

Fake Review Detection Using Hybrid Transformer and Behavioral Analysis

**Janagan A.¹, Madhan P.², Leo Aandrew G.³,
Mohammed Azardeen M.⁴**

¹Assistant Professor, ^{2,3,4} Student, Department of Computer Science and Engineering,
NPR College of Engineering and Technology, Natham, Dindigul, India.

E-mail: ¹janagana@nprcolleges.org, ²madhanpitchaimani@gmail.com, ³andrewleo173@gmail.com,

⁴mohamedazardeen7867@gmail.com

Abstract

Fake review detection is challenging when the deceptive information is semantically plausible while reviewer level and rating information are unusual. This paper presents Review Verity, a multi-modal fake review detection system that uses context, behavior and language representations in a single classification model. The language representation is created using static BERT-base-uncased model to generate 768-dimensional vector and account age, verified-purchase and multiple IP usage create three behavioral representation features. Review depth, subjectivity, Flesch readability, rating extremeness, sentiment agreement and rating disagreement create six linguistic representation features that combine with the vectors to produce 777-dimensional feature representation set. The feature set is normalized and rebalanced using SMOTE and then training is performed using 100 trees Random Forest classifier. SHAP and LIME models interpret the features and instances and the final model is deployed using FastAPI and React frameworks.

Keywords: Fake Review Detection, BERT Contextual Embeddings, Behavioral Feature Fusion, Linguistic Authenticity Signatures, SMOTE-Based Feature Balancing, Explainable Random Forest, SHAP–LIME Attribution

1. Introduction

Product and service reviews that are posted online have become one of the most important information sources affecting consumers' purchasing decisions. Nevertheless, an increase in the use of fake, incentivized, and automated reviews affects the credibility of review websites. Inaccurate reviews may be created to replicate actual customer experiences, making it hard to detect their inaccuracy if decisions are made using text analysis alone. It becomes even more challenging due to variances in the behaviors and access patterns of reviewers.

The Existing techniques used for detection concentrate on either text properties or reviewer information. The former can be effective at identifying linguistic features, but it will not work when there is semantic plausibility of the review. Conversely, behavioral methods can detect suspicious reviewing practices, though they might fail to differentiate between legitimate and fake writing styles. Consequently, an effective fraud detection mechanism would entail the combination of contextual semantics, reviewer's behavior, and linguistically identifiable features.

This research proposes the Review Verity framework, a hybrid fake-review detection combines a set of features including contextual features obtained through BERT model pre-training, behavioral features of reviewers, and linguistic signatures. After fusion, the resulting feature vector is then fed through a classification algorithm based on machine learning, where normalization and balancing at the training stage are used. In order to increase interpretability, the use of SHAP and LIME for interpreting classification decisions in a general case and in individual instances is considered. Moreover, in order to measure the impact of the linguistic feature module, a comparative experiment with BERT-behavioral versus the whole multimodal set is carried out. Finally, the presented framework is embedded into the FastAPI and React application.

2. Related Works

Fake review detection has evolved beyond traditional approaches of detecting review fraud to transformer-based learning, behavioral learning, and explainable learning approaches. Initial studies highlighted the significance of reviewer's behavior and reputation for the detection of fraudulent reviews [2]. The development of transformer-based language representations made it possible to develop context-aware review representation using BERT [1]. Further, SHAP and

LIME provided explaining the results of machine-learning models [3], [4].

Several recent researches have started to incorporate both textual and structured for improved detection. Credibility analysis has shown the benefits of merging review features and semantic aspects together [5]. Multi-view learning has additionally provided evidence that it is useful to utilize complementary feature representation in detecting and explaining fakes reviews [6]. The relationship between target and non-target features has also been considered in fraudulent reviewer detection [7]. Transformer models based on RoBERTa have proved the value of contextual embeddings for detecting deceptive reviews [8]. Interpretability in e-commerce review analysis has shown the importance of integrating review features with explainable predictive models [9]. In more recent times, the combination of RoBERTa with behavioral features has revealed the importance of combining reviewer features with transformer embeddings [10].

