ReviewGuard: Real-Time Detection of Fake Online Reviews

[1]Giriprasath K*, [2]Atharv Bandekar, [3]Tejas Kavanthankar, [4]Shubham Govekar and [5]Akshay Shetye

[1, 2, 3, 4, 5]Department of Computer Science and Engineering (AI-ML), Finolex Academy of Management and Technology, Ratnagiri, Maharashtra, India
[1]r220016@famt.ac.in
[2]r220161@famt.ac.in
[3]r220249@famt.ac.in
[4]r220009@famt.ac.in
[5]akshay.shetye@famt.ac.in

Abstract. Online reviews strongly influence purchasing choices in today's e-commerce platforms. However, due to an increase in promotional, spammy, and deceptive reviews, it has often become difficult for a consumer to get a correct idea about the quality of a product. To mitigate this issue, this paper proposes a solution in the form of a browser extension called ReviewGuard, which can identify fake reviews during real-time web surfing. The system extracts live review text from the Document Object Model (DOM) and conducts real-time analysis using a fine-tuned DeBERTa v3-small transformer model, which was initially pretrained on English texts. The model focuses on identifying manipulative language, vague promotional content, and spam-like text patterns, instead of simply flagging AI-generated text. To enhance reliability, the system combines the confidence scores from the neural model with contextual platform indicators, such as the Verified Purchase badge on e-commerce sites like Amazon. The framework was tested using the Amazon Fake Reviews dataset, which includes 40,432 reviews, with an 80–20 split for training and testing. Experimental results show strong performance, with approximately 97% overall accuracy and an F1-score of 0.9696 on the test dataset, demonstrating balanced detection capability (0.9701 F1 for deceptive reviews and 0.9691 for genuine reviews). In addition, the system includes an explainable AI component (XAI) that generates short explanations within the browser interface, showing why the review is flagged.

Keywords: Fake Review Detection, DeBERTa v3, Browser Extension, XAI, E-commerce Reviews.

1 Introduction

Online product reviews decide people's opinion about a particular product they want to buy. More and more fake or made-up reviews are showing up, which makes people

B. Singh et al. (eds.), *Proceedings of the International Conference on Advances in Computing Technology and Artificial Intelligence (COMPUTATIA 2026)*, Atlantis Highlights in Intelligent Systems 18,
https://doi.org/10.2991/978-94-6239-713-2_40

worry a lot about whether online feedback is trustworthy. Fake reviews can influence customer views. They also reduce marketplace credibility. Due to this issue, automated fake review detection has emerged as a crucial research area in natural language processing and online platform security by the authors [1][2]. Early fake review detection methods relied on manually designed linguistic features and rule-based analysis. Recent studies have examined advanced machine learning and deep learning techniques to enhance detection accuracy. Graph learning models, convolutional neural networks, and transformer-based architectures have shown good results in detecting fake reviews by the authors [3][5][7][8]. On online shopping websites, there should be a quick process of checking reviews. Precautions should be taken to avoid false labeling of real reviews as artificial intelligence (AI)-generated. Machine learning predictions can be hard for normal people to understand. This shows why we need explainable AI methods that offer clear, easy-to-understand reasons for the decisions made by automated systems, by the authors [10][11]. A fine-tuned DeBERTa transformer model is used in the proposed system. It also includes post-processing that takes metadata into account to make the detection more reliable. The proposed system also includes AI explanations, which are easy to understand. It turns out the model's predictions into clear and readable explanations for people, with the help of some post-processing. It enhances trust in automated review analysis.

The main contributions of this work are as follows:

1. Development of ReviewGuard, a real-time browser extension that detects misleading reviews as they are being written.
2. A combination of a DeBERTa transformer model with metadata-aware post-processing to minimize false positives.
3. Integration of an explainable AI component that gives clear textual explanation for predictions.
4. Experimental evaluation demonstrating the effectiveness of the system for fake review detection.

2 List of Abbreviations

Table 1. List of Abbreviations used in the paper

Abbreviation	Description
AI	Artificial Intelligence
NLP	Natural Language Processing
ML	Machine Learning
DL	Deep Learning
BERT	Bidirectional Encoder Representation from Transformer
RoBERTa	Robustly Optimized BERT Pretraining Approach
DeBERTa	Decoding Enhanced Bidirectional Encoder Representation from Transformer
BiLSTM	Bidirectional Long Short-Term Memory
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network

XAI	Explainable Artificial Intelligence
DOM	Document Object Model
API	Application Programming Interface
GPU	Graphical Processing Unit

Table 1 lists the abbreviations used throughout in this study. These terms are commonly used while discussing about transformer architectures.

