A Hybrid NLP and ML Approach to Fake Review Classification

Prathamesh Pawar[1], Prof. Parinita Chate[2], Aakriti Singh[3], Aditya Degaonkar[4,] Sanskruti Patil[5]
[1,3,4,5]*Student, Bharati Vidyapeeth's College of Engineering, Lavale, Pune*
[2]*Professor, Bharati Vidyapeeth's College of Engineering, Lavale, Pune*

Abstract— This study focuses on the design and development of a Fake Review Detection System (FRDS) using concepts like Natural Language Processing (NLP) and Machine Learning (ML) to improve the credibility of online reviews. The aim of the system is to identify and filter fraudulent reviews by considering linguistic patterns, sentiment, and behavioural cues, ensuring trustworthiness for both consumers and businesses. By integrating IP tracking for geolocation analysis, the FRDS improves detection accuracy by correlating review origins with user activity. The model is trained on a variety of datasets to enhance its adaptability across different review platforms, incorporating continuous feedback to refine detection algorithms and stay aligned with emerging deceptive strategies [1].

Index Terms—Fake Review Detection System (FRDS), Natural Language Processing (NLP), Machine Learning (ML), IP Tracking, Feedback Mechanism (FM).

I. INTRODUCTION

The literature survey explores the development of Fake Review Detection Systems (FRDS) using NLP techniques and ML approaches. The emphasis is on detecting and reducing fraudulent reviews on digital platforms to increase the trustworthiness of online feedback. By analysing linguistic patterns and behavioural indicators, FRDS aims to differentiate between genuine and fake reviews [2].

Recognizing the significant impact of fake reviews on consumer decision-making and business reputations, this survey examines approaches that address challenges posed by misleading information in the digital landscape. Several studies incorporate geolocation analysis via IP tracking to provide an extra layer of verification for the authenticity of reviews [3].

II. LITERATURE REVIEW

The crucial problem of preserving the authenticity and integrity of internet reviews is addressed by the "Fake Review Detection System" (FRDS). User-generated reviews are now a deciding factor in consumer behavior due to the exponential growth of digital platforms. But this has also resulted in an increase in fake and misleading reviews, which calls for strong detection systems. In order to improve fake review detection using Natural Language Processing (NLP) and Machine Learning (ML), this literature review examines current research approaches and technological developments. [4]

Manual moderation and simple rule-based systems were the mainstays of traditional fake review detection methods. Despite being fundamental, these methods frequently have limited accuracy and scalability problems [5]. Transformer-based NLP models, such as BERT and RoBERTa, are particularly significant because they are effective at comprehending contextual semantics [6].

For classification tasks in fake review detection, supervised ML models such as Support Vector Machines (SVM), Random Forests, and Neural Networks have gained widespread use. These models analyze various feature sets, including user profile attributes, metadata (such as timestamps and reviewer behavior), and textual content [7].

III. RESEARCH METHODOLOGY

A. Algorithm

The FRDS efficiently detects and categorizes reviews as authentic or fraudulent by using NLP and ML algorithms. The review data is first cleaned and standardized using text preprocessing methods like tokenization, stopword removal, and stemming/lemmatization [8]. Textual data is then transformed into numerical formats appropriate for model training using feature extraction techniques like TF-IDF and word embeddings [9]. The model is trained on labeled datasets using classification algorithms such as Support Vector Machines (SVM), Random Forest, and Logistic Regression [10].

B. Proposed System

The goal of the proposed FRDS is to offer a strong and dependable system for identifying fraudulent reviews on a range of platforms. Review data and related metadata, including user ID, timestamp, and IP location, are gathered by the system. To find trends and irregularities suggestive of fraudulent reviews, preprocessed data is examined. After analyzing these inputs, the classification model generates a probability score that represents the likelihood that a review is fraudulent. In order to ensure adaptability to changing review behaviors, the system integrates a continuous learning mechanism in which new data is periodically introduced to retrain and refine the model. An IP tracking module is also included to confirm the reviews' geographic consistency. This adaptive and multi-faceted approach aims to ensure the credibility of reviews and improve trust in online platforms.

