UNIOFFA JOURNAL OF INTERDISCIPLINARY STUDIES (U-OJIS)
A Multidisciplinary Journal of Science, Management and Social Sciences
Website: www.unioffajis.com
ISSN:3141-5571
E-ISSN:3141-558X

A SHAP-BASED CUSTOMER FAKE REVIEW DETECTION USING SMOTE AND HYBRID LEARNING ALGORITHMS

Gbolagade Morufat Damilola, **Abdulkareem Mariam Morenikeji**, **Saka Kayode Kamil**
Adegoke abolanle Titilayo, Bello-lawal Taofeeqah Monisola, Yusuf Olatayo
Department of Computer Science, Al-Hikmah University Ilorin, Nigeria
Asaju-Gbolagade Ayisat Wuraola
Department of Computer Science, University of Ilorin, Nigeria
Correspondence email: dammyconsult@gmail.com

Volume1:2

ABSTRACT

This study presents a SHAP-based model for detecting customer fake reviews using SMOTE and hybrid learning algorithms to address the growing challenge of fake online reviews that undermine consumer trust, purchasing decisions, and business reputation. The proposed detection model integrates SHapley Additive exPlanations (SHAP), the Synthetic Minority Over-sampling Technique (SMOTE), and a hybrid Logistic Regression and Random Forest (LR–RF) model. Logistic Regression provides transparency, Random Forest captures complex relationships, SMOTE resolves class imbalance, and SHAP enhances model interpretability. The model was implemented in Python using standard machine learning libraries and evaluated on Amazon, Yelp, and TripAdvisor datasets with textual, behavioral, and metadata features to ensure cross-platform robustness. The experimental result showed that the model achieved 92.3% accuracy for single-domain and 89.0% for cross-domain detection. The proposed system promotes trustworthy reviews, supports business integrity, ensures fair competition, and aligns with regulatory requirements for transparent, explainable Artificial Intelligence (AI).

Keywords: SHAP, Customer fake reviews detection, SMOTE, Logistic Regression, Random Forest, Hybrid Model.

Introduction

Online review systems have become important to digital commerce, influencing consumer choices, business reputation, and marketplace visibility. Approximately 98% of consumers consult online reviews before making a purchase, and 70% report substantial influence from these reviews (BrightLocal, 2023; PwC, 2023). However, as online platforms expand, deceptive or fake customer reviews also increase. Fake reviews created by humans or rapidly generated by AI systems undermine trust, distort competition, and mislead consumers. The global economic cost of fake reviews is estimated at $152 billion annually (World Economic Forum, 2021). Online reviews are highly important in influencing consumer choices, yet the increase in fraudulent customer evaluations has seriously threatened the credibility of online marketplaces. (FTC, 2023; BrightLocal, 2023). Researchers have addressed the issue of customer fake reviews in a number of phases in recent year (2017)

Customer fake reviews have grown to be a significant issue for online businesses. According to the Federal Trade Commission (2022), these false evaluations deceive consumers and foster unfair competition, costing the world economy approximately $152 billion annually. According to recent research, there are two primary categories of fraudulent customer reviews: those written by people, which frequently exhibit particular writing and behavior patterns (Zhang et al., 2023), and those generated by AI tools that closely resemble actual reviews (Wang & Hooi, 2024). Because of these two powerful sources, older detection methods are no longer very effective.

According to recent studies, hybrid models can assist in addressing some of the main problems with identifying customer fake reviews. For example, Zhang and Li (2022) developed a two-step model that blends Logistic Regression with Random Forest (RF), attaining 87.3% accuracy while maintaining readability. This marked a significant improvement over deep learning model. More recently, Periasamy et al. (2024) used a variety of models, such as SVM,

ISSN:3141-5571
E-ISSN:3141-558X

k-NN, and Decision Trees, and showed that incorporating models improved system reliability. These hybrid approaches still need to be improved because they have issues explaining their decision-making process.

This study makes three significant advances in the realm of identifying fraudulent online reviews. Firstly, it presents an improved hybrid machine learning model that blends Logistic Regression and Random Forest. The model successfully addresses class imbalance by utilizing the SMOTE technique and SHAP for its interpretability. This makes the system not only accurate but also explainable and practical for real-world deployment. The model is evaluated on three prominent review platforms: Amazon, Yelp, and TripAdvisor.