3 Literature Survey

3.1 Overview

Detecting fraudulent and AI-generated reviews has become an increasingly important research challenge due to the rapid growth of online shopping platforms and automated content generation tools. Various research methods that previous studies have considered involve traditional machine learning methods, deep learning models, hybrid models of both, as well as optimization methods on top of transformer models by the authors [1], [2], [4].

3.2 Transformer Models for Fake Review Detection

With advancements in natural language processing, the effectiveness of transformer models for deceptive text analysis has been well established. The authors [5] introduced an optimized DeBERTa-based framework that integrates Monarch Butterfly Optimization to improve convergence and robustness in deceptive review detection. The method was evaluated on several benchmark datasets such as Amazon reviews, Deceptive Opinion Spam, and other e-commerce review collections, achieving performance close to 98% accuracy. The authors cited a number of critical limitations despite such promising results. Heavy computational cost due to transformer depth, along with optimization cost, the necessity of a large labeled dataset, and difficulty in generalization of the domain are hurdles in making this a practical browser extension by the authors [5].

3.3 Recurrent Neural Networks with Attention Mechanisms

Authors [6] introduced a PosAtt-BiLSTM model to better capture the meaning of long textual reviews. The approach extends the traditional BiLSTM architecture by integrating positional attention and NN-SPE encoding. These additions help the model represent contextual relationships within long review sequences. The model was evaluated on several benchmark datasets, including Spam, Yelp Hotel, and Yelp Restaurant reviews. The results showed that the proposed model handled long review texts more effectively than conventional bidirectional LSTM or RNN models. The positional encoding mechanism improved the model's ability to capture semantic patterns. However, the architecture increased computational complexity. Training such models requires more resources, which may limit scalability. In addition, RNNs combined with

attention mechanisms can be slower during inference. This makes them less suitable for real-time detection tasks in live browsing sessions.

The authors [6] later introduced a hybrid architecture that combines RoBERTa embeddings with LSTM-based sequence modeling. In this approach, RoBERTa provides contextual word representations, while the LSTM layer captures sequential relationships in the review text. The model achieved strong performance on datasets such as OpSpam and Deception.

3.4 Linguistic and Heuristic Approach for Review Detection

The authors **[7]** proposed a CNN-based method for detecting deceptive reviews by combining word embeddings with emotion analysis and bootstrap cosine similarity. The approach was evaluated on the Amazon 2018 dataset across several product categories and achieved nearly 96% accuracy. The model also attempted to address the imbalance between Verified Purchase (VP) and Non-Verified Purchase (NVP) reviews. However, methods that rely heavily on handcrafted linguistic and heuristic signals may struggle to adapt to evolving review patterns, particularly with the rise of AI-generated content, which limits their generalization across platforms.

3.5 Resource-Efficient Transformer Approaches

To reduce the computational cost issue for the model, the authors [9] proposed a lightweight model called DenyBERT. The architecture is based on a distilled transformer design that integrates BERT and the De Light layers and knowledge distillation techniques by the authors [9]. The goal is to preserve the language understanding ability of larger transformers. This makes them lighter and faster. However, training still demands high computational resources. Another limitation is the lack of interpretability, since the model's predictions are not easy to understand.

3.6 Key Research Gaps in Existing Works

Based on the literature survey, several important research gaps can be identified:

1. **No Real-Time Deployment:** Solutions tackling fake reviews detection are mostly available in controlled offline environments. No solutions providing real-time service are deployed in a real-world setting.
2. **Absence of Explainability:** There is no clear understanding behind how the classification is done in case of many current approaches.
3. **Over-Dependence on Model Predictions:** The majority of existing approaches pay very little attention to the use of metadata, such as verified purchase indicators.
4. **High False Positive Rates:** As a result of insufficient use of heuristics or contextual clues, some methods misclassify genuine reviews as fake.
5. **Limited Consideration of Client-Side Integration:** Many practical browser-level challenges are not accounted for by existing solutions, dynamic content loading is a challenge to name one.

4 Methodology

ReviewGuard is a browser-level extension that detects fake and AI-generated reviews in real time during normal browsing sessions. The overall methodology is illustrated in Fig. 1.

