Leveraging Machine Learning for Authentic Review Verification

Priyanshu Singh[1], Priyanka Yadav[2], Nidhi Saxena
Department of Computer Applications Babu Banarsi Das University Lucknow,India

***Abstract*— E-commerce sites are growing fast, bringing with them an explosion in user-generated content such as online reviews that shape the purchases people make. Unfortunately, fake reviews – deliberately misleading feedback intended to influence opinions – are becoming an increasingly serious problem. They destroy trust and hurt businesses. Traditional approaches, such as checking reviews manually or through simple rule-based systems, are no longer adequate to cope with the volume and complexity of such fake reviews. The core of this study will involve the use of ML and NLP techniques for an efficient review-fake detector based on the Fake reviews Dataset. Comparative analysis of various models including SVMs, Naive Bayes, and even complex deep learning architecture such as LSTM is provided with a view to understand what features distinguish linguistic and textual patterns between authentic and deceitful reviews. In addition to this, additional metadata features, such as reviewer behavior and temporal trends, enhance the accuracy of detection. The findings suggest the possibility of using machine learning to combat fake reviews and call for strong, scalable solutions to maintain trust and integrity in the online ecosystem.**

***Keywords*— component, formatting, style, styling, insert (key words)**

INTRODUCTION

Online reviews are crucial in today's digital economy as they help consumers decide on things such as hotels, restaurants, and even online shopping. They give information about product quality, prices, and reliability of the seller. As online shopping increases, reviews play an even greater role in influencing consumer behavior and market trends. However, there is a problem with fake reviews. Some are written by paid individuals who have no experience with real products, either praising products or criticizing competitors. Others are generated in bulks by automated systems using highly advanced text-generation technology making them even harder to detect. [1] The pandemic accelerated the shift to online shopping as lockdowns and safety concerns made e-commerce the go-to option for consumers. The convenience of home delivery and the ability to buy everything from groceries to electronics further cemented this trend. [2] However, the rise in online shopping brought a surge in fake reviews.

Dishonest sellers and malicious actors exploited the growing reliance on reviews to manipulate ratings, misleading buyers and distorting perceptions. These fake reviews, whether overly positive or unfairly negative, harm businesses, damage brand reputations, and disrupt market dynamics. As a result, addressing this issue has become critical for maintaining trust in online marketplaces. Detecting and mitigating fake reviews is now a major focus, with researchers exploring advanced methods and tools to identify fraudulent content. The ultimate goal is to safeguard consumer trust and ensure that e-commerce platforms remain reliable and transparent. The growth of ecommerce sales worldwide forecast from pre-covid to post-covid can be seen through fig.1 below-:

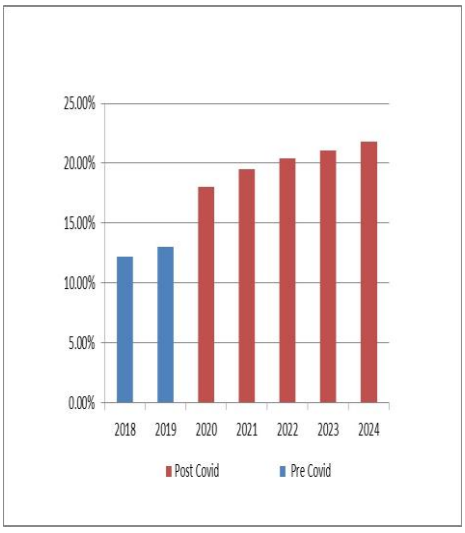
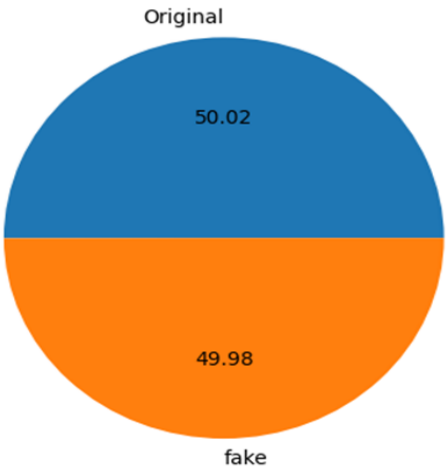


Fig.1 GRAPH: E-commerce Share of Total Global Retail

About 81% of consumers trust online reviews, making them a key factor in purchasing decisions. These reviews are highly visible on e-commerce sites, social media, and other platforms, encouraging people to rely on others' experiences without

verifying their authenticity. Unfortunately, some businesses exploit this trust through "review manipulation." They post fake positive reviews to promote their products or fake negative ones to harm competitors, all for financial gain. These deceptive practices mislead buyers, erode trust, and compromise the credibility of online shopping platforms, making it harder to maintain authenticity. [1]