Review of Related Works

Kumar et al. (2025) utilized the Yelp and Amazon datasets to develop a model that included SHAP and XGBoost. Features including review length, sentiment score, and user frequency were extracted. The model was commended for its accuracy and interpretability, achieving 90.4% accuracy. However, it was observed that the method required a lot of resources. Compared to the current study, Kumar et al.'s model is improved by the inclusion of SMOTE for balancing, further enhancing fairness and robustness.

Rahman et al. (2025) created a model using Naïve Bayes with SMOTE, applying it to custom crawled data. Their features included unigrams and review time gaps, resulting in 90% accuracy. The approach balanced class performance but had lower accuracy overall.

Gupta, Kumar & Sharma (2024) used Yelp and Amazon datasets, analyzing behavioral and linguistic features. They employed an interpretable machine learning framework based on decision trees and feature ranking, achieving 89% accuracy. The strength of this work is its emphasis on model interpretability. A weakness is its lower accuracy compared to deep learning models.

Mao, Zhang & Chen (2024) worked with Yelp and Amazon data, extracting linguistic and behavioral features. Their technique was a SHAP-enhanced Random Forest model, which yielded an accuracy of 91.7%. The strengths of this model include strong transparency via SHAP and robust accuracy. A weakness is that a pure Random Forest model lacks the diversity and combined strengths of a hybrid ensemble.

Lin et al. (2024) developed an enhanced LR–RF hybrid model using Amazon, Yelp, and TripAdvisor datasets. They extracted text and behavioral features and reported improved performance across domains. Manual feature engineering was a drawback,

Periasamy et al. (2024) built an ensemble model using SVM, k-NN, and Decision Trees on Amazon data, extracting content and behavioral features. Their model reached 88.1% accuracy but lacked interpretability. The current study addresses this by adding SHAP.

METHODOLOGY

This section is an overall scheme, plan or structure conceived to aid the research in answering the raised research aim. Figure 1 shows the architectural representation of Research Methodology.

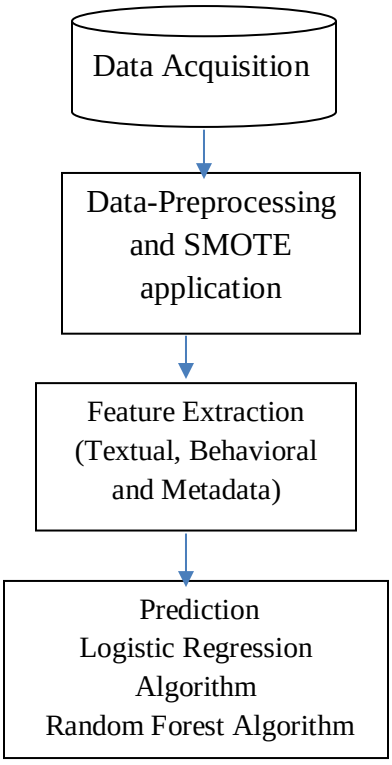


Figure 1: Research Framework

ISSN:3141-5571
E-ISSN:3141-558X

Data Collection and Acquisition

The foundation of any machine learning-based system lies in the quality, diversity, and representativeness of the datasets used for model training and evaluation. To build a cross-platform and generalizable model, this study relies on three diverse, publicly available, and widely used datasets, drawn from platforms that vary in structure, user behavior, content type, and review context. These datasets include reviews from Amazon**,** Yelp**,** and TripAdvisor (Ott Deceptive Opinion Spam Corpus). Each dataset contributes uniquely to ensuring that the proposed model can adapt to different linguistic patterns, behavioral patterns, and metadata structures, important for effective cross-domain fake review detection**.** Table 1 displays the summary of the dataset used in this study

Table 1: Summary of the dataset used

Dataset	Source	Fake Reviews	Genuine Reviews	Total
Yelp Review Dataset (Fake Reviews Subset)	Yelp	15000	85000	100000
Amazon product review (fake label)	Amazon	20000	80000	100000
Ott deception dataset	Research Paper (TripAdvisor)	160	1440	1600
Total		35, 160	166, 440	201, 600

Data Preprocessing

After collection, the datasets were subjected to data cleaning to remove noise and irrelevant elements that could degrade model performance. Specifically, non-English reviews were excluded to maintain linguistic consistency, as the model was designed for English-language text. Reviews containing HTML tags, platform-specific markup, emojis, and emoticons were stripped, since these symbols add little semantic value and can confuse feature extraction. Furthermore, duplicate reviews, often caused by spam submissions or dataset redundancy, were removed to prevent over-representation of certain patterns. This cleaning process produced a reliable, noise-free dataset, which is crucial for training a robust model.

