International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

EDL-DRD: AN ENHANCED DEEP LEARNING MODEL FOR DECEPTIVE REVIEWS DETECTION WITH ATTENTION MECHANISM

Abeer Hassan[1*,2*], Fahad Alotaibi[3]

[1*]Faculty of Computer and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia, aasiri0434@stu.kau.edu.sa
[2*]Department of Information Systems, King Khaled University, Abha, Saudi Arabia, aaassiry@kku.edu.sa
[3]Faculty of Computer and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia, fmmalotaibi@kau.edu.sa

Abstract

Detecting deceptive online reviews is essential to maximize customers' trust and protect organizations' profit. While neural network models have achieved progress in this field, many approaches still rely on surface-level textual features. The interpretability of these models remains limited, as little attention has been given to highlighting linguistically motivated cues through attention mechanisms. To address these challenges, this study introduces the EDL-DRD framework, a hybrid deep learning model that combines textual representations with linguistic and behavioral features. The model employs a CNN–LSTM structure enhanced with a cue mask and an interpretable attention layer. Experiments on a balanced benchmark dataset show that the proposed framework consistently outperforms baseline models across evaluation metrics.

Keywords— Deep learning; EDL-DRD; Deceptive Reviews; Hybrid Model; Attention Mechanism**.**

1. INTRODUCTION

Online reviews play a major role in how people make purchasing choices. They give customers quick, experience-based insights from others who have already used a product or service [1]. Platforms such as e-commerce sites, travel services, and restaurant directories rely heavily on these user contributions to help potential buyers make decisions. For many consumers, reviews have become a primary source of information that supports timely and informed choices [2]. A deceptive review is content meant to mislead customers by misrepresenting the actual product or service experience [3]. The deceptive reviews pose a threat to the integrity of e-commerce systems. This type of content is often designed to distort perceptions for promotional or financial reasons. It can inflate product ratings, damage competitors, and damage consumer trust [4].

As a result, detecting deceptive reviews has become a significant research area in natural language processing (NLP) and cybersecurity. Early work on this topic started around 2007 with initial studies on review spam. Since then, interest in this area has expanded, and deep learning approaches have shown strong potential [5]. The main advantages of DL include its ability to handle large-scale data and complex classification problems, making it effective for identifying misleading content [6]. However, many detection methods rely only on textual features, overlooking other behavioral attributes that could provide complementary signals, potentially limiting their robustness and interpretability across domains.

This work presents a deep learning framework that combines an attention mechanism with a hybrid deep learning model. The goal is to improve accuracy and enable the model to detect deception cues perfectly. We also introduce what we term a cue mask, which is a binary vector aligned with the tokenized text, where each position is set to (1) if the corresponding word matches a deception cue (e.g., first-person pronouns, generality markers, exaggerated emotional words) and (0) otherwise.

While deception theories described a set of deception cues, such as sentiment, usage of some words, and the percentage of pronouns[7], our approach embeds this theoretical knowledge directly into the

Received: August 06, 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

model pipeline. Furthermore, the proposed model includes an attention layer, which plays a key role by identifying which part of the input sequence is more valuable in a classification task where not all words equally contribute to the detection process. For instance, words such as "always perfect" or " terrible" may provide a stronger sign of deceptive intent compared to neutral terms. These methods allow the model to concentrate on the most valuable parts of the dataset.
The models are trained on a publicly available dataset [8] consisting of real and deceptive reviews, enriched with linguistic and behavioral features. In addition, we apply dataset augmentation techniques to address class imbalance and expand the dataset. The comparative analysis aims to identify the model with the highest accuracy and to assess the added value of a proposed approach compared to other models. The study also outlines the limitations of the experiments and suggests future directions to advance deceptive review detection research.

This study provides several notable contributions, which can be outlined as follows:
1- It proposed EDL-DRD, a novel deep learning model that integrates metadata, behavioral features, and text-based neural representations to provide predictive accuracy and theoretical interpretability.
2- It applied augmentation techniques on textual and numerical data to balance the dataset and improve robustness.
3- It introduced a cue mask mechanism that emphasizes specific linguistic categories (hedges, intensifiers, superlatives, vague terms, and sensory descriptors) within the attention layer, guaranteeing the model's ability to capture deceptive linguistic cues.

The structure of the paper is as follows: Section 2 discusses related literature, Section 3 explains the proposed methodology, Section 4 details the experimental setup and evaluation matrices, Section 5 reports the results, Section 6 highlights limitations and potential directions for future research, and finally, Section 7 concludes the work.

