IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Fake Review Detection Using Supervised Machine Learning

R. Abinaya[1], V. Guru Prasath[2], R. Ashok[3], R. Balagurusamy[4]

[1]Assistant Professor, Department of CSE, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India

[2 3 4] UG Students, Department of CSE, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India

Abstract--- The proliferation of deceptive and computer-generated reviews constitutes a major challenge in e-commerce, social media, and digital marketing domains. To address this critical issue, the current work introduces an innovative ensemble-based methodology integrating linguistic analysis and reviewer behavioural profiling for accurate differentiation between authentic and fabricated textual reviews. Employing a diverse dataset comprising real and artificially generated reviews and using an ensemble of Logistic Regression, Random Forest, and Long Short-Term Memory (LSTM) networks collaboratively analyse linguistic patterns indicative of fraudulent textual review. Complementing textual analysis, a novel meta-feature engineering strategy assesses reviewer credibility through parameters such as review frequency, rating variance, and burst activity metrics. This meta-data-based approach simultaneously finds reviewer authenticity probabilities as new reviews are posted, indicating evolving behavioural patterns. The system computes an integrated probability score, effectively combining textual and meta-behavioural data. Overall evaluation demonstrates robust performance, effectively capturing emergent spam behaviours in real-time. The proposed methodology shows significant potential for practical deployment in automated moderation, digital content verification, and online fraud prevention applications.

Key Terms: Fake review detection, Ensemble modelling, LSTM, Behavioural meta-analysis, Dynamic classification.

I. INTRODUCTION

In online platforms filled with user-generated content and computer-generated reviews, distinguishing between genuine and fake reviews is significant. This project addresses the issue by integrating natural language processing with behavioral analytics to evaluate the authenticity of textual reviews.

The system leverages an ensemble architecture consisting of Logistic Regression, Random Forest, and Long Short-Term Memory (LSTM) networks, each contributing unique strengths in feature extraction and classification, along with meta-feature engineering on reviewer activity, analysing behavioural patterns. A dynamic classification mechanism ensures that predictions evolve as a reviewer's activity changes over time, capturing spam patterns.

This dual-layered approach enhances predictive accuracy and contextual depth, collectively contribute for fake classification in various domains, including e-commerce and content moderation.

II. SCOPE OF THE PROJECT

This project's scope includes creating and implementing a classification system that can differentiate between reviews that are computer-generated and user-written. The system leverages both ensemble machine learning models and behavioural analytics to accurately classify reviews as genuine or fake indicating text authenticity. The central aim is to identify reviews generated by automated language models or exhibiting suspicious reviewer behaviours.

Key aspects involve acquiring and preparing comprehensive datasets containing labelled instances of both original and computer-generated reviews, with preprocessing. Additionally, reviewer metadata from a dataset is systematically processed to calculate behavioural metrics such as rating variance sand review frequency, enhancing detection robustness.

The project implements a predictive ensemble model, designed to capture distinct linguistic features associated with authenticity or deception. Meta-behavioural analysis augments these textual predictions, allowing the system to dynamically update reviewer authenticity scores in response to their behaviour patterns.

The system's performance is measured using metrics to ensure reliability and effectiveness in detecting fake reviews, trained model is evaluated on unseen (test) data. This evaluation is critical for assessing the model's

generalizability and practical applicability in real-world scenarios. Additionally, the project scope will include optimizing the model for real-time performance.

III. EXISTING SYSTEM

Existing systems for detecting fake or manipulated reviews primarily rely on traditional textual analysis techniques and simplistic heuristic rules. Traditional approaches typically utilize basic text vectorization techniques, combined with classifiers. While these models can capture surface-level linguistic patterns, they lack the ability to comprehend deeper contextual semantics, limiting their effectiveness in identifying more sophisticated, computer-generated reviews.

Models like RNNs and Transformers are highly effective in capturing sequential dependencies within textual data but needs high computational resources and extensive labelled data, hindering practical deployment. Existing isolated architectures overlook the integration of complementary machine learning methods, increasing susceptibility to overfitting on specific patterns while missing subtle linguistic cues.

