Fake Review Identification using Hybrid Fusion of Machine Learning and Natural Language Processing Techniques

Chitti Reddy Veda[1]
Department of Information Technology,
Vardhaman College of Engineering (Autonomous),
Hyderabad, India.
chittireddyveda@gmail.com

Namburu Apoorva[2]
Department of Information Technology,
Vardhaman College of Engineering (Autonomous),
Hyderabad, India.
apoorva.namburu@gmail.com

N. Vishnu[3]
Department of Information Technology,
Vardhaman College of Engineering (Autonomous),
Hyderabad, India.
nagapurvishnu39@gmail.com

Muni Sekhar Velpuru[4]
Department of Information Technology,
Vardhaman College of Engineering (Autonomous),
Hyderabad, India.
munisek@vardhaman.org

Hammikolla Akshaya[5]
Department of CSE,
Vardhaman College of Engineering (Autonomous),
Hyderabad, India
hammikaakshaya04@gmail.com

Sai Prakhyath Siripuram[6]
Department of ECE,
Vardhaman College of Engineering (Autonomous),
Hyderabad, India
saiprakhyath37@gmail.com

***Abstract*— The proliferation of online shopping platforms has brought about a surge in user-generated product reviews, making it susceptible to the infiltration of fake reviews. For these platforms to continue to be dependable and reliable, it is necessary to identify and mitigate the impact of fake reviews. When depending on reviews for the product present on various web pages and applications, the rate of false reviews has been growing in the e-commerce sector. The goal is to anticipate and identify fraudulent reviews on e-commerce sites, namely Amazon, by using a hybrid model that combines classic machine learning (ML) with natural language processing (NLP). The proposed hybrid approach that has been suggested seeks to improve detection accuracy and interpretability by utilizing the combined abilities of ML and NLP technologies. Our method combines the power of Bidirectional Encoder Representations from Transformers (BERT), a state-of-the-art language model and Bag of Words (BoW) with traditional ML algorithms like Random Forest and XGBoost. To enhance model performance, we employ stacking ensemble method with logistic regression as the meta learner. The machine learns complex linguistic patterns, contextual information, and cooperative behaviors suggestive of fake reviews through training on many datasets. The outcomes of the experimental evaluations demonstrate the effectiveness of the hybrid model, surpassing existing methods in accuracy and robustness. This research contributes to a reliable solution, poised to enhance the trustworthiness of online product reviews and fortify consumer decision-making processes which guarantees continued safety and assurance in online shopping environments.**

Keywords*— *Bidirectional Encoder Representations from Transformers (BERT), Bag of Words (BoW), Machine Learning, Random Forest, XGBoost, Natural Language Processing, Hybrid Stacking model

I. INTRODUCTION

In the digital age, the proliferation of online platforms has transformed the way consumers interact with products and services. E-commerce websites, social media platforms, and online review forums have become essential channels for consumers to seek information and make informed purchasing decisions. These platforms act as a marketplace on the internet enabling users to buy, sell and give reviews for the products [1]. However, amidst the wealth of user-generated content lies a pervasive issue which is the prevalence of fake reviews. Reviews are important evaluations that shape market perceptions of goods, services, and experiences as well as customer choices [2]. The integrity of online information ecosystems is seriously threatened by fake reviews, which are purposefully created to mislead or influence customers. They weaken trust in online platforms and corrupt people's opinions on the quality of the products. Negative reviews can hurt revenues, whereas positive ratings might result in huge financial gains. Businesses are increasingly shaped by customer opinions, which lead to improvements in products and services. These reviews are often composed for financial benefit, product promotion, defaming the reputations of competitors, or even other motives. Addressing this issue necessitates advanced techniques that can distinguish genuine user reactions from fraudulent ones.

