ISSN: **3107-6513**

Fake Review Detection Using Transformer-Based Models With Explainable Artificial Intelligence and Machine Learning Techniques

Mbuyi Cecile Ngoie

Maters' Student, Department of Computer Science, Rathinam College of Arts and Science (Autonomous), Coimbatore, Tamil nadu, India.

cecilembuyin@gmail.com

***Abstract*—**Online reviews play a major role in influencing consumer behavior in online shopping environments. However, there has been a noticeable growth of fake or misleading reviews, which cause a serious problem as it represents consumers and affects the credibility of online business. This research is centered on the development of a machine learning and deep learning-based on system to detect fake reviews.This research employs classical machine algorithms,including Logistic Regression, Support Vector Machine(SVM), and Random Forest along with TF-IDF for feature extraction to identify reviews as genuine or fake. Additionally transformer-based models such as BERT are employed to capture and interpret the textual information in context to better the performance of the model. Furthermore, explainable AI methods like LIME and SHAP were used to give explanations for the predictions made by the model and to increase interpretability.The main goal of the study is improve the overall performance of the fake review detection system while being transparent and interpretative without any compromises. The outcome of the study shows that the use of classical machine learning techniques together with transformer-based architectures strengthen the efficiency of the fake review detection systems.

Keywords—TF-IDF, Explainable Artificial Intelligence(XAI), Support Vector Machine (SVM), Bidirectional Encoder Representations from Transformers(BERT), Local Interpretable Model-Agnostic Explanations(LIME), SHapley Additive exPlanations(SHAP).

I. INTRODUCTION

Online reviews play a crucial role in regard to customer's decisions of buying good or services in the modern digital era. Customers understand the quality and performance of goods and services before making a purchasing decision with the help of reviews. There is now a big amount of customer-generated reviews because of the ever-growing use of online markets and platforms. Meanwhile, one of the biggest challenges in this matter is the problem of misleading or fake reviews. Fake reviews are a form deception or manipulation used to wrongfully promote or discredit products or services, thus misleading consumers perception and affecting buying decisions.Furthermore, they sabotage the reliability of online platforms and businesses. The issue is more increased with the ability of creating large volumes of similar reviews, rendering manual detection inefficient, expensive, and less practical.

Automated detection systems based on machine learning have been broadly implemented in order to address the issue of misleading or fake reviews. Algorithms like Logistic Regression, Support Vector Machines (SVM), and Random Forest are often used, mostly depending on feature extraction techniques like TF-IDF to transform text into numerical form. Despite their effectiveness, these approaches may face difficulties in capturing complex linguistic patterns and contextual relationships. Deep learning methods, in particular transformer- based models such as BERT (Bidirectional Encoder Representations from Transformers), address these challenges by effectively modeling contextual relationships and and detecting nuanced patterns in a text.Moreover, explainability is essential in such systems. Explainable such as LIME and SHAP provide interpretability of model prediction fostering greater transparency, reliability and trust.

II. LITERATURE REVIEW

Several studies have investigated fake review detection using both conventional machine learning and deep learning.Ealy approaches employed algorithms like Support Vector Machine (SVM), Naïve Bayes, and Logistic Regression,using handcrafted features including patterns and sentiment analysis [1],[2].Nervelessness, these approaches usually fail to capture contextual relationships in text. With recent advancements, transformer-based models like BERT have gained great attention due to their ability to capture contextual and semantic information,resulting in improved classification performance [3].In addition, hybrid approaches that mix conventional machine learning and deep learning methods have been to further improve accuracy and robustness [4].Moreover, Explainable Articial Intelligence (XAI) methodes, like LIME and SHAP, have been introduced to improve model interpretability by detecting key features that influencepredictions [5],[6].Although there are progress,obstacles like data imbalance and the detection of subtle deceptive patterns continue.