2. RELATED WORK

Deceptive review detection: it has been an active research area since 2007. Jindal and Liu [9] were among the first to propose solutions for this issue. They created a publicly available dataset of manually labeled reviews. They identified outliers and duplicated content as deceptive, laying the foundation for subsequent studies. Ott et al. [10] advanced this line of work by generating deceptive reviews through Amazon Mechanical Turk (MTurk) and pairing them with other reviews collected from the Amazon website. Although their early studies relied only on linguistic features and the review dataset classification did not define realistic rules, they demonstrated the feasibility of automatic deceptive review detection and motivated further exploration using advanced techniques.

Deep learning techniques: Early studies on deceptive reviews depended on traditional ML techniques such as Support Vector Machines and logistic regression. These approaches achieved good performance but were limited in capturing complex semantic and contextual patterns. With the rapid advancement of deep learning, researchers applied neural network architecture for this task. DL enabled automatic feature extraction and allowed models to learn deeper representations of deceptive cues, resulting in more robust detection systems. Deep learning has changed the way text classification is handled. Instead of relying on manually crafted features, these methods can quickly uncover useful patterns from the data [11].

Among the different architectures, convolutional neural networks (CNNs) and long short-term memory networks (LSTMs) have played a central role in improving detection deception. They are particularly good at spotting both hierarchical and sequential patterns in textual information[5]. CNNs began to be applied to textual data around 2016, and since then, they have become a standard

Received: August 06, 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

tool for identifying n-gram structures and other layered patterns within reviews [12]. LSTM networks, on the other hand, are better suited for handling longer dependencies in language, allowing the model to account for word order and semantic relations [13].

Several have explored these models in different ways to enhance detection accuracy. For instance, the method proposed by P. Hajek and J.-M. Sahut [14] focuses on extracting information from text using CNNs and a bag-of-words approach. The CNN model uses concurrent convolutional layers that take n-gram representations as input to extract richer feature embeddings. Additionally, the authors' deceptive review classification technique combines both language-based textual features and non-textual aspects of reviewers' behavior. Three models were integrated utilizing ensemble techniques in the research that was presented. The results of the studies demonstrated that adding local sequences of the word enhances the performance of the model. The study also showed how the concurrent convolution layer yields helpful information that is incorporated to improve performance. Hajek et al. [15] proposed a classification technique for detecting deceptive reviews by combining review and reviewer-centric information. They started by extracting text features that depended on sentiment. Reviews were categorized into two sentiment classes (positive and negative)by fine-tuning the embedding matrix weight using the CNN classifier. Behavioral patterns are merged into the CNN output.

According to the study, combining linguistic and behavioral characteristics is crucial for achieving better results. Mohawesh et al. [16] used a Multiview Ensemble DL technique to extract explicit and implicit features of the review. The work mixes word-level and sentence-level using a Bi-LSTM architecture. It incorporates CNN techniques to extract review features, uses CNN to extract product-level features, and detects deceptive reviews.

Hybrid Models: Studies have shown that the hybrid deep learning techniques can improve deception detection. Jacob et al. [17] suggested a hybrid model combining LSTM and CNN. Their experiment results indicate that this hybrid model enhances the ability to identify deceptive reviews. The LSTM part boosts classification performance by capturing sequential patterns across different products and emotion polarities. Zhang et al. [18] proposed a model that combines a Recurrent Neural Network with a Convolutional Neural Network. Their results exceed those of current leading techniques.
Wang et al. [19] introduced a system for detecting deceptive reviews using dictionary-based features and long short-term memory (LSTM). This model has three layers: an LSTM input layer, a hidden layer, and an output layer. Their model surpasses SVM performance, achieving an accuracy of 89.4% in detecting deceptive reviews. The proposed model ignored other crucial elements like metadata and behavioral traits that could improve performance. Liu, Jing, and Li [20] enhanced LSTM performance through feature fusion, incorporating word embeddings, first-person pronouns, and parts of speech. The model exhibited high accuracy in individual domains. The hotel, restaurant, and healthcare domains had respective rates of 83.9%, 85.8%, and 83.8%, surpassing state-of-the-art approaches of mixed domains (83.9% accuracy).
Some studies introduced a model based on semantic features, as the experiment introduced by Alawadh et al. [21] using a balanced hotel review dataset. This work was implemented on web portals to test real-time deceptive detection. It shows that applying deep learning techniques with semantic features can effectively detect deceptive reviews. This study does not fit the needs of the neural network techniques as they used a small dataset.

The attention mechanism: it is a technique used for interpretability and to maximize the performance of machine learning models in Natural Language Processing tasks like summarization, classification, and translation [22] [23]. Previous studies confirm that attention provides improved

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

accuracy and decision interpretability. It was originally introduced in neural machine translation [24] as a means of dynamically weighting input tokens according to their relevance to the prediction task. This technique can be implemented in fake detection models to increase the ability of the models by focusing more on a set of features. Different studies incorporated this technique to improve the DL models and have proven its efficiency across various fields. For instance, the work proposed by Wang et al. [25] that used a CNN model with an attention technique to detect deceptive reviews based on language, behavior, or both. They used CNN to extract linguistic features and a multi-layer perception to obtain behavioral features. They tested the model using two different domain datasets, hotel and restaurant. Compared with state-of-the-art techniques, the suggested approach achieved superior performance.