Nearly 70% of people have switched towards utilizing the internet for daily necessities and accessories as a result of its accessibility. When making purchases on these e-commerce sites, consumers rely only on the product ratings and reviews that these platforms offer to influence their judgments. The purpose of these fictitious evaluations is to mislead consumers and persuade them to purchase or avoid a particular product. These spams are able to appear first since there isn't a strong system in place to distinguish between genuine and false evaluations [3]. The sophisticated techniques used by those who create false reviews, such as making use of AI-generated content and many fake accounts, represent one of the main barriers to separating fake reviews from real ones.

The transparent nature of product review websites makes it extremely difficult to identify fraudulent individuals disguising themselves as real users. Certain companies compensate spammers for posting false reviews [4]. Typically, spammers create fake negative comments with the intention of criticizing competing goods. However, the evaluations that contain critical comments, which are true can't be classified as spam because they represent the views of actual customers [5]. Furthermore, with little labelled data available for training and an unbalanced ratio of real to fraudulent reviews, datasets used to train detection algorithms frequently suffer from data sparsity and imbalance. This imbalance may limit the effectiveness of detection systems in real-world situations, causing biased or poor performance. Making use of user data, new features can improve the detection of fraudulent activity by addressing weaknesses like the inability to read voice, facial expressions, or body language. In order to successfully identify misleading activities, the focus is still on analyzing platform access patterns, written reviews, and rating frequency over time [6]. Furthermore, in order to accurately detect fraudulent reviews, one must possess a profound understanding of language and semantics as they frequently contain subtle linguistic indicators and contextual ambiguities. Existing detection systems may struggle to capture these nuances effectively, resulting in false positives or false negatives. Another issue is scalability, since the amount of online reviews is increasing at an exponential rate and requires robust systems and algorithms for real-time identification. To effectively prevent fraudulent behaviors in the online review ecosystem, detection systems that are resilient, scalable, and dynamic must be developed.

In recent years, the domain of ML and NLP have emerged as powerful allies in the fight against fake reviews. ML algorithms offer the capacity to examine enormous volumes of data and extract meaningful patterns, while NLP techniques enable the understanding of human language at scale, particularly models like BERT, have shown remarkable results. By employing the strength of BERT for capturing semantic nuances and the effectiveness of BoW for managing big datasets, we hope to develop a robust detection system. The hybrid model improves overall performance by combining the outputs of different classifiers using a stacking ensemble technique. Our approach attempts to increase the accuracy and durability of fake review detection systems by utilizing these cutting-edge techniques. This research not only provides a reliable method for detecting fake reviews but also advances the more general goal of preserving the trustworthiness and integrity of online platforms. In doing so, it offers a more competitive and transparent marketplace for both consumers and businesses, thus promoting the continual growth of e-commerce and the digital economy.

II. Related Work

There are typically a lot of different ways to identify phoney evaluations, but it's challenging to match modern technologies while maintaining accuracy. This paper focuses on the SVM method to find fake reviews. It incorporates sentiment analysis to divide reviews into real and fake groups, filtering out false ones and recommending genuine products to users. Future improvements could involve enhancing the SVM algorithm's robustness and exploring additional features such as user behavior analysis to further improve the accuracy of fake review detection [7]. T. -Y. Lin, B. Chakraborty and C. -C. Peng proposed a framework for detecting fake reviews on platforms like Amazon and Yelp, emphasizing sentiment, topic, and readability features. Readability features are highlighted as particularly effective, and combining them with topics improves performance. Future research could enhance feature engineering, experiment with data augmentation and develop real-time detection methods for timely insights to businesses and consumers [8].

Addresses the growing concern of fake reviews in the E-commerce industry by employing NLP techniques and ML models to detect and eliminate them. It highlights the importance of trust in product reviews and the need for large platforms like Flipkart and Amazon to combat fake reviewers and spammers. Built using Naive Bayes and Random Forest methods, the proposed model enables instant detection of spam reviews, aiding website owners in taking necessary actions [9]. This research examines how fraud affects app rankings in the mobile market, making it harder for users to trust reviews and pushing developers to game the system. It discusses three main ways to spot fraud: looking at where an app ranks, its ratings, and the reviews it gets. Of these, reviews from real users are seen as the best guide, especially if they're already signed in. The study also looks at ways to catch fraud and stresses how important reviews are for students when choosing apps [10]. This paper uses a method to assess online review credibility without explicit labelling. It identifies key characteristics indicating authenticity and applies them in a classification model. The approach offers a generalized solution applicable across industries. Ultimately, it enhances trust in online reviews, aiding consumers in making informed decisions [11].