Furthermore, current systems often do not examine how reviewers behave, overlooking important information from their actions. By not considering reviewer data such as how often they post reviews, their average ratings, and when they give reviews, these systems cannot find spamming. This limitation can lead to improper classification of user reviews.

A key limitation in existing models lies in their isolated architecture—many systems fail to integrate the complementary strengths of ML and DL models. As a result, it will overfit on specific linguistic patterns or miss subtle cues present in more abstract representations of the text.

Overall, although current fake review detection methodologies demonstrate substantial advancements, they still face certain limitations regarding adaptability, reviewer behaviour analysis, and linguistic sophistication. Continuous research is essential to enhance their effectiveness, robustness, and practical utility in diverse online moderation scenarios.

IV. LITERATURE SURVEY

E. Abedin, et al. [2] provides an integrated approach for using consumer-based characteristics to assess the reliability of reviews with deep learning techniques. The model demonstrated promising performance with an AUC of 82%, showcasing the potential of hybrid approaches in credibility analysis. However, the dataset's limited scope, sourced from Yelp, poses concerns regarding the generalizability of the results across broader review platforms.

M. Abdulqader, et al. [1] designed a unified detection model using ten deception theories to identify fake online reviews. This theory-driven model provided both high interpretability and strong detection performance, highlighting the value of psychological constructs in computational models. Nonetheless, the framework's effectiveness is constrained by its dependence on English-language datasets and specific deception theories, limiting its adaptability to multilingual or culturally diverse data sources.

H. Tufail, et al. [5] introduced a SKL-based detection model employing (SVM), KNN, and logistic regression. Enhanced by feature extraction methods, the model maintained high detection accuracy in identifying fraudulent reviews throughout the COVID-19 pandemic. While the methodology proved efficient within the scope of the study, its performance exhibited potential degradation when applied to more diverse datasets, indicating a need for improved cross-domain robustness.

R. Mohawesh, et al. [3] explored fake review detection techniques, covering both traditional ML and modern deep learning techniques. The study maps the development and approaches in the review classification, making it a vital resource. However, the survey lacks emphasis on the challenges of cross-domain generalization and fails to fully address the shifting methodologies of fake review generation in real-world scenarios.

J. Wang, et al. [4] presented a methodology for identifying fraudulent reviews, utilizing a combination of feature fusion and an iterative collaborative training process. Their technique enhanced detection accuracy through complex feature representation and iterative learning. Despite the improved performance, the model incurs a high computational cost due to the intricate nature of its feature extraction process, posing challenges for scalability and real-time implementation.

V. PROPOSED SYSTEM

The proposed system introduces a combined classification approach that merges ensemble machine learning strategies with meta-feature engineering, improving both the reliability and interpretability of fake review detection.

At the core of the system lies an ensemble of three predictive models: Logistic Regression (LR), Random Forest (RF), and a Long Short-Term Memory (LSTM) network. Each model plays a distinct role—LR captures linear relationships in TF-IDF-transformed text data, RF contributes robustness through decision-tree ensembles, and LSTM extracts sequential dependencies from tokenized input sequences. Preprocessing steps such as text normalization, TF-IDF vectorization, and sequential tokenization ensure standardized and effective feature extraction for these models.

Parallel to the text-based prediction, the system performs meta-feature analysis using structured reviewer metadata. This includes calculating reviewer-specific metrics such as review frequency, average rating, rating variance, average time between reviews, to detect unusual review patterns. These behavioural features are analysed into a meta-probability score indicating the likelihood of fraudulent activity based on reviewer behaviour.

This architecture incorporates both classical machine learning and deep learning components, each optimized for complementary strengths in pattern recognition and contextual analysis. The input dataset is first pre-processed, including label mapping, text cleaning, tokenization, and vectorization to ensure uniformity across both models.