Bahdanau et al. [26] proposed a machine translation system with an attention mechanism using neural machine translation models. Hao et al. [24] introduced an end-to-end neural network model for question answering, where an attention mechanism was applied to capture the similarity between questions and answers. Their findings demonstrated the effectiveness of this approach.

Despite the demonstrated effectiveness of deep learning models, few studies have adopted theory-driven frameworks combined with attention mechanisms for deceptive review detection. Moreover, limitations such as small and imbalanced datasets have restricted model performance. To address these gaps, this study proposes an enhanced model that integrates a hybrid CNN–LSTM architecture with an attention mechanism and theory-based feature fusion. Data augmentation techniques are also employed to balance and expand the dataset.

3. METHODOLOGY

This section outlines the methodology adopted in this study. As illustrated in Figure 1, the workflow consists of dataset augmentation and preprocessing, model implementation, and results evaluation.

The following subsections describe each stage in detail.

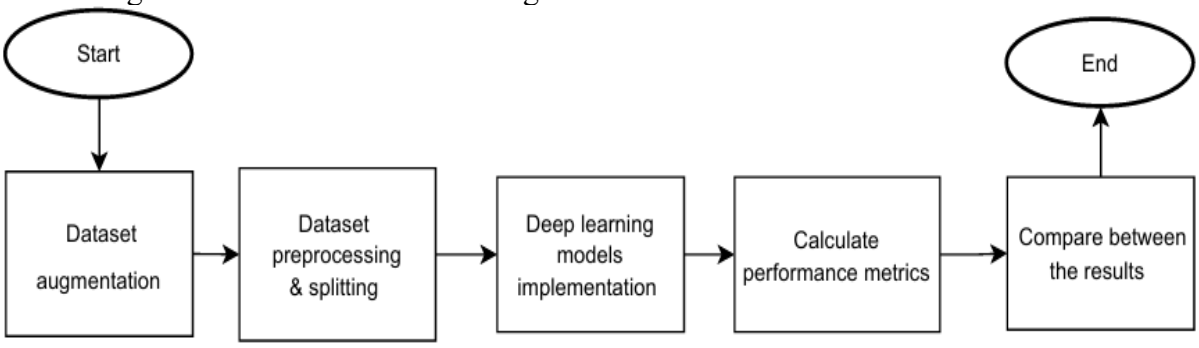

Figure 1 Study workflow

A. Dataset Description

The dataset used in this work is a publicly available, labeled dataset of Google Maps restaurant reviews[8]. It is imbalanced, comprising a total of 21,476 reviews, 12,583 of which are labeled as truthful and 8,893 as deceptive. The dataset includes 11 engineered features capturing behavioral characteristics such as the total number of reviews per user, the total number of media attachments, and the reviewer account type. It also incorporates linguistic features such as sentiment scores, repeated word ratios, and word counts. A binary label indicates whether each review is deceptive (1) or genuine (0).

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

B. Dataset Augmentation

Because the dataset was skewed toward real reviews, we applied augmentation methods to create a more balanced distribution between deceptive and genuine classes. This step was necessary to make the training process more stable and to avoid bias toward the majority class. In NLP, data augmentation refers to generating synthetic samples to strengthen model performance without manual labeling effort [21].

Our dataset contained only 8,893 deceptive reviews, so we focused the augmentation on this class to match the number of real reviews. Several approaches were used to expand the data in its textual and numerical parts. For the text, we relied on techniques such as synonym substitution, swapping the order of words, and back-translation. These changes altered the wording or structure while keeping the original meaning intact. For the numeric features, we added small random perturbations to fields like sentiment scores and review lengths to mimic natural variation [29].

After augmentation, the dataset size increased to 25,166 records. The augmentation process followed the pipeline illustrated in Figure 2, which consisted of the following steps:

- For textual data, we used parallel back-translation to speed up the process. We also applied synonym replacement, random deletion, and word swapping to create different text variations.
- For Numerical data, we created a function to add small random numbers to numeric features to mimic minor changes in values.
- Finally, we merged the augmented text with its corresponding numeric features to form the final dataset.