The pie chart in Fig.2 illustrates the composition of a dataset used to train and test machine learning and deep learning models. It shows that the dataset consists of two nearly equal categories: Original Reviews (50.02%): These are genuine, user-generated reviews, forming a slight majority. Fake Reviews (49.98%): These are artificially created or manipulated reviews intended to deceive readers. This balanced dataset (close to 50-50) is ideal for training machine learning models because it helps prevent class imbalance issues. Class imbalance often causes models to favor the majority class, reducing overall performance. Here, the balanced ratio ensures that the models learn to differentiate between fake and original reviews effectively. The models were trained and tested using this dataset with different machine learning models like Naive Bayes, Random Forest, Logistic Regression and Voting Classifier and then deep learning model LSTM, to evaluate their ability to distinguish between fake and original reviews, with the aim of improving detection accuracy and ensuring reliable outcomes. [3]



II. RELATED WORK

The detection of fake reviews using machine learning models has been a research topic since 2007. Generally, the problem of fake review detection relies on two main categories of features: textual and behavioral features. [4] Textual features pertain to the content and language used in the reviews, focusing on aspects such as sentiment, grammar, word choice, and phrasing. On the other hand, behavioral features are features which concern the behavior of reviewers themselves, such as style, frequency of reviewing, and emotional expressions in the text. Though textual features have already been studied extensively and proven necessary, it is not possible to underestimate the significance of behavioral features, which drastically impact the performance of detection models for fake reviews.

Several studies have explored the use of textual features for detecting fake reviews:

1. In [5], supervised machine learning techniques such as SVM, Naive Bayes , KNN, K-Star, and Decision Tree were applied to labeled datasets of movie reviews ranging from 1,400 to 10,662 reviews.
2. In [6], classifiers like Naive Bayes, Decision Tree, SVM, Random Forest, and Maximum Entropy were used to detect fake reviews from a dataset of 10,000 negative tweets about Samsung products.
3. In [7], SVM and Naive Bayes were employed on a dataset of 1,600 reviews from 20 popular hotels in Chicago to identify fake reviews.
4. In [8], deep learning models, including CNN, RNN, GRNN, and Bi-directional GRNN, were used to detect deceptive opinion-based spam in reviews related to hotels, restaurants, and doctors.

While these studies achieved promising results, they primarily focused on textual features, with limited emphasis on analyzing behavioral features in greater depth.

III. IDENTIFICATION OF PROBLEM

In today's fast-paced world, technology and online marketing play a crucial role. People increasingly prefer online shopping because it is cost-effective, time-saving, and offers a wide variety of choices. However, to boost sales, some marketers create fake reviews, misleading consumers and making it difficult for them to judge products accurately. Detecting fake reviews is essential to help consumers make informed purchasing decisions. This work focuses on using supervised machine

learning models to identify fake reviews. Online reviews heavily influence consumer choices, but fake reviews, also called opinion spam, have become a major issue for online platforms. Early detection methods relied on models like Support Vector Machines (SVM) and Naive Bayes, using linguistic features such as n-grams and part-of-speech tags. Later advancements added temporal patterns like review burstiness and metadata (e.g., reviewer activity and rating behavior) for better accuracy. Recent research explores deep learning methods, including FakeGAN with adversarial training and FRD-LSTM, which use deep word representations and dimensionality reduction. Despite their success, these models face challenges in adapting to diverse datasets and large-scale real-world applications. This research aims to build a more robust and scalable model by combining linguistic, temporal, and behavioral features, while exploring unsupervised and semi-supervised learning techniques to enhance detection accuracy for real-world scenarios.

IV. PROPOSED METHODOLOGY

The detection of fake reviews is a critical task in ensuring the reliability of online platforms. To address this challenge, we propose a comprehensive machine learning-based methodology that combines traditional machine learning classifiers with advanced deep learning techniques. This approach leverages the strengths of various algorithms and culminates in the application of Long Short-Term Memory (LSTM) networks, which provide superior performance in identifying fake reviews.

Dataset Preparation

Dataset Acquisition: A dataset of reviews labeled as real or fake is collected from trusted sources.

Data Preprocessing:

- Text cleaning: Removal of stop words, punctuation, special characters, and converting text to lowercase.
- Tokenization: Splitting text into individual words or tokens.
- Feature extraction: Using methods like Term Frequency-Inverse Document Frequency (TF-IDF) to convert textual data into numerical form.
- Train-test split: Dividing the dataset into training and testing sets, e.g., 80%-20%.