In this research, many DL based models for the spotting of false reviews are examined, emphasizing the significance of reliable online reviews in consumer decision-making. RoBERTa emerges as a standout performer, showcasing superior accuracy in detecting fake reviews. However, challenges such as group spammers detection, model interpretability, one-class classification, cross-domain detection, and multilingual analysis endure. Future study should concentrate on tackling these issues to enhance robustness and reliability [12]. This study uses an ensemble method based on multi-view learning to suggest a novel solution to the ubiquitous problem of false internet reviews. In contrast to conventional bag-of-words models that have issues with sparsity and handling unfamiliar words, the authors present a model that brings together CNN with bag-of-n-grams. With the addition of non-textual data about reviewer behavior, this integration

facilitates the solid representation of features and the fast retrieval of local context from review text [13].

This paper addresses a substructure for spotting fake reviews on web sites using sentiment analysis with the help of GRNN, LSTM and Bi-LSTM and also activation functions to separate fake from real reviews [14]. This paper highlights individual and group spam detection algorithms while providing an extensive evaluation of approaches for identifying fraudulent reviews in e-commerce. It talks about the different attributes that these techniques use, the datasets for products, reviews, and updates. This study evaluates existing methods that require labelled datasets, especially those based upon supervised machine learning techniques. In the end, it defines important research questions and outlines potential directions for further study in the field of false review identification [15]. The paper explores the effectiveness of Naive Bayes classifiers that shows the difference between true and fake reviews based on their characteristics. While Naive Bayes offers quick training and compatibility with text data, it may not identify all fake reviews and faces challenges in data pre-processing [16].

III. Materials and Methods

We utilized publicly available fake reviews dataset which consists of above 40000 reviews which are computer generated reviews and original reviews. Original reviews are those that were given by the reviewers themselves, while computer generated reviews are fake reviews. The reviews were collected from the internet and contain features such as category, review text, rating, and a label column which indicates whether a review is true or fake review. Python, NLTK, and scikit-learn are the software tools used for ML and NLP respectively. A variety of preprocessing techniques were used in the study to clean and get the data ready for analysis. These steps included tokenizing the text, eliminating stop words, removing out HTML tags, special characters, and numbers, and lemmatizing words to their most basic forms. The textual data was converted into numerical representations suitable for ML models by using feature extraction techniques like Bag of Words (BoW) and BERT embeddings. The cleaned data was used to train a number of models, such as Random Forest and XGBoost. We blended the features from BERT and BoW into a hybrid model to take advantage of the benefits of both feature extraction methods. By combining the results of the Random Forest and XGBoost models, this model created a stacking ensemble by feeding the logistic regression meta-learner's output. Accuracy, precision, recall, F1 score, and area under the ROC curve (AUC) are the metrics used to assess the models. By evaluating the model's capacity to reliably categorise reviews, reduce false positives and negatives, and generalise across various data subsets, these metrics offer an all-encompassing overview of its performance.

IV. Proposed Method

This paper's methodology section describes the specific actions required to make a robust system for spotting counterfeit reviews with the help of a hybrid fusion. Fig. 1 describes the flowchart of our proposed model.

We began by preprocessing the textual data to improve its quality and uniformity. This involved removing HTML tags, special characters, and digits using regular expressions, and converting all text to lowercase for consistency. After that, we eliminated stop words and tokenized the content into individual words, using the Natural Language Toolkit (NLTK) library. To standardize word forms, we applied lemmatization with NLTK's WordNetLemmatizer.

