Detecting Fake Online Reviews using Random Forest with Generative Adversarial Networks

Deepa.S[1], Kanishka R[2], Anandhi A[3], Sathiyavani M[4], Thirisha V[5]

[1]*Assistant Professor, Department of Computer Science and Engineering, Maha Barathi Engineering College (Affiliated to Anna University), Chinnasalem (Tk), Kallakurichi (Dt)-606 201.*

[2,3,4,5]*UG Student, Department of Computer Science and Engineering, Maha Barathi Engineering College (Affiliated to Anna University), Chinnasalem (Tk), Kallakurichi (Dt)-606 201.*

***Abstract-* User reviews have emerged as essential in customer selection and business image building because of e-commerce development, social media, and online service growth. The widespread presence of deceptive review manipulation creates considerable problems when we aim to guarantee their trustworthiness. The detection of fraudulent reviews has become a crucial research problem because they lead customers astray while degrading business product rankings and causing financial losses. Traditional fake review detection methods, such as rule-based approaches, sentiment analysis, and conventional machine learning algorithms, face multiple disadvantages in their operation. Many present algorithms face difficulties working within extensive feature domains, which causes computational challenges and decreases their interpretability level. This study proposes integrating Regular Expression Matching (REM) for preprocessing together with Principal Component Analysis (PCA) for feature selection to use as components of a novel Random Forest with Generative Adversarial Networks (RF-GAN) classification model. REM functions to standardize textual information by eliminating extraneous characters, symbols, and normalization errors. PCA's dimensional reduction techniques maintain fundamental patterns from the data, so both computational efficiency and interpretability benefits become possible. RF-GAN merges GAN-based artificial data creation with RF ensemble performance to create a solution that handles data imbalance issues and advances classification outcomes. Results show that RF-GAN produces 95.2% accuracy as a solution for detecting fake reviews in dynamic online environments.**

***Keywords-* e-commerce, user review, fake review, sentiment analysis, REM, PCA, RF-GAN**

I. INTRODUCTION

The digital era depends heavily on customer reviews because they determine how customers view products, their buying choices, and how businesses maintain their trustworthiness. User-generated content is a vital tool for business platforms, hospitality services, and online marketplaces because it establishes trust between service providers and customers. The increasing online consumer base relies on reviews to evaluate product standards, service quality, and complete satisfaction with their purchases. The need for businesses to preserve excellent digital profiles and customer feedback investigations before purchasing has become essential for companies. Online reviews constitute today's digital commercial foundation because consumers use them to collect user-submitted data that guides their purchasing decisions [1].

Online reviews have significant disadvantages due to the widespread existence of fake reviews designed to control buyer preferences during their purchase journey. Multiple reasons drive users to post false reviews to manipulate product reputations artificially, damage competitor reputations, or create artificial market trends. Deceptive review practices substantially affect businesses and consumers because they create review authenticity problems, decrease online platform trust, and result in financial losses. Multiple research findings show that spam reviewers have become more advanced while bot-written content grows alongside incentive programs for fabricating reviews, thus making standard detection techniques less successful. Rule-based and machine learning detection models face weaknesses when processing shifting deceptive strategies while working in multi-dimensional feature domains and unbalanced information alongside each other. Therefore, their operational results prove inadequate. The language characteristics of fake reviews align with natural ones so that standard supervised learning detection techniques prove inefficient [2,3].

The study develops a novel solution that applies Random Forest with Generative Adversarial Networks (RF-GAN) for detecting fake online reviews. The RF-GAN model combines automatic GAN technology and data synthesis to correct class distribution problems and create stronger classifiers. The RF classifier receives training on the augmented dataset to identify complex structures in phony reviews. Experimental data reveal that the proposed method boosts accuracy levels and total classification ability at efficient computing costs [4,5]. The research provides an efficient and reliable approach to detect fake online reviews by overcoming the limitations of current methods to strengthen digital platform fake review prevention.

