Received: 22 December 2025 | Revised: 9 February 2026 | Accepted: 21 April 2026 | Published online: 13 May 2026

RESEARCH ARTICLE

FinTech and Sustainable Innovation
2026, Vol. 00(00) 1–8
DOI: 10.47852/bonviewFSI62028859

BON VIEW PUBLISHING

Metadata-Enhanced Hybrid Fusion Architecture: Commercial Fake Reviews Detection Model Using Transformer Embeddings

Hisham AbouGrad[1,*] and Fiza Riaz[2]

[1] *Department of Computer Science and Digital Technologies, University of East London, UK*

Abstract: Commercial fake reviews have become an important issue for online businesses and e-commerce platforms, as they affect customer choices and decisions and present fake product quality. To enhance the identification of misleading commercial reviews, this study investigates a hybrid approach that combines machine learning and deep learning algorithms. Amazon fake reviews and Yelp reviews are two distinct datasets that have been utilized. Text cleaning, information extraction, sentiment analysis, and a special sentiment rating mismatch feature are all part of the comprehensive preparation pipeline used in this research study. While the DistilBERT technique is used to extract deeper contextual meaning from the text, traditional machine learning models, such as Random Forest, naive Bayes, logistic regression, and support vector machine, are trained using term frequency–inverse document frequency characteristics. Also, a hybrid fusion model has been developed by integrating DistilBERT embeddings with metadata variables, such as sentiment, rating, and text length. By obtaining 93% accuracy on Amazon and 91% on Yelp datasets, the study's results found that DistilBERT outperforms conventional models. The emotion rating mismatch technique also assists considerably by recognizing behavioral anomalies commonly found in fraudulent reviews. Overall, the research findings showed that combining semantic understanding with behavioral indications gives a more accurate and trustworthy architecture for detecting fraudulent commercial fake reviews in real-time online business contexts.

Keywords: hybrid fusion model, feature-driven intelligent system, commercial fake reviews detection, DistilBERT embeddings, sentiment mismatch

1. Introduction

Customer online reviews have a significant impact on consumer decisions across e-commerce platforms and financial technology (FinTech) systems, which influence perceptions of product reliability, quality, and general trust [1]. The quantity and significance of user-generated reviews have grown dramatically as online retailers, such as Amazon, continue to expand. Recent studies, however, indicated that such FinTech systems and platforms are becoming more susceptible to manipulated or false reviews intended to affect customer behavior and product ratings. According to several studies, fake reviews can be created by automated bots, paid persons, or complex language models, which makes them challenging to identify [2].

Further, public safety, market fairness, and consumer trust are all seriously threatened by the increasing sophistication of false commercial reviews. Manipulated reviews can trick customers into buying dangerous or useless products in industries such as dietary supplements and medical supplies by spreading false information on social media and in online marketplaces [3]. Also, by deceptively promoting poor products or harming the reputation of reputable companies, fake commercial reviews skew competition.

Traditional detection methods based on rule matching or lexical patterns struggle to keep up with increasing deception strategies, as fraudsters change swiftly to escape detection. To reliably identify fraudulent content and maintain review integrity, researchers contend that scalable, automated, and intelligent systems, especially those utilizing machine learning (ML) and deep learning (DL), are crucial [4].

This research study aims to utilize and compare ML and DL models for identifying fraudulent business evaluations across two distinct datasets to address such a problem. The efficiency of five popular models, which are Random Forest, naive Bayes, logistic regression, support vector machine (SVM), and the DistilBERT model, in identifying the specific patterns of deception in review data was assessed. A hybrid model has been formulated by combining the contextual embeddings from DistilBERT with selected by importance metadata features, such as rating, sentiment, and text length, to improve the detection process, in response to earlier research showing the value of linguistic and behavioral

***Corresponding author:** Hisham AbouGrad, Department of Computer Science and Digital Technologies, University of East London, UK. Email: h.abougrad@uel.ac.uk

indicators for identifying deceptive content [5]. The research findings are intended to contribute to the development of a robust, scalable architecture that strengthens trust in online marketplaces and FinTech systems and protects consumers from misleading and false digital information. Such intelligent fake review detection systems are especially supportive of FinTech-enabled digital marketplaces, where trust, transparency, and sustainable commercial platform growth depend on reliable automated decision-making. This hybrid approach enables the model to utilize both linguistic features from the text and behavioral insights from the metadata by providing a thorough understanding of review authenticity.

2. Literature Review

In today's digital marketplace, the validity of online reviews plays an important role in determining consumer trust, brand reputation, and purchase decisions [6]. The exponential growth of user-generated information on websites and apps, such as Amazon, Yelp, and Naver, has resulted in misleading methods aimed at altering customer perception. To address this issue, a growing corpus of research has concentrated on finding linguistic, behavioral, and computational indicators of fake reviews and developing intelligent detection models.

