International Journal of Electrical and Computer Engineering (IJECE)
Vol. 16, No. 2, April 2026, pp. 991~1001
ISSN: 2088-8708, DOI: 10.11591/ijece.v16i2.pp991-1001

Multimodal machine learning framework for fake review detection

Rashmi R., Shobha T., Dhanushree C. S., Gayatri S. Santi, Jeevita S. Devadig, Harshitha L. V.
Department of Information Science and Engineering, B.M.S. College of Engineering, Bengaluru, India

Article Info

Article history:

Received Nov 1, 2024
Revised Jan 1, 2026
Accepted Jan 16, 2026

Keywords:

Ensemble models
Explainability
Fake review detection
Machine learning
Multimodal features
Shapley additive explanations

ABSTRACT

Online reviews significantly influence consumer decision-making, yet their credibility is increasingly undermined by the rise of fake and manipulated content. This study addresses the growing challenge of detecting deceptive online reviews by developing a highly accurate, robust, and explainable machine learning framework that supports trust and reliability in digital marketplaces. The proposed multimodal framework integrates textual, behavioural, temporal, and network-based features to enhance detection performance. Textual characteristics are extracted using term frequency-inverse document frequency (TF-IDF) and sentiment analysis, while behavioural and temporal attributes model reviewer activity patterns. Network-oriented features capture suspicious reviewer interactions. To mitigate class imbalance, synthetic samples are generated using the synthetic minority over-sampling technique (SMOTE). Several machine learning models—including logistic regression, decision trees, XGBoost, and a stacking ensemble—are trained and evaluated. Experimental findings show that XGBoost and the stacking ensemble deliver strong balanced performance, achieving an F1-score of approximately 0.87 and an accuracy of 0.94. Decision Trees exhibit high precision (0.98), albeit with comparatively lower recall. To ensure transparency and interpretability, Shapley additive explanations (SHAP) are used to analyse model predictions. Results indicate that reviewer connectivity, co-reviewer counts, and sentiment–rating inconsistencies are among the most influential features. Overall, the proposed framework enhances detection accuracy and provides meaningful, explainable insights, making it well-suited for deployment in real-world digital marketplaces. Future work will focus on extending the framework to multilingual datasets and incorporating adaptive learning mechanisms to address evolving deceptive behaviour.

Corresponding Author:

Rashmi R.
Department of Information Science and Engineering, B.M.S. College of Engineering
Bull Temple Road, Bengaluru-560019, KA, India
Email: rashmir.ise@bmsce.ac.in

1. INTRODUCTION

Every time a consumer or business looks online, they see that online reviews have become an essential part of their research process when choosing what products to buy. When it comes to online retail and hospitality services, about 90% of today's consumers look at reviews before deciding on a particular product or service [1]. Recent studies have shown that 93% of consumers rely on online reviews as a basis for deciding whether to purchase a product and that they trust these reviews just as much as they would trust personal recommendations from friends or family [2]. In the hospitality industry, online reviews play an even

more pivotal role in influencing consumer behaviours, given that most of the products (*e.g.*, hotels and restaurants) are intangible and cannot be experienced until after the purchase has been made. Unfortunately, many consumers are having their views of hotels and restaurants distorted by the proliferation of fake reviews. Unfortunately, the lack of strict purchasing verification processes among many platforms that allow for consumer product rating (*i.e.*, Yelp and Tripadvisor) has added another layer of vulnerability to independent hotels and restaurants that rely heavily on positive online reviews to establish themselves within the marketplace.

Several studies have demonstrated the effectiveness of supervised machine learning algorithms, such as decision trees, random forests, and support vector machines, in identifying fake reviews using textual and reviewer-based features [3]. Recent work has also shown that combining multiple classical machine learning models can improve fake review detection accuracy, even when labelled data is limited [4]. Early approaches to fake review detection primarily relied on textual analysis, focusing on linguistic patterns, sentiment polarity, and writing styles. Although these methods achieved reasonable performance, they are limited by their inability to capture non-textual cues, such as reviewer behaviour, temporal posting patterns, and coordinated review activities. More recent studies have explored behavioural and network-based signals; however, many of these approaches still rely on a single feature modality or lack interpretability, making their adoption in real-world platforms challenging. With advances in natural language processing, transformer-based models such as DeBERTa and large language models have demonstrated strong performance by capturing deeper semantic and implicational characteristics of deceptive reviews [5]. However, despite high accuracy, such models often face challenges related to interpretability and real-world deployment [6].

The objective of this study is to design a reliable, accurate, and interpretable framework for detecting fake reviews by overcoming the limitations of traditional text-based and single-modality methods. To achieve this, the study integrates multiple complementary feature types while maintaining practicality for real-world deployment on online platforms. The proposed framework focuses on identifying fake reviews in domains such as e-commerce and hospitality using a stacking machine learning approach. It aggregates the predictions of several base models and employs XGBoost as the final classifier to enhance detection performance.

The approach incorporates textual features (derived through natural language processing (NLP) and term frequency-inverse document frequency (TF-IDF)), behavioural attributes (based on reviewer activity patterns, including frequency and consistency), and time-series characteristics (from datasets such as Yelp). This multimodal feature integration enables high accuracy, precision, and recall across both small and large datasets [7]. Furthermore, the inclusion of Shapley additive explanations (SHAP) within the methodology enhances model interpretability, offering deeper insight into feature contributions and potential bias. Overall, the proposed framework not only improves detection accuracy but also provides meaningful explainability, making it well-suited for implementation in real-world digital marketplaces.