II. LITERATURE SURVEY

The authors in [6] presented an Integrated Approach (IA) that merged linguistic together with behavioural attributes and statistical elements to estimate review authenticity. The improved classification performance came with a drawback because redundant features caused additional computational complexity. The Hybridized Approach (HA) presented in [7] combines transformer-based models with traditional machine learning classifiers to boost the detection precision. The strategy experienced limitations because it required labeled information to function effectively within semi-supervised situations.

The Deep Learning and Aspect Features Fusion Model (DL-AFFM), introduced in [8] to assess e-commerce authenticity by analyzing aspect-based sentiments, faced issues with short, ambiguous reviews, thus reducing its ability to generalize. A Data Resampling and Feature Pruning-based Ensemble Model (DRFP-EM) presented in [9] performed data imbalance management alongside feature selection and parameter tuning. Still, its rigorous preprocessing demands elevated computational requirements, which complicate scalability.

The author [10] introduced a Sentiment Majority Voting Classifier with Transfer Learning-Based Feature Engineering (SMVC-TLFE) as a deepfake tweet analysis system, which performed well at sentiment classification while having an extensive manual feature selection process that increased the labour requirements. The researchers in [11] developed a Multilingual Spam Review Detection Model (MSRDM) using pre-trained word embeddings and Weighted Swarm Support Vector Machines (WS-SVMs). Still, the model faced challenges due to computational demands and sensitivity to noisy review data.

The combination of a Deep Neural Network with Emotion Mining and Word Embeddings (DNN-EMWE) was developed in [12] to detect fraudulent consumer reviews through emotional context analysis, even though its training focused on specific domains that limited cross-dataset applicability. During the development process, [13] created the ML-SRPM Machine Learning-Based Sentiment and Rating Prediction Model for tourist reviews using supervised learning. However, it faced difficulties working with biased training data that limited its practical viability.

This domain benefits from detection models designed to check fake news, where researchers developed a Multiple Features-Based Fake News Detection Model (MF-FNDM) in [14] using deep learning, but demonstrated performance diminishment when processing adversarial created fake reviews. The authors in [15] demonstrated an Ensemble Machine Learning Model with Effective Feature Extraction (EFE-EML) for fake news detection through multiple classifiers combined with optimized features. Yet, it required enormous computational power incompatible with real-time scenarios. These investigative studies provide essential knowledge about fake review detection. Yet, they demonstrate ongoing difficulties involving computations running slowly, duplicate features, uneven data distribution, and limited capability to adapt across different domains.

III. PROPOSED METHOD

In this section, we briefly describe the performance of the proposed method for more accurately detecting fake online reviews. The process has three phases: preprocessing, feature selection, and classification. The Amazon review dataset from the Kaggle website performed those three phases. In the first phase, the REM was deployed, followed by the PCA, and in the third phase, the RF-GAN method was deployed.

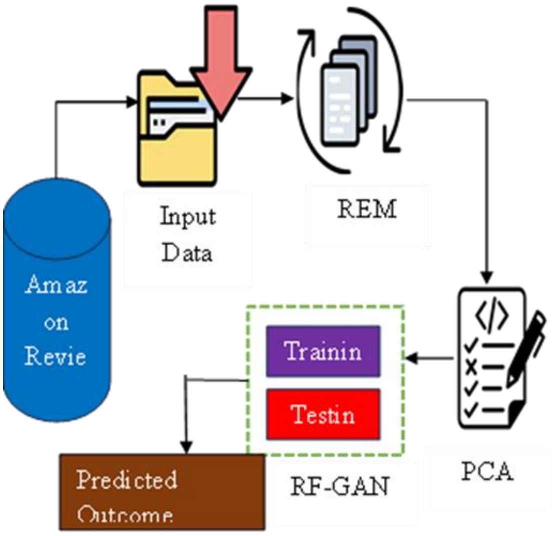

Fig. 1 Architecture diagram of the proposed method

Figure 1 shows the fake online review detection workflow using the Amazon Review Dataset with the RF-GAN approach. The Amazon platform provides raw data stored as an input dataset before collection. The text passage proceeds through a preprocessing task with REM that eliminates strange characters alongside removing unimportant formatting elements and tokens. The refined dataset moves to PCA for feature choice before distribution, keeping only the vital aspects needed for model preparation. After selecting essential features, the dataset is separated into training and testing partitions for developing and validating the classification system. The RF-GAN classification engine runs at this stage. Combining Random Forest ensemble learning strength with GANs' generative capacity allows the model to maximize its classification potential for dealing with unbalanced or scarce labeled data sets. Following training and testing, the RF-GAN model predicts whether the review is real or fake.