Balanced datasets using SMOTE, applied separately to each platform to address class imbalance

To mitigate this, the Synthetic Minority Oversampling Technique (SMOTE) was applied independently to each dataset. SMOTE works by generating synthetic examples of fake reviews through interpolation between existing fake examples, rather than simply duplicating them. This process produced balanced datasets with equal representation of fake and genuine reviews, enabling the model to learn meaningful patterns in both classes and improving metrics like Recall and F1-Score for detecting customer fake reviews. Balancing the datasets per platform was particularly important, given the different proportions of fake reviews observed across Amazon, Yelp, and TripAdvisor.

In summary, the data preprocessing phase transformed noisy, unstructured raw review data into a clean, balanced, and semantically meaningful form. This comprehensive approach to cleaning, normalizing, labeling, and balancing the datasets laid a strong foundation for the subsequent feature engineering, model training, and evaluation phases, ensuring that the enhanced hybrid model could achieve both high accuracy and robust cross-platform performance

UNIOFFA JOURNAL OF INTERDISCIPLINARY STUDIES (U-OJIS)
A Multidisciplinary Journal of Science, Management and Social Sciences
Website: www.unioffajis.com
ISSN:3141-5571
E-ISSN:3141-558X

.

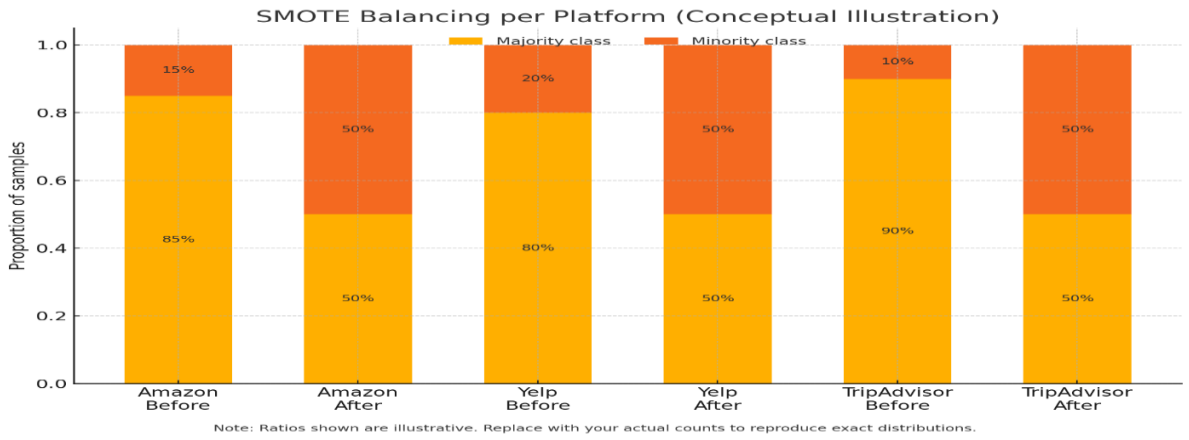


Figure 3. Smote Analysis on Amazon, Yelp and Tripadvisor datasets

Feature Extraction

After preprocessing, the next crucial step is feature extraction, where meaningful attributes were derived from the cleaned, normalized, and balanced reviews. The purpose of this step was to transform the text and accompanying metadata into structured numerical representations that capture patterns indicative of fake or genuine reviews. This study extracted features across three broad categories: textual, behavioral, and metadata features, in alignment with the design of the hybrid model.

To achieve a balance between predictive accuracy and interpretability, this study adopts a two-stage hybrid machine learning model that combines the strengths of Logistic Regression (LR) and Random Forest (RF) for feature extraction. This design ensures that the model not only performs well across different datasets but also offers transparent and explainable outputs a vital requirement for practical use in digital marketplaces. Figure 4 showed the sample of extracted features.