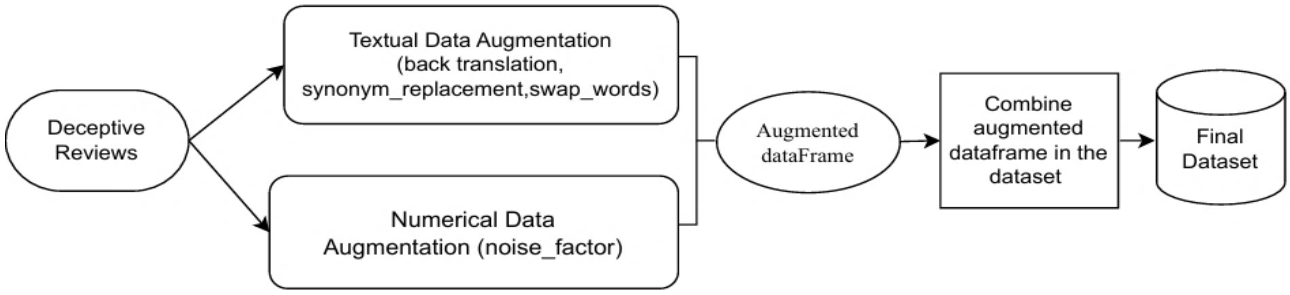

Figure 2 Augmentation Pipeline

C. Dataset Preprocessing

The dataset passed a series of preparation steps to enhance the quality of the data and classification performance.

In order to address missing values, any rows in the review text or numeric feature columns containing null values were eliminated. To ensure compatibility with binary classification tasks, the review labels were then normalised by converting them to integer format (1 for deceptive and 0 for real). The review column then passed several text-cleaning procedures. Text was changed to lowercase to remove case sensitivity. All non-alphabetic elements, such as punctuation marks, digits, and special characters, were eliminated. Popular English stop words such as articles, prepositions, and pronouns were eliminated to reduce noise and enhance the model's focus on content words indicative of deceptive language. Lemmatization was applied to reduce words to their root, minimizing vocabulary sparsity and aligning semantically similar terms. The cleaned reviews were then tokenized using Keras Tokenizer, which mapped each word to an integer index based on word frequency. The generated sequences were padded to a fixed maximum length of 150 tokens to consistent input dimensions across models, enabling efficient batch processing. In parallel, the 15 engineered numeric linguistic features were standardized using the StandardScaler. This

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

normalization centered all features around a mean of zero with the unit standard deviation, facilitating faster and more reliable convergence during training.
We adopted a data-splitting strategy where 70% of the samples were used for training, 15% for validation, and 15% for testing using stratified sampling to maintain original proportions. We selected this partition to adjust model hyperparameters to optimize the model performance and avoid overfitting. It also balances training signals with reliable validation, supports early stopping and learning-rate scheduling, and avoids the computational overhead of k-fold cross-validation.

D. Proposed Model Architecture

As early linguistic studies showed, deceptive opinions frequently differ systematically from truthful communication. The literature on deception detection supported the inclusion of deception cues in various computational models. For instance, George et al. [30] demonstrated that liars employ more generic or abstract phrases and fewer first-person pronouns. Levine [31] emphasized that features such as emotion words, exclusivity markers, and self-references are strong indicators of deception. Our approach embeds this theoretical knowledge directly into the model pipeline.
We proposed an enhanced deep learning model for deceptive reviews detection (EDL-DRD) as a theory-driven framework that combines CNN and LSTM models, provides a cue mask, and an attention layer to recognize deceptive reviews.
As shown in Figure 3, the EDL-DRD model consists of two parallel branches: one focuses on extracting semantic and sequential information from textual reviews, while the other encodes metadata like the behaviourally motivated features. The integration of these branches serves as input to the final classification layer. The following sections discuss the proposed framework.

- **Text Branch**

The text branch is responsible for processing the review text to learn meaningful linguistic patterns. First, each review is tokenized and mapped into a sequence of embeddings, which gives a dense representation of words in a continuous vector space. These embeddings reflect syntactic and semantic similarities, allowing the model to recognize word relationships beyond their surface form. After mapping reviews into embeddings, a cue mask is applied at the token level. This mask is derived from deception theories and marks textual markers that tend to occur more frequently in deceptive writing[32]. Examples of such markers include the use of hedges (e.g., maybe, probably), exaggerated expressions (e.g., always perfect, amazing), vague or generalized statements, intensifiers, and personal pronouns. We present the generation of the cue mask in the following Algorithm 1.