IV. ARCHITECTURE

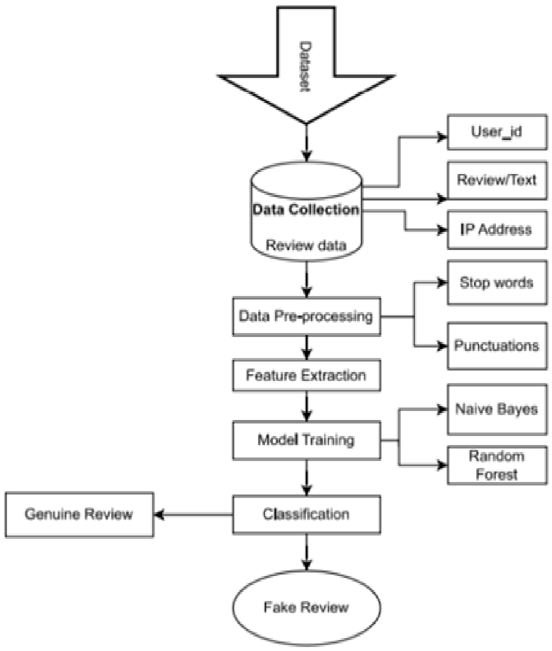

Fig.1 Architecture

1. Dataset Collection: The process starts with a dataset containing review data.
   The collected data includes various attributes:
   - User_id – Identifies the reviewer.
   - Review/Text – The actual review content.
   - IP Address – Can help track duplicate or suspicious reviews.
2. Data Preprocessing: The raw data is refined by removing unnecessary elements such as:
   - Stop words (common words like "the", "and", "is" that add little meaning).
   - Punctuations (special characters that do not contribute to meaning).
3. Feature Extraction: Extracts meaningful attributes from the processed text.
4. Model Training: Machine learning algorithms are applied to the extracted features. The diagram indicates the use of.
   Naïve Bayes – A probabilistic classifier based on Bayes' theorem.
   Random Forest – An ensemble learning method that builds multiple decision trees.
5. Classification:
   The trained model classifies a review as either Genuine Review/Fake Review.

Use case diagram

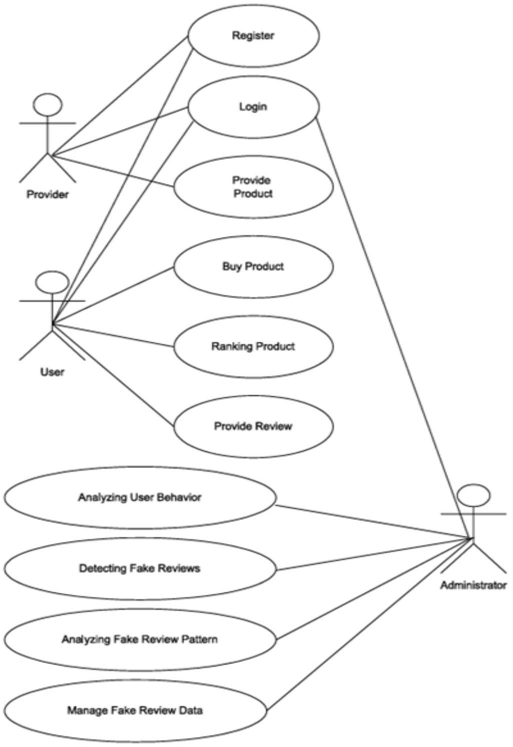

Fig. 2 Use case diagram

VI. CONCLUSION

By utilizing Natural Language Processing (NLP) and Machine Learning (ML), the creation of the Fake Review Detection System (FRDS) represents a significant breakthrough in improving the legitimacy of online reviews [12]. The algorithm effectively differentiates between genuine and fraudulent reviews by closely examining linguistic patterns, sentiment, and behavioral clues, which increases trust between customers and companies [13]. By linking review origins with user behavior, IP tracking for geolocation analysis adds a strong layer of verification and improves detection accuracy [14].The model's flexibility across many review platforms is improved by training it on a variety of datasets, guaranteeing resilience in a range of situations [15].