Despite these developments, integration of contextual transformer representations along with reviewer behavioral features and designed language signatures is an important research path in light of these advances. The introduced Review Verity system takes care of such an integration along with including SHAP and LIME techniques in its analysis pipeline.

3. Proposed Methodology

3.1 Review Verity Framework and Dataset Description

The proposed Review Verity framework is presented as an integrated pipeline approach for detecting fake reviews that involves contextual text representation, reviewer's behavior details and linguistic features. As can be seen from Figure 1, the framework starts from review collection and preprocessing, and next the parallel feature extraction, feature fusion, normalization, class balancing, classification, and explainability tasks are performed. The framework is evaluated based on the Amazon Product Reviews Dataset provided by Kaggle [11]. The initial data set includes 21,058 review records. For the evaluation purposes, 500 reviews were randomly chosen with `random_state = 42`.

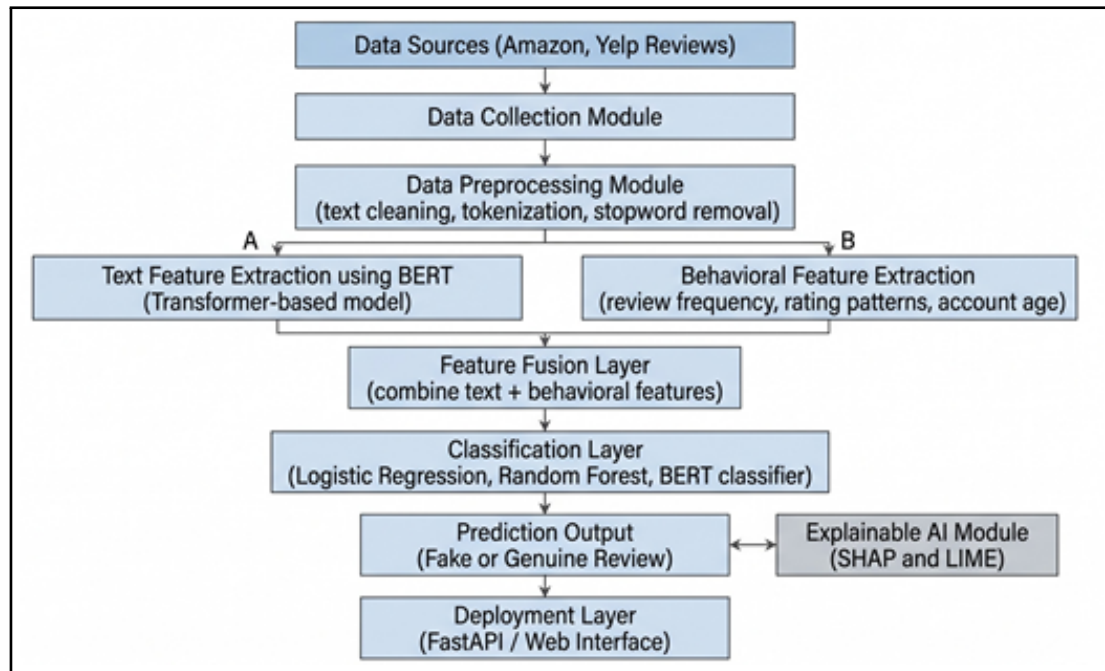


Figure 1. Architecture of the review verity hybrid fake review detection framework

The selected review data set is split into 80% training data and 20% test data, which constitutes 400 training reviews and 100 test reviews, respectively. The training set is used to perform feature extraction and construct classifiers, while the test set remains for model evaluation.

After dataset preparation, each selected review will go through the preprocessing step and then represented using three streams of features that complement each other. In particular, the semantic stream involves the use of a pre-trained BERT model to construct contextualized review features, and the behavioral stream involves the attributes of the reviewer, while the linguistic stream uses the structured features from the review text and rating. These streams' output is then merged to produce one feature representation to be used for classification. Figure 1 shows the processing flow.