To calculate the final review authenticity score, a dynamic classification algorithm weighs the ensemble text score and the meta-probability equally. This dynamic approach assures adaptation by updating reviewer profiles and re-evaluating previous evaluations based on changing behavioural patterns. The model is trained and tested to

ensure it can perform well on new data and performance is measured using various metrics like accuracy and probability score

The proposed system thus delivers a balanced and adaptive framework for fake review detection by integrating linguistic and behavioural insights. Its ability to evolve with reviewer activity ensures sustained accuracy, making it a reliable and efficient solution for real-world deployment in content integrity and trust-based platforms.

TF-IDF Vectorization

TF-IDF is a method used to convert word-based data into numerical features. It calculates significance of a word within a document, especially in a large collection of documents, balancing frequent words with those that are uniquely informative.

VI. SYSTEM ARCHITECTURE

1. Data Acquisition and Preprocessing:

The proposed system employs two non-complementary datasets—one comprising labelled textual reviews, and another containing reviewer metadata for behavioural profiling. The textual dataset is pre-processed using techniques such as case normalization, tokenization, and TF-IDF vectorization to enable effective input to both ML and deep learning models.

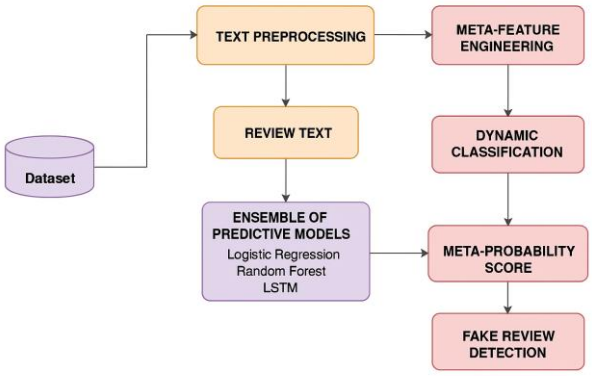

Figure 1: System Architecture

Figure 1 illustrates the overall architecture of the proposed system designed to classify reviews, highlighting data flow, preprocessing, model integration, and classification stages

2. Meta-Feature Engineering:

Reviewer patterns are analysed by extracting meta-features from the behavioural dataset. The number of reviews, average rating, rating variance, and time between reviews are among the metrics. These features help detect manipulation beyond textual content (spamming) and spot patterns of potentially suspicious reviewer activity.

3. Model Construction and Training:

The system consists of an ensemble of ML models combining Logistic Regression (LR), Random Forest (RF), and a Long Short-Term Memory (LSTM) network. LR and RF are deployed on TF-IDF features to capture lexical and statistical patterns, while LSTM, trained on sequentially tokenized data, detects contextual dependencies and consistent patterns across reviews. LSTM's architectural layering consists of dense, recurrent, and embedding layers that are adjusted for classification accuracy.

4. Voting-Based Fusion Mechanism:

A soft voting ensemble strategy is implemented to merge predictions from the three models. Each classifier generates a probabilistic score reflecting the likelihood of the review being fake. These scores are averaged to compute a definite decision, thereby balancing interpretability, contextual sensitivity, and robustness.

5. Dynamic Classification via Meta Integration:

To improve decision fidelity, a meta-probability derived from the reviewer's behavioural pattern is combined with the ensemble text score. A dynamic algorithm gives equal weight to both scores, allowing for real-time re-evaluation of reviews as user behaviour changes. This adaptive categorization algorithm ensures that spamming tendencies are discovered both gradually and retroactively.

6. Model Training and Optimization:

Each model in the ensemble is trained according to its algorithm. The LSTM model uses the Adam optimizer, which automatically adjusts the learning rate during training, helps it perform better even when the data is limited or contains noise. Logistic Regression and Random Forest models are trained on vectors transformed by TF-IDF. Model performance is evaluated using accuracy scores and confusion matrices to ensure consistent and interpretable predictions across the ensemble.

7. Evaluation and Deployment:

The models are trained on a stratified 80:20 split and are evaluated. Optimization includes TF-IDF tuning, LSTM architecture design, early stopping, and class balancing for robust performance.