4.1 System Architecture Overview

This section outlines the ReviewGuard framework, which we developed to flag deceptive product reviews in real time during active e-commerce browsing [1]. Instead of a monolithic architecture, we adopted a modular design. The process begins at the level of the Document Object Model (DOM), where the raw review text is extracted. We preprocess and tokenize this content before feeding it into a transformer-based classification model for authenticity prediction. Neural models often focus mainly on linguistic patterns and may overlook behavioural context [3]. Because of this limitation, our pipeline incorporates a metadata-aware hybrid post-processing layer that adjusts these initial predictions based on rating histories and structural anomalies. To improve transparency, the explainable AI component generates short textual explanations describing the system's reasoning.

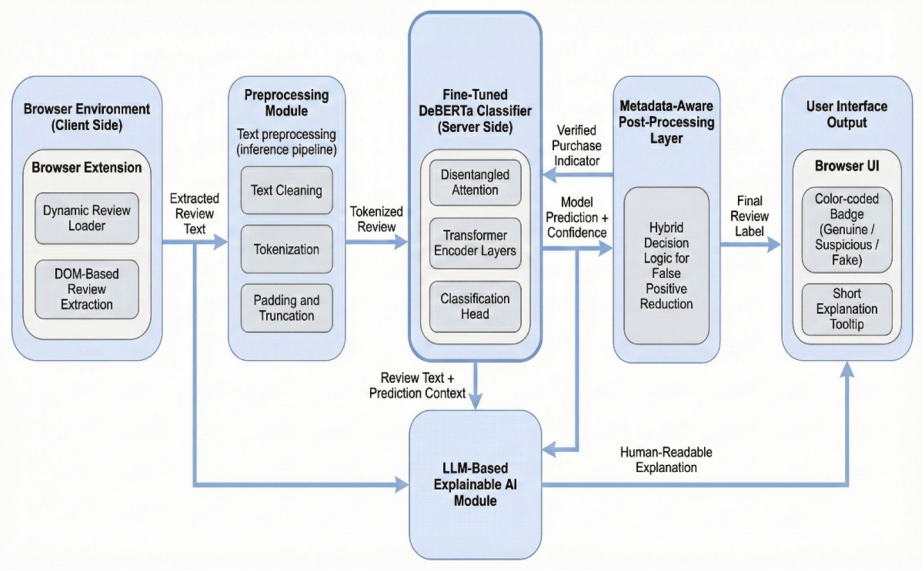

Fig. 1. The overall architecture of the proposed ReviewGuard framework

4.2 Browser Environment and Review Extraction

In our implementation, ReviewGuard is a lightweight browser extension. It monitors live webpages for review elements in the background. It uses standard DOM parsing

techniques. The extension collects reviews directly from the webpages which will be stored in DOM. It then extracts review text and checks whether the verified batch is present in the DOM. The extracted text is then sent to the backend without slowing the browser activity using an API.

4.3 Text Preprocessing Module

Before feeding the extracted reviews into the model, the text is first preprocessed. Any leftover HTML tags or unwanted extraction artifacts are removed at this stage. The text is then normalized by converting it to lowercase and standardizing characters to maintain consistency. After this, tokenization is performed using the DeBERTa tokenizer so that the input matches the model's vocabulary. Unnecessary symbols and noise are removed, so that the processed text is suitable for accurate classification.

4.4 Classification using Fine-Tuned DeBERTa

We used a fine-tuned DeBERTa model for the classification component by the authors **[5]**. By using disentangled attention to separate content from positional embeddings, this extension upgrades standard transformers to map word relationships more effectively. The system stacks multiple encoder layers where multi-head self-attention dynamically builds contextual representations in different stages. Shallow layers lock onto basic lexical syntax, and deeper layers isolate complex, deceptive patterns. This allowed DeBERTa to outperform baseline BERT models in our experiments by the authors [9]. Finally, a fully connected dense network layer caps the pipeline, and we apply a SoftMax activation function to compute the probability of deception.

4.5 Metadata-Aware Hybrid Post-Processing Layer

The neural model mainly relies on textual features while making predictions. However, fake reviews often show highly unusual behavior. To handle this, a post-processing step is introduced using metadata by the authors [8]. It combines the model's output with signals like review length and repeated content. These signals help identify suspicious activity. As a result, the final prediction becomes more accurate and reduces false positives.

4.6 LLM-Based Explainable AI Module

The model generates small explanations that help the user understand why the review was tagged as fake. The model doesn't directly represent the weights; instead, it finds patterns that look like anomalies to it. This becomes the base for an easy explanation of the question 'why,' rather than only providing the main answer, i.e., true or false. Suspicious phrasing or rating anomalies are visually indicated within the text. Presenting users with the actual reasoning behind a flag makes the framework's logic transparent and much easier to trust.