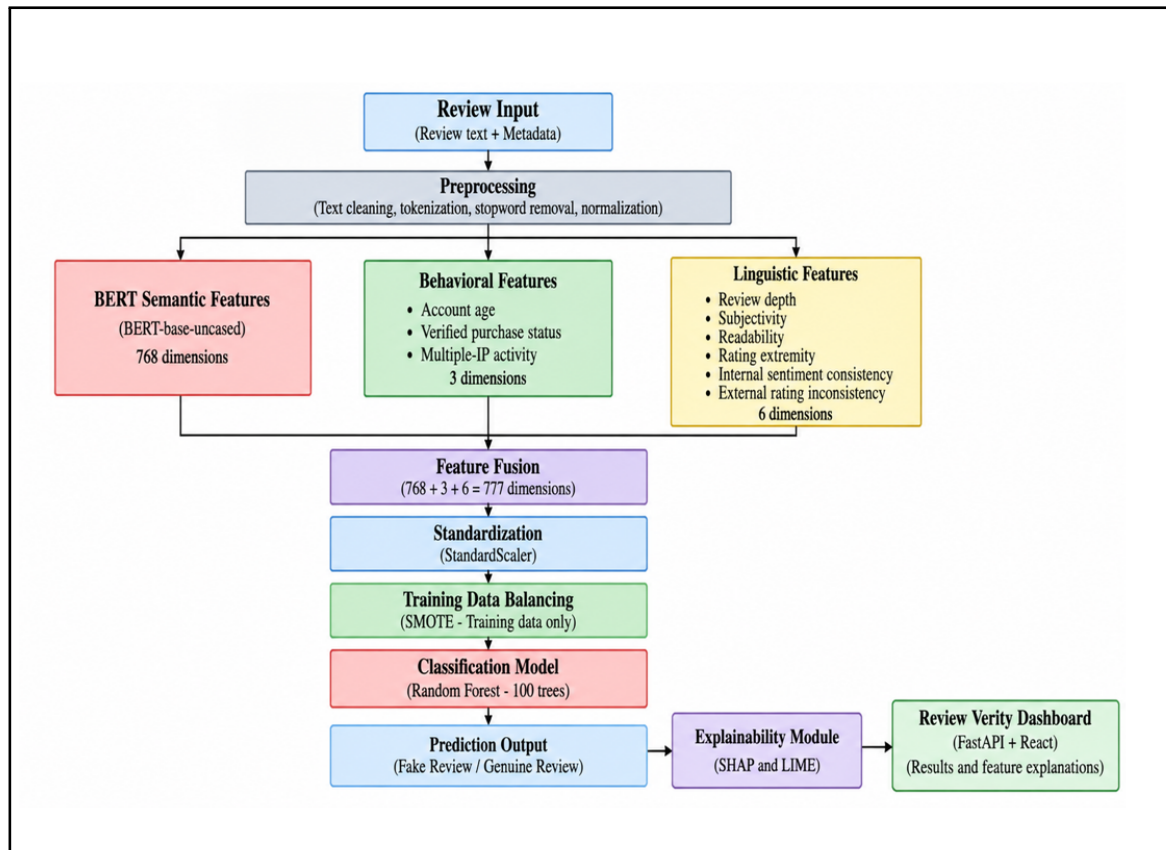


Figure 2. Review verity workflow for multimodal fake review detection and explainability

3.2 Review Preprocessing and Multimodal Feature Extraction

Each review is initially normalized to obtain an identical representation for further processing. Preprocessing takes care of the problem of missing data, standardizes the reviews, and splits it into its textual component, rating information, and reviewer features. Representation of the processed review is done using three different feature streams that include context semantic features, behavior features, and linguistic features, as shown in Figure 2. Software used for preprocessing, feature extraction, classification, and explainability is presented in Table 1.