A. System Design

As the number of fake reviews on online platforms continue to grow, the need for need for effective detection techniques has become very important. This study presents an hybrid framework to detect fake reviews by using machine learning models, transformer-based techniques, and explainable AI. A diagram illustrating a system design architecture is presented. The primary contributions of this research are as follow:
The application of text preprocessing tools such as cleaning, tokenization, and normalization to enhance the quality of data.

i. The application of TF-IDF for conventional models and BERT for generating contextual embaddings from from texual data.
ii. The implementation of classifiers such as Logistic Regression, SVM, Random Forest, and BERT to improve classification.
iii. The implimentation of Explainable AI techniques such as LIME and SHAP, to ensure classification of model prediction.
iv. The evaluation of models performance through
v. Accuracy, precision, recall, F1-score, and confusion matrix.

1. Dataset Collection

More than 400 textual reviews dataset was obtained from publicly available platforms like Kaggle for fake review detection. The dataset is labeled into two classes: Fake and Genuine (Real).The dataset was developed into training and testings sets, with approximately 250 reviews used for training and 150 for testing. This division allows effective learning and reliable evaluation on unseen data. Enabling models to capture patterns from training data and evaluate their performance on unseen data. The dataset comprises different ranges of review texts with different writing styles and sentiment expressions enhancing the robustness of performance and generalization of the proposed models.

2. Pre-processing

Pre-processing is a crutial step that improves the quality of textual data by eliminating noise and preparing it for effective model training. The pre-processing steps adopted in this study include:

1. Text cleaning to erase special characters, punctuation, and unrelated symbols.
2. Transformation of text to lowercase to guarantee uniformity
3. Tokenization to divide text into indivdual words.
4. Erasing stop words to exclude commonly employed but insignificant terms.
5. The application of Lemmatization to change words to their base or root form.

These steps reduce noise, enhance data consistency, and better the quality of inputs features. The effective prepossessing enables upgraded feature extraction and better the overall performance of models' classification

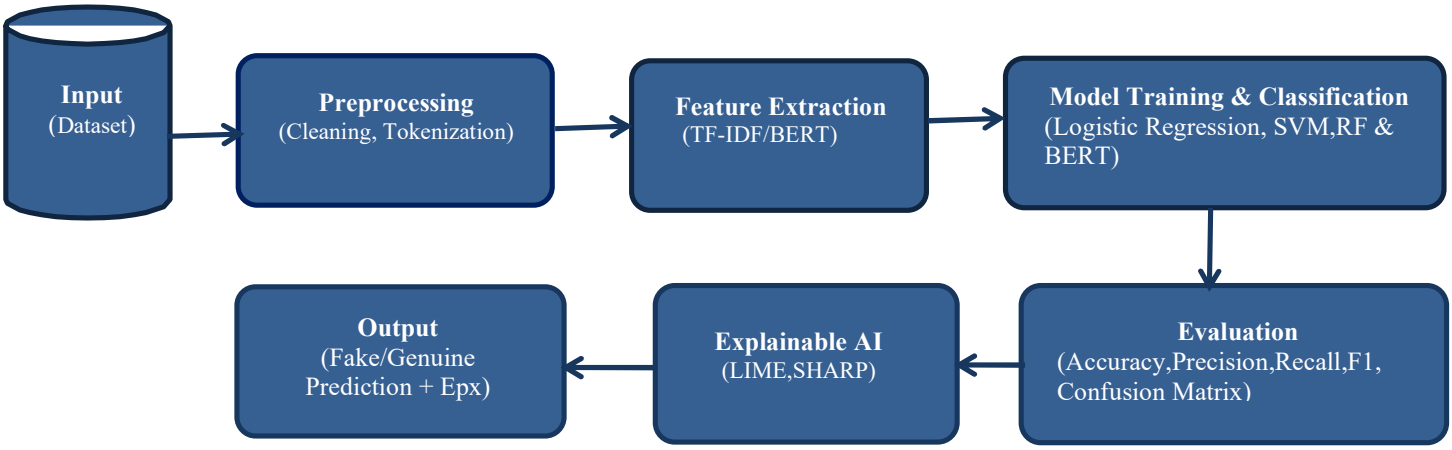


Fig. 1 System architecture of fake review detection

 ISSN: **3107-6513**

3. TF-IDF(Term Frequency–Inverse Document Frequency)

This method transforms text into feature vectors by assigning weights according to their performances across the dataset.This approach commonly applies traditional machine learning models like Logistic Regression, SVM, and Random Forest.