2. LITERATURE REVIEW

Fake review detection is becoming a popular field of study with respect to the use of technology in online marketplaces and the potential impact on consumer confidence within those markets. Historically, researchers have focused their efforts on identifying deceptions based on the use of language and text analysis techniques. For example, Ott *et al.* [8] were able to show that there were many different types of linguistic features associated with deceptive reviews, including sentiment polarity and bag-of-words approaches. Meanwhile, Jindal and Liu [9] developed models to analyse opinion spam based on rule-based systems and reviewer behaviour. While these early research studies did provide some valuable insights into identifying deceptive reviews, they were limited in their ability to detect coordinated behaviour and behaviour-driven fraud because they relied heavily on textual-based features. As researchers learned more about how consumers interacted with and wrote reviews, they began using alternative methods to better understand the behaviours and relationships between consumers. For example, Mukherjee *et al.* [10] conducted a study on how consumers behave and develop relationships on review sites like yelp to better understand how review sites filter out fraudulent reviews. They found that a better understanding of those behaviours and relationships could improve the accuracy of detecting fraudulent reviews. Today, researchers are increasingly using network-based approaches to identify fraudulent activity. For example, He *et al.* [11] developed a model for identifying products purchased with fraudulent reviews on Amazon using network-level structural attributes such as degree centrality, PageRank, and clustering coefficients. This method works very well but is limited to identifying the fraudulent activities of products on Amazon since it requires data from that platform. Natural language processing (NLP) has been making great strides due to rapid advancements in deep learning and transformer-based models. Transformer-based architectures have been

used by Salminen *et al.* [12] to classify deceptive reviews. Similarly, Chen *et al.* [13] used bidirectional-long short-term memory (BiLSTM) models with attention mechanisms to capture semantic dispersion between different review texts. Several studies have validated the use of DeBERTa [14] and RoBERTa [15] as effective models for detecting fake reviews. Further, Su *et al.* [16] worked on methods to identify AI-generated (uses pretrained language models such as BERT and DeBERTa). Although these types of models have been found to achieve high levels of accuracy, they remain text-based, heavily computational, and lack interpretability as they generally do not capture coordinated human fraud and reviewer collusion.

The use of ensemble and multi-feature learning approaches has emerged as a way to improve robustness. For example, Cao *et al.* [17] proposed a deceptive review detection system which separated multi-feature learning and classification phases to create an effective deception detection model by improving generalization performance. Mohawesh *et al.* [18] presented an explainable ensemble of multiview DL models which combined textual and behavioural features of users to enhance transparency while providing a method to improve explainability. The use of stacking-based ensemble frameworks [1] have shown improvements in model performance through the combination of multiple classifiers; however, these approaches often emphasized the architecture of models rather than providing thorough integration of features.

In addition to text-based and ensemble-based approaches, alternative techniques are being explored for identifying misleading online product reviews. For example, Shahariar *et al.* [19] have created a benchmark dataset of fictitious Bengali reviews that can be used to evaluate language-specific detection systems due to the absence of multilingual datasets. Birim *et al.* [20] used topic models to identify the unique quality of deceptive review patterns by identifying unusual or unexpected distributions among the topics of the reviews, while Yao *et al.* [21] developed a graph-based approach to performing contrastive learning to model hierarchical contractions of reviewer behaviour. On a broader scale, Walther *et al.* [22] explored how people use different factors to determine whether or not reviews are authentic, while Jabeur *et al.* [23] summarised the major research trends in this area and outlined possible future directions for investigation through bibliometric analysis.

While many studies have produced high levels of performance modelling fake reviews with text-based methods, deep learning techniques, and graph/community-based approaches, there are still several challenges associated with these approaches, including their ability to be interpretable (*i.e.*, easily understood), their capacity for scaling (*i.e.*, growing in size and number), and their vulnerability to being exploited for coordinated fraudulent activities. To solve the challenges stated above, this work proposes a unified solution that integrates four different types of information that might signal the presence of a fake review (*e.g.*, textual, behavioural, temporal, and network-based indicators) using a stacking ensemble modelling architecture. Additionally, by applying Shapley additive explanations (SHAP), we provide a means of improving model interpretability for decision-makers. By using a combined multimodal and explainable framework to detect fake reviews, our solution can be deployed independently across multiple e-commerce platforms to protect consumers from being misled by fraudulent or deceptive reviews.

3. METHODOLOGY

3.1. Pre-processing

The raw review dataset undergoes standard preprocessing steps to ensure data quality and consistency. Textual content is cleaned by removing punctuation, stop words, and special characters, followed by tokenisation and normalisation. Missing values in non-textual attributes are handled appropriately, and categorical variables are encoded where required.

3.1.1. Data balancing

Fake review datasets are naturally imbalanced because genuine reviews significantly outnumber fraudulent ones. To address this issue, we used the synthetic minority over-sampling technique (SMOTE), which generates new, synthetic examples for the minority class rather than simply duplicating existing samples. SMOTE works by identifying the nearest neighbours of minority-class samples and creating new points along the line that connects them. This helps expand the decision boundary for the minority (fake) class and reduces misclassification. As a result, the classifier becomes more sensitive and accurate in identifying fake reviews.