The discussion starts with the study of Kim et al. [7], in which they used the interpersonal deception theory (IDT) to analyze the differences in writing styles and intentions between fake and authentic reviewers. This research study applied more than a million verified and unverified movie reviews from Naver.com and discovered that early in a product's lifecycle, fraudulent reviewers usually publish overly enthusiastic and extremely informative reviews. The study by Kim et al. [7] has provided one of the first extensive evaluations connecting linguistic deception to user intention by combining textual and behavioral indicators, including timing and sentiment polarity. The theoretical groundwork for further computational research was established by Kim et al.'s [7] study and its IDT. Building on this theoretical foundation, further understanding of deception detection has been built through computer experiments. They tried many supervised learning models, including SVM, naive Bayes, logistic regression, and Random Forest, to distinguish between genuine and fraudulent reviews using Amazon's extensive review datasets. Their findings confirmed that review length has a direct impact on detection performance and showed that SVM outperformed other models, obtaining up to 93% accuracy in sentiment analysis and 88% accuracy in fake review classification.

While shorter, sentimental remarks were more likely to be dishonest, longer reviews tended to seem more sincere [8]. The assertion that message length and grammatical complexity are reliable markers of review authenticity was empirically supported by this study. Further, expanded into linguistic and psychological aspects by combining transformer architectures such as the BERT model with psycholinguistic cues from linguistic inquiry and word count (LIWC) analysis. Indeed, fake reviewers frequently exhibit abnormal pronoun usage, heightened emotions, and cognitive strain, which reflect the stress patterns of dishonest communications [9]. This work demonstrated how deceit may be algorithmically represented while staying rooted in human linguistic behavior by fusing psychological understanding with deep contextual embeddings. Likewise, structural topic modeling is used to examine the behavioral and persuasive effects of phony reviews. The results showed that even if they are frequently well written, false reviews lose their persuasiveness with time and, once discovered, can potentially damage a brand's reputation [10]. This study made a crucial connection between detection research and consumer psychology by highlighting the possibility that phony evaluations could momentarily divert consumers' attention but eventually undermine long-term trust. To set a performance benchmark for feature-driven detection intelligent systems in the direction of traditional machine learning techniques. The Random Forest model obtained an accuracy of 79% using Amazon datasets and linguistic and behavioral variables, confirming the potential of conventional algorithms, while also highlighting their shortcomings in managing contextual and semantic depth [11].

Further, a complementary perspective has been presented as several traditional machine learning methods, such as Random Forest, logistic regression, SVM, and XGBoost, are used to identify fake Amazon commercial evaluations. According to their research, ensemble techniques, such as Random Forest and XGBoost, perform better than linear models because they are better at capturing nonlinear relationships within textual characteristics. The study's emphasis on computational efficiency, feature selection, and classification accuracy highlights the ongoing significance of traditional machine learning techniques, particularly in contexts where DL models are not practical because of resource constraints [12]. The author proposed a semi-supervised adversarial learning technique to enhance model generalization by generating deceptive-like data in order to get over these limitations. This dual discrimination design was one of the first uses of generative adversarial networks (GANs) in textual deception detection, and it effectively improved classification accuracy in situations with limited labeled data [13].

Furthermore, a study by Mohawesh et al. [14] proved that DL improves by combining bidirectional long short-term memory (LSTM) layers with RoBERTa embeddings. By capturing both contextual and sequential dependencies, this hybrid approach revealed an over 93% detection accuracy and found that the language signals contributed the most to classification using SHapley Additive exPlanations (SHAP) interpretability.

Moreover, the model's performance has been improved by fine-tuning the DeBERTa model to reach up to 98% accuracy and maintain stability even against hostile text changes using an optimization strategy inspired by nature, which led to revealing the cutting edge of transformer optimization in review authenticity research [15].

While the preceding studies emphasized model sophistication, the study of Phukon et al. [16] proposed a logic-aware method for deception detection. Using graph convolutional networks to link product attributes, such as "battery life" and "design," with associated attitude polarity, the study was able to identify contradictions within reviews, such as when a review praises a product overall but dislikes each component. This approach emphasizes the significance of semantic coherence as a differentiator between authentic and fake reviews. The growing realism of AI-generated content inspired the startling discoveries in Meng et al. [17], which proved that it is difficult for both humans and detection systems to distinguish between authentic and AI-generated reviews, especially when negative sentiment is present. Also, large language models (LLMs) are a new class of dishonest agents that may produce contextually coherent but false opinions, rendering conventional detection methods outdated if they are not combined with more robust verification techniques. Recognizing that algorithmic detection alone cannot ensure review authenticity, the author presented a blockchain-based governance paradigm in which people prove their identities through secure authentication while remaining anonymous. This hybrid paradigm offers a

useful framework for combining technical AI-based detection with digital trust controls, which prevent the formation of large-scale phony accounts and automated spam. The paper's focus on integrating human accountability with computer intelligence is an innovative approach to long-term online integrity [18].