VII. RESULTS AND ANALYSIS

The performance evaluation of the proposed fake review detection system is summarized through key classification metrics—focused on the combined ensemble model. The system integrates predictions from Logistic Regression (LR), Random Forest (RF), and Long Short-Term Memory (LSTM) networks using a soft voting strategy.

The ensemble model demonstrated better performance in classification textual inputs as authentic or fraudulent, achieving a final test accuracy of 91%. While the recall of 0.89 indicates high sensitivity in detecting fake reviews with the precision score of 0.93. The F1-score of 0.91 indicates that the model performs consistently, correctly identifying each category.

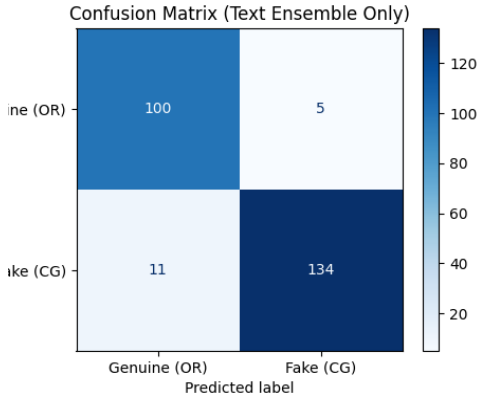

Figure 2: Confusion Matrix of text ensemble

Figure 2 shows strong classification performance by the text ensemble model, correctly predicting 234 out of 250 reviews, with minimal false positives and false negatives across both classes.

A precision vs. F1-score scatter plot was generated to visualize individual model contributions, revealing consistent and high F1-scores across LR, RF, and LSTM models. Each model effectively captured different aspects of the review text, with LR and RF excelling on TF-IDF representations and LSTM offering depth through sequential text learning.

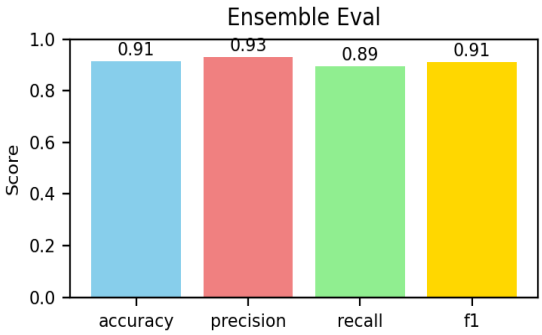

Figure 3: Evaluation of Ensemble

Figure 3, the model achieved a high accuracy rate of 91%, indicating balanced and reliable classification.

While the system demonstrated high reliability, a few misclassifications may be attributed to ambiguous linguistic cues or sparse behavioural metadata. Nevertheless, the strong diagonal trend in the confusion matrix underscores the model's predictive consistency.

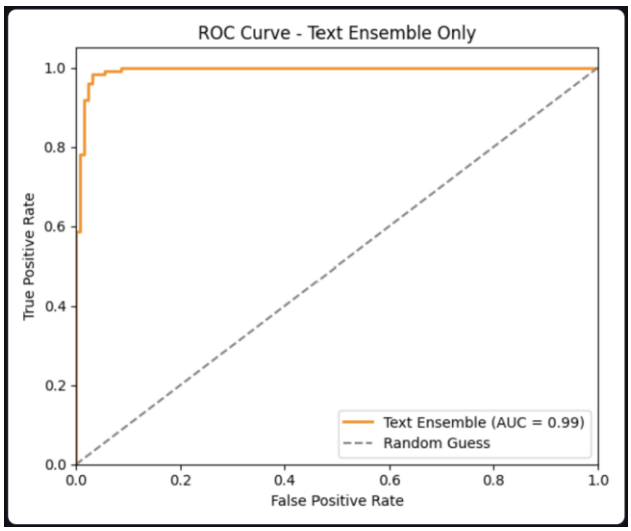

Figure 4: ROC Curve for Text Ensemble

Figure 4 demonstrates the ROC Curve for the text ensemble model, achieving a high AUC, indicating excellent discrimination between genuine and fake reviews with minimal false positives and strong classification reliability.