3.2.1 Contextual Semantic Representation

Textual review is represented using the BERT-base-uncased pretrained model. Tokenization is done with maximum sequence length set at 128 tokens where both truncation and padding are utilized to ensure a consistent size of the input. The BERT layer parameters are kept static while extracting features, and [CLS] token representation is chosen as the semantic representation.

$$h_{\text{BERT}} = \text{BERT}_{\theta} T_{\text{CLS}}, h_{\text{BERT}} \in \mathbb{R}^{768} \quad (1)$$

where T denotes the preprocessed review text and h_{BERT} represents its 768-dimensional contextual embedding. This representation contains semantic and contextual knowledge that cannot be properly captured using manual engineering of textual statistics.

3.2.2 Behavioral Feature Extraction

The features for the reviewer level are captured by three characteristics, namely account age, verified purchase, and multiple IP behavior. The reviewer's account age captures the time since the creation of the review account, while verified purchase captures the verification of the purchase of the product being reviewed. Multiple IP behaviors capture another characteristic that can be used to detect any abnormal behavior from the reviewer.

3.2.3 Linguistic Feature Engineering

A total of six interpretive features are extracted from the review text and corresponding rating: review depth, subjectivity, readability, extreme rating, consistency of internal sentiment and consistency of external rating. Review depth is measured in terms of number of tokens separated by white spaces:

$$D_r = |\text{split}T| \quad (2)$$

Subjectivity is obtained using TextBlob, readability is determined based on the Flesch readability formula. Rating extremeness is determined in relation to the neutral rating value:

$$E_r = |r - 3| \quad (3)$$

where r denotes the review rating. Internal consistency is derived from the variance of sentence-level polarity scores:

$$V_{\text{int}} = \text{Var}(\{p_i\}_{i=1}^n) \quad (4)$$

where p_i represents the polarity of the i -th sentence. External consistency is measured through the difference between the rating given by the review and the average product rating. These linguistic signatures will be helpful in establishing whether or not there is an extremely

consistent review.

Table 1. Technology stack

Component	Technology
Semantic Representation	BERT (bert-base-uncased)
Text Processing	TextBlob, TextStat
Classification	Random Forest
Class Balancing	SMOTE
Feature Scaling	StandardScaler
Explainability	SHAP, LIME
Backend	FastAPI, Uvicorn
Frontend	React, Vite, Axios, Lucide-React

3.3 Multimodal Feature Fusion and Training Preparation

The three feature streams are then integrated to form an integral feature representation for detecting fake reviews. The 768-dimensional BERT representation is combined with the three behavioral features and six engineered language-based features to form a 777-dimensional fused feature representation:

$$F = [h_{\text{BERT}} \parallel B \parallel L] \in \mathbb{R}^{777} \quad (5)$$

where $B \in \mathbb{R}^3$ denotes the behavioral feature vector and $L \in \mathbb{R}^6$ denotes the linguistic feature vector.

Before classifier training, the merged features are standardized using StandardScaler. These parameters for the feature scaling are estimated solely from the training set and then applied to the test set to ensure consistency in transformation without any input from the test samples.

Since the training partition may contain unequal class distributions, SMOTE technique is used on the normalized training data. The synthetic samples are created in the feature space to better represent the minority class samples, but the original test set stays untouched to avoid any biased testing results.

3.4 Random Forest Classification and Explainable Prediction

Normalized and SMOTE-balanced feature vector representations are then classified through a RF ensemble classifier that consists of 100 decision trees. For training consistency purposes, a fixed `random_state = 42` is used in the classifier. Two setups are explored, both trained with the same setup: the first setup is the baseline model that is made up of BERT and behavioral features, while the second one is the proposed setup that is further enhanced by the addition of the six linguistic features.

Given any particular review, the RF classifier sums the output classes of each decision tree and chooses the one that gets maximum votes. Then, the prediction goes to the explainability layer. Figure 2 presents the entire process from feature fusion up to the final prediction.