3. The Methodology and Proposed Model

This research project looks to develop and evaluate an effective model to detect commercial fraud and fake reviews by combining classic ML models with advanced DL architectures. The main objective is to use different datasets, especially the Yelp and Amazon reviews databases, to reliably identify commercial fake reviews. To accomplish this, a hybrid fusion model is used, as shown in Figure 1, which combines the advantages of both conventional machine learning methods and cutting-edge DL models. It also incorporates sentiment analysis and metadata to improve classification accuracy.

Figure 1 explains each stage from data preprocessing to model training and shows how the combination of DistilBERT embeddings and metadata characteristics leads to a comprehensive architecture for detecting fake reviews.

3.1. Datasets and preprocessing

To assess the performance of the proposed model, two distinct datasets are used, which are publicly available, customer-generated commercial reviews, and no human subjects were involved. The first dataset is Amazon fake reviews, which includes Amazon product reviews [19]. Either Computer Generated or Original Reviews labels are applied to the reviews. The Amazon data aids in determining the range of bogus and genuine reviews across various product categories. The second is the Yelp reviews dataset, which is made up of Yelp reviews, where each category is either a fake as −1 or a genuine as 1 [20]. The Yelp data is frequently used to assess models in the context of company reviews by offering more information on identifying fraudulent reviews on a platform that focuses on services. Both datasets consist of customer-generated commercial reviews, reflecting authentic consumer opinions and deceptive review behavior on real-world platforms.

Raw text is preprocessed to train machine learning models. The Amazon and Yelp datasets are subjected to the following steps [21]:

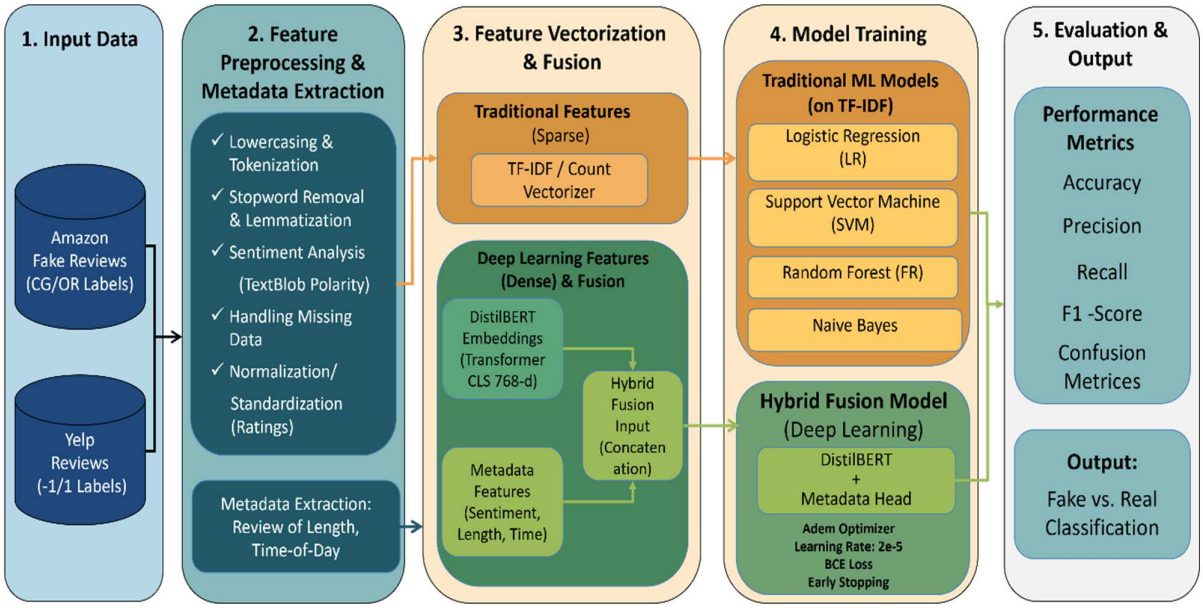
1) Lowercasing: Used to convert text to lowercase to eliminate case differences;
2) Stopword Removal: Used to remove common, non-informative words to reduce noise, for example, "the," "is";
3) Tokenization: Splits the reviews into individual words/tokens for easier processing;
4) Lemmatization: Reduces words to their root form, for example, "running" becomes "run";
5) Sentiment Analysis: To calculate sentiment polarity scores using TextBlob and identify sentiment and rating mismatches, which indicate potential fake reviews;
6) Handling Missing Data: This removes or imputes missing reviews or ratings to ensure data integrity;
7) Normalization/Standardization: This scales numeric features, such as ratings, to improve the model performance.