4. Feature Classification using Machine Learning and Transformer Models

Traditional machine learning including logistic regression,Support vector machine (SVM),and Random Forest, are trained using TF-IDF features,whereas BERT is used to capture contextual information,thus bettering classification performance.Feature classification involves categorizing reviews based on either fake or Genuine classes based on the extracted features.

5. BERT Embeddings

Birectional Encoder Representations from Transformers(BERT)is employed to model contextual and semantic relationships with textual data. Unlike conventional methods, BERT accounts for word context within sentences, resulting in better classification accuracy.

Performance Metrics Evaluation

Precision Recall (Pr)

Pr = Tp / Tp + Fn

Recall (Rc)

Rc=Tp/Tp+FnTp

Accuracy (Acc)

Acc = Tp + Tn/ Tp + Tn + Fp + Fn

F1-Measure (F-M)

F-M = 2 × Pr × Rc / Pr + Rc

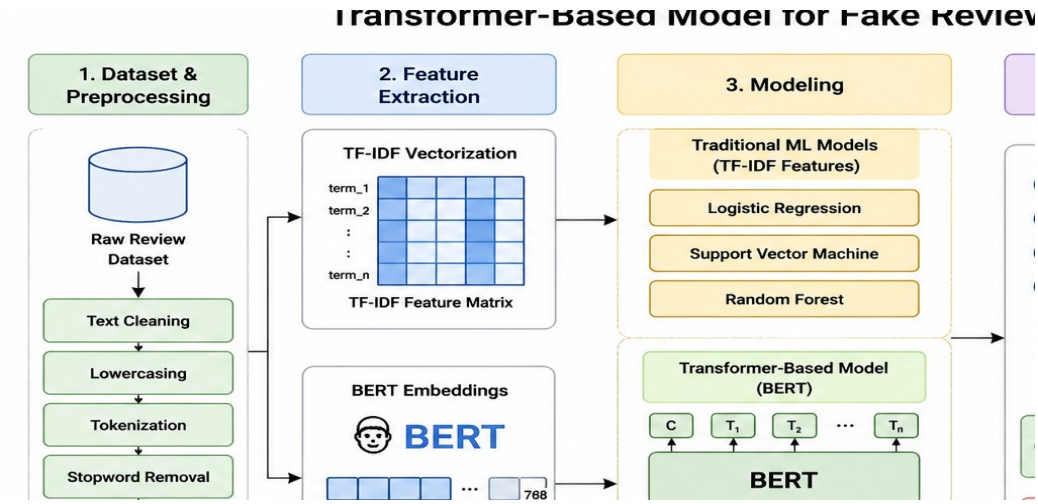
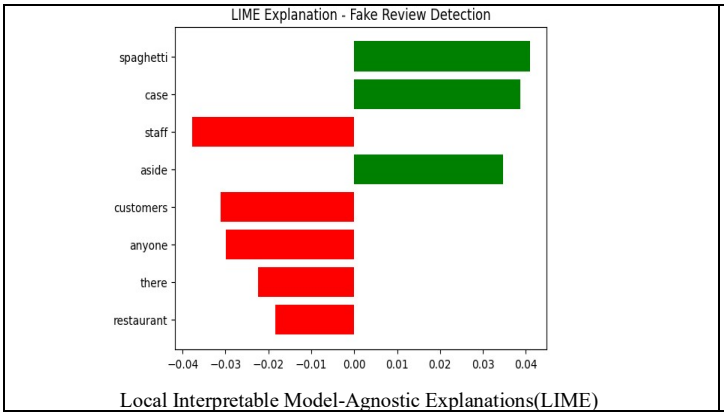
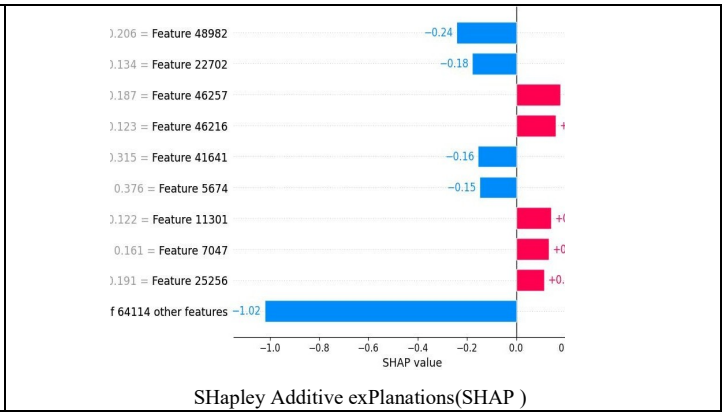


Fig. 2 Illustrates the architecture of the proposed hybrid machine learning and transformer-based model for fake review detection.