SHAP is utilized to analyze the model's behavior by determining the global feature attributions that help in decision making by the classifier. Meanwhile, LIME provides local explanations to the classifier's prediction at a particular instance of review. Components used in this step for the process of implementation and explanations are provided in Table 1. Explanations are provided using the same feature representation used by the classifier.

3.5 System Implementation

The proposed Review Verity approach is realized in the form of a web application, which serves as an interface for interaction between the trained detection algorithm and the review analysis interface. The backend is implemented with the help of FastAPI and Uvicorn frameworks, allowing API-based access to review classification, URL analysis, statistics on models, and cross-platform comparison functionality. The frontend is implemented via React with Vite, using Axios for the connection between the interface and backend services.

During review analysis, the submitted review goes through the same preprocessing pipeline as was performed for the purpose of building the model. The output vector of 777 dimensions obtained from the above step is transformed with the help of the stored training scaler and provided to the trained Random Forest model. The resultant class prediction along with its explanation is provided back to the frontend.

The system design also provides for URL analysis as well as cross-platform comparison functionalities. In the present implementation, such functionalities work as simulated models

and not live scraping services for the reviews on the platforms. The URLs are analyzed using deterministic hashing and representative review information in order to show how the process is supposed to take place. Consequently, these functionalities provide for application-level illustration of the system design.

4. Results and Discussion

4.1 Comparative Model Performance

Table 2. Performance evaluation of the proposed review verity model

Metric	Integrated Model
Accuracy	99.8%
AUC-ROC	1.000
F1-Score	0.996
Precision	≈ 1.00
Recall	≈ 1.00

The proposed Review Verity framework is evaluated using accuracy, AUC-ROC, F1-score, precision, and recall to assess its classification performance and ability to distinguish genuine reviews from fake reviews. The evaluation considers the integrated feature representation comprising 768-dimensional BERT embeddings, three behavioral attributes, and six engineered linguistic features, forming a unified 777-dimensional feature space. The Random Forest classifier is used to generate predictions, while the selected metrics characterize overall classification performance, class separability, and the balance between precision and recall.

As reported in Table 2, the proposed model exhibits strong classification performance across the evaluated metrics. The AUC-ROC indicates effective discrimination between the two review classes, while the F1-score reflects a strong balance between precision and recall.

4.2 Contribution of Linguistic Feature Signatures

The feature composition shown above is indicative of the complementary relationship between the proposed multimodal representation. There are 768 features coming from BERT

model, while account age, verified purchase, and multiple IP addresses contribute three additional reviewer-level features. The rest of the six features include depth, subjectivity, readability, rating extremeness, and consistency.

This combination is significant in terms of feature fusion since the three feature sets have different levels of information abstraction. BERT captures contextual semantics, the behavioral features capture the behavior of the reviewers, while the linguistic signatures capture measurable characteristics of review creation and rating process. The integrated model does not only depend on semantic similarity because classification is performed based on heterogeneous information within a single feature space

4.3 Explainability and Feature-level Interpretation

The integrated representation is further analysis using SHAP and LIME. SHAP is utilized for calculating the importance of the individual features to the Random Forest decision-making, while LIME gives local explanations for the individual predictions. The original implementation notes that SHAP gives directions to the contributions of the integrated representation, while LIME produces local feature importance by perturbed samples.

The explainability layer is particularly relevant to the multimodal design because the classifier combines dense transformer embeddings with explicit behavioral and linguistic attributes. Consequently, the latter features provide a more directly interpretable component of the prediction, allowing suspicious reviewer behavior or atypical linguistic characteristics to be examined alongside the classification outcome. The explanation should therefore be interpreted as feature-level model attribution, rather than as complete semantic interpretability of the BERT embedding dimensions.

4.4 Application-level Evaluation

The model that has been trained is included in the Review Verity program using a FastAPI server and React framework. The REST API provides features for review prediction, URL input, cross-platform comparison, and model statistics. These functions are summarized in Table 3.