4.7 UI Design and Output Visualization

The final output is displayed directly within the browser interface. Reviews are visually annotated using color-coded badges indicating whether they are classified as genuine, suspicious, or fake. Additionally, users can view short explanatory tooltips generated by the explainable AI module. This presentation minimizes disruption to browsing while providing useful insights for evaluating reviews.

4.8 Tools and Technologies Used

The tools and technologies used in this research are mentioned in Table 2.

Table 2. Tools and Technologies used in the ReviewGuard System

Component	Technology Used
Frontend	Browser Extension
Backend	Python REST Services
Data Processing	NLP- based text preprocessing
Database	In – memory caching
Transformer Model	DeBERTa v3 Small
AI/ML Framework	Deep Learning for NLP
Explainability Module	LLM – based natural language exploration

5 System Implementation

ReviewGuard is implemented as a lightweight browser extension that will detect fake reviews in real-time in a regular browsing session. A fine-tuned DeBERTa-v3-small transformer works as the REST-based inference engine that processes the extracted batch of review text along with metadata such as verified purchase badge in real-time. Our system first grabs the review content from the Document Object Model (DOM), including the dynamically loaded elements, and sends the data to the backend for analysis. Results obtained after classification are rendered on the webpage in the form of a badge so that they can be viewed by users without having to leave the page, as shown in Fig. 2. These badges show whether the model thinks a review is genuine or fake, or if the model is uncertain about the review. The explainable AI (XAI) module provides short, easy-to-understand explanations. The interface is designed with ease of use of a non-technical user in mind, ensuring a smooth and pleasant experience, improving trust and usability in online shopping.

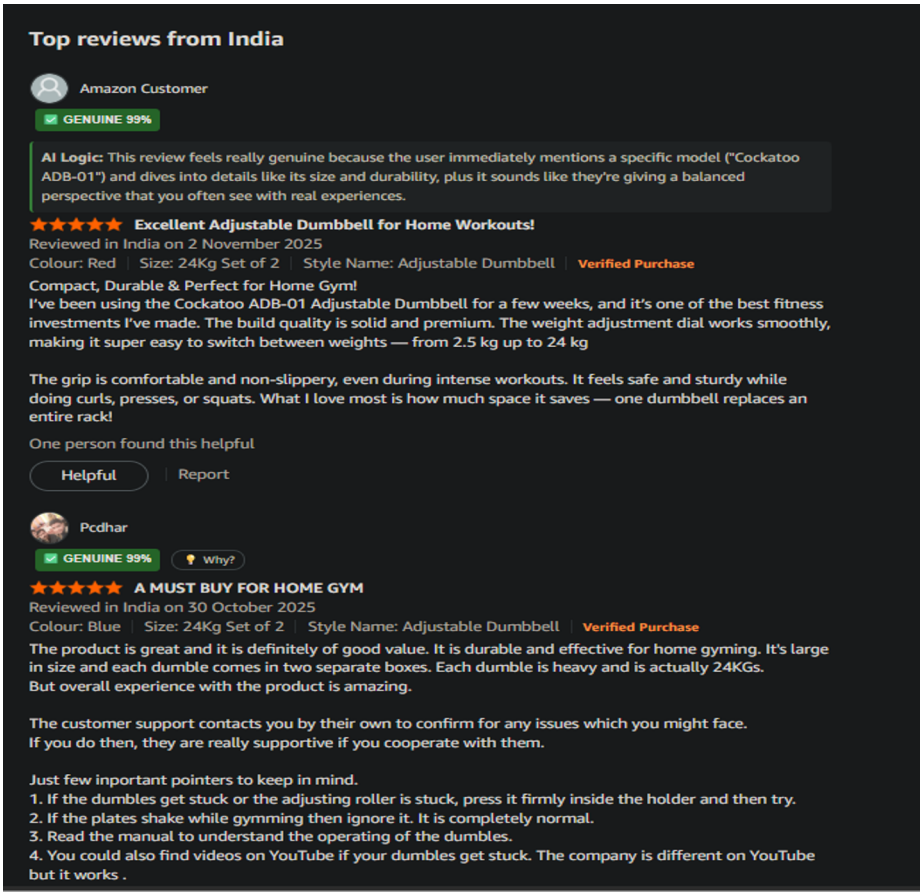

Fig. 2. Demonstration of the ReviewGuard system deployed on a live e-commerce platform for real-time review analysis.