Algorithm 1: Cue Mask Generation

Input: tokenized review T = [t1, t2, ..., tn], cue lexicon L
Output: cue mask M = [m1, m2, ..., mn]

```
for each token ti in T do
   if ti ∈ L then
      mi ← 1
   else
      mi ← 0
   end if
end for
return M
```

By integrating this mask into the representation process, the model assigns higher importance to theoretically motivated cues before extracting the local and sequential patterns through

Received: August 06, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

convolutional and recurrent layers. After this processing, a hybrid CNN-LSTM model is used; the convolutional layer with multiple filters of sizes 3, 4, and 5 is applied to extract local structures, such as common word combinations or short deceptive phrases. This step highlights n-gram patterns that may signal exaggeration, vagueness, or strong sentiment. In parallel, the LSTM layer examines the model's entire sequence for long-term dependencies and contextual flow. This helps detect cues of deception that rely on word order and sentence progression. An attention mechanism is added to highlight words or phrases that are more important for telling apart deceptive and genuine reviews [33]. This mechanism does not treat all tokens equally but instead creates a weighted representation where informative cues receive stronger emphasis. To further refine this representation, two simple scaling parameters (β (beta) and γ (gamma)) are applied. Here, β acts like a small offset that slightly shifts the attention scores, while γ works like a magnifier that scales the importance of specific cues. This simple adjustment helps the model highlight deception-related signals more clearly before merging with the numeric feature branch. The output of this branch is thus a rich textual representation that combines semantic information, sequential dependencies, and focused attention on key deception indicators [33], [34]. The output generated from the Interpretable Attention layer is transferred to the fusion module, where it is concatenated with the numeric feature representation. The fused vector is then processed by Dense layers for final classification. The detailed steps of the interpretable attention mechanism are outlined in Algorithm

Algorithm 2: Interpretable Attention Mechanism

Input: Hidden states H = [h1, h2, ..., hn]
Output: Context vector C

1: for each hi in H do
2: ui ← tanh(Wa · hi + ba) # compute intermediate representation
3: αi ← exp(ui) / Σj exp(uj) # normalize into attention weights
4: end for
5: C ← Σi αi · hi # weighted sum to get context vector
6: **return** C

- **Features Branch**

The feature branch is designed to incorporate other features in the dataset that include metadata and behavioral aspects of review. These features cover elements that are not always learned effectively by text analysis alone, such as the spatiotemporal data, the type of reviewer account, the ratio of punctuation marks in reviews, the availability of useful votes, and whether the review contains media or not. This numerical data often provides contextual evidence of deceptive activity. Before entering the neural network, these heterogeneous features are normalized and encoded to ensure comparability across different scales. Then the processed features are passed through fully connected layers that reduce dimensionality and learn higher-level abstractions. This allows the model to transform raw numerical signals into compact representations that complement the semantic information from the text branch.

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- **Concatenation Stage**

A concatenation step is a process of combining textual and numerical representations before the final classification. This method enhances classification interpretability by focusing on the most important elements and helps the model to detect local lexical cues, long-range dependencies, and reviewer behavioral signals together. Concatenation ($\oplus$) combines the text representation from the Interpretable Attention and the feature representation from the numeric branch into a single joint vector. After concatenation, the fused representation is passed through dense and dropout layers to reduce overfitting, followed by a final sigmoid layer to produce the binary output.