3.1.2. Feature engineering

Our methodology integrates a broad set of features and multiple machine learning models to capture the various patterns present in both genuine and fake reviews. The approach combines temporal, behavioural, network, textual, and synthetic signals, all extracted or engineered from the dataset fields: review_id, user_id, business_id, rating, review_text, date, and flag.

a. Temporal features

Using the review timestamp, we generated time-based behavioural patterns to understand when the review was posted. The following features were extracted: Year, Month, Day, Weekday, Hour of review posting, Is_weekend (Saturday/Sunday), Is_night (posted between 10 pm and 6 am), Is_business_hours (posted during 9 am–5 pm) These features are useful because fraudulent activity often clusters around unusual hours or specific days, indicating abnormal posting behaviour.

b. Behavioural features

To capture reviewer behaviour, we computed several user-level statistics: user_total_reviews, user_avg_rating, user_rating_std, user_min_rating, user_max_rating, reviews_per_day, and rating_consistency (how often a user gives the same rating). These features suggest whether the user behaves like a typical reviewer or shows signs of automated, repetitive, or suspicious rating behaviour.

c. Network features

We constructed a user–business interaction graph using NetworkX to understand relationships and detect patterns of collusion. From this graph, we extracted:

- user_network_size — number of connected users
- user_network_density — connections per review
- co-reviewer count — users reviewing the same businesses
- degree centrality — how influential the user is in the network

These network signals help reveal suspicious clusters of reviewers posting together, which is common in coordinated fake review campaigns. Figure 1 illustrates the reviewer–business interaction network, highlighting influential reviewers and densely connected clusters that may indicate coordinated fake review behaviour.

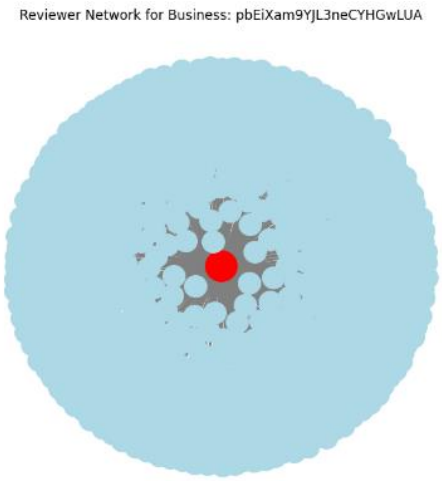

Figure 1. Network analysis of features

d. Textual features and text embedding

To create representations of review data, we analysed review length, word count, and sentiment (via a BERT based sentiment scoring model), as well as TF-IDF embedding values (limited to a maximum of 5000 features). The TF-IDF method allows for the identification of key terms in documents by giving low weights to commonly used words, while the sentiment score, which indicates the emotional valence/quality of the review's content, highlighted inconsistencies within the data if there are very negative reviews with 5-star ratings, thereby allowing potential identification of deceptive reviews. TF-IDF was applied against the text of all reviews, producing numeric values that were used in combination with behavioural and network features to enable model input of both unstructured (review text) and structured (numerically represented) data. Figure 2 shows the feature importance derived from the Decision Tree model, indicating that network centrality and reviewer activity metrics play a dominant role in fake review detection.

e. Synthetic behavioural features and data processing

The dataset was enhanced with synthetic behavioural proxy features derived exclusively from publicly available review metadata. These proxies approximate real-world reviewer activity patterns—such as inferred activity regularity, interaction consistency, reviewing speed, and account longevity—without

relying on any platform-internal logs or private user information. The original flag attribute was converted into a binary label (0: genuine review, 1: fake review). In addition, a composite suspiciousness score, termed fake_score, was computed to capture abnormal behavioural patterns including unusually high review frequency, rating homogeneity, rapid review posting, and posting during atypical time periods. Missing values were handled using linear interpolation, numerical features were standardized using StandardScaler, and the dataset was split into 80% training and 20% testing sets. All feature engineering steps are deterministic, reproducible, and documented in the public GitHub repository.