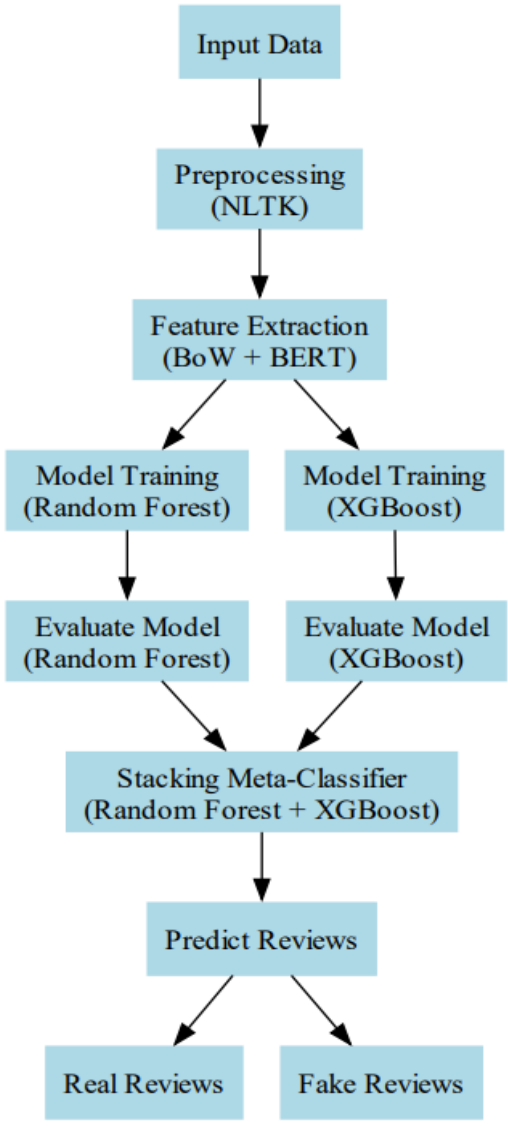

Fig. 1 Flowchart of the Proposed Model

To capture the semantic richness of the reviews, we employed multiple feature extraction techniques. The text is then converted to a sparse matrix of token counts with the help of a CountVectorizer function from the sklearn library. Tokenizing the text and compiling a list of the

terms that are used most frequently are required for this. A feature matrix is made, with a term from the vocabulary represented by each column and a row representing a review. This matrix's values correspond to term frequencies.

Furthermore, we also generated embeddings that capture more complex semantic information using BERT (Bidirectional Encoder Representations from Transformers). Because BERT models can comprehend a word's context inside a sentence, a more complex representation of the text is possible. These embeddings are obtained using the transformers package and utilized as input features for the models. BERT can be modified on a labelled review dataset to guarantee that the model gains the ability to encode meaningful representations unique to the task of classifying review authenticity. By combining these methods, we concatenate the feature vectors from the BERT and BoW embeddings to produce a hybrid representation. By combining the best aspects of contemporary and conventional text representations, this hybrid feature set improves model performance by offering a thorough understanding of the text data.

Three different models are used in the model training phase those are Random Forest, XGBoost, and a stacking classifier. The Random Forest model constructs several decision trees and combines their outputs to produce precise predictions. It is renowned for its resilience and capacity to handle high-dimensional data. To fix mistakes made by earlier models, XGBoost, a scalable and effective gradient boosting implementation, creates decision trees one after the other. Using the hybrid feature set derived from the BoW and BERT embeddings, the model is initially trained on the training group.