Local Interpretable Model-Agnostic Explanations(LIME)



SHapley Additive exPlanations(SHAP)

 ISSN: **3107-6513**

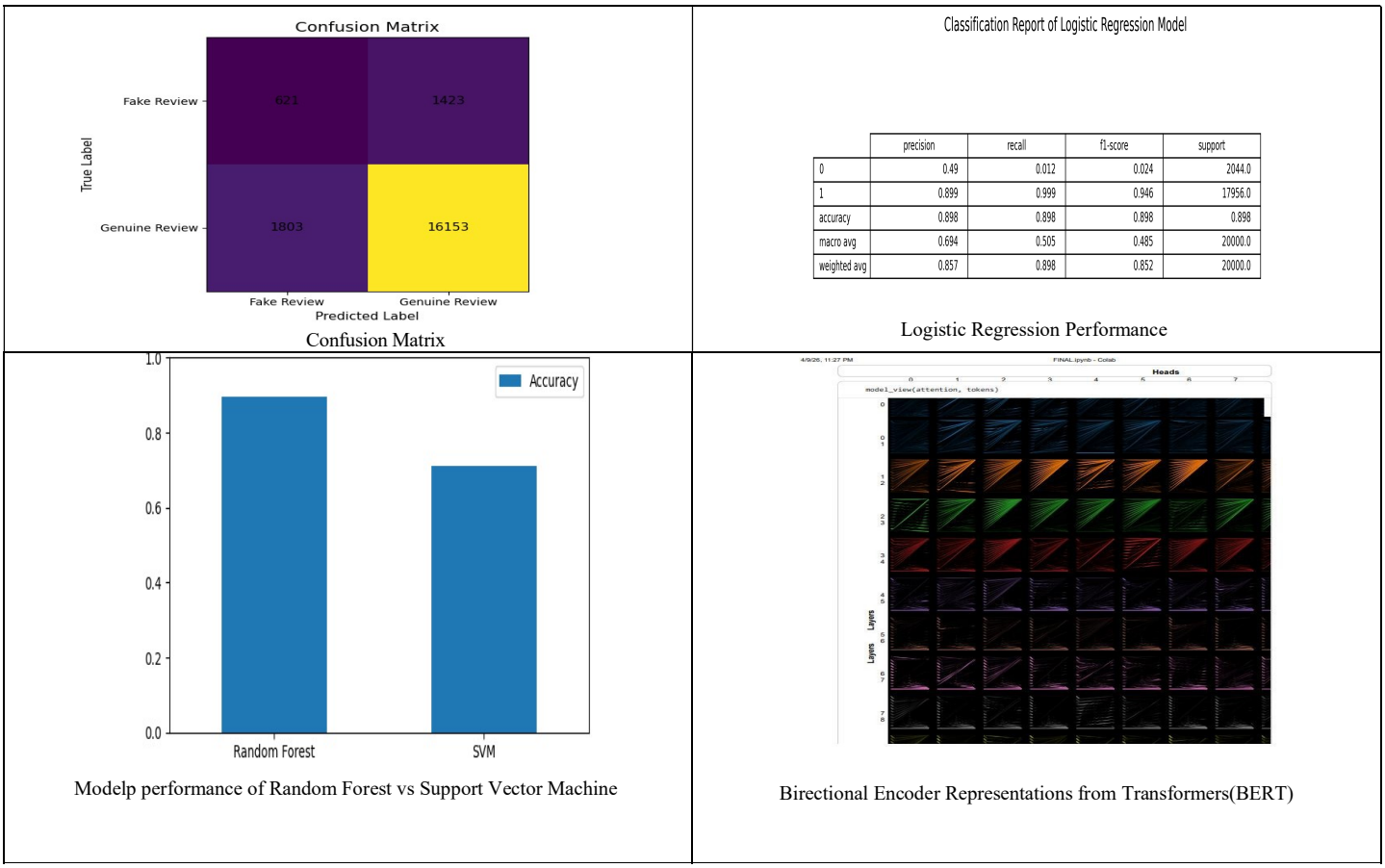


Fig. 3 Outcomes from inputs text from the dataset

Metrics for Assessing Performance	Logistic Regression	SVM	Random Forest	BERT (Proposed Hybrid)
Accuracy (%)	84.00	71.00	89.62	91.50
Precision (%)	91.90	75.00	89.86	92.30
Recall (%)	89.95	70.00	99.68	93.80
F-measure (%)	90.92	72.40	94.52	93.00

Fig. 4 Forecast the overall performance of skin cancer classification schemes

III. PREPARE YOUR PAPER BEFORE STYLING

The proposed hybrid fake review detection framework was evaluated using different models like Logistic Regression, Support Vector Machine (SVM), Random Forest, and as well as the transformer-based BERT model. The database underwent preprocessing and was divided into training and testing sets, with a remarkable class imbalance favoring genuine reviews. Random Forest achieved highest accuracy(89.62%), whereas Logistic Regression and SVM displayed a competitive performance. Nevertheless, all models were biased toward the majority class, achieving a better performance on genuine reviews, emphasizing the importance of recall and F1- score over accuracy.

Furthermore, the BERT model enhanced performance by effectively capturing contextual and semantic representation of the text. Overall, the hybrid approach perform better than individual models by blending machine learning and deep learning techniques. The incorporation of LIME and SHAP futher increase interpretability, thereby making the system more the detection of fake review more reliable.

IV. OUTCOMES AND CLOSURE

The proposed hybrid fake review detection framework was evaluated using different models like Logistic Regression, Support Vector Machine (SVM), Random Forest, and as well as the transformer-based BERT model. The database underwent preprocessing and was divided into training and testing sets, with a remarkable class imbalance favoring genuine reviews. Random Forest achieved highest accuracy(89.62%), whereas Logistic Regression and SVM displayed a competitive performance. Nevertheless, all models were biased toward the majority class, achieving a better performance on genuine reviews, emphasizing the importance of recall and F1- score over accuracy.

