Robust Statistical Models for Identifying Inauthentic Online Reviews

1st D.Swapna
Department of Computer Science and Engineering
BVRIT HYDERABAD College of Engineering for Women
Hyderabad, India
swapna.d@bvrithyderabad.edu.in

2nd M.Shanmuga Sundari
Department of Computer Science and Engineering
BVRIT HYDERABAD College of Engineering for Women
Hyderabad, India
sundari.m@bvrithyderabad.edu.in

3rd G.Khadhyothi Sreeja
Department of Computer Science and Engineering
BVRIT HYDERABAD College of Engineering for Women
Hyderabad, India
20wh1a0508@bvrithyderabad.edu.in

4th S.Sriya Varma
Department of Computer Science and Engineering
BVRIT HYDERABAD College of Engineering for Women
Hyderabad, India
20wh1a0512@bvrithyderabad.edu.in

5th M.Rajeshwari
Department of Computer Science and Engineering
BVRIT HYDERABAD College of Engineering for Women
Hyderabad, India
21wh5a0505@bvrithyderabad.edu.in

***Abstract*—Review authenticity is a key factor in determining trustworthiness in the internet marketplace. However, the threat of fake reviews remains constant, endangering both fair competition and customer confidence. The rise of fake reviews undermines the credibility of these platforms and poses a significant challenge to users seeking genuine feedback. In response, this study proposes a machine learning (ML) approach for detecting fake reviews.The research begins by collecting a dataset comprising both genuine and fake reviews from various online platforms. Features are extracted from the textual content of these reviews, including linguistic characteristics, sentiment analysis, and metadata such as review length and rating distribution. Several ML models are employed for classification, such as Naive Bayes, Random Forest, and Support Vector Machines. These algorithms are assessed using measures like accuracy, precision, recall, and F1-score after being trained on the dataset. In addition, methods like hyperparameter tuning and cross-validation are used to improve the performance of models.Overall, the findings suggest that ML-based approaches hold promise for mitigating the problem of fake reviews, thereby enhancing the reliability of online review platforms and empowering consumers to make informed decisions.**

***Index Terms*—Fake detection, Reviews, Machine Learning, Natural language processing**

I. Introduction

Review sites are a potent instrument in the age of digital media for influencing customer behavior, shaping purchasing decisions and brand perceptions. However, the proliferation of fake reviews presents a formidable challenge to the integrity of online platforms. Detecting fake reviews is crucial for maintaining trust and transparency within online communities and safeguarding consumers from deceptive practices.Fake reviews, often crafted to manipulate public opinion, blur the line between genuine feedback and deceptive endorsements, posing a threat to the integrity of online review ecosystems. As such, the need for robust fake review detection mechanisms has become imperative to maintain the credibility of online platforms and safeguard consumer trust. Leveraging advancements in natural language processing (NLP) and machine learning (ML) [5] technologies, the aim of this research project is to develop and enhance algorithms capable of accurately [7] identifying fake reviews amidst the vast sea of online feedback.

II. Literature Survey

The paper [4] undertakes a pre-labeled dataset by Ott et al., containing 1280 reviews related to 20 Chicago hotels. Various technologies were employed, including TF-IDF vectorization, machine learning models (such as Random Forest , SVM, Logistic Regression, and Naive Bayes), and deep learning with neural networks. The achieved accuracies ranged from 71.09 percent to 87 percent, with deep learning showing the best results after parameter tuning. The implementation involved NLP techniques, ensemble methods, and grid search for parameter optimization. In conclusion, the research highlightsThe possibilities of deep learning and machine learning in distinguishing between real and fake reviews, with future work focusing on expanding datasets and refining models.

The paper [2] examines the use of a dataset of hotel reviews in a supervised machine learning strategy to identify fraudulent internet evaluations.The dataset was created by Ott

et al. and includes 800 reviews that are accurate and 800 that are dishonest, with a balanced mix of positive and negative feedback. TF-IDF, Empath categories, and sentiment polarity are among the technologies utilized for feature extraction, and classifiers such as logistic regression, Naive Bayes, and SVM are employed.The best accuracy achieved was 88.75 percent with the SVM classifier.The conclusion suggests that supervised learning effectively classifies fake reviews and future work could include user behavior analysis for improved results.

The paper [3] discusses various datasets and technologies used for fake reviews detection.It mentions datasets like Yelp CHI, Yelp NYC, and Yelp ZIP, which are labeled by Yelp's spam filter, and others created through crowd sourcing platforms like Amazon Mechanical Turk (AMT). Technologies such as metadata features, Part of Speech (POS), Bag of Words (BoW), Linguistic Inquiry and Word Count (LIWC), and word embedding methods like Word2Vec, GloVe, and FastText are used. The paper highlights that combining features yields better performance, with accuracies ranging from 67.8 percent to 89.6 percent depending on the dataset. The conclusion emphasizes the significance of using behavioral and text features to improve fake reviews detection models.