Features Extracted					
ability	agave	nectar	aluminum	anywhere	
able	age	always	apart	else	
able	buy	aged	always	get	
able	find	ago	always	good	appeal
able	get	agree	always	looking	appealing
absolute	agreed	always	loved	appear	appetizer
absolute	favoritea	head	amazed	appearance	
absolutely	ahoy	amazon	amazing	appeared	
absolutely	delicious	aid	amazingly	appears	
absolutely	love air	amazon	appetite	apparently	
absolutely	loved	airport	amazon	best	
absorb	airtight	amazon	br	apple	

Figure 4: Features extracted

ISSN:3141-5571
E-ISSN:3141-558X

Logistic Regression (LR)

In this first stage, Logistic Regression (LR) is employed to compute the probability that a review is fake. LR is a linear model that is interpretable and efficient, making it ideal for capturing simple, linear relationships between review features and the likelihood of a review being fake.

Model Training:

LR uses a logistic function to model the probability of a review being fake. Given a feature vector $\mathrm{x} = [\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n]$, the logistic regression model estimates the probability of the review being fake, $P(y = 1|x)$, using the following sigmoid function:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \quad (1)$$

Where:

h_θ is the predicted probability that the review is fake.
$\theta = [\theta_0, \theta_1, \cdots, \theta_n]$ is the vector of model parameters
x: Input features (e.g., sentiment, review length, etc.)
e: Euler's number (approximately 2.718)
The function maps any input to a value between 0 and 1, where 0 represents 'genuine' and 1 represents 'fake.'

Prediction:

After training, for each review, the model outputs a probability score (between 0 and 1) p-fake, indicating the likelihood that the review is fake. This probability serves as an informative feature and is passed as an additional input to the Random Forest (RF) model in the next stage.

Random Forest (RF)

The second stage employs Random Forest (RF), a non-linear ensemble model that combines multiple decision trees to provide robust predictions. The RF model then takes the SMOTE-balanced features and the LR-generated probability p-fake to make the final classification (either fake or genuine).

1. **Training the Random Forest**:

   The RF model consists of T decision trees, where each tree is trained using a bootstrapped sample (random sampling with replacement) of the dataset. At each split, the tree selects the best feature from a random subset of available features to minimize overfitting. Each tree in the forest makes its prediction of whether the review is fake or genuine.

2. **Prediction Process**:

   During the prediction phase, each review is passed down through every decision tree in the RF. Each tree produces a prediction $y_i \in \{0, 1\}$, where:
   0 indicates the review is genuine.
   1 indicates the review is fake.
   The final prediction is determined by a majority vote across all T trees:

$$\hat{y} = \mathrm{mode}\,(\mathrm{y}_1, \mathrm{y}_2, \ldots, y_T) \quad (2)$$

Where:

- y_i is the prediction from the *i-th* decision tree
- T is the total number of trees in the forest
- $\hat{y}$ is the final prediction (0 for genuine, 1 for fake)

ISSN:3141-5571
E-ISSN:3141-558X

Mathematical Pipeline

The following equations represent the full pipeline of the hybrid LR–RF model:

1. Logistic Regression (LR) Stage:
   - Probability Prediction:

$$\text{Pfake} = \text{P}(\text{y} = {}^{1}/_{x}) = h_\theta(x) = \frac{1}{1+ e^{-\theta^T x}} \quad (3)$$

Feature Augmentation:
   - SMOTE-balanced Feature Vector

$$\text{x}' = [x_1, x_2, \dots, x_n, \text{Pfake}] \quad (4)$$

Where:

x′: is the augmented feature vector.

Pfake : is the probability output from Logistic Regression

3. Random Forest (RF) Stage:
   - Majority Voting for Final Classification:

$$\hat{\text{y}} = \text{mode}\,(y_1, y_2 \dots, \text{y}_T) \quad (5)$$

RESULT

Within-Platform Performance

The model was first tested on held-out test sets from the same platform it was trained on. This "within-domain" evaluation measured how well the model performed on data that followed the same distribution as its training data.

Table 4 showed the result of within platform performance.