Preprocessed data is separated into sets for training and testing. Typically, 20% of the preprocessed data is used for testing and 80% is used for training. The BERT embeddings and BoW are used as feature inputs in the Random Forest and XGBoost classifiers to train individual instances. With this combination, classification accuracy is increased as the classifiers are able to identify intricate patterns and dependencies in the review text. The base models (Random Forest and XGBoost) predict the fake reviews using this hybrid feature set, and their outputs are blended using logistic regression as a meta classifier in a stacking ensemble model. Performance using this ensemble method is generally superior to that of using a single model alone. We did comprehensive tests and analyzed our model's performance on multiple metrics, such as rate of accuracy, amount of preciseness, recall, and F1-score, to assess how efficient they are. The results show us the Stacking Classifier consistently surpasses other models, achieving the highest accuracy. Stacking Classifier then predict the reviews and differentiate them as real and fake reviews. This hybrid fusion approach leverages the strengths of different text representation techniques and ML algorithms, providing a comprehensive solution for the reliable detection of fake reviews. The proposed methodology improves fraudulent review detection systems but also contributes to the broader field of sentiment analysis and natural language understanding.

V. Simulation and Result Analysis

Our experiments were conducted on the fake reviews dataset which is available on internet that has computer generated and original reviews, where we applied multiple feature extraction techniques, such as BoW, BERT and trained several ML models to identify fake reviews. We assessed our model's performance using F1-score, accuracy, precision, and recall metrics. Table1 summarizes the accuracies for all models.

TABLE I. Accuracy of each model

Model	**Accuracy**
Random Forest (BoW)	80.66
Random Forest (BERT)	79.26
XGBoost (BoW)	84.44
XGBoost (BERT)	81.08
Stacking Classifier (Hybrid)	86.45

The Stacking Classifier with BoW, BERT and traditional ML models achieved the precise accuracy of 86.45%, surpassing other techniques. A detailed classification report for the Stacking Classifier is shown in table2. The weighted average metrics reflect the overall robust performance of this model, making it the most effective for fake review detection.

	precision	recall	f1-score	support
CG	0.86	0.87	0.86	4016
OR	0.87	0.86	0.86	4071
accuracy			0.86	8087
macro avg	0.86	0.86	0.86	8087
weighted avg	0.86	0.86	0.86	8087

Fig. 2. Classification report of Stacking Classifier

The confusion matrix for the Stacking Classifier with BERT embeddings further illustrates the performance of the model in fig. 2, highlighting the distribution of predicted and actual classes, which displays a balanced distribution of true positives and true negatives with fewer false positives and false negatives indicating model's robustness. The model correctly identified 3483 out of 4016 actual negative samples and 3508 out of 4071 actual positive samples, demonstrating its effectiveness in distinguishing between fake and genuine reviews.

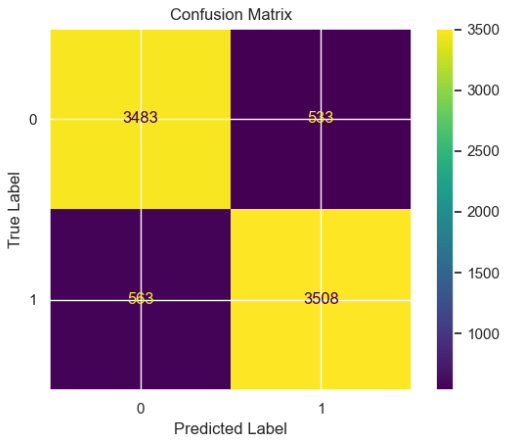

Fig. 2. Confusion matrix for Stacking Classifier

The effectiveness of each model in separating between genuine and fraudulent reviews is shown by the ROC curves. The stacking ensemble shows its superior capacity to differentiate between the classes by achieving the highest AUC of 0.95 in fig. 3. The stacking ensemble's ROC curve continuously beats the individual model's, demonstrating its superior ability to strike a balance between sensitivity and specificity. The high AUC value indicates the reliability in finding false reviews, making it an important tool for e-commerce websites to preserve the legitimacy and honesty of online reviews.