A.Regular Expression Matching (REM)

Regular Expression Matching (REM) is the fundamental processing method that preprocesses data extracted from the Amazon Review Dataset. Since raw review data collection, REM facilitates systematic text cleaning through pattern detection to format the data. The system removes all additional punctuation, special characters and numbers, stop words, and URLs, and significant blank spaces that naturally appear in user-generated content. REM attains structured and uncluttered text through pattern detection algorithms to upscale the data quality for the following feature selection stage. The initial phase takes a raw Amazon dataset as input, which is presented by equation 1 during fake review detection.

$$D = \{r_1, r_2, r_3, \dots, r_n\} \quad (1)$$

The D set contains the user-related reviews and the individual unprocessed review documents indicated as r_i. The reviews contain numerous noise sources, including HTML tags, emojis, special characters, numbers, and inconsistent formatting. Each review must undergo REM-based cleaning to create machine learning usable inputs. The REM function applies its symbols to data processing through Equation 2.

$$\mathrm{r}_\mathrm{i}' = \mathrm{REM}(\mathrm{r}_\mathrm{i}) \quad (2)$$

The application of REM results r_i in the cleaned review output r_i'. The REM function removes unwanted tokens by running different regular expression pattern operations during processing. The cleaning operation uses a pattern through Equation 3 to eliminate special characters and punctuation.

$$\mathrm{r}_\mathrm{i}' = \mathrm{r}_\mathrm{i} \backslash \mathrm{Regex}([\backslash \mathrm{W}]+) \quad (3)$$

is used; to eliminate numbers, the pattern,

$$\mathrm{r}_\mathrm{i}' = \mathrm{r}_\mathrm{i} \backslash \mathrm{Regex}(\backslash \mathrm{d}+) \quad (4)$$

is applied; and to compress excessive whitespace into a single space, the pattern,

$$\mathrm{r}_\mathrm{i}' = \mathrm{r}_\mathrm{i} \backslash \mathrm{Regex}(\backslash \mathrm{s}+) \quad (5)$$

is performed. The expressions serve as a framework for ordered text cleaning to deliver consistent results without distortions. The dataset becomes available after the cleaning operation runs on every review,

$$\mathrm{D}' = \{\mathrm{r}_1', \mathrm{r}_2', \mathrm{r}_3', \dots, \mathrm{r}_\mathrm{n}'\} \quad (6)$$

After refinement, the newly separated dataset D' exhibits standardized information, leading to better analytical performance.

B. Principal Component Analysis (PCA)

The Amazon review dataset following REM preprocessing results in a clean dataset $D' = \{r_1', r_2', r_3', \dots, r_n'\}$ where each review r_i' contains cleaned text without noise elements, including punctuation and numbers, as well as excessive whitespace. Textual reviews get transformed into numerical feature vectors using TF-IDF or word embeddings, which results in an $X \in R^{n \times m}$ matrix. The n represents reviews (1,800,000) and m represents features (terms or tokens). The matrix contains dimensions that exceed the number of samples, so Principal Component Analysis (PCA) serves to decrease its dimensionality before analysis. The first PCA procedure requires computing the mean-centered feature matrix $\bar{X}$ by performing column-wise mean subtraction on features. Next, in equation 7 we calculate the covariance matrix,

$$C = \frac{1}{n-1}\bar{X}\wedge T\bar{X} \qquad (7)$$

C represents the $R^{n\times m}$ matrix elements that indicate the relationship strength between dataset features. C contains elements c_{ij} demonstrating the covariance relationships between the i^{th} and j^{th} features throughout all Amazon reviews. The eigendecomposition performed on matrix C through PCA produces equation 8.

$$\mathrm{Cv} = \lambda\mathrm{v} \qquad (8)$$

The eigenvectors v represent principal components, while the eigenvalues λ indicate the captured variance at each principal component. The eigenvalues guide the top k components' selection process, which preserves the maximum data variability. We create the projection matrix $P \in R^{n\times k}$ from selected k eigenvectors at the top of the list. The new feature matrix X_{PCA} of reduced dimension can be computed through formula 9.