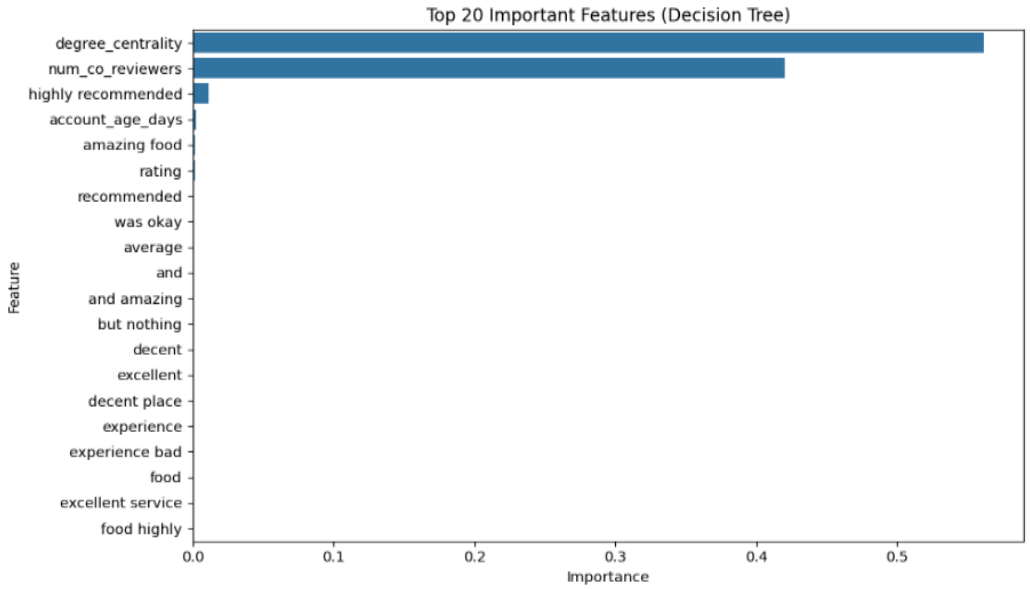

Figure 2. Feature importance for the decision

f. Fake review detection models

A framework that uses various methods for creating a machine learning environment to identify possible fraudulent or misleading reviews is presented in this paper. With logistic regression, decision tree, XGBoost, and stacking ensembles, these models are trained using stratified sampling to maintain the balance of classes. Logistic regression acts as a baseline model to see if linear separability exists with the trained data and to be able to interpret the effect of features on the trained data. XGBoost utilizes complex non-linear relationships within the structured features of this data set, which allows it to provide superior predictive capabilities over all other models. Stacking ensembles combine logistic regression and XGBoost by way of logistic regression being the meta-classifier, allowing for improved robustness and decreased classification error. The decision tree model adds to the overall interpretability of this model by allowing users to identify the most important features influencing the classification of a fraudulent review. Figure 3 presents the Decision Tree structure, demonstrating how key features such as reviewer count, degree centrality, and account age influence classification decisions.

3.2. Overview of the proposed framework

A machine learning based approach was developed and tested to detect fraud/false reviews on the internet. The goal of this system was to make consumers more comfortable when utilising online review sites. The methodology used to develop the machine learning based model involved extensive and complex data preparation. Different types of machine learning models were created to analyse online review data; these models included logistic regression, XGBoost, and stacking ensembling. The combination of these multiple models has yielded strong results as they capture multiple dimensions of fraudulent writing behaviours and trends. Additionally, performance measures, strengths and limitations of each model are discussed in the following subsections; it is stressed that robust approaches are critical to preventing the manipulation of reviews, as well as to better understand the fundamental aspects of online deception to create more reliable e-commerce environments. Figure 4 presents the proposed end-to-end framework for fake review detection, illustrating the preprocessing pipeline, feature extraction stages, base classifiers, stacking ensemble, and evaluation metrics.