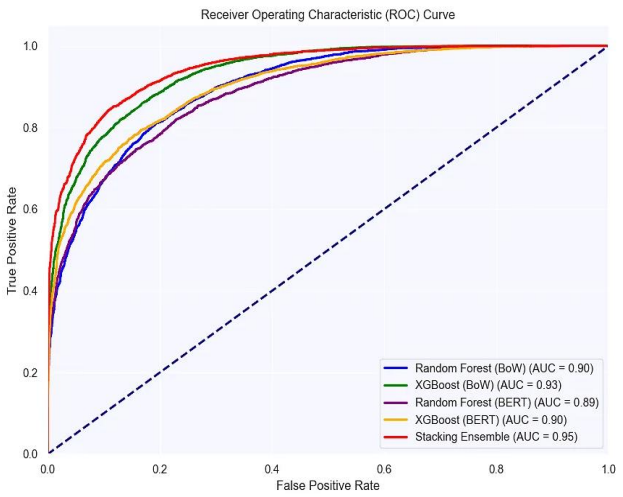

Fig. 3. ROC Curve of models

The experimental results confirm our prediction that advanced feature extraction techniques like BERT and BoW when combined with ensemble approaches, particularly stacking classifiers, greatly improve fake review detection systems performance which is shown in fig. 4. Particularly, the stacking classifier outperformed the others in terms of accuracy, proving the value of fusing traditional and BERT based features which shows the value of contextualised word representations and the effectiveness of mixing several models.

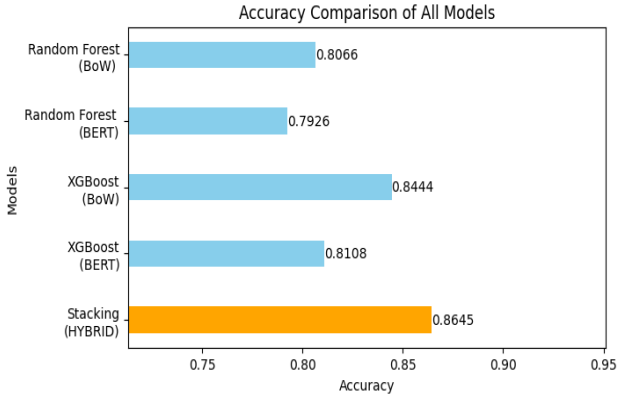

Fig. 4. Accuracy comparison of different models

VI. Discussion

The findings of our research demonstrate the value of combining ensemble learning models and various feature extraction techniques to analyze false reviews and detect them. The proposed methodology, where BoW captures word frequency effectively combines BERT embeddings with a stacked ensemble of Random Forest and XGBoost classifiers has demonstrated superior performance in identifying fake reviews. The high accuracy and reliability of the proposed model have significant implications for practical applications in e-commerce and beyond. Additionally, the methodology can be modified to fit other fields that require sentiment study findings, like customer feedback analysis and social media monitoring.

VII. Conclusion

In conclusion, our method illustrates the efficiency of a hybrid fusion combining NLP and ML models to enhance the accuracy and reliability of automated false review identification systems. Our research trained multiple machine learning (ML) classifiers, such as Random Forest and XGBoost and combined BoW and BERT embeddings. Our findings suggest that leveraging both word frequency and contextual embeddings can significantly increase the detection of fake reviews when combined using a stacking meta-classifier. Comparing our methodology against individual classifiers and baseline methods, the experimental results showed how successful it is in enhancing classification efficiency and dependability. The higher performance metrics were a result of the synergy between the predictive power of ensemble learning and BERT's semantic knowledge. By achieving high accuracy in identifying fake reviews, our system helps to preserve the reliability and trustworthiness of online review sites, empowering consumers to make more informed decisions.

VIII. References

[1] M. A. Hadiwijaya, F. P. Pirdaus, D. Andrews, S. Achmad and R. Sutoyo, "Sentiment Analysis on Tokopedia Product Reviews using Natural Language Processing," 2023 International Conference on Informatics, Multimedia, Cyber and Informations System (ICIMCIS), Jakarta Selatan, Indonesia, 2023, pp. 380-386, doi: 10.1109/ICIMCIS60089.2023.10348996.