6 Results and Evaluation

The evaluation results show that the DeBERTa-v3 small-based ReviewGuard system performs effectively for fake review detection in real time browsing sessions, as shown in Fig. 3.

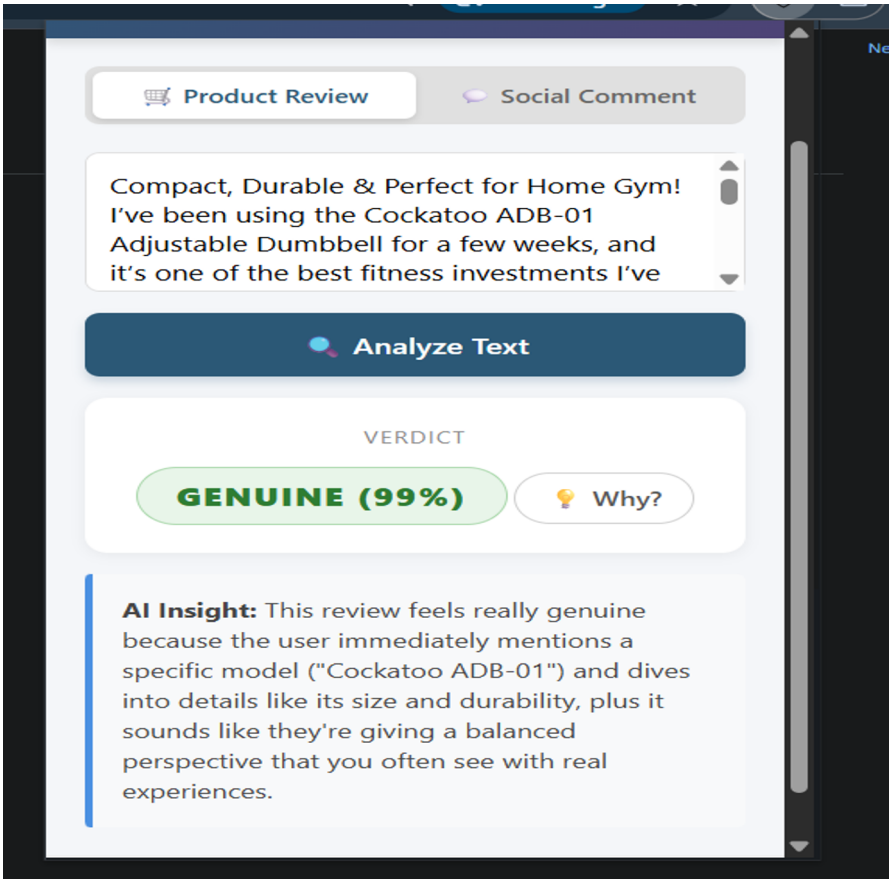

Fig. 3. ReviewGuard user interface displaying real-time prediction results along with explainable insights for detected reviews

6.1 Performance Evaluation of DeBERTa-v3 Small

The DeBERTa-v3 small model an accuracy of 96.04% on the evaluation dataset, demonstrating its effectiveness for fake review detection. The precision and recall values indicate that the model can detect fake reviews while reducing misclassification of genuine ones, as shown in Table 3.

Table 3. Performance Metrics of DeBERTa v3 Small

Model	Accuracy	Precision	Recall	F1-score
DeBERTa v3 small	96.04	0.97	0.96	0.96

6.2 Confusion Matrix-Based Analysis

Fig. 4 shows the confusion matrix, which gives a better understanding of how the model performs. The predictions are fairly balanced for both genuine and fake reviews. It also tends to be slightly conservative.

