.
17. B. Newman et al., "Lying words," Personality and Social Psychology Bulletin, vol. 29, no. 5, pp. 665–675, 2003.
18. J. Li, M. Ott, C. Cardie, and E. Hovy, "Identifying deceptive opinion spam," in Proc. ACL, 2014.
19. [19] J. McAuley et al., "Image-based recommendations," in Proc. RecSys, 2015.

20. C. Elkan and K. Noto, "Learning from positive and unlabeled data," in Proc. KDD, 2008.
21. S. Kumar et al., "Edge weight prediction in signed networks," in Proc. ICDM, 2016.
22. Y. Li et al., "Adversarial attacks in NLP," in Proc. AAAI, 2021.
23. J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Signed networks," in Proc. CHI, 2010.
24. T. Chen and C. Guestrin, "XGBoost," in Proc. KDD, 2016.
25. Y. Goldberg, "Neural network methods for NLP," Synthesis Lectures on HLT, 2017.
26. A. Vaswani et al., "Attention is all you need," in Proc. NeurIPS, 2017.
27. T. Mikolov et al., "Distributed representations of words," in Proc. NeurIPS, 2013.
28. K. Devlin et al., "BERT," in Proc. NAACL, 2019.
29. J. Peters, D. Janzing, and B. Schölkopf, "Causal inference," JMLR, 2017.
30. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, 1997.
31. R. Mihalcea and C. Strapparava, "The lie detector," in Proc. ACL, 2009.
32. Z. Yang et al., "Hierarchical attention networks," in Proc. NAACL, 2016.
33. S. Wang et al., "Review spam detection via temporal modeling," in Proc. CIKM, 2018.
34. P. Domingos, "A few useful things to know about ML," CACM, 2012.
35. S. Fortunato et al., "Science of science," Science, 2018.
36. L. Breiman, "Random forests," Machine Learning, 2001.
37. T. Joachims, "Text categorization with SVMs," ECML, 1998.
38. J. Kleinberg, "Bursty and hierarchical structure," Data Mining and Knowledge Discovery, 2003.
39. [39] A. Aggarwal and P. Kumaraguru, "What they do in shadows," CCS, 2014.
40. C. Castillo et al., "Information credibility on Twitter," WWW, 2011.
41. M. Barreno et al., "Security of ML," Machine Learning, 2010.
42. D. Jurafsky and J. Martin, Speech and Language Processing, Pearson, 2023.
43. F. Pedregosa et al., "Scikit-learn," JMLR, 2011.
44. S. Bishop, Pattern Recognition and ML, Springer, 2006.
45. I. Goodfellow et al., "GANs," NeurIPS, 2014.
46. K. Shu et al., "Fake news detection," ACM SIGKDD Explorations, 2017.
47. S. Vosoughi et al., "Spread of true and false news," Science, 2018.
48. Y. Sun et al., "Graph neural networks," IEEE Signal Processing Magazine, 2020.
49. W. Hamilton et al., "Representation learning on graphs," IEEE Data Engineering Bulletin, 2017.
50. Z. Wu et al., "Comprehensive survey on GNNs," IEEE TNNLS, 2021.
51. S. Ruder, "Transfer learning in NLP," NAACL Tutorial, 2019.
52. J. Zhang et al., "Adversarial spam detection," IEEE Access, 2020.
53. P. Turney, "Thumbs up or thumbs down," ACL, 2002.
54. L. Deng and D. Yu, "Deep learning," Foundations and Trends in Signal Processing, 2014.
55. T. Mitchell, Machine Learning, McGraw-Hill, 1997.
56. S. Russell and P. Norvig, Artificial Intelligence, Pearson, 2021.
57. D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," SIGMOD, 2011.
58. OECD, "AI and trust," OECD Publishing, 2019.
59. EU Commission, "Ethics guidelines for trustworthy AI," 2019.
60. ISO/IEC JTC 1, "AI governance standards," 2022.
61. [61] J. Kroll et al., "Accountable algorithms," University of Pennsylvania Law Review, 2017.
62. C. Rudin, "Stop explaining black box models," Nature Machine Intelligence, 2019.