Here, v are the eigenvectors (also called principal components), and λ are the eigenvalues representing the amount of variance captured by each principal component. These eigenvectors are sorted in descending order of their corresponding eigenvalues to select the top k components that retain the most variance in the data. We construct the projection matrix $P \in R^{n\times k}$ using the top k eigenvectors. The dimensionality-reduced dataset is then computed in equation 9,

$$X_{PCA} = X.P \qquad (9)$$

The new feature matrix X_{PCA} consists of k principal components per review and has dimensions $n \times k$ instead of the original $m \times m$ features. Using this transformation process, Amazon reviews can store the most informative patterns of linguistic structure and specific word usage, which signals fake or genuine reviews and simultaneously eliminates and reduces redundancy.

C. Random Forest with Generative Adversarial Networks (RF-GAN)

An optimized feature matrix $X_{PCA} \in R^{n\times k}$ results from PCA dimensionality reduction and becomes the Random Forest with Generative Adversarial Networks (RF-GAN) classification model input. The proposed RF-GAN framework creates an advanced detection system by combining RF solid footing with GANs generation methods to boost fake online review detection. The RF component is a strong decision-maker consisting of various decision trees $T_1, T_2, \ldots, T_M$. The training process of each tree happens with a sample selection from the dataset through bootstrap aggregation (bagging). The prediction $h_m(x_i)$ emerges from each decision tree when analyzing feature vector x_i in Amazon reviews, but the final class label $\hat{y}_i$ emerges through majority voting according to equation 10.

$$\hat{y}_i = mode\big(h_1(x_i), h_2(x_i), \ldots, h_M(x_i)\big) \qquad (10)$$

$\hat{y}_i$ represents the classifier's output, which is either 0 for genuine reviews or 1 for fake reviews. Integrating a GAN with the classifier enhances its generalization ability and capability to process complex or restricted labeled data sets. As the fundamental part of a GAN setup, users can find two fundamental artificial neural networks: The Generator (G) and the Discriminator (D). The generator $G(z;\theta_g)$ accepts random input noise z distributed according to $p_z(z)$ to generate imitations $G(z)$ of actual reviews. The discriminator model $D(x;\theta_d)$ tries to identify genuine reviews from the dataset and synthetic reviews that G has produced. The goal of GAN training is achieved through equation 11, expressing a min-max game objective function.

$$\min_G \max_D V(D,G) = E_{x\sim pdata(x)}[\log D(x)] + E_{z\sim p_z(z)}\Big[\log\Big(1 - D\big(G(z)\big)\Big)\Big] \qquad (11)$$

The discriminator D from RF-GAN utilizes the Random Forest classifier as an enhancement to conduct partial validation by processing both real and synthetic samples. Using this hybrid system improves generalizable pattern recognition while enhancing detection through its ability to force the generator toward creating higher-quality review-like fake samples that evaluate RF classification capabilities. Training extends the ability of the generator to produce data that looks real and builds the RF's capability to detect phony patterns simultaneously. When implemented on the Amazon reviews dataset, this approach accurately separates genuine from deceptive content. The RF ensures high interpretability and stability, and the GAN helps solve data imbalance by generating more training data.

IV. RESULT AND DISCUSSION

In this section, we evaluate the performance of the RF-GAN using metrics such as accuracy, and time complexity. This evaluation is compared against existing methods, including HA, WS-SVM, and DNN-EMWE, to identify reliable and accurate fake review detection.