The predict endpoint provides direct access to the model pipeline by providing review text, rating, account age, purchase validation, and multi-IP information and returns predictions of the predicted class, confidence score, linguistic features, and XAI results. The URL and

cross-platform functionality are, however, to be viewed only as prototypes because the present implementation is based on deterministic simulation and not real marketplace reviews.

Table 3. Review verity rest API endpoints

Endpoint	Method	Input	Function
/	GET	None	Health check
/predict	POST	Review input	Individual review analysis
/scan_url	POST	Product URL	URL-based analysis
/compare_platforms	POST	Product URL	Cross-platform comparison
/model_stats	GET	None	Model-performance information

4.5 User Interface Evaluation

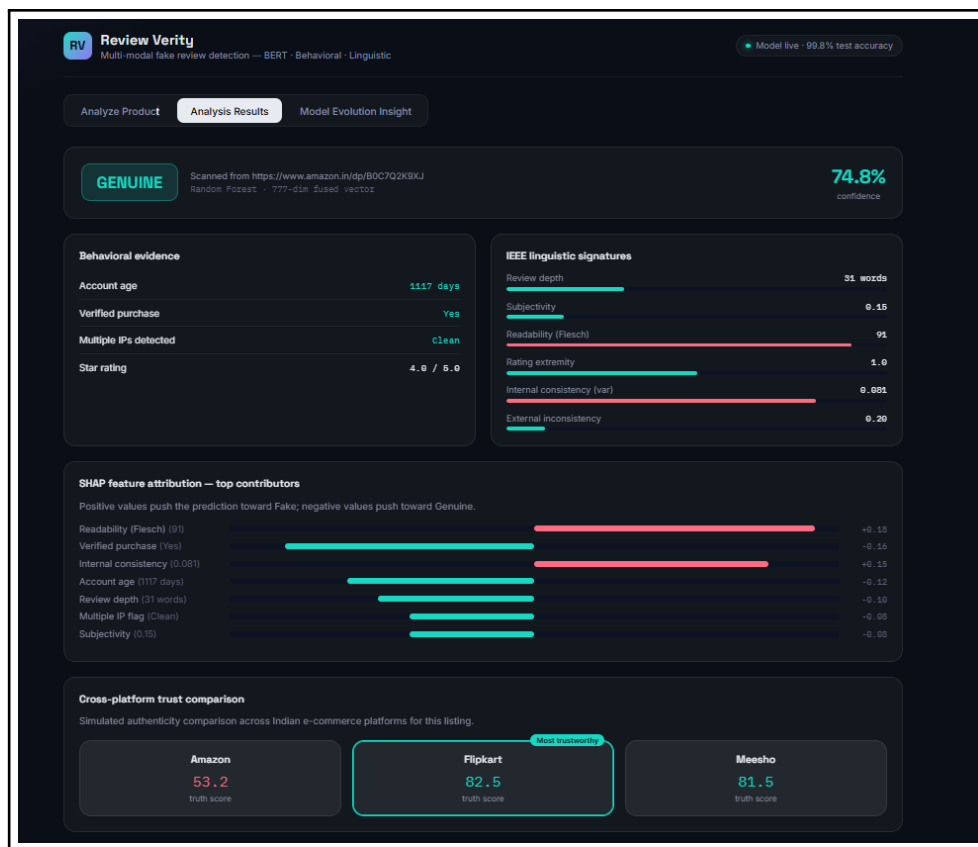


Figure 3. Review verity interface for explainable fake review analysis

The proposed Review Verity design includes a unified interface that incorporates review authenticity, behavioral data, linguistic features, and model reasoning. Figure 3 shows the Analysis Results screen, where the predicted class of the review and its probability are presented along with reviewer-specific information, which includes account age, verified purchase status,

multiple IPs use, and star rating. Linguistic analysis screen includes review depth, subjectivity, readability, star rating extremeness, internal consistency, and external inconsistency.