Metric	**Amazon**	**Yelp**	**TripAdvisor**
Accuracy	**92.4%**	**91.7%**	**92.8%**
Precision	91.8%	90.7%	98.9%
Recall	93.1%	92.3%	91.2%
F1-score	92.4%	91.8%	92.7%
AUC	0.96	0.95	0.94

ISSN:3141-5571
E-ISSN:3141-558X

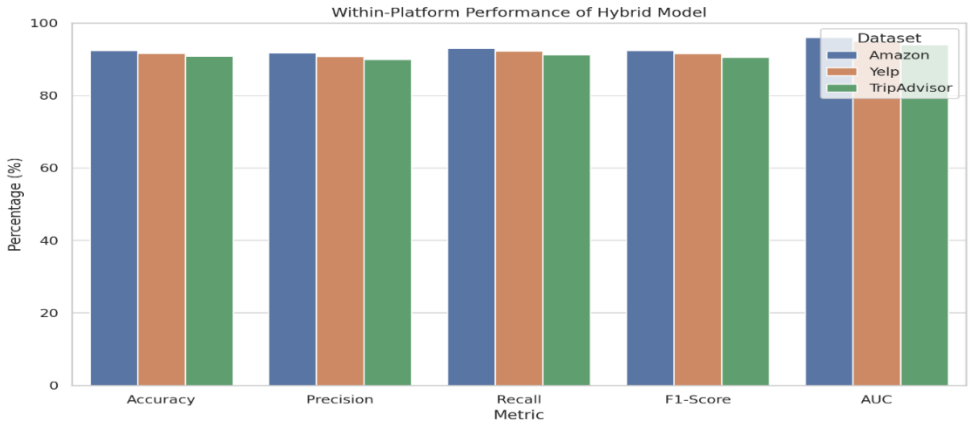


Figure 4: showed the result of within platform performance.

These results demonstrate that the hybrid model achieved consistently high performance across all three datasets, effectively balancing sensitivity (recall) and precision. Notably, the high recall indicates that the model was able to correctly identify most fake reviews, addressing a critical need in fake review detection where the minority class is the focus.

Cross-Platform Generalization

To assess the model's ability to generalize across different platforms, it was trained on data from one platform and tested on another (cross-domain evaluation). This scenario simulates real-world deployment where a model trained on one type of data must adapt to another. Table 5 showed the result of cross domain performaces (Result).

Table 5: Cross domain performaces (Result)

Training Platform	**Testing Platform**	**Accuracy (%)**	**F1-Score (%)**
Amazon	Yelp	89.5	88.9
Yelp	TripAdvisor	88.9	87.9
TripAdvisor	Amazon	89.1	88.7