III. Proposed Model

Our objective is to meticulously compare each model's performance on our dataset, examining metrics like accuracy and speed to understand their capabilities and limitations. Through this process, we aim to identify the most effective machine learning model for our project's specific requirements, considering factors such as predictive accuracy, computational efficiency, and scalability. We'll also assess each model's robustness to real-world data challenges, enhancing our confidence in our chosen approach. Additionally, we'll consider qualitative aspects like interpretability, robustness, and generalization capabilities [8] to ensure reliable insights across diverse scenarios, driving informed decision-making from our data analysis efforts.

IV. Architecture

A. Random Forest

Random Forest constructs multiple decision trees, each trained on random subsets of data and features, for fake review detection. Features like word frequencies and sentiment scores are utilized. Aggregating predictions through majority voting mitigates overfitting and improves model robustness. Leveraging collective wisdom and randomness, Random Forest effectively captures complex patterns, yielding reliable predictions.

B. Naive Bayes

A probabilistic categorization called Naive Bayes algorithm, is favored for fake review detection due to its simplicity and efficiency. It relies on Bayes' theorem, assuming feature independence. Despite this simplification, it performs well in text classification tasks like fake review detection. By modeling word or feature frequency distributions from labeled data, it calculates posterior probabilities for classifying new reviews based on observed features, yielding effective predictions.

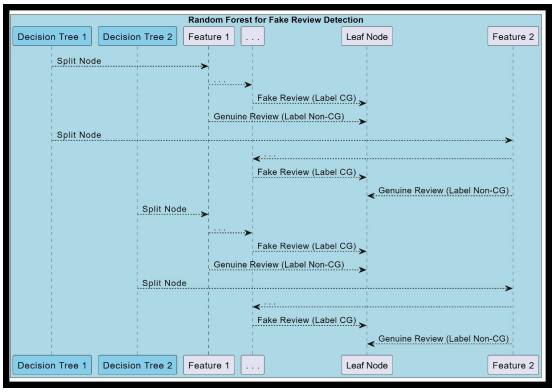

Fig. 1. Random Forest Architecture Diagram

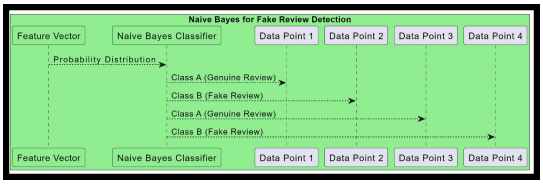

Fig. 2. Naive Bayes Architecture Diagram

C. Support Vector Machines

Support Vector Machines (SVM) [6]excel in fake review detection for their adeptness in handling high-dimensional feature spaces and effectively segregating data with a hyperplane.By transforming text reviews into numerical feature vectors using techniques like TF-IDF, SVM [1]identifies the optimal hyperplane during training to maximize the margin, aiming to best separate genuine and fake reviews. It can handle non-linear data through kernel tricks, ensuring precise classification of new reviews.

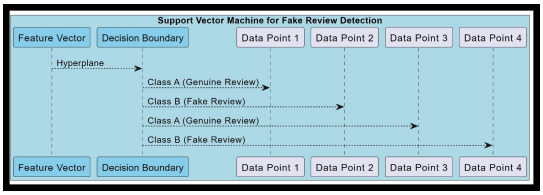

Fig. 3. Support Vector Machine Architecture Diagram

V. Experimental Results

A. Random Forest

1) Confusion Matrix : The matrix contains four outcomes, namely true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). The TP value indicates that the model accurately predicted 3653 instances as positive, while the FP value shows that the model incorrectly predicted

363 instances as positive. The TN value indicates that the model correctly predicted 3150 instances as negative, while the FN value reveals that the model incorrectly predicted 921 instances as negative.

2) ROC Curve : The disparity between the TPR and FPR is displayed by the curve, where increasing the TPR typically results in an increase in the FPR. The curve is used to evaluate the overall performance of the classifier by calculating the AUC, or area under the curve, has a range of 0 to 1. Better performance is indicated by a larger AUC, with a value of 1 denoting a perfect classifier. The AUC in this instance is 0.93, indicating that the classifier is performing well.