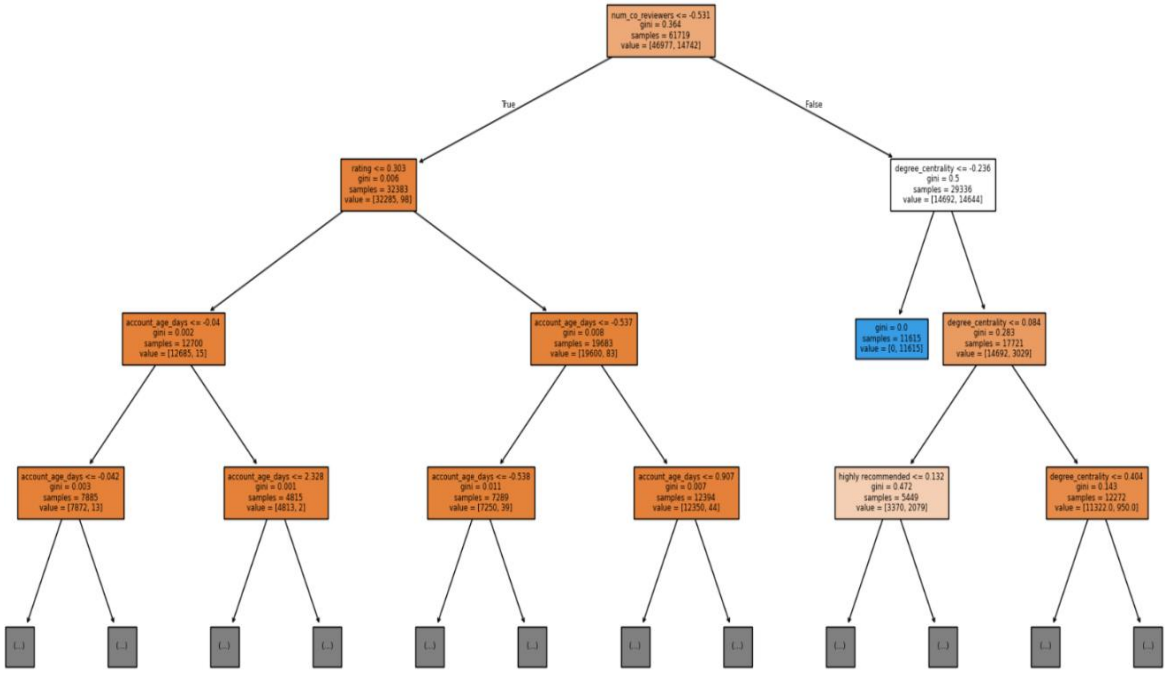

Figure 3. Decision tree

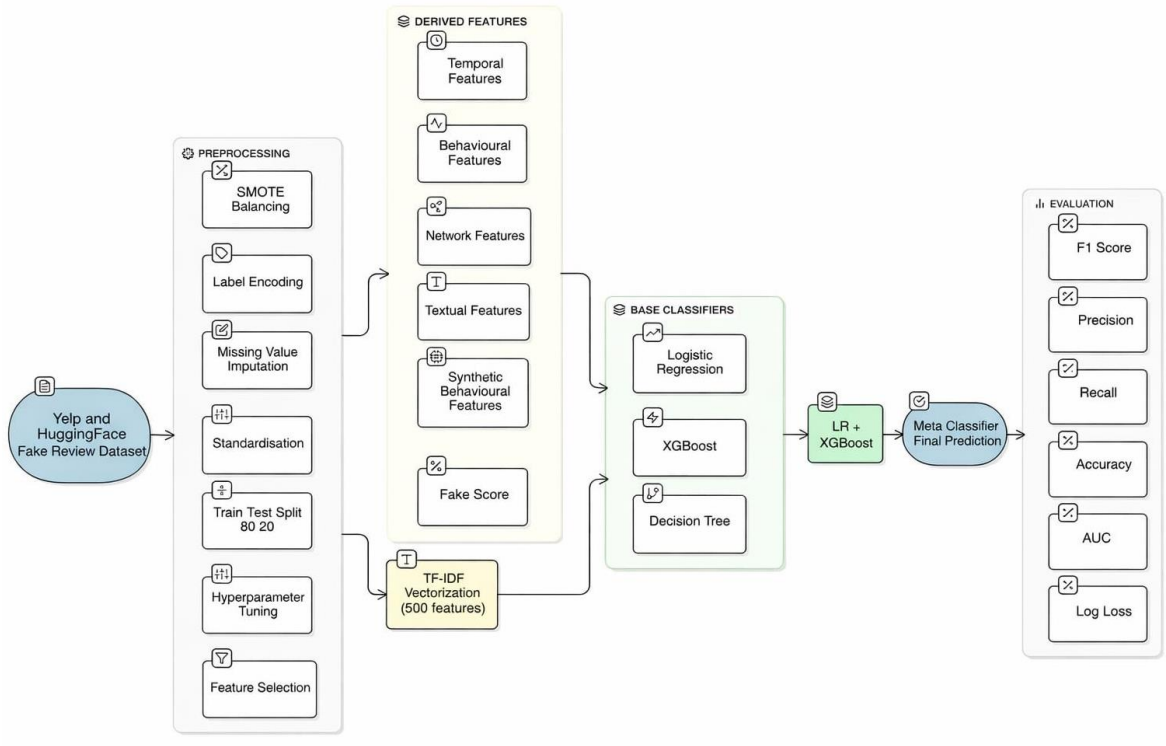

Figure 4. Proposed framework for fake review detection

4. RESULTS

This section presents the experimental evaluation of the proposed multimodal fake review detection framework. The models are assessed using accuracy, precision, recall, and F1-score to provide a balanced evaluation under class imbalance conditions as provided in (1), (2), (3) and (4).

4.1. Performance evaluation

To assess the effectiveness of our fake review detection system, several standard classification metrics were used. These include accuracy, precision, recall, and the F1-score. Performance is based on a binary classification setting, where reviews were labelled as either "True" or "Fake". Table 1 presents the confusion matrix of the best-performing model

Table 1. Confusion matrix of the best-performing model

Actual/Predicted	Truthful	Fake
Truthful	True Positive (TP)	False Negative (FN)
Fake	False Positive (FP)	True Negative (TN)

The confusion matrix indicates a reduced number of false negatives, which is critical in fake review detection scenarios where undetected fraudulent reviews may directly affect consumer trust. The balanced distribution of false positives and false negatives demonstrates that the proposed framework achieves reliable performance suitable for real-world deployment. The evaluation metrics are computed as (1), (2), (3), (4):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$\text{F1 Score} = \frac{2 \text{ x Precision x Recall}}{\text{Precision+Recall}} \quad (4)$$

These metrics collectively provide a comprehensive assessment of model effectiveness, particularly under imbalanced class distributions.

4.2. Model-wise performance comparison

Table 2 presents the comparative performance of different machine learning models trained using multimodal features. Logistic regression shows moderate performance due to its linear decision boundaries, which limit its ability to model complex interactions among textual, behavioural, temporal, and network-based features. Decision Trees achieve the highest precision (0.98), indicating strong confidence in identifying fake reviews; however, their lower recall suggests that a considerable number of fraudulent reviews remain undetected. In contrast, XGBoost and the stacking ensemble achieve the most balanced performance across all metrics. Their ability to model non-linear relationships and leverage feature interactions enables improved detection of coordinated and subtle deceptive behaviours. The ensemble model further enhances robustness by combining the strengths of individual classifiers.