These steps ensure clean, consistent, and relevant data for model training.

3.2. Feature engineering methods

For machine learning models to use raw data as inputs, feature engineering is important. The Amazon and Yelp reviews datasets are used in this research study with a variety of feature engineering techniques. Term frequency–inverse document frequency (TF–IDF) and Count Vectorizer are used to transform text into numerical features for conventional ML models, such as Random Forest, naive Bayes, logistic regression, and SVM [22]. These features capture the significance of phrases and their frequency. Word embeddings capture semantic and contextual links by representing words as dense vectors for DL models such as DistilBERT. Text-based features are combined with

Figure 1
Architecture of proposed hybrid fusion model



metadata features, including review length, sentiment scores from TextBlob, and time of day features. These characteristics improve the model's capacity to identify fraudulent reviews by offering insights into the emotional tone, specifics, and posting patterns of the review.

Based on the study experiments, the proposed model with its chosen algorithms can produce a thorough representation of reviews by integrating textual and metadata elements, which increases its precision in identifying fraudulent content. Algorithm 1 breaks down the Multi-Model Fake Review Detection into a step-by-step framework to explain how the proposed model operates. Preprocessing the review data, extracting significant features such as a sentiment rating mismatch indicator, and training both conventional machine learning models and the sophisticated hybrid fusion model are all crucial processes in this model.

Algorithm 1: Hybrid Model for Fake Reviews Detection

Input: Dataset D containing Reviews (X), Ratings (R), and Labels (Y)
Output: Performance evaluation of 5 classification models
Initialize Baseline Models M_base = {LR, NB, SVM, RF}
Initialize Deep Learning Components (DistilBERT Tokenizer T and Model B)
Phase 1: Preprocessing & Feature Engineering
for each review sample (x_i, r_i) in D **do**
    x'_i ← Preprocess(x_i) {Lowercase, Lemmatize, Remove Stopwords}
s_i ← Compute Sentiment(x'_i) using TextBlob
m_i ← 0 {Initialize Mismatch Flag}
if (r_i ≥ 4 and s_i < 0) or (r_i ≤ 2 and s_i > 0) **then**
        m_i ← 1 {Behavioral Inconsistency Detected}
else
v_meta ← [r_i, s_i, m_i, WordCount(x'_i)]
end
Phase 2: Baseline Model Evaluation
V_tfidf ← Extract TF-IDF features from all cleaned texts X'
for each model m in M_base do
Train m using (V_tfidf, Y)
Record Accuracy, Precision, Recall, and F1-Score
end
Phase 3: Hybrid Fusion Model Training
(T_ids, T_mask) ← T. tokenize (X', max_len = 128)
Construct Hybrid Network Architecture:
Text Branch: h_text ← B (T_ids, T_mask) -> Dense (128) -> Dropout
Meta Branch: h_meta ← Dense (v_meta, 32)
Fusion: h_fused ← Concatenate (h_text, h_meta)
Output: y_hat ← Sigmoid (Dense (h_fused, 64))
Train Hybrid Model using Adam Optimizer (2e-5) and Binary Cross-Entropy Loss
Return Compare results of M_base Vs. Hybrid Model

The research study systemically evaluates the performance of all five classifiers by adhering to this structured architecture, which ensures a comprehensive and equitable assessment of their efficiency.

3.3. Model training

During the model training phase, the preprocessed dataset is used to train both the hybrid fusion model and conventional machine learning models. Sentiment polarity, review rating, text length, and the sentiment rating mismatch indicator are the elements derived from the data that are used to differentiate between authentic and fraudulent reviews. To provide a fair comparison and assess the efficacy of the advanced hybrid fusion model, traditional machine learning models are used as baseline models under comparable experimental settings.

TF–IDF characteristics, which capture the significance and frequency of words in each review, are used to represent the reviews during the first training phase. The feature set also includes metadata features, such as emotion scores, review length, and word count, in addition to text-based features. Each of the conventional machine learning models is trained using such a feature matrix, which is formulated by combining these features.

Hyperparameter adjustment is done for each baseline model to maximize its performance. For instance, to prevent overfitting and improve generalization, the regularization strength (C) in logistic regression is changed. Numerous metrics, including accuracy, F1-score, precision, and recall, are used to assess the performance of naive Bayes models, which are trained using a probabilistic method. Grid search is used in SVM to optimize the kernel function and the penalty parameter (C), allowing the model to identify the optimal decision boundary to differentiate between genuine and fraudulent reviews. To increase classification accuracy for Random Forest, hyperparameters, such as the number of trees in the forest, the minimum samples needed for splitting, and the maximum depth of trees, are adjusted. Afterward, a cross-validation is used to assess each model to ensure that it is not overfitting to the training set and can effectively generalize to new data. Standard classification metrics, including accuracy, F1-score, precision, and recall, are used to train and assess the models. These indicators aid in evaluating how well each baseline model performs in differentiating between authentic and fraudulent reviews. These models' outcomes serve as a basis for comparison with such a sophisticated hybrid model.