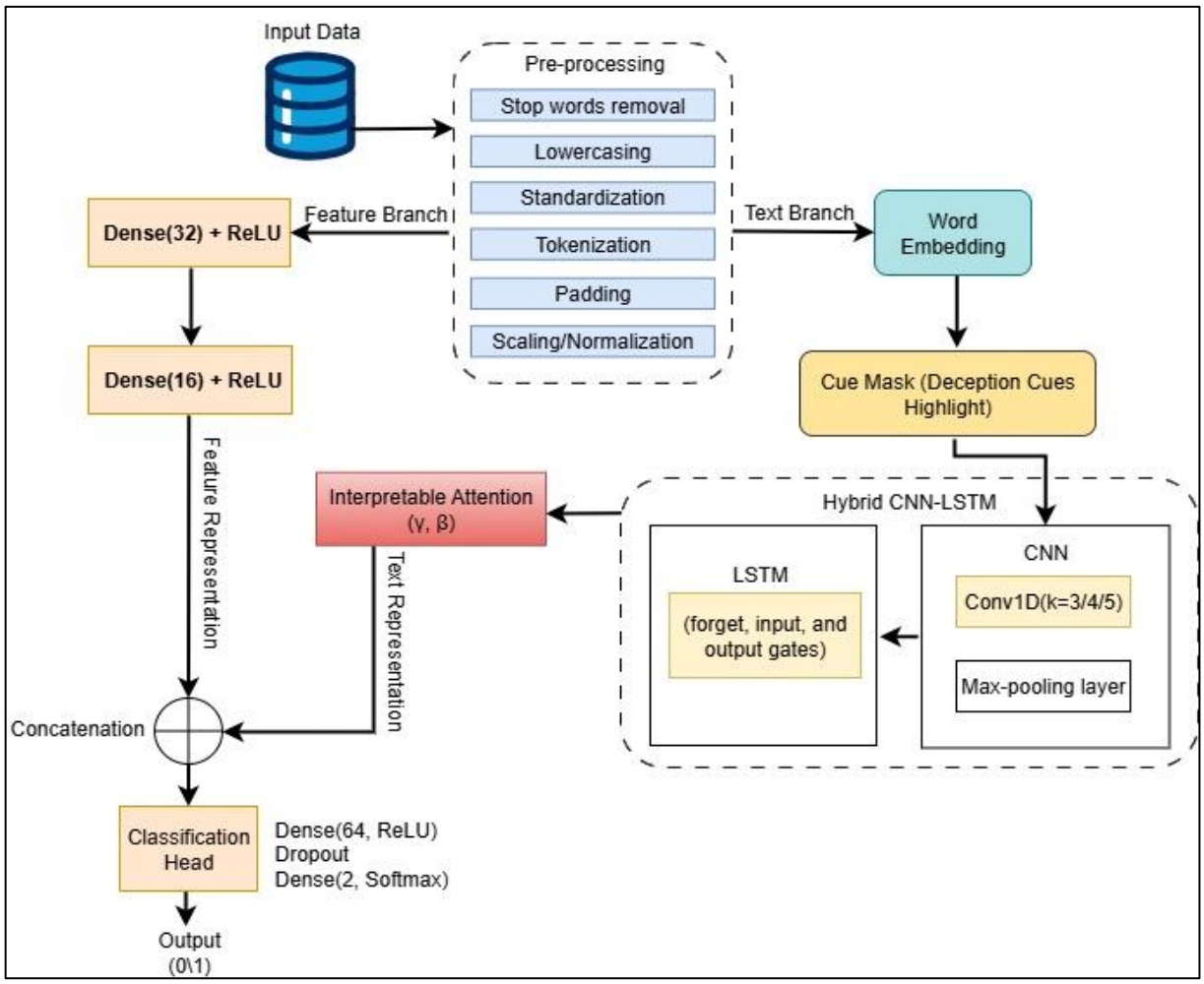

Figure 3 Proposed Model Architecture

E. Baseline Models

To assess the EDL-DRD efficiency, we contrast its performance with the following standard baseline models:

- **CNN**: A convolutional neural network that extracts local n-gram features from word embeddings. It is effective in capturing short patterns such as exaggerated expressions or repeated terms that are common in deceptive reviews.
- **LSTM**: A recurrent model that learns long-range dependencies within review texts. It is designed to capture sequential patterns and contextual flow that may reveal subtle deception cues spread across sentences.

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- **CNN+LSTM**: A hybrid model that integrates convolutional filters to extract local features with LSTM layers for sequential and long-range relations. This integration aims to balance short-term and long-term dependencies, providing a stronger baseline for deception detection.

These baseline models provide a fair comparison and allow us to highlight the added value of integrating theoretical cues and behavioral features in the EDL-DRD model.

4. EXPERIMENTAL SETUP AND EVALUATION METRICS

All experiments were conducted using Python 3.10.12 on the Google Colab Pro platform with GPU acceleration to enable efficient deep learning model training. The NLTK library was employed for natural language preprocessing, including lemmatization, stop-word removal, and WordNet integration. Pandas and NumPy were used for data manipulation and preprocessing.

The review texts were cleaned using regular expressions and tokenized using the Keras Tokenizer. Model development was carried out with TensorFlow 2.12.0 and Keras 2.12.0. The dataset included textual reviews, behavioral features, and binary labels (0 = real, 1 = deceptive). Before training, the dataset underwent cleaning, augmentation, and balancing procedures. All models, including CNN, LSTM, and CNN+LSTM baselines, as well as the proposed EDL-DRD framework, were trained on the same augmented dataset to ensure fair comparison. All models were trained for 10 epochs with a batch size of 64, using the Adam optimizer and binary cross-entropy loss. Model performance was evaluated using accuracy, precision, recall, F1-score, and confusion matrices. To ensure stability and reproducibility, experiments were repeated with different random seeds.

The evaluation of both the proposed model and the baseline models used standard binary classification metrics, including accuracy, precision, recall, and F1-score. These metrics together provide a thorough assessment of model effectiveness.

- **Accuracy** measures the overall percentage of reviews that are classified correctly. It gives a general idea of model performanc**e.**
- **Precision** assesses the proportion of reviews predicted as deceptive that are deceptive. This reflects the model's ability to reduce false alarms.
- **Recall** shows the percentage of real deceptive reviews that are correctly identified. This is essential for lowering the chance of overlooking fraudulent content.
- **F1-score** is the harmonic mean of precision and recall. It offers a balanced measure that takes both false positives and false negatives into account.

In addition to these metrics, we examine the confusion matrices to provide a detailed breakdown of false positives, false negatives, true positives, and true negatives. These matrices visualize the types of model errors.

5. RESULTS AND DISCUSSION

The experiment results are summarized in Table 1 and Figure 4, showing clear differences in performance across the tested models. The proposed EDL-DRD framework outperformed all baselines, achieving an accuracy of 90% and F1-score of 0.89, with a balanced precision of 0.91 and recall of 0.88. This indicates that combining CNN and LSTM with an attention mechanism provides more robust detection for deceptive reviews.