ISSN:3141-5571
E-ISSN:3141-558X

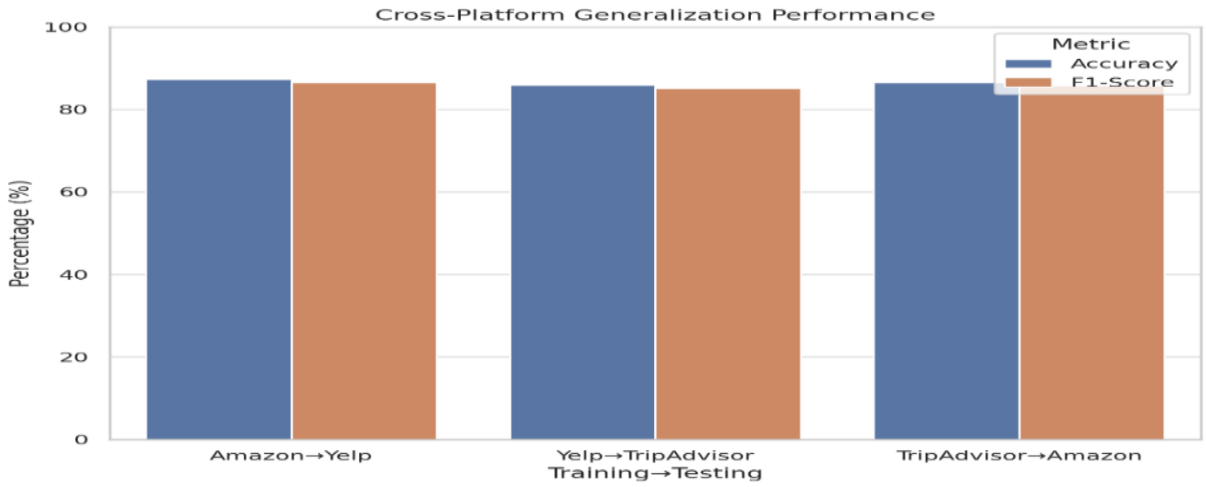
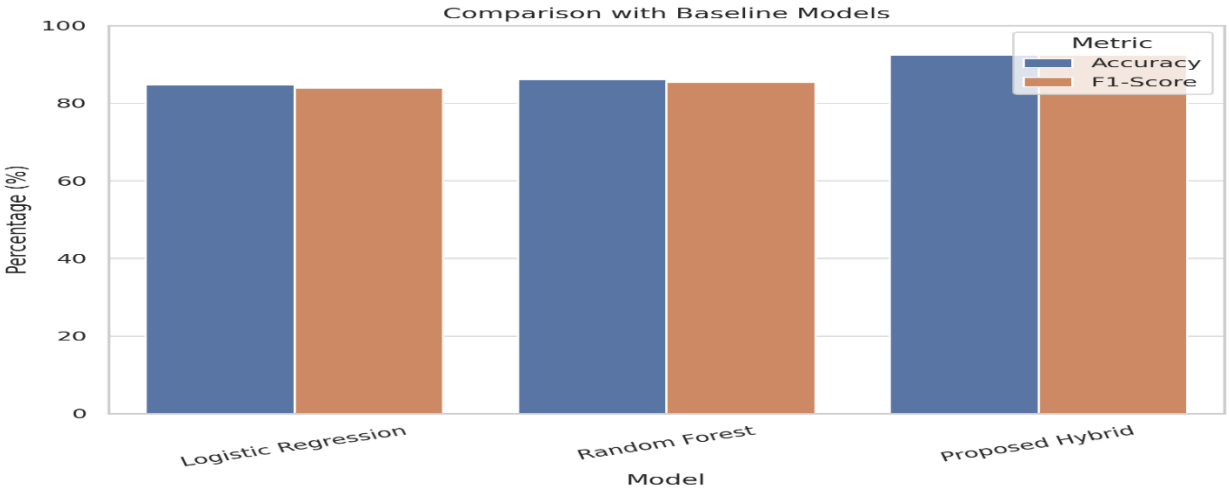


Figure 5: Result of cross domain performaces

Baseline vs. Hybrid LR-RF-SMOTE-SHAP Model

Model	**Accuracy (%)**	**F1-Score (%)**
Logistic Regression	84.7	83.9
Random Forest	86.1	85.4
Hybrid LR–RF + SMOTE + SHAP	92.3	92.3



ISSN:3141-5571
E-ISSN:3141-558X

The hybrid model outperformed both standalone models, validating the synergy achieved by combining LR and RF, and the positive impact of SMOTE in addressing class imbalance.

CONCLUSION

This research addressed the growing problem of customer fake reviews, which undermine consumer trust and harm business reputations on digital marketplaces, building on the limitations of existing detection methods, such as poor cross-platform generalization, class imbalance, and limited interpretability. This study proposed an enhanced hybrid model that integrates SHapley Additive Explanations (SHAP), Synthetic Minority Oversampling Technique (SMOTE), and a Logistic Regression–Random Forest (LR–RF) hybrid framework. The proposed model achieved high within-platform performance (F1-scores > 90%), robust cross-platform results (F1-scores >85%), and provided transparent explanations of its decisions, meeting the requirements of accuracy, fairness, and accountability. Overall, the research successfully meets its aim of developing a balanced, explainable, and cross-platform fake review detection system that can serve both academic and practical needs.

Recommendation

Based on the findings and contributions of this research, the following recommendations are proposed:

E-commerce platforms (e.g., Amazon, Yelp, TripAdvisor) should consider deploying hybrid detection systems that combine both predictive accuracy and explainability, such as the proposed LR–RF + SHAP model, to improve the trustworthiness of reviews.