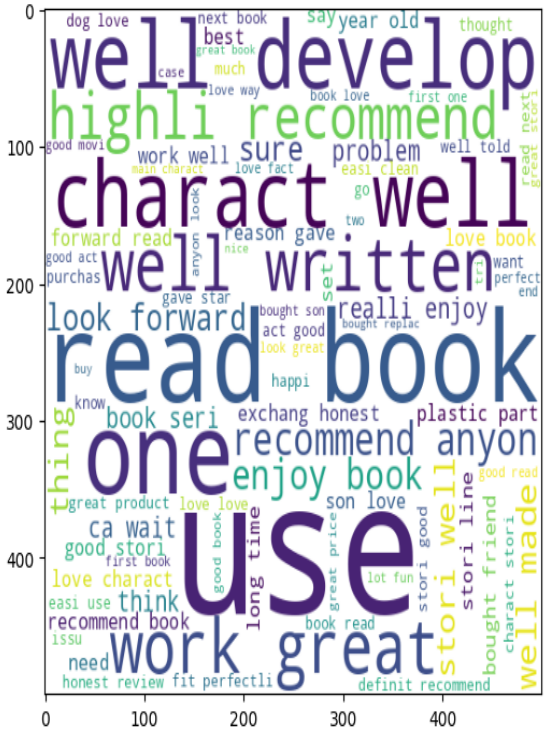


Fig. Fake WC

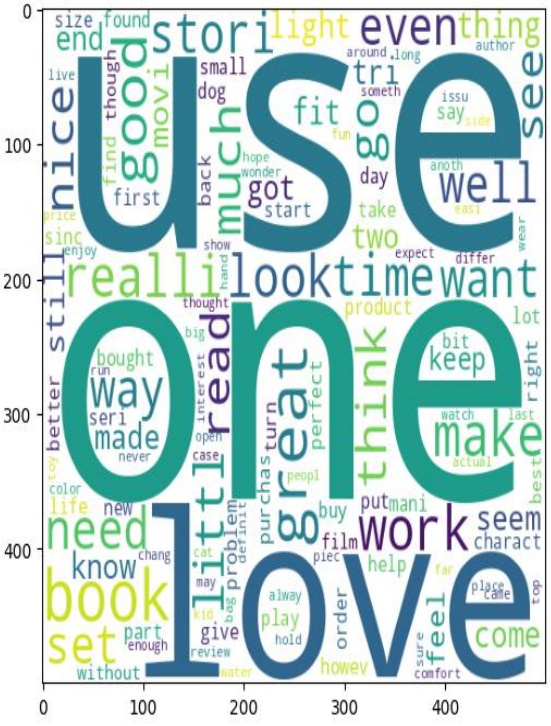


Fig. Original WC

Model Development and Training

1.TRADITIONAL MACHINE LEARNING MODELS: SEVEN MACHINE LEARNING CLASSIFIERS ARE EMPLOYED FOR INITIAL EVALUATION.K-NEIGHBORS

CLASSIFIER: A DISTANCE-BASED CLASSIFICATION ALGORITHM.

- ❖ Multinomial Naive Bayes (Multinomial NB): Suitable for text classification tasks.
- ❖ Decision Tree Classifier: A tree-based model capturing non-linear relationships.
- ❖ Logistic Regression: A linear classifier often used for binary classification tasks.
- ❖ Random Forest Classifier: An ensemble method combining multiple decision trees for robust performance.
- ❖ Ada Boost Classifier: Boosting technique improving weak classifiers.
- ❖ XGB Classifier: Gradient boosting model known for high accuracy and speed.

2. ENSEMBLE LEARNING WITH VOTING CLASSIFIER

To enhance performance, a Voting Classifier is constructed by combining the predictions of Logistic Regression, Multinomial NB, and Random Forest Classifier. This approach utilizes the strengths of individual models to achieve higher overall accuracy.

3. DEEP LEARNING WITH LSTM

LSTM networks, a type of Recurrent Neural Network (RNN), are employed to capture the sequential nature of text data. The steps include:

- ❖ Embedding Layer: Converts words into dense vector representations.
- ❖ LSTM Layer: Processes the sequential data to capture contextual dependencies.
- ❖ Dense Layer: Outputs the final classification result.The LSTM model is trained using: Optimizer: Adam optimizer for efficient learning. Loss Function: Binary

cross-entropy for binary classification. Metrics: Accuracy, precision, recall, and F1-score for performance evaluation.