The DistilBERT tokenizer is used to tokenize the preprocessed reviews for the proposed model by transforming the text into subwords, known as "tokens," which the DistilBERT model can interpret [21]. A transformer-based architecture called DistilBERT is renowned for its capacity to extract profound contextual and semantic information from text. Contextual embeddings, which function as each review's textual representation, are obtained by running the tokenized reviews through the DistilBERT model. Emotional tone, hyperbole, and subtle indicators of dishonesty are just a few of the intricate patterns in the language that these embeddings record.

Apart from the text embeddings, metadata features are also extracted, including sentiment scores, review ratings, and word count. A different dense layer consisting of 32 units is used to process these metadata attributes. The textual embeddings from the DistilBERT model are concatenated with the output from the metadata layer to produce a composite feature vector. A dropout layer of 0.3 to avoid overfitting and a final sigmoid activation function to have a binary classification result, fake or real, come after this fused representation is run through more dense layers in the model. The Adam optimizer with a learning rate of 2e-5 and binary cross-entropy loss is used to train the hybrid fusion model, which is perfect for binary classification problems. The model is trained for a set number of epochs, usually two, using a batch size of 16n. To track the validation loss and avoid overfitting, early stopping is used during training. The model is assessed using the same metrics as conventional machine learning models, such as accuracy and confusion matrices for misclassification rate analysis. The training procedure of the proposed model makes

use of the DistilBERT model's ability to capture complex semantic associations in the text, and the metadata features give more context, which enhances the model's capacity to identify fraudulent reviews. Deep contextual embeddings and behavioral insights from the metadata are combined to provide the model a thorough understanding of review authenticity. Also, hyperparameter adjustment is essential for the hybrid fusion model training. To reach the best results, important hyperparameters, such as batch size, learning rate, and number of units in the dense layers, are adjusted. To find the ideal values for these hyperparameters and make sure the model performs effectively when applied to new data, grid search or random search can be utilized.

In conclusion, the preprocessed and feature-engineered data are used to train both the proposed model and conventional machine learning models. The hybrid fusion model, which combines the strength of DistilBERT embeddings and metadata characteristics, seeks to provide better performance in identifying fraudulent reviews, while the standard models function as baseline classifiers. The models are then assessed using key performance indicators (KPIs) to ascertain their efficiency and determine the best model.

3.4. Model evaluation

As Table 1 exhibits, important measures, such as accuracy, precision, recall, and F1 score, were used to evaluate the fake review detection models' performance on the Yelp and Amazon datasets. These KPI measurements offer a thorough comprehension of the pros and cons of each model across various datasets.

Logistic regression performed well on a structured dataset by reaching 89% accuracy on Amazon. Nevertheless, it only attained 77% accuracy on Yelp, suggesting that although the model does a good job of generalizing to Amazon data, it finds it difficult to adjust to the varied and informal phrases used in Yelp reviews. This performance disparity shows that there is an area of improvement in handling evaluations with more varied language elements. Conversely, Random Forest proved stable performance across both datasets with 86% accuracy on Amazon and 85% accuracy on Yelp. Its ability to sustain comparable performance across the two datasets implies that ensemble techniques, such as Random Forest, can identify significant patterns in the data, even when challenged with a variety of writing styles.

SVM showed the highest accuracy on Amazon at 90%, making it one of the top performers for the more structured dataset. However, it struggled with Yelp reviews, dropping to a lower 77% accuracy. This highlights the SVM's limitation in handling noisy and informal text data, which is a common characteristic of Yelp reviews. Despite this, SVM performed well in terms of precision with 90% accuracy on Amazon, but the recall on Yelp was significantly lower with 57% accuracy, which indicates that it missed a considerable number of fake reviews. The naive Bayes algorithm performed similarly to Random Forest on the Amazon dataset by reaching 86% accuracy. However, its performance on Yelp was less robust, with a dramatic drop in recall to 53%. This suggests that while naive Bayes is effective in identifying fake reviews on structured datasets such as Amazon, it struggles to detect fake reviews in datasets with diverse review styles and more ambiguous linguistic cues.