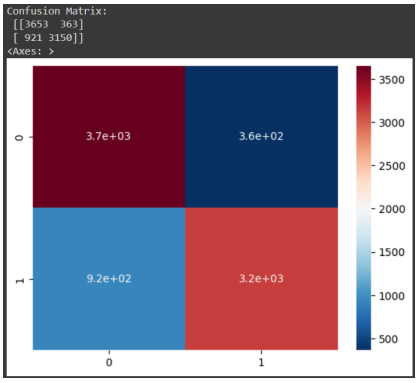

Fig. 4. Confusion Matrix

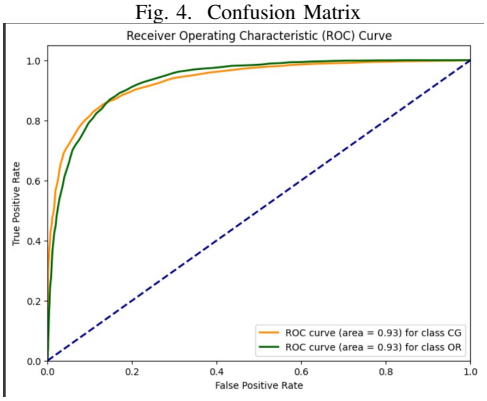

Fig. 5. ROC Curve

B. Naive Bayes

1) Confusion Matrix : The provided image demonstrates a confusion matrix, which is a table arrangement that illustrates how well a classification model performs. The table contains four outcomes produced by the classifier, including true positives (3754), false positives (262), false negatives (905), and true negatives (2006).The axes represent the predicted and actual classes, and the values indicate the number of instances in each category.

2) ROC Curve : The ROC curve depicts a binary classifier's performance across varied discrimination thresholds, illustrating how true positive rate (TPR) and false positive rate (FPR) are traded off. AUC of 0.96, indicating superior classification, the curve demonstrates the classifier's adeptness in distinguishing positive and negative instances. As the threshold decreases (right to left on the curve), TPR increases while FPR also rises but remains relatively low, affirming the classifier's effective discrimination.

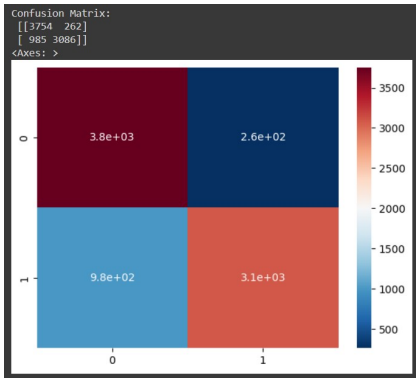

Fig. 6. Confusion Matrix

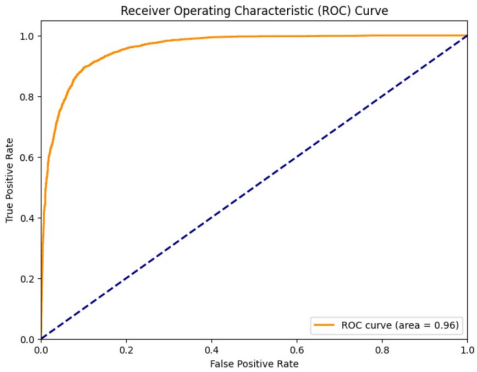

Fig. 7. ROC Curve

C. Support Vector Machine

1) Confusion Matrix : The given context presents a Confusion Matrix for a binary classification problem.The number of true positive (3612), false positive (404), false negative (431), and true negative (3640) incidents is displayed in the confusion matrix.The values of the matrix are displayed as a grid with anticipated classes (positive on the top, negative on the left) and actual classes (negative on the left). The image shows a total of 7,012 instances, with the majority being correctly classified as negative or positive.

2) ROC Curve : The provided context displays a Receiver Operating Characteristic (ROC) curve, with its corresponding rates of False Positives (FPR) and True Positives (TPR) at various cutoff points. The area under the curve (AUC) is 0.96, indicating an excellent classifier, as AUC values close to 1 imply better performance. In this case, as the threshold decreases (moving from right to left on the curve), TPR increases while FPR also increases, but remains relatively low, highlighting the classifier's ability to distinguish positive and negative instances effectively.