The interface also incorporates a SHAP feature-attribution panel that visualizes the impact of individual features on the predicted class. The positive and negative directions of attribution are clearly defined, and thus, the impact of linguistic and behavioral features on individual predictions can be analyzed. Moreover, the Cross-platform Trust Comparison panel provides an authentic comparison perspective for the given product listing. Such an integration allows the combination of prediction, visualizations, explainability, and comparative analysis in unified analytical panel.

5. Conclusion and Future Work

The proposed Review Verity framework presents an integrated solution for detecting fake reviews using contextual representations of text, reviewer behavioral features, and linguistically salient markers all together in unified classification pipeline. The use of a multimodal representation allows for complementary evidence about the semantics of text, reviewer behavior, and ratings to be used to assess authenticity. The integration of SHAP and LIME contributes to the easy analysis of decisions made by the model at feature and individual-review level. Additionally, the approach has been realized as a web application which combines the detection process with REST services and interactive analysis interface. The future research will involve the use of large amounts of data reviews along with their diversity, gathering of authentic review history data, multilingual transformer representation, and real-time collection of reviews. The research of reviewer-product relationship, temporal activities, and fine-tuning of end-to-end transformers will help enhance robustness to manipulation strategies.

References

- [1] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In Proceedings of the conference of the North American chapter of the association for computational linguistics: Human Language Technologies, 2019, vol. 1: 4171-4186.
- [2] Luca, Michael, and Georgios Zervas. "Fake it Till You Make it: Reputation, Competition, and Yelp Review Fraud." *Management Science*, 2016, vol. 62, no. 12: 3412-3427.

- [3] Lundberg, Scott M., and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems*, 2017, vol. 30: 1-10.
- [4] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why Should I Trust you? Explaining the Predictions of Any Classifier." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, 135-144.
- [5] Abedin, Ehsan, Antonette Mendoza, Pouria Akbarighatar, and Shanika Karunasekera. "Predicting Credibility of Online Reviews: an Integrated Approach." *IEEE Access*, 2024, vol. 12: 49050-49061.
- [6] Mohawesh, Rami, Shuxiang Xu, Matthew Springer, Yaser Jararweh, Muna Al-Hawawreh, and Sumbal Maqsood. "An Explainable Ensemble of Multi-View Deep Learning Model for Fake Review Detection." *Journal of King Saud University Computer and Information Sciences*, 2023, vol. 35, no. 8: 101644.
- [7] Wang, Qiang, Wen Zhang, Jian Li, and Zhenzhong Ma. "Complements or Confounders? A Study of Effects of Target and Non-Target Features on Online Fraudulent Reviewer Detection." *Journal of Business Research*, 2023, vol. 167: 114200.
- [8] Mohawesh, Rami, Haythem Bany Salameh, Yaser Jararweh, Mohannad Alkhalaileh, and Sumbal Maqsood. "Fake Review Detection using Transformer-Based Enhanced LSTM and Roberta." *International Journal of Cognitive Computing in Engineering*, 2024, vol. 5: 250-258.
- [9] Das, Ronnie, Wasim Ahmed, Kshitij Sharma, Mariann Hardey, Yogesh K. Dwivedi, Ziqi Zhang, Chrysostomos Apostolidis, and Raffaele Filieri. "Towards The Development of an Explainable E-Commerce Fake Review Index: An Attribute Analytics Approach." *European Journal of Operational Research*, 2024, vol. 317, no. 2: 382-400.
- [10] Liu, Jinhao, Pei Quan, and Wen Zhang. "A Study on Fake Review Detection Based on Roberta and Behavioral Features." *Procedia Computer Science*, 2024, vol. 242: 1323-1330.
- [11] Kaggle, "Amazon Product Reviews Dataset," <https://www.kaggle.com/datasets/yasserh/amazon-product-reviews-dataset>