The DL DistilBERT model has reached 93% accuracy on Amazon and 91% accuracy on Yelp, outperforming other models. This proves how transformer-based models, such as the DistilBERT model, can effectively capture the contextual and semantic complexities seen in review language. The model's robustness in handling various writing styles is demonstrated by its ability to perform consistently across both datasets. Also, the integration of the sentiment rating mismatch feature as a crucial model input substantially boosted the model's capacity to identify bogus reviews. It has identified differences between review sentiment and ratings, which are generally symptoms of behavioral anomalies in phony commercial reviews.

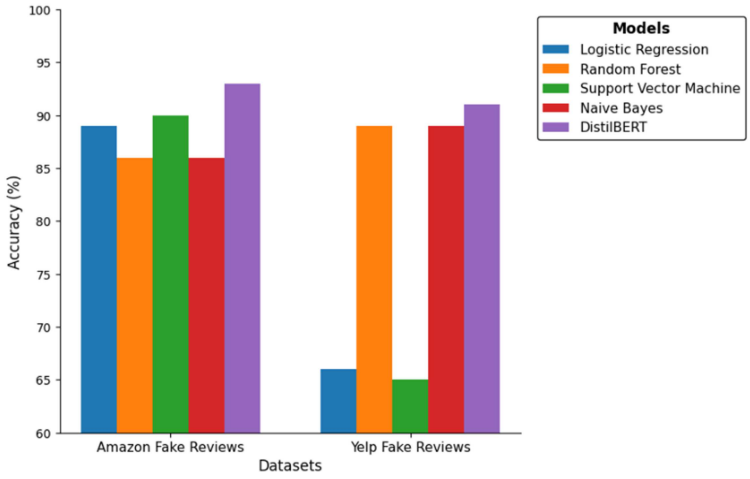
The evaluation of the proposed model provides a comprehensive overview of the model's performance across both datasets by utilizing a variety of evaluation criteria, such as accuracy, F1 score, precision, and recall. These metrics revealed information about the models' capacity to reduce false positives and false negatives, in addition to evaluating how well they can distinguish between authentic and fake reviews. As demonstrated in the results, a hybrid approach that blends DL with metadata features such as emotion and rating mismatch proved highly effective, especially in recognizing subtle patterns of fraud.

The evaluation of the study findings reported in Table 1 shows a comparison of the performance of the classic machine learning models versus the advanced DL-based DistilBERT model. Traditional models are a good starting point, but their effectiveness varies greatly among datasets. However, the DistilBERT model consistently produces the highest and most balanced results, proving its ability in recognizing complicated language patterns and its effectiveness in detecting bogus reviews. These findings underline the need to use advanced DL algorithms for accurate, scalable false detection. On the Amazon fake reviews dataset, SVM consistently outperformed other traditional

Table 1
Evaluation metrics for fake reviews detection

Models	Datasets	Accuracy (%)	F1-score (%)	Precision (%)	Recall (%)
Logistic Regression	Amazon fake reviews	89	89	89	89
	Yelp fake reviews	66	67	53	53
Random Forest	Amazon fake reviews	86	86	86	86
	Yelp fake reviews	89	90	86	85
Support Vector Machine	Amazon fake reviews	90	90	90	90
	Yelp fake reviews	65	57	62	57
Naive Bayes	Amazon fake reviews	86	86	87	86
	Yelp fake reviews	89	85	81	90
DistilBERT	Amazon fake reviews	93	93	93	93
	Yelp fake reviews	91	91	91	91

Figure 2
Model performance comparison across different models



classifiers with 90% accuracy, while logistic regression achieved 89% accuracy with balanced precision and recall.

Naive Bayes and Random Forest have been followed closely at 86%, revealing that Amazon evaluations contain distinct textual patterns that classical TF–IDF-based models can capture efficiently. The DistilBERT model has rich contextual knowledge and capacity to capture semantic variations in user-generated text, which allowed it to outperform all models with 93% accuracy and similarly high F1-scores. Such results show that transformer-based models can detect subtle, misleading patterns, such as emotional exaggeration or unusual language, which are hard for standard models to interpret using fixed statistical variables. Conversely, because of its varied writing styles, colloquial language, and less organized narratives, the Yelp dataset showed a difficult classification environment (Figure 2).

As a result, both SVM and logistic regression saw a sharp decline in performance, attaining only 66% and 65% accuracy, respectively, underscoring their shortcomings in managing extremely varied linguistic patterns. However, Random Forest and naive Bayes outperformed the others, both achieving 89% accuracy, which indicates that their probabilistic and ensemble processes are more suited to Yelp's nonlinear text features (Figure 2). DistilBERT once again gave the highest results, reaching 91% accuracy across all evaluation metrics, proving its robustness across all selected datasets.