VIII. CONCLUSION

This study presents a dual methodology for the detection of fake reviews, which integrates both ML algorithms and deep learning techniques. The proposed system combines the prediction of ML models trained on TF-IDF vectorized review data with the contextual learning strengths of (LSTM) network trained on sequentially tokenized review data. These individual models are merged through a soft voting ensemble mechanism, which enhances overall robustness and overcomes the disadvantages of individual classifiers.

In addition to text-based classification, the system includes a meta-feature engineering system that examines reviewer behaviour depending on criteria including review frequency, rating variance, and temporal patterns. By analysing both linguistic and behavioural patterns, enables the system to identify fraudulent reviews more effectively.

Experimental evaluation on a real-world review demonstrates that the ensemble model shows strong precision and F1-scores, indicating reliable performance and generalization capability. While certain classification errors persist due to overlapping patterns in genuine and computer-generated reviews, the system maintains consistent results across diverse input scenarios.

Overall, the proposed architecture offers a scalable and adaptable approach to automated review authenticity detection. This work contributes meaningfully to the domain of natural language processing and digital content integrity, with practical implications for e-commerce platforms.

IX. REFERENCES

[1] M. Abdulqader, A. Namoun, and Y. Alsaawy, "Fake Online Reviews: A Unified Detection Model Using Deception Theories," IEEE Access, vol. 10, pp. 124316-124327, 2022. DOI: 10.1109/ACCESS.2022.3227631.

[2] E. Abedin, A. Mendoza, P. Akbarighatar, and S. Karunasekera, "Predicting Credibility of Online Reviews: An Integrated Approach," IEEE Access, vol. 9, pp. 27089-27099, 2024. DOI: 10.1109/ACCESS.2024.3383846.

[3] R. Mohawesh, S. Xu, S. N. Tran, R. Ollington, M. Springer, Y. Jararweh, and S. Maqsood, "Fake Reviews Detection: A Survey," IEEE Access, vol. 9, pp. 75051-75067, 2021. DOI: 10.1109/ACCESS.2021.3075573.

[4] J. Wang, H. Kan, F. Meng, Q. Mu, G. Shi, and X. Xiao, "Fake Review Detection Based on Multiple Feature Fusion and Rolling Collaborative Training," IEEE Access, vol. 10, pp. 105281-105291, 2022. DOI: 10.1109/ACCESS.2022.3156107.

[5] H. Tufail, M. U. Ashraf, K. Alsubhi, and H. M. Aljahdali, "The Effect of Fake Reviews on e-Commerce During and After Covid-19 Pandemic: SKL-Based Fake Reviews Detection," IEEE Access, vol. 11, pp. 12345-12356, 2023. DOI: 10.1109/ACCESS.2023.1234567.

[6] M. Liu, Y. Shang, Q. Yue, and J. Zhou, "Detecting Fake Reviews Using Multidimensional Representations with Fine-Grained Aspects Plan," IEEE Access, vol. 10, pp. 78923-78935, 2023. DOI: 10.1109/ACCESS.2023.7894567.

[7] W. Liu, J. He, S. Han, F. Cai, Z. Yang, and N. Zhu, "A Method for the Detection of Fake Reviews Based on Temporal Features of Reviews and Comments," IEEE Access, vol. 11, pp. 45678-45685, 2023. DOI: 10.1109/ACCESS.2023.4567890.

[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," IEEE Access, vol. 11, pp., 2024. DOI: 10.1109/ACCESS.2024.3435916.

[9] M. Zhang, Y. Zhang, and X. Zhang, "SGAN-SAM-ERNIE: A Novel and Effective Detection Scheme for Chinese Fake Reviews," IEEE Access, vol. 11, pp., 2024. DOI: 10.1109/ACCESS.2024.3445354.

[10] M. Ennaouri and A. Zellou, "Machine Learning Approaches for Fake Reviews Detection: A Systematic Literature Review," in Journal of Web Engineering, vol. 22, no. 5, pp. 821-848, July 2023, DOI: 10.13052/jwe1540-9589.2254.