In contrast, the standalone CNN and LSTM models performed noticeably worse. CNN achieved 72% accuracy and an F1-score of 0.76, showing its limitation in handling sequential dependencies. The LSTM model produced the lowest results overall, with accuracy reaching 70% and an F1-score of 0.73. This suggests that sequential modeling alone was insufficient, especially given the relatively short review texts where deceptive cues are often localized.

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The hybrid CNN-LSTM model slightly improved over other standalone models, with an accuracy of 73% and an F1-score of 0.75. However, its performance remained below that of the proposed EDL-DRD, highlighting the added value of the attention mechanism.

Table 1 Model Performance Comparison with Baseline Models

Model	Accuracy	Precision	Recall	F1-Score
EDL-DRD	0.90	0.91	0.88	0.89
CNN	0.72	0.66	0.87	0.76
CNN+LSTM	0.73	0.69	0.83	0.75
LSTM	0.70	0.67	0.80	0.73

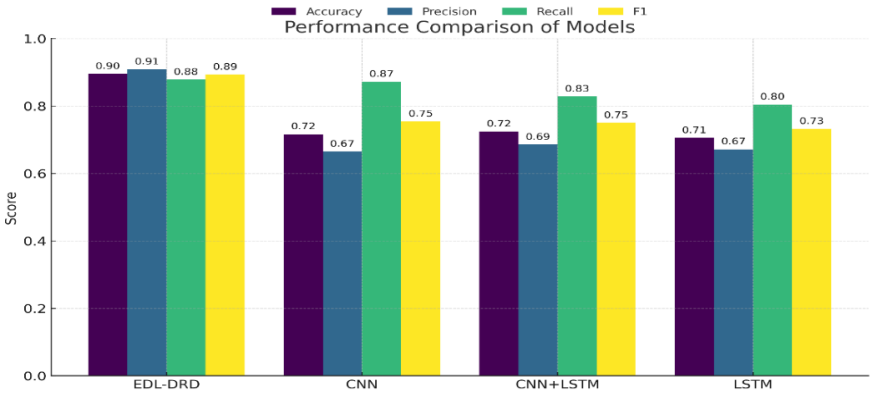

Figure 4 Models Performance Comparison

As shown in Figure 5, the confusion matrices show that EDL-DRD demonstrates a higher number of true positives and true negatives compared to the baseline models, indicating a stronger ability to correctly identify both deceptive and genuine reviews.

In contrast, the CNN and LSTM baselines show a relatively larger proportion of misclassifications, either by labeling genuine reviews as deceptive (false positives) or by failing to detect deceptive ones (false negatives). The hybrid CNN+LSTM model performs better than its single-component counterparts, but still falls short of the accuracy achieved by EDL-DRD. The confusion matrix analysis confirms that integrating deception cues and behavioral features significantly improves classification reliability.

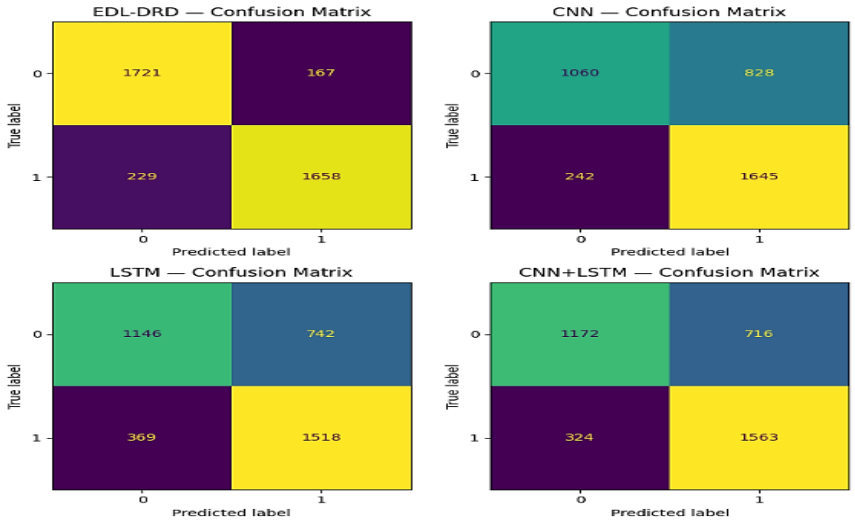

Figure 5 Confusion Matrices

To evaluate the EDL-DRD model's effectiveness, we compared it with some previous studies in the literature that applied hybrid deep learning for deceptive review detection, as shown in Table 2.