Furthermore, the BERT model enhanced performance by effectively capturing contextual and semantic representation of the text. Overall, the hybrid approach perform better than individual models by blending machine learning and deep learning techniques. The incorporation of LIME and SHAP futher increase interpretability, thereby making the system more the detection of fake review more reliable.

V. CONCLUSION

In the proposed framework, review texts are preprocessed and TF-IDF are used to extract features by transforming raw data into structured numerical form using unigram and bigram representations. For classification such as machine learning models like Logistic Regression, Random Forest, and Support Vector Machine (SVM) were used. Notably Support Vector Machine (LinearSVM) demonstrated strong performance on textual data. In addition, a BERT model was incorporated to capture deeper contextual and semantic presentation. Standard metrics like precision, recall, accuracy and F1-score on a labeled dataset were used to evaluate the models. The outcomes indicated that approaches achieves effective and reliable performance.

Future work will aim to use larger datasets, strengthen model robustness, and use explainable AI approaches to achieve better interpretability.

REFERENCES

[1] J. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 309–319, 2011.

[2] N. Jindal and B. Liu, "Opinion spam and analysis," *Proceedings of the International Conference on Web Search and Data Mining (WSDM)*, pp. 219–230, 2008.

[3] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic detection of fake reviews," *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pp. 3391–3401, 2018.

[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

[5] T. Mikolov et al., "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[6] H. Li, Z. Chen, A. Mukherjee, B. Liu, and J. Shao, "Analyzing and detecting opinion spam on a large-scale dataset via temporal and linguistic patterns," *Proceedings of ICWSM*, 2014.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," *Proceedings of the 22nd ACM SIGKDD Conference*, pp. 1135–1144, 2016. (LIME)

[8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4765–4774, 2017. (SHAP)

[9] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[10] Kaggle, "Fake and real review dataset," [Online]. Available: https://www.kaggle.com/