The sentiment rating mismatch feature, which effectively captured behavioral abnormalities frequently found in fake reviews, such as excessively positive phrasing paired with low ratings or vice versa, was a noteworthy addition to both datasets. By offering clues beyond textual content, this supplementary feature reinforced classic models and improved the interpretability of results. Overall, the comparative results prove that transformer-based models, such as the DistilBERT model, consistently provide better performance and generalization, making them the most dependable method for fake review detection across real-world platforms, even though classical models can still be competitive under specific dataset conditions.

4. Conclusion

This research study explored how different machine learning and DL models perform in detecting false reviews across two prominent datasets. Transformer-based models, particularly DistilBERT, which regularly achieved the best accuracy on both Amazon and Yelp reviews datasets, clearly have an edge, according to the data. Traditional ML models, such as naive Bayes, logistic regression, Random Forest, and SVM, performed well on the Amazon dataset, where the writing style is more structured, but their accuracy declined considerably on the more diversified and informal Yelp reviews. The hybrid fusion model has combined DistilBERT with metadata characteristics, such as sentiment, rating, and text length, which outperformed in both scenarios and was effective in detecting subtle signals of deception.

A crucial aspect of this study is the use of the sentiment rating mismatch feature, which assisted in detecting reviews where the emotional tone did not match the assigned rating, a common flaw in phony reviews. These extra behavioral cues improved the overall accuracy of the intelligent detection system and increased its dependability.

In conclusion, the research study proved that merging classical linguistic analysis with modern DL and metadata elements offers a robust and scalable approach for spotting commercial fake reviews. By demonstrating how validation hybrid fusion architectures may integrate semantic intelligence with behavioral metadata to provide reliable, scalable, and sustainable digital review and FinTech systems, this study contributes to the corpus of knowledge beyond simply increasing detection accuracy.

5. Future Work

Future research can improve upon the findings of this research study by expanding the training corpus to include bilingual, cross-platform, and domain-specific review datasets, which enables the model to generalize more effectively across varied online settings. Although the hybrid fusion model performed well, further improvements could be made by incorporating more sophisticated transformer architectures, such as RoBERTa, DeBERTa, or LLM-based encoders, as well as investigating dynamic fusion approaches that assign adaptive weights to metadata cues such as reviewer behavior, posting frequency, and sentiment rating inconsistencies. Also, future research work should study real-time deployment methodologies using model compression, distillation, and edge-optimized inference to enable large-scale e-commerce platforms. Incorporating explainable AI

algorithms is also vital to promote transparency by assisting stakeholders in understanding whether linguistic clues, emotional patterns, or behavioral anomalies contribute largely to fake review identification. Lastly, working with industry partners can make it easier to obtain richer behavioral logs and actual fraud patterns, which will increase the robustness of the suggested intelligent detection model along with its generalizability and practicality.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com/datasets/mexwell/fake-reviews-dataset and https://www.kaggle.com/datasets/abidmeeraj/yelp-labelled-dataset.

Author Contribution Statement

Hisham AbouGrad: Conceptualization, Validation, Resources, Writing – review & editing, Visualization, Supervision, Project administration. **Fiza Riaz:** Conceptualization, Validation, Software, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Visualization.

References

[1] Wu, Y., Ngai, E. W. T., Wu, P., & Wu, C. (2020). Fake online reviews: Literature review, synthesis, and directions for future research. *Decision Support Systems*, *132*, 113280. https://doi.org/10.1016/j.dss.2020.113280

[2] Sree, T. R., & Tripathi, R. (2023). Fake review detection using evidential classifier. In *2023 Second International Conference on Advances in Computational Intelligence and Communication*, 1–5. https://doi.org/10.1109/ICACIC59454.2023.10435343

[3] Kannal, N. M., Asmathunnisa, N., & Kallimani, J. S. (2024). Impacts of fake reviews on dietary supplements and healthcare products in social media. In *2024 15th International Conference on Computing Communication and Networking Technologies*, 1–4. https://doi.org/10.1109/ICCCNT61001.2024.10725718

[4] Sharma, R., Sharma, V., Vashishth, T. K., Shashi, S., Pandey, A., & Chaudhary, S. (2025). Revealing the reliability of Amazon products via innovative fake review detection using machine learning. In *2025 6th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks*, 217–221. https://doi.org/10.1109/ICICV64824.2025.11086089

[5] Cardoso, E. F., Silva, R. M., & Almeida, T. A. (2018). Towards automatic filtering of fake reviews. *Neurocomputing*, *309*, 106–116. https://doi.org/10.1016/j.neucom.2018.04.074

[6] Salminen, J., Kandpal, C., Kamel, A. M., Jung, S.-G., & Jansen, B. J. (2022). Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services*, *64*, 102771. https://doi.org/10.1016/j.jretconser.2021.102771