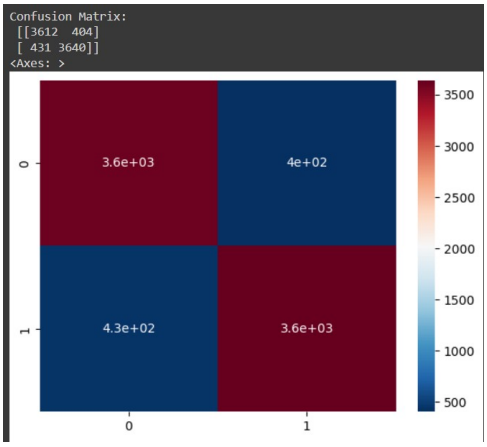

Fig. 8. Confusion Matrix

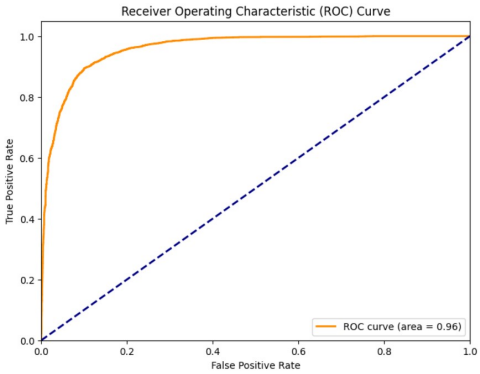

Fig. 9. ROC Curve

VI. Comparative Analysis:

Machine Learning Models	Accuracy	Precision	Recall	F1-score	ROC
Random Forest	84%	CG – 0.8 OR – 0.9	CG – 0.91 OR – 0.78	CG – 0.85 OR – 0.84	CG – 0.93 OR – 0.93
Naïve Bayes	84%	CG – 0.79 OR – 0.92	CG – 0.93 OR – 0.76	CG – 0.86 OR – 0.83	0.96
Support Vector Machine (SVM)	89%	CG – 0.89 OR – 0.9	CG – 0.90 OR – 0.89	CG – 0.90 OR – 0.90	0.96

Fig. 10. Accuracy comparison table

VII. Conclusion

The deployed system represents a sophisticated solution for combating fake reviews within online platforms, marking a significant advancement in ensuring the integrity of digital feedback ecosystems. Utilizing a combination of natural language processing (NLP) and machine learning (ML) techniques, the system adeptly analyzes and classifies reviews, distinguishing between genuine and deceptive content with precision.

The system's incorporation of advanced technologies, including feature extraction and sentiment analysis, underscores its commitment to modernizing traditional review validation methods. Through functionalities such as data collection, model training, and real-time review analysis, the system offers a comprehensive approach to detecting and addressing fraudulent reviews in online platforms. Moreover, its ability to handle various aspects of review management, including data storage and result visualization, enhances its usability and practicality for businesses and consumers alike.By leveraging a combination of industry-standard libraries and frameworks, such as NLTK, Scikit-learn, the implementation ensures scalability, reliability, and efficiency in fake review detection processes. The system's structured organization and adherence to best practices underscore its reliability and robustness in addressing the complexities of online review validation.

References

[1] Murali Krishna Doma, Kayal Padmanandam, Sunil Tambvekar, Keshav Kumar, Bilal Abdualgalil, and RN Thakur. Artificial intelligence-based breast cancer detection using wpso. *International Journal of Operations Research and Information Systems (IJORIS)*, 13(2):1–16, 2022.

[2] Rakibul Hassan and Md Rabiul Islam. A supervised machine learning approach to detect fake online reviews. In *2020 23rd international conference on computer and information technology (ICCIT)*, pages 1–6. IEEE, 2020.

[3] Rami Mohawesh, Shuxiang Xu, Son N Tran, Robert Ollington, Matthew Springer, Yaser Jararweh, and Sumbal Maqsood. Fake reviews detection: A survey. *IEEE Access*, 9:65771–65802, 2021.

[4] Aaryan Rustagi, Vajraang Padisetti, and Suresh Subramaniam. Fake review detection using machine learning. *Journal of Student Research*, 11(1), 2022.

[5] M Shanmuga Sundari, M Dyva Sugnana Rao, and Ch Anil Kumar. Effective prediction analysis for cardiovascular using various machine learning algorithms. In *Proceedings of Fourth International Conference on Computer and Communication Technologies: IC3T 2022*, pages 641–650. Springer, 2023.

[6] M Shanmuga Sundari and Vijaya Chandra Jadala. Neurological disease prediction using impaired gait analysis for foot position in cerebellar ataxia by ensemble approach. *Automatika*, 64(3):540–549, 2023.

[7] M Shanmuga Sundari and Rudra Kalyan Nayak. Efficient tracing and detection of activity deviation in event log using prom in health care industry. In *2021 Fifth international conference on I-SMAC (IoT in social, mobile, analytics and cloud)(I-SMAC)*, pages 1238–1245. IEEE, 2021.

[8] Shanmuga Sundari, Yeluri Divya, KBKS Durga, Vidyullatha Sukhavasi, M Dyva Sugnana Rao, and M Sudha Rani. A stable method for brain tumor prediction in magnetic resonance images using finetuned xceptionnet. *International Journal of Computing and Digital Systems*, 15(1):67–79, 2024.