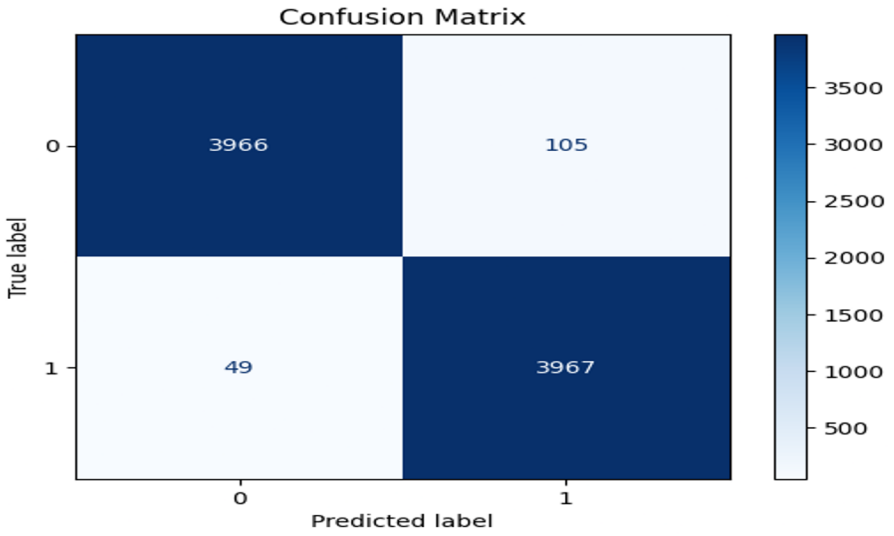

Fig. 4. Confusion matrix of the DeBERTa-v3 small model

6.3 Impact of Hybrid Heuristic Post-Processing

The use of heuristic rules based on platform metadata also adds to the reliability of classification. Features such as verified purchase metadata and abnormal review patterns were incorporated into this model to better utilize the predictions made by the neural model. This hybrid model for making inferences corrects some of the bias that is present in purely neutral predictions.

7 Conclusion

This paper presents ReviewGuard, a real-time browser extension for identifying fake reviews on online shopping platforms. The system uses a fine-tuned DeBERTa-v3 small model along with supporting metadata and an AI that can explain its decisions to make the detection more accurate and easier for users to understand. Tests show the transformer model works well compared to other models that use attention in a repeating way. Also, using a mixed method after processing helps cut down on incorrect results that are common in models that rely only on neural networks. By functioning within the user's browsing session and offering clear explanations for predictions, the system enhances usability for end users. The results show that the proposed framework

is a good method for detecting fake reviews and can help in using it in real e-commerce settings in the future.

Acknowledgments. The authors would like to thank the academic mentors and reviewers for their valuable guidance and feedback during the development of this research.

Disclosure of Interests. The authors confirm that there is no conflict of interest associated with this work.

References

[1] R. Gupta, V. Jindal, and I. Kashyap, "Recent state-of-the-art of fake review detection: a comprehensive review," The Knowledge Engineering Review, vol. 39, 2024, doi: 10.1017/s0269888924000067.

[2] N. V. Khairnar, S. L. Mankar, M. R. Pandav, H. Kotecha, and M. Ranjanikar, "A Survey Paper on Fake Review Detection System," New Frontiers in Communication and Intelligent Systems. Soft Computing Research Society, pp. 625–634, 2021. doi: 10.52458/978-81-95502-00-4-64.

[3] S. Yu, J. Ren, S. Li, M. Naseriparsa, and F. Xia, "Graph Learning for Fake Review Detection," Front. Artif. Intell., vol. 5, Jun. 2022, doi: 10.3389/frai.2022.922589.

[4] S. Geetha, E. Elakiya, R. S. Kanmani, and M. K. Das, "High performance fake review detection using pretrained DeBERTa optimized with Monarch Butterfly paradigm," Sci Rep, vol. 15, no. 1, Mar. 2025, doi: 10.1038/s41598-025-89453-8.

[5] S. Geetha, E. Elakiya, R. S. Kanmani, and M. K. Das, "High performance fake review detection using pretrained DeBERTa optimized with Monarch Butterfly paradigm," Sci Rep, vol. 15, no. 1, Mar. 2025, doi: 10.1038/s41598-025-89453-8.

[6] J. Chen, T. Zhang, Z. Yan, Z. Zheng, W. Zhang, and J. Zhang, "Attention-based BiLSTM with positional embeddings for fake review detection," J Big Data, vol. 12, no. 1, Apr. 2025, doi: 10.1186/s40537-025-01130-9.

[7] P. Hajek, A. Barushka, and M. Munk, "Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining," Neural Comput & Applic, vol. 32, no. 23, pp. 17259–17274, Feb. 2020, doi: 10.1007/s00521-020-04757-2.

[8] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network," IEEE Access, vol. 10, pp. 21517–21525, 2022, doi: 10.1109/access.2022.3152828.

[9] L. Q. Thao et al., “Designing a deep learning-based application for detecting fake online reviews,” Engineering Applications of Artificial Intelligence, vol. 134, p. 108708, Aug. 2024, doi: 10.1016/j.engappai.2024.108708.

[10] A. Barredo Arrieta et al., “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” Information Fusion, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[11] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, Eds., Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer International Publishing, 2019. doi: 10.1007/978-3-030-28954-6.