Table 1. Simulation Parameters

Parameters	Values
Name of the Dataset	Amazon reviews
No of Values	34,686,770
Training Samples	1,800,000
Testing Samples	200,000
Programming Language	Python
Development Tool	Jupyter Notebook

The dataset employed for this study is the Amazon Reviews collection with 34,686,770 values. The dataset underwent division for training purposes, using 1,800,000 units, and testing requires 200,000. The system operates through the Python programming language, while Jupyter Notebook is the primary development environment for all data processing steps and analysis, and model training tasks.

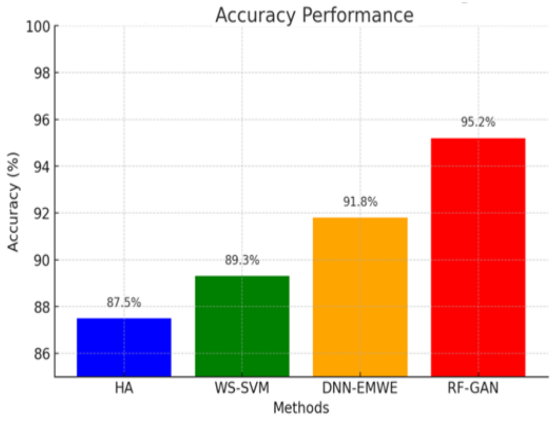

Fig. 2 Performance Analysis of Accuracy in %

The figure 2 compares the effectiveness of different fake review detection methods applied to the Amazon reviews dataset. The proposed RF-GAN method achieved the highest accuracy of 95.2%, surpassing previous approaches such as HA 87.5%, WS-SVM, and DNN-EMWE 91.8%. The RF-GAN model enhances classification by generating synthetic review data to address class imbalance, allowing the RF classifier to learn robust patterns distinguishing genuine from fake reviews. This performance improvement highlights the efficacy of hybrid models in handling deceptive reviews, reducing false positives, and adapting to evolving fraud tactics in e-commerce platforms.

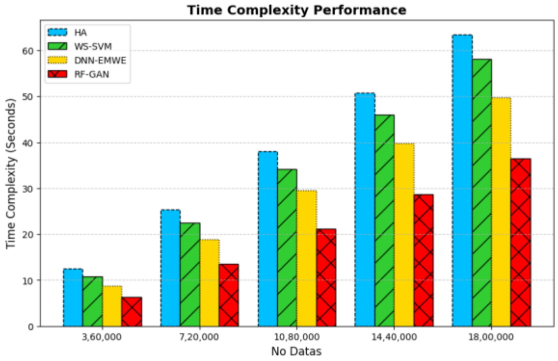

Fig. 3 Performance Analysis of Time Complexity

The figure 3 represents the computational efficiency of different methods HA, WS-SVM, DNN-EMWE, and the proposed RF-GAN when applied to the Amazon Reviews dataset. It confirms that RF-GAN achieves better computational efficiency, making it more suitable for large-scale fake review detection.

V. CONCLUSION

In conclusion, the study presented an RF-GAN model that successful integrates fake review detection capabilities. The proposed workflow collects the Amazon reviews dataset, which contains genuine and fake feedback from the dataset as its initial process. The REM preprocessing step functions to normalize text data by removing stopwords, symbol interruptions, and special characters for a consistent input format. The subsequent step involves using Principal Component Analysis to maintain crucial information within dimensional data while reducing its high dimensions for classification purposes. The RF-GAN receives processed data, which allows GANs to create synthetic data to resolve class imbalance, while RF uses learned patterns to classify reviews. The integrated approach provides better detection performance and bias minimization in the classification process. The Amazon reviews dataset promoted an exceptional detection accuracy of 95.2% when the presented model underwent experimental testing. The proposed system demonstrated excellent capabilities in working with unbalanced data while learning evolving deceptive patterns and minimizing false alerts. According to this study, supervised learning techniques deliver effective solutions to maintain authenticity and trust for consumer feedback in online platforms.