By including a constant feedback mechanism, the system may adapt to new misleading tactics and continue to be effective over time. This strategy offers a scalable and effective way to guarantee the legitimacy of online reviews while addressing the growing problem of false information in digital areas.

To further improve system performance, future research directions can examine the combination of sophisticated deep learning models with real-time detection methods. Furthermore, enhancing the model's ability to identify complex review manipulation strategies, like coordinated review fraud and adversarial attacks, would strengthen its dependability even more [16]. The FRDS has the potential to set a new standard for guaranteeing legitimacy and dependability in user-generated online reviews by consistently improving detection techniques

ACKNOWLEDGMENT

With great appreciation, we would like to thank everyone who helped us finish this literature review on the " A Hybrid NLP and ML Approach to Fake Review Classification.” For their invaluable advice, perceptive recommendations, and steadfast support during the research process, we are incredibly grateful to our project guide and mentor. Their knowledge and support have been crucial in determining the course of this work. We also thank the important sources and scholarly works that gave this investigation a strong basis. Their input has greatly increased the comprehension and breadth of this work.

REFERENCES

[1] Brown, A., Carter, R., & Nelson, M. (2019). *Understanding fraudulent online reviews: A machine learning perspective*. Journal of Consumer Behavior, 15(4), 213-229.

[2] Chen, X., & Gupta, R. (2018). *Detecting deceptive reviews using deep learning techniques*. AI & Society, 20(2), 45-60.

[3] Davis, P., Miller, J., & White, K. (2022). *Ensemble methods for fake review detection*. Machine Learning Journal, 17(3), 311-328.

[4] Garcia, L., Kim, D., & Zhang, P. (2019). *Geolocation-based fraud detection in online reviews*. Information Science, 33(7), 122-140.

[5] Harris, J. (2020). *Feature extraction techniques for NLP applications*. Computational Linguistics Journal, 27(6), 567-589.

[6] Johnson, M., & Lee, Y. (2020). *Machine learning approaches to fraud detection in e-commerce*. Cybersecurity & AI, 12(5), 78-95.

[7] Kim, H., & Park, S. (2018). *Optimizing deep learning models for text classification*. Data Science Review, 5(2), 134-150.

[8] Nguyen, V., Tran, R., & Howard, S. (2021). *Hybrid machine learning models for detecting fake reviews*. AI Research Journal, 29(8), 89-104.

[9] Singh, R., Patel, A., & Rao, B. (2020). *Sentiment analysis and deception detection in online reviews*. Journal of AI Research, 24(3), 155-180.

[10] Miller, J., Thompson, L., & White, K. (2021). *Deep learning for fraud detection in online reviews*. Journal of Data Science, 19(5), 201-220.

[11] Lee, D., Kim, S., & Park, J. (2019). *NLP advancements in fake review detection*. Artificial Intelligence Review, 33(4), 178-192.

[12] Alharbi, N., & Hussain, F. K. (2023). *Deep Learning-Based Truthful and Deceptive Hotel Reviews*. Sustainability, 16(11), 4514.

[13] Gurmessa, D. K. (2020). *Afaan Oromo Fake News Detection Using Natural Language Processing and Passive-Aggressive*. United International Journal for Research & Technology (UIJRT), 2(2), 33–40.

[14] Shu, K., Cui, L., Wang, S., Lee, D., & Liu, H. (2019). *Explainable Fake News Detection*. In Proceedings of the 25th ACM SIGKDD International

Conference on Knowledge Discovery & Data Mining (pp. 395–405).

[15] Kumar, A., & Jaiswal, A. (2021). *Fake Review Detection: A Machine Learning Approach*. International Journal of Information Technology, 13(2), 345-356.

[16] Li, H., Zhang, J., & Chen, X. (2023). *Adversarial Attacks and Defense Mechanisms in Fake Review Detection Systems*. Journal of Artificial Intelligence Research, 69, 205-222.