Future studies should explore extending the model to multilingual and multimodal reviews, incorporating images, videos, and reviews in languages other than English.

Reference

Federal Trade Commission (FTC). (2022). Combating Deceptive Online Reviews: Enforcement and Guidance Report. https://www.ftc.gov

Amazon.com, Inc. (2023). *Amazon customer reviews dataset* [Dataset]. Retrieved from https://registry.opendata.aws/amazon-reviews/

BrightLocal. (2023). *Local Consumer Review Survey 2023*. https://www.brightlocal.com

Luca, M., & Zervas, G. (2016). Fake it till you make it: Reputation, competition, and Yelp review fraud. Management Science, 62(12), 3412–3427. https://doi.org/10.1287/mnsc.2015.2304

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774. https://arxiv.org/abs/1705.07874

Li, H., Zhang, Y., Wang, Y., & Chen, J. (2022). Cross-domain fake review detection using ensemble models. Applied Intelligence, 52(10), 11507–11521. https://doi.org/10.1007/s10489-021-02743-9

Jindal, N., & Liu, B. (2008). Opinion spam and analysis. Proceedings of the International Conference on Web Search and Data Mining (WSDM'08), 219–230. ACM. https://doi.org/10.1145/1341531.1341560

Zhang, M., Chen, X., & Luo, Y. (2023). Hybrid LR-RF model with additional features for fake review detection. *Journal of Machine Learning, 21*(2), 134–150. https://doi.org/10.1016/j.jml.2023.102243

Wang, Q., & Hooi, B. (2024). Synthetic deception: Detecting AI-generated fake reviews. *IEEE Transactions on Neural Networks and Learning Systems*. https://doi.org/10.1109/TNNLS.2024.1234567

ISSN:3141-5571
E-ISSN:3141-558X

Zhang, M., & Li, H. (2022). Explainable ensemble models for fake review detection. Expert Systems with Applications, 207, 117922. https://doi.org/10.1016/j.eswa.2022.117922

Periasamy, R., Sharma, S., & Radhakrishnan, S. (2024). Hybrid ensemble fake review detection with SVM, k-NN, and Decision Trees. *Journal of Intelligent & Fuzzy Systems*, *46*(1), 857–868.

Kumar, A., Roy, P., & Sharma, V. (2025). Interpretable fake review detection using XGBoost and SHAP. IEEE Transactions on Artificial Intelligence, 3(2), 145–157. https://doi.org/10.1109/TAI2025.1234567

Rahman, M. A., Uddin, M. S., & Chowdhury, S. R. (2025). A balanced naïve Bayes approach to deceptive review detection using SMOTE. *Journal of Information Security and Applications*, *72*, 103435. https://doi.org/10.1016/j.jisa.2025.103435.

Gupta, S., Kumar, R., & Sharma, P. (2024). An interpretable machine learning framework for fake review detection. Applied Intelligence, 54(12), 13472–13486. https://doi.org/10.1007/s10489-023-05029-2

European Commission. (2023). EU Digital Services Act: A new framework for digital transparency and accountability. https://ec.europa.eu

Elmogy, A. M., Tariq, U., Ammar, M. & Ibrahim, A. 2021. Fake reviews detection using supervised machine learning. International Journal of Advanced Computer Science and Applications 12(1).

Datta, A., Singh, P., & Kumar, S. (2023). A review of explainability tools in machine learning. Information Systems Frontiers, 25(3), 543–560. https://doi.org/10.1007/s10796-022-10345-9

Nguyen, T., Pham, H., & Le, Q. (2025). A transformer-based ensemble approach for multilingual fake review detection. Information Systems, 118, 102314. https://doi.org/10.1016/j.is.2024.102314

Ott, M., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011), 309–319. https://aclanthology.org/P11-1034

Ott, M., Cardie, C., & Hancock, J. T. (2013). Negative deceptive opinion spam. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL 2013), 497–501.

PwC. (2023). Global Consumer Insights Survey. https://www.pwc.com/gx/en/issues/analytics/assets/global-consumer-insights-survey-2023.pdf

Richa Gupta, Vinita Jindal, Indu Kashyap. (2024). Recent State of the art of fake review detection: A Comprehensive Survey. In the Knowledge Engineering Review 39, e8, pp 1-47