V. EXPERIMENTAL RESULTS

Traditional classifiers, such as Random Forest Classifier and Logistic Regression, demonstrate Competitive accuracy.

S.N	Algorithm	Accuracy
0	LR	0.85
1	NB	0.83
2	RF	0.83
3	XGB	0.81
4	AdaBoost	0.74
5	DT	0.59
6	KN	0.51

The Voting Classifier enhances performance by combining the strengths of multiple models, including Logistic Regression, Multinomial Naive Bayes, and Random Forest Classifier, achieving an accuracy of 0.86. [22]

LSTM achieves the highest accuracy of 0.94, surpassing traditional methods by effectively capturing the semantic and sequential nuances of the text. [22]

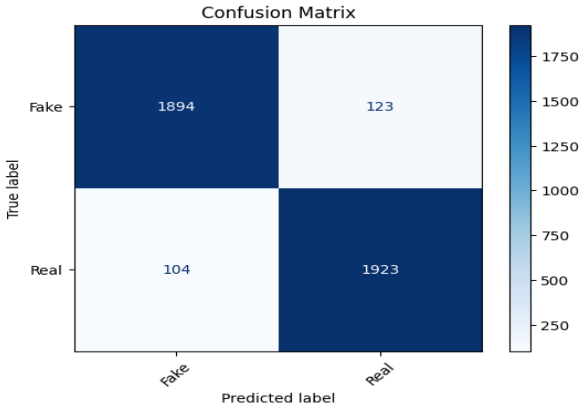


Fig. Confusion Matrix

This confusion matrix evaluates the performance of an LSTM model for detecting fake reviews. Here's a breakdown:

Matrix Components:

- True Labels: Rows represent the true labels:
  - Fake (Row 1): True label is "Fake."
  - Real (Row 2): True label is "Real."
- Predicted Labels: Columns represent the predictions made by the model:
  - Fake (Column 1): Predicted as "Fake."
  - Real (Column 2): Predicted as "Real."

Values:

1. True Negatives (1894): Fake reviews correctly classified as "Fake."
2. False Positives (123): Fake reviews incorrectly classified as "Real."
3. False Negatives (104): Real reviews incorrectly classified as "Fake."
4. True Positives (1923): Real reviews correctly classified as "Real."

Insights: The high values of true positives and true negatives suggest that the LSTM performs well in distinguishing between fake and real reviews. The relatively small false positive and false negative counts indicate that the model makes few misclassifications, but there is room for improvement, particularly in refining the boundary between real and fake reviews.

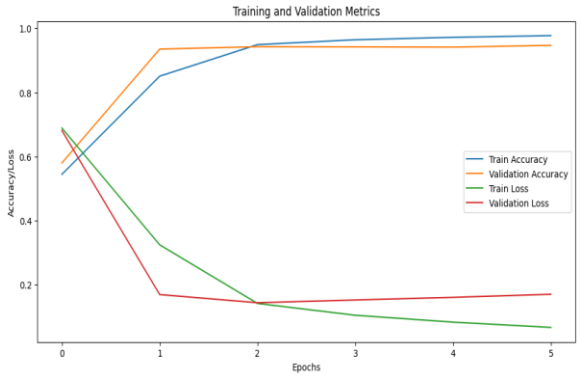


Fig. Training and Validation Metrics

This diagram illustrates the training and validation metrics (accuracy and loss) of an LSTM model used for detecting fake reviews, measured across epochs. Here's an explanation of its components: [22]

Axes:

- X-axis (Epochs): Represents the number of training iterations.
- Y-axis (Accuracy/Loss): Measures the respective metrics:
    - Accuracy: How well the model classifies fake and real reviews.
    - Loss: The model's error in prediction.

Lines:

1. Train Accuracy (blue):
    - Shows the improvement in the model's performance on the training dataset as the epochs increase.
    - Starts low and reaches close to 1.0 (100%), indicating the model fits the training data well.
2. Validation Accuracy (orange):
    - Tracks the model's performance on unseen data (validation set).
    - Follows a similar pattern to train accuracy but plateaus or increases slower, reflecting generalization.
3. Train Loss (green):
    - Indicates the error on the training set.
    - Decreases rapidly and approaches zero, suggesting the model learns effectively on the training data.
4. Validation Loss (red):
    - Tracks the error on the validation set.
    - Initially decreases but stabilizes or slightly increases toward the end, which could suggest minor overfitting.

Insights:

Overfitting: The gap between training accuracy and validation accuracy, along with the stabilization or increase in validation loss, might indicate overfitting after a certain number of epochs.

Convergence: The model achieves high accuracy, suggesting it successfully distinguishes between fake and real reviews.