[2] P. Shetgaonkar, J. T. Rodrigues, S. Aswale, V. L. K. Gonsalves, J. C. Rodrigues and A. Naik, "Fake Review Detection Using Sentiment Analysis and Deep Learning," 2021 International Conference on Technological Advancements and Innovations (ICTAI), Tashkent, Uzbekistan, 2021, pp. 140-145, doi: 10.1109/ICTAI53825.2021.9673375.

[3] V. P. Sumathi, S. M. Pudhiyavan, M. Saran and V. N. Kumar, "Fake Review Detection Of E-Commerce Electronic Products Using Machine Learning Techniques," 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Coimbatore, India, 2021, pp. 1-5, doi: 10.1109/ICAECA52838.2021.9675684.

[4] D. Jain, S. Kumar and Y. Goyal, "Fake Reviews Filtering System Using Supervised Machine Learning," 2022 IEEE International Conference on Data Science and Information System (ICDSIS), Hassan, India, 2022, pp. 1-4, doi: 10.1109/ICDSIS55133.2022.9915878.

[5] Singh, M. Memoria and R. Kumar, "Deep Learning Based Model for Fake Review Detection," 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT), Gharuan, India, 2023, pp. 92-95, doi: 10.1109/InCACCT57535.2023.10141826.

[6] M. Abdulqader, A. Namoun and Y. Alsaawy, "Fake Online Reviews: A Unified Detection Model Using Deception Theories," in IEEE Access, vol. 10, pp. 128622-128655, 2022, doi: 10.1109/ACCESS.2022.3227631..

[7] R. Poonguzhali, S. F. Sowmiya, P. Surendar and M. Vasikaran, "Fake Reviews Detection using Support Vector Machine," 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 2022, pp. 1509-1512, doi: 10.1109/ICSCDS53736.2022.9760747.

[8] T. -Y. Lin, B. Chakraborty and C. -C. Peng, "A Study on Identification of Important Features for Efficient Detection of Fake Reviews," 2021 International Conference on Data Analytics for Business and Industry (ICDABI), Sakheer, Bahrain, 2021, pp. 429-433, doi: 10.1109/ICDABI53623.2021.9655845.

[9] S. M. Anas and S. Kumari, "Opinion Mining based Fake Product review Monitoring and Removal System," 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2021, pp. 985-988, doi: 10.1109/ICICT50816.2021.9358716.

[10] K. Manoj, T. S. Sandeep, N. Sudhakar Reddy and P. M. D. Alikhan, "Genuine ratings for mobile apps with the support of authenticated users' reviews," 2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT), Bangalore, India, 2018, pp. 217-221, doi: 10.1109/ICGCIoT.2018.8753068.

[11] R. Mothukuri, A. Aasritha, K. C. Maremalla, K. N. Pokala and G. K. Perumalla, "Fake Review Detection using Unsupervised Learning," 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 2022, pp. 119-125, doi: 10.1109/ICSCDS53736.2022.9760908.

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

[13] Ashraf, F. Rehman, H. Sharif, H. Kirn, H. Arshad and H. Manzoor, "Fake Reviews Classification using Deep Learning," 2023 International Multi-disciplinary Conference in Emerging Research Trends (IMCERT), Karachi, Pakistan, 2023, pp. 1-8, doi: 10.1109/IMCERT57083.2023.10075156.

[14] J. C. Rodrigues, J. T. Rodrigues, V. L. K. Gonsalves, A. U. Naik, P. Shetgaonkar and S. Aswale, "Machine & Deep Learning Techniques for Detection of Fake Reviews: A Survey," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 2020, pp. 1-8, doi: 10.1109/ic-ETITE47903.2020.063.

[15] R. Agarwal and D. K. Sharma, "Detecting Fake Reviews using Machine learning techniques: a survey," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022, pp. 1750-1756, doi: 10.1109/ICACITE53722.2022.9823633.

[16] P. Kalaivani, V. D. Raj, R. Madhavan and A. P. Naveen Kumar, "Fake Review Detection using Naive Bayesian Classifier," 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS), Coimbatore, India, 2023, pp. 705-709, doi: 10.1109/ICSCSS57650.2023.10169838