REFERENCES

[1] Roobini, M. S., Chowdary, B. N., Chowdary, J. M., Aruna, J., & Ponraj, A. (2020). Detection of Fake Online Reviews Using Semi-Supervised and Supervised Learning. Journal of Computational and Theoretical Nanoscience, 17(8), 3577-3580.

[2] 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).

[3] R. Mohawesh et al., "Fake Reviews Detection: A Survey," in IEEE Access, vol. 9, pp. 65771-65802, 2021, doi: 10.1109/ACCESS.2021.3075573.

[4] Ahmed, A. A., Aljabouh, A., Donepudi, P. K., & Choi, M. S. (2021). Detecting Fake News Using Machine Learning : A Systematic Literature Review. ArXiv. https://arxiv.org/abs/2102.04458

[5] M. F. Mridha, A. J. Keya, M. A. Hamid, M. M. Monowar and M. S. Rahman, "A Comprehensive Review on Fake News Detection With Deep Learning," in IEEE Access, vol. 9, pp. 156151-156170, 2021, doi: 10.1109/ACCESS.2021.3129329.

[6] E. Abedin, A. Mendoza, P. Akbarighatar and S. Karunasekera, "Predicting Credibility of Online Reviews: An Integrated Approach," in IEEE Access, vol. 12, pp. 49050-49061, 2024, doi: 10.1109/ACCESS.2024.3383846.

[7] S. Xu, H. Cuan, Z. Yin and C. Yin, "A Hybridized Approach for Enhanced Fake Review Detection," in IEEE Transactions on Computational Social Systems, vol. 11, no. 6, pp. 7448-7466, Dec. 2024, doi: 10.1109/TCSS.2024.3411635.

[8] S. M. Abd-Alhalem, H. A. Ali, N. F. Soliman, A. D. Algarni and H. S. Marie, "Advancing E-Commerce Authenticity: A Novel Fusion Approach Based on Deep Learning and Aspect Features for Detecting False Reviews," in IEEE Access, vol. 12, pp. 116055-116070, 2024, doi: 10.1109/ACCESS.2024.3435916.

[9] J. Yao, Y. Zheng and H. Jiang, "An Ensemble Model for Fake Online Review Detection Based on Data Resampling, Feature Pruning, and Parameter Optimization," in IEEE Access, vol. 9, pp. 16914-16927, 2021, doi: 10.1109/ACCESS.2021.3051174.

[10] M. Khalid et al., "Novel Sentiment Majority Voting Classifier and Transfer Learning-Based Feature Engineering for Sentiment Analysis of Deepfake Tweets," in IEEE Access, vol. 12, pp. 67117-67129, 2024, doi: 10.1109/ACCESS.2024.3398582.

[11] A. M. Al-Zoubi, A. M. Mora and H. Faris, "A Multilingual Spam Reviews Detection Based on Pre-Trained Word Embedding and Weighted Swarm Support Vector Machines," in IEEE Access, vol. 11, pp. 72250-72271, 2023, doi: 10.1109/ACCESS.2023.3293641.

[12] Hajek, P., Barushka, A. & Munk, M. Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining. Neural Comput & Applic 32, 17259–17274 (2020). https://doi.org/10.1007/s00521-020-04757-2

[13] Puh, K. and Bagić Babac, M. (2023), "Predicting sentiment and rating of tourist reviews using machine learning", Journal of Hospitality and Tourism Insights, Vol. 6 No. 3, pp. 1188-1204. https://doi.org/10.1108/JHTI-02-2022-0078

[14] Sahoo, S. R., & Gupta, B. (2021). Multiple features based approach for automatic fake news detection on social networks using deep learning. Applied Soft Computing, 100, 106983. https://doi.org/10.1016/j.asoc.2020.106983

[15] Hakak, S., Alazab, M., Khan, S., Gadekallu, T. R., Maddikunta, P. K. R., & Khan, W. Z. (2021). An ensemble machine learning approach through effective feature extraction to classify fake news. Future Generation Computer Systems, 117, 47-58. https://doi.org/10.1016/j.future.2020.11.022