[7] Kim, J. M., Park, K. K.-C., Mariani, M., & Wamba, S. F. (2024). Investigating reviewers' intentions to post fake vs. authentic reviews based on behavioral linguistic features. *Technological Forecasting and Social Change*, *198*, 122971. https://doi.org/10.1016/j.techfore.2023.122971

[8] Tabany, M., & Gueffal, M. (2024). Sentiment analysis and fake Amazon reviews classification using SVM supervised machine learning model. *Journal of Advances in Information Technology*, *15*(1), 49–58. https://doi.org/10.12720/jait.15.1.49.-58

[9] Salminen, J., Mustak, M., Jung, S.-G., Makkonen, H., & Jansen, B. J. (2025). Decoding deception in the online marketplace: Enhancing fake review detection with psycholinguistics and transformer models. *Journal of Marketing Analytics*. Advance online publication. https://doi.org/10.1057/s41270-025-00393-8

[10] Kumar, R., Mukherjee, S., & Rana, N. P. (2024). Exploring latent characteristics of fake reviews and their intermediary role in persuading buying decisions. *Information Systems Frontiers*, *26*(3), 1091–1108. https://doi.org/10.1007/s10796-023-10401-w

[11] Mane, K., Dongre, S., & Madankar, M. (2025). Fake review detection using random forest classifier. In *2025 IEEE International Students' Conference on Electrical, Electronics and Computer Science*, 1–6. https://doi.org/10.1109/SCEECS64059.2025.10940605

[12] Akshara, S., Shiva, S., Kubireddy, S., Arun, T., & Kanthety, V. L. (2023). A small comparative study of machine learning algorithms in the detection of fake reviews of Amazon products. In *2023 6th International Conference on Contemporary Computing and Informatics*, 2258–2263. https://doi.org/10.1109/IC3I59117.2023.10398096

[13] Aghakhani, H., Machiry, A., Nilizadeh, S., Kruegel, C., & Vigna, G. (2018). Detecting deceptive reviews using generative adversarial networks. In *2018 IEEE Security and Privacy Workshops* (pp. 89–95). https://doi.org/10.1109/SPW.2018.00022

[14] Mohawesh, R., Salameh, H. B., Jararweh, Y., Alkhalaileh, M., & Maqsood, S. (2024). Fake review detection using transformer-based enhanced LSTM and RoBERTa. *International Journal of Cognitive Computing in Engineering*, *5*, 250–258. https://doi.org/10.1016/j.ijcce.2024.06.001

[15] Geetha, S., Elakiya, E., Kanmani, R. S., & Das, M. K. (2025). High performance fake review detection using pretrained DeBERTa optimized with Monarch Butterfly paradigm. *Scientific Reports*, *15*(1), 7445. https://doi.org/10.1038/s41598-025-89453-8

[16] Phukon, P., Potikas, P., & Potika, K. (2025). Detecting fake reviews using aspect-based sentiment analysis and graph convolutional networks. *Applied Sciences*, *15*(7), 3771. https://doi.org/10.3390/app15073771

[17] Meng, W., Harvey, J., Goulding, J., Carter, C. J., Lukinova, E., Smith, A., . . . , & Nica-Avram, G. (2025). *Large language models as' hidden persuaders': Fake product reviews are indistinguishable to humans and machines*. *arXiv Preprint: 2506.13313*

[18] Shukla, A. D., & Goh, J. M. (2024). Fighting fake reviews: Authenticated anonymous reviews using identity verification. *Business Horizons*, *67*(1), 71–81. https://doi.org/10.1016/j.bushor.2023.08.002

[19] Mexwell. (2024). *Fake reviews dataset [Data set]*. *Kaggle*. https://www.kaggle.com/datasets/mexwell/fake-reviews-dataset

[20] Meraj, A. (2019). *Yelp Labelled dataset [Data set]. Kaggle*. https://www.kaggle.com/datasets/abidmeeraj/yelp-labelled-dataset

[21] AbouGrad, H., & Shabarshov, A. (2024). AI-framework to detect ecommerce fake reviews: A hybrid neural network machine learning model. In *Artificial Intelligence and Computational Technologies: Innovations, Usage Cases, and Ethical Considerations*.

[22] AbouGrad, H., Chakhar, S., & Abubahia, A. (2023). Decision making by applying machine learning techniques to mitigate spam SMS attacks. In *Key Digital Trends in Artificial Intelligence and Robotics: Proceedings of 4th International Conference on Deep Learning, Artificial Intelligence and Robotics*, 154–166. https://doi.org/10.1007/978-3-031-30396-8_14

How to Cite: AbouGrad, H., & Riaz, F. (2026). Metadata-Enhanced Hybrid Fusion Architecture: Commercial Fake Reviews Detection Model Using Transformer Embeddings. *FinTech and Sustainable Innovation*. https://doi.org/10.47852/bonviewFSI62028859