Table 2. Comparative performance of machine learning models using multimodal features

Model	Precision	Recall	F1-score	Accuracy
Logistic regression	0.64	0.95	0.77	0.86
XGBoost	0.83	0.93	0.87	0.94
Stacking ensemble	0.81	0.93	0.87	0.93
Decision tree classifier	**0.98**	**0.81**	**0.88**	**0.95**

Overall, these results highlight the advantage of multimodal feature integration in fake review detection. By combining textual, behavioural, temporal, and network-based signals, the proposed framework captures both content-level deception and coordinated reviewer activity that single-modality approaches may overlook. When compared with recent text-centric deep learning models reported in the literature, which typically achieve F1-scores in the range of 0.85–0.90, the proposed approach demonstrates competitive performance (F1 ≈ 0.87) while requiring significantly lower computational resources and offering enhanced interpretability. Furthermore, SHAP-based analysis confirms that network connectivity, co-reviewer interactions, and behavioural consistency play a critical role in improving detection effectiveness, reinforcing the importance of multimodal learning for practical deployment.

4.3. Comparison with existing and state-of-the-art methods

Research on early detection of fake reviews began by looking at textual characteristics and using manually created language-based cues. For example, Ott *et al.* [8] demonstrated that deceptive reviews have distinct characteristics through sentiment polarity and bag-of-words characteristics, while Jindal and Liu [9] researched opinion spam using rule-based techniques to model reviewer behaviour. While these studies provided important foundational knowledge, both of these techniques had their limitations regarding their ability to identify coordinated reviewer behaviours and detect time anomalies.

In order to expand upon the findings of these studies, Mukherjee *et al.* [10] proposed two new forms of analytical frameworks for detecting fraudulent reviews based on behavioural and relational methods. Mukherjee *et al.* provided empirical evidence showing how sites such as Yelp filter out fraudulent reviews using the interactions of reviewers or networks of review authors, thus leading to better detection results; however, the methods discussed in this study are still platform-specific and does not easily adapt to emerging spamming strategies. Over the past few years, there has been an increased focus on utilising deep learning and transformer-based architectures in various fields. For instance, Li *et al.* [24] used BERT to create a contextual semantic representation of reviews from which they could successfully classify the severity of deceiving reviews. Likewise, Ren and Ji [25] have utilised convolutional neural networks (CNNs) and long short-term memory (LSTM)-type NN models to perform deceptive opinion spam detection. More recently, Zhang *et al.* [26] incorporated the utilisation of transformer-based models into their methodology for detecting fake reviews by providing deeper semantic representations of the reviews being evaluated.

Despite the improved predictive capability provided by these types of deep learning techniques, the models are predominantly based on text data, require substantial amounts of compute and are difficult to interpret. In contrast, the proposed approach includes a multimodal model that integrates textual, behavioural, chronological, and network/computerised feature sets to produce competitive results through improved robustness and interpretability of the multimodal model.

4.4. Practical deployment implications

From a deployment perspective, the proposed approach offers significant advantages over deep learning-based methods. Transformer models typically require substantial computational resources, large-scale data, and frequent retraining. In contrast, the presented framework is lightweight, interpretable, and adaptable, making it suitable for real-time deployment in large-scale review platforms. The explainable nature of the predictions further supports trust, accountability, and regulatory compliance in automated review moderation systems.

4.4.1. Model explainability (SHAP analysis)

Shapley additive explanations (SHAP) were used to interpret model predictions. The analysis revealed that the following features had the greatest influence: degree_centrality, num_co_reviewers, account_age_days, rating, and sentiment_score.SHAP plots showed that higher reviewer connectivity, unusual reviewing patterns, and mismatches between sentiment and rating significantly increased the likelihood of a review being classified as fake. Figure 5 illustrates the SHAP-based explanation for an individual prediction, showing how key features such as the number of co-reviewers, degree centrality, and review length contribute positively or negatively to the model's decision. Importance: Explainability increased trust in the system and ensured that the models were learning meaningful patterns rather than noise or bias.

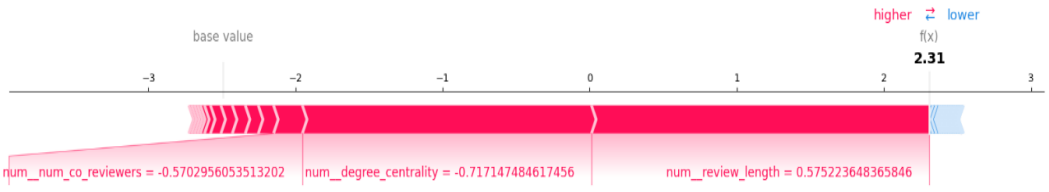

Figure 5. Feature contribution for the review

5. CONCLUSION

This study proposes a multimodal machine learning framework for fake review detection by integrating behavioural, temporal, textual, network, and synthetic features. Experimental evaluation shows that ensemble models, particularly XGBoost and the stacking ensemble, achieve consistent and balanced

performance across precision, recall, and F1-score, while SHAP-based analysis improves interpretability by highlighting key feature contributions. Beyond strong predictive performance, the framework has practical deployment potential for real-world e-commerce and hospitality platforms, where it can support large-scale review moderation, reduce fraudulent activity, and enhance consumer trust. However, the current approach is limited to English-language datasets and remains sensitive to evolving deceptive practices. Future work should focus on extending multilingual capabilities, incorporating adaptive learning to address emerging manipulation strategies, and validating the framework through integration into operational platforms to improve the reliability of user-generated content at scale.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Rashmi R.	✓	✓	✓		✓	✓		✓	✓	✓			✓	
Shobha T.		✓				✓			✓	✓		✓		
Jeevita S. Devadig	✓		✓	✓			✓			✓	✓		✓	✓
Gayatri S. Santi	✓			✓				✓				✓		
Dhanushree C. S.					✓		✓			✓		✓		✓
Harshita L. V.		✓				✓			✓				✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available in below repositories:

1. Hugging Face Dataset: *https://huggingface.co/datasets/debojit01/fake-review-dataset*
2. Yelp Dataset: GitHub - *https://github.com/chiragdaryani/fake-review-detection*
3. Final Dataset: GitHub - *https://github.com/Dhanu0746/MLG-Dataset.git*

REFERENCES

[1] S. A. Ashraf, A. F. Javed, S. Bellary, P. K. Bala, and P. K. Panigrahi, "Leveraging stacking framework for fake review detection in the hospitality sector," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 19, no. 2, p. 1517, Jun. 2024, doi: 10.3390/jtaer19020075.

[2] W. H. Asaad, R. Allami, and Y. H. Ali, "Fake review detection using machine learning," *Revue d'intelligence artificielle*, vol. 37, no. 5, Oct. 2023, doi: 10.18280/ria.370507.

[3] M. Lee, Y. Song, L. Li, K. Lee, and S.-B. Yang, "Detecting fake reviews with supervised machine learning algorithms," *Service Industries Journal*, vol. 42, no. 13–14, pp. 1101–1121, Oct. 2022, doi: 10.1080/02642069.2022.2054996.

[4] A. H. Alshehri, "An online fake review detection approach using famous machine learning algorithms," *Computers, Materials & Continua*, vol. 78, no. 2, p. 2767, Jan. 2024, doi: 10.32604/cmc.2023.046838.

[5] Z. Wang, A. Yao, G. Xu, and M. Ren, "A large language model-based approach for fake review detection: the implicit characteristics perspective," *Information Processing & Management*, vol. 63, no. 1, p. 104352, Aug. 2025, doi: 10.1016/j.ipm.2025.104352.

[6] J. Salminen, C. Kandpal, A. M. Kamel, S. Jung, and B. J. Jansen, "Creating and detecting fake reviews of online products," *Journal of Retailing and Consumer Services*, vol. 64, p. 102771, 2022, doi: 10.1016/j.jretconser.2021.102771.

[7] R. Gupta, V. Jindal, and I. Kashyap, "Recent state-of-the-art of fake review detection: a comprehensive review," *Knowledge Engineering Review*, vol. 39, p. e8, Nov. 2024, doi: 10.1017/S0269888924000067.

[8] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," in *Proceedings of 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2011, pp. 309–319.