Application: This graph shows the LSTM model learns effectively, but fine-tuning (e.g., early stopping or regularization) might help reduce over fitting and improve generalization further.

VI. CONCLUSION

The proposed method effectively combines traditional machine learning techniques with deep learning to detect fake reviews. The inclusion of LSTM provides a significant improvement in accuracy, making it the most reliable approach. This methodology can be further extended by exploring advanced neural architectures, such as Transformers or attention mechanisms, which have shown promise in natural language processing tasks. Additionally, incorporating larger and more diverse datasets, including cross-domain data, can enhance the model's generalization and robustness. Future work could also investigate techniques for interpretability and explainability to provide insights into the decision-making process of the models, ensuring their application in real-world scenarios is transparent and trustworthy.

VII. REFERENCES

[1] Qayyum, Huma, et al. "FRD-LSTM: a novel technique for fake reviews detection using DCWR with the Bi-LSTM method." Multimedia Tools and Applications 82.20 (2023): 31505-31519.

[2] Tabany, Myasar, and Meriem Gueffal. "Sentiment analysis and fake amazon reviews classification using SVM supervised machine learning model." Journal of Advances in Information Technology 15.1 (2024): 49-58.

[3] 3.https://www.kaggle.com/datasets/mexwell/fake-reviews-dataset

[4] Elmogy, Ahmed M., et al. "Fake reviews detection using supervised machine learning." International Journal of Advanced Computer Science and Applications 12.1 (2021).

[5] Elmurngi, Elshrif, and Abdelouahed Gherbi. "Detecting fake reviews through sentiment analysis using machine learning techniques." IARIA/data analytics (2017): 65-72.

[6] Molla, Alemu, Yenewondim Biadgie, and Kyung-Ah Sohn. "Detecting negative deceptive opinion from tweets." Mobile and Wireless Technologies 2017: ICMWT 2017 4. Springer Singapore, 2018.

[7] Shojaee, Somayeh, et al. "Detecting deceptive reviews using lexical and syntactic features." 2013 13th International Conference on Intellient Systems Design and Applications. IEEE, 2013.

[8] Ren, Yafeng, and Donghong Ji. "Neural networks for deceptive opinion spam detection: An empirical study." Information Sciences 385 (2017): 213-224.

[9] Attri, Vikas, Isha Batra, and Arun Malik. "Enhancement of fake reviews classification using deep learning hybrid models." Journal of Survey in Fisheries Sciences 10.4S (2023): 3254-3272.

[10] Aghakhani, Hojjat, et al. "Detecting deceptive reviews using generative adversarial networks." 2018 IEEE security and privacy workshops (SPW). IEEE, 2018.

[11] Mir, Abrar Qadir, Furqan Yaqub Khan, and Mohammad Ahsan Chishti. "Online fake review detection using supervised machine learning and BERT model." arXiv preprint arXiv:2301.03225 (2023).

[12] Alsubari, S. Nagi, et al. "Data analytics for the identification of fake reviews using supervised learning." Computers, Materials & Continua 70.2 (2022): 3189-3204.

[13] 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, 309-319.

[14] Mukherjee, A., Liu, B., & Glance, N. (2013). Spotting fake reviewer groups in consumer reviews. Proceedings of the 22nd International Conference on World Wide Web, 191-200.

[15] Rayana, S., &Akoglu, L. (2015). Collective opinion spam detection: Bridging review networks and metadata. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 985-994

[16] Aghakhani, Hojjat, et al. "Detecting deceptive reviews using generative adversarial networks." 2018 IEEE security and privacy workshops (SPW). IEEE, 2018.

[17] Qayyum, Huma, et al. "FRD-LSTM: a novel technique for fake reviews detection using DCWR with the Bi-LSTM method." Multimedia Tools and Applications 82.20 (2023): 31505-31519.

[18] Tabany, Myasar, and MeriemGueffal. "Sentiment analysis and fake amazon reviews classification using SVM supervised machine learning model." Journal of Advances in Information Technology 15.1 (2024): 49-58.

[19] Attri, Vikas, IshaBatra, and Arun Malik. "Enhancement of fake reviews classification using deep learning hybrid models." Journal of Survey in Fisheries Sciences 10.4S (2023): 3254-3272.

[20] Graves, Alex, and Alex Graves. "Long short-term memory." Supervised sequence labelling with recurrent neural networks (2012): 37-45.

[21] Devlin, Jacob. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[22] https://colab.research.google.com/drive/1lXurqSLIMrCuagc3s2yUQ6In9_8T2f48?pli=1#scrollTo=C0l75EfJigjv