[9] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proceedings of the 1st ACM International Conference on Web Search and Data Mining (WSDM)*, 2008, pp. 219–230, doi: 10.1145/1341531.1341560.
[10] A. Mukherjee, V. Venkataraman, B. Liu, and N. S. Glance, "What Yelp fake review filter might be doing?," in *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2013, pp. 409–418.
[11] S. He, B. Hollenbeck, G. Overgoor, D. Proserpio, and A. Tosyali, "Detecting fake-review buyers using network structure: direct evidence from Amazon," *Proceedings of the National Academy of Sciences*, vol. 119, no. 47, p. e2211932119, 2022, doi: 10.1073/pnas.2211932119.
[12] J. Salminen, M. Mustak, S. Jung, H. Makkonen, and B. J. Jansen, "Decoding deception in the online marketplace: enhancing fake review detection with psycholinguistics and transformer models," *Journal of Marketing Analytics*, Mar. 2025, doi: 10.1057/s41270-025-00393-8.
[13] J. Chen, T. Zhang, Z. Yan, Z. Zheng, W. Zhang, and J. Zhang, "Attention-based BiLSTM with positional embeddings for fake review detection," *Journal of Big Data*, vol. 12, no. 1, Apr. 2025, doi: 10.1186/s40537-025-01130-9.
[14] S. Geetha, E. Elakiya, R. S. Kanmani, and M. Das, "High performance fake review detection using pretrained DeBERTa optimized with Monarch Butterfly paradigm," *Scientific Reports*, vol. 15, no. 1, p. 7445, 2025, doi: 10.1038/s41598-025-89453-8.
[15] L. Zhang, M. Lee, S. Patel, and J. Seo, "A study on fake review detection based on RoBERTa," *Procedia Computer Science*, vol. 242, pp. 1323–1330, 2024, doi: 10.1016/j.procs.2024.08.131.
[16] Z. Su, M. Yang, Q. Zhai, K. Guo, Y. Huang, and Y. Cong, "A multigrained preference analysis method for product iterative design incorporating AI-generated review detection," *Scientific Reports*, vol. 15, no. 1, p. 2528, Jan. 2025, doi: 10.1038/s41598-025-86551-5.
[17] N. Cao, S. Ji, D. K. W. Chiu, and M. Gong, "A deceptive reviews detection model: separated training of multi-feature learning and classification," *Expert Systems with Applications*, vol. 187, p. 115977, Jan. 2022, doi: 10.1016/j.eswa.2021.115977.
[18] R. Mohawesh, S. Xu, M. Springer, Y. Jararweh, M. Al-Hawawreh, and S. Maqsood, "An explainable ensemble of Multiview deep learning model for fake review detection," *Journal of King Saud University -- Computer and Information Sciences*, vol. 35, no. 8, p. 101644, Sep. 2023, doi: 10.1016/j.jksuci.2023.101644.
[19] G. M. Shahariar, M. T. R. Shawon, F. M. Shah, M. S. Alam, and M. S. Mahbub, "Bengali fake reviews: a benchmark dataset and detection system," *Neurocomputing*, vol. 592, p. 127732, 2024, doi: 10.1016/j.neucom.2024.127732.
[20] S. Ö Birim, I. Kazancoglu, S. K. Mangla, A. Kahraman, S. Kumar, and Y. Kazancoglu, "Detecting fake reviews through topic modelling," *Journal of Business Research*, vol. 149, pp. 884–900, Oct. 2022, doi: 10.1016/j.jbusres.2022.05.081.
[21] J. Yao, L. Jiang, C. Shi, and S. Yan, "Fake review detection with label-consistent and hierarchical-relation-aware graph contrastive learning," *Knowledge-Based Systems*, vol. 302, p. 112385, 2024, doi: 10.1016/j.knosys.2024.112385.
[22] M. Walther, T. Jakobi, S. J. Watson, and G. Stevens, "A systematic literature review about the consumers' side of fake review detection – Which cues do consumers use to determine the veracity of online user reviews?," *Computers in Human Behaviour Reports*, vol. 10, p. 100278, May 2023, doi: 10.1016/j.chbr.2023.100278.
[23] S. B. Jabeur, H. Ballouk, W. B. Arfi, and J.-M. Sahut, "Artificial intelligence applications in fake review detection: bibliometric analysis and future avenues for research," *Journal of Business Research*, vol. 158, p. 113631, 2023, doi: 10.1016/j.jbusres.2022.113631.
[24] J. Li, M. Huang, X. Yang, and X. Zhu, "Exploiting BERT for fake review detection," *IEEE Access*, vol. 7, pp. 150000–150010, 2019, doi: 10.1109/ACCESS.2019.2945002.
[25] Y. Ren and D. Ji, "Neural networks for deceptive opinion spam detection," *Expert Systems with Applications*, vol. 113, pp. 140–151, Jan. 2019, doi: 10.1016/j.eswa.2018.06.027.
[26] Y. Zhang, G. Zhong, J. Chen, and H. Li, "Fake review detection using transformer-based models," *Information Processing & Management*, vol. 58, no. 3, p. 102472, May 2021, doi: 10.1016/j.ipm.2021.102472.

BIOGRAPHIES OF AUTHORS

Rashmi R. received her B.E. degree in computer science from Appa Institute of Engineering and Technology, Karnataka, in 2007 and her M.Tech. degree in computer networks and engineering from B.M.S. College of Engineering, Bengaluru, in 2019. She is currently an assistant professor in the Department of Information Science and Engineering at B.M.S. College of Engineering. Her research interests include machine learning and deep learning. She has over 10 years of industry experience in software development and contributes to this study through her expertise in intelligent systems. She can be contacted at rashmir.ise@bmsce.ac.in.

Shobha T. received her B.E. degree in information science and engineering in 2004 and the M.Tech. degree in computer science and engineering in 2010, both from Visvesvaraya Technological University, Karnataka. She completed her Ph.D. in machine learning from Visvesvaraya Technological University, Belagavi, in 2022. She is currently an associate professor in the Department of Information Science and Engineering at B.M.S. College of Engineering, Bengaluru. Her research interests include data mining, machine learning, and artificial intelligence. She has published extensively in Scopus and Web of Science–indexed journals and conferences. She can be contacted at shobha.ise@bmsce.ac.in.

Dhanushee C. S. is pursuing a bachelor's degree in information science and engineering at B.M.S. College of Engineering, Bengaluru. Her academic interests include machine learning, deep learning, and data analytics. In this study, she contributed to conceptualization, methodology, software development, and the original draft preparation. She can be contacted at dhanushreecs.is23@bmsce.ac.in.

Gayatri S. Santi is currently pursuing her B.E. in information science and engineering at B.M.S. College of Engineering, Bengaluru. Her fields of interest include software engineering, cloud computing, and intelligent systems. She contributed to visualisation, resource management, review and editing, and project administration for this paper. She can be contacted at gayatrisadashiv.is23@bmsce.ac.in.

Jeevita S. Devadig is an undergraduate student of information science and engineering at B.M.S. College of Engineering. Her research interests include machine learning, network security, and data-driven application development. She contributed to supervision support, experimentation, writing–review, validation, and final refinement of results. She can be contacted at jeevitasubray.is23@bmsce.ac.in.

Harshita L. V. is an undergraduate student in the Department of Information Science and Engineering at B.M.S. College of Engineering. Her areas of interest include artificial intelligence, information security, and data mining. She contributed to investigation, data curation, validation, and formal analysis for this research work. She can be contacted at harshitalv.is23@bmsce.ac.in.