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Previous studies reported accuracies of 0.86 – 0.88 on Amazon and Ott's restaurant datasets. A CNN-LSTM with attention on the JD dataset achieved 88% accuracy and an F1-score of 0.71**.** In this experiment, the proposed CNN–LSTM with interpretable attention achieved an F1-score of 0.89 and an accuracy of 0.90, outperforming previously reported results.

Table 2 Comparison with previous studies

Ref.	Used techniques	Dataset	accuracy	F-1
[35]	CNN-LSTM	Amazon product datasets	0.86	0.86
[36]	CNN-LSTM with attention technique	JD product dataset	0.85	0.71
[18]	CNNs-RNNs	Ott et al. restaurant dataset	0.87	0.86
EDL_DRD	CNN-LSTM with attention technique	Google map restaurant dataset	0.90	0.89

6. LIMITATION AND FUTURE WORK

Although the findings demonstrate the effectiveness of attention-based deep learning models for deceptive review detection, several limitations remain, highlighting opportunities for future research and enhancement. This study relied on a single, domain-specific dataset, which may not fully capture the diversity and subtleties of deceptive behaviors observed in other contexts. Future research should validate the models on larger, naturally occurring datasets from multiple domains and platforms to enhance generalizability. Although the models incorporated both textual and engineered numeric features, they did not leverage contextual language representation models such as BERT or other transformer-based models. Integrating such architecture could be an avenue for performance improvement.

The models used here assume binary classification (deceptive vs. real). Deceptive content may exist on a spectrum or include gray areas such as exaggeration, partial truth, or manipulation for marketing. Future efforts could move toward fine-grained, multi-class deception classification or even unsupervised anomaly detection approaches. The current models were trained using static hyperparameters and architectural settings. Although early stopping and dropout were used for regularization, further exploration of automated hyperparameter tuning or model ensembling could improve robustness and adaptability across domains. Real-world deployment was not considered in this study. Future work could focus on real-time detection, user interface design, and API deployment, allowing the proposed models to be integrated into live review monitoring systems.

CONCLUSION

In this study, we presented the EDL-DRD framework, a novel approach for deceptive review detection that integrates CNN and LSTM architectures with an interpretable attention mechanism. Before model training, the dataset underwent thorough preprocessing and data augmentation to enhance quality and balance. The experimental evaluation demonstrated that the proposed framework consistently outperforms standard baseline models, underscoring the effectiveness of incorporating attention to focus on salient cues within review texts. Among the baselines, the CNN model achieved higher performance than LSTM, suggesting that local lexical patterns, such as repeated expressions and linguistic markers, play a critical role in identifying deception in short textual content. Overall, the findings confirm that combining multiple feature types through the EDL-DRD framework provides measurable gains over conventional hybrid deep learning models. Future work may build on this by integrating contextual embeddings, multimodal signals, and real-world deployment scenarios to develop more accurate, interpretable, and scalable detection systems.

REFERENCES

[1] A. Costa, J. Guerreiro, S. Moro, and R. Henriques, "Unfolding the characteristics of incentivized online reviews," *J. Retail. Consum. Serv.*, vol. 47, pp. 272–281, Mar. 2019, doi: 10.1016/j.jretconser.2018.12.006.

Received: August 06, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[2] R. A. Duma *et al.*, "Fake review detection techniques, issues, and future research directions: a literature review," *Knowl. Inf. Syst.*, vol. 66, no. 9, pp. 5071–5112, Sep. 2024, doi: 10.1007/s10115-024-02118-2.

[3] K. S. Desale, S. Shinde, N. Magar, S. Kullolli, and A. Kurhade, "Fake Review Detection with Concept Drift in the Data: A Survey," in *Proceedings of Seventh International Congress on Information and Communication Technology*, X.-S. Yang, S. Sherratt, N. Dey, and A. Joshi, Eds., Singapore: Springer Nature Singapore, 2023, pp. 719–726.

[4] S. He, B. Hollenbeck, G. Overgoor, D. Proserpio, and A. Tosyali, "Detecting fake-review buyers using network structure: Direct evidence from Amazon.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 119, no. 47, p. e2211932119, Nov. 2022, doi: 10.1073/pnas.2211932119.

[5] C. Cao, S. Li, S. Yu, and Z. Chen, "Fake Reviewer Group Detection in Online Review Systems," in *2021 International Conference on Data Mining Workshops (ICDMW)*, Dec. 2021, pp. 935–942. doi: 10.1109/ICDMW53433.2021.00122.

[6] W. M. Lim, R. Agarwal, A. Mishra, and A. Mehrotra, "The Rise of Fake Reviews: Toward a Marketing-Oriented Framework for Understanding Fake Reviews," *Australas. Mark. J.*, vol. 33, no. 2, pp. 178–198, May 2025, doi: 10.1177/14413582241283505.

[7] J. Y. Thomas and D. P. Biros, "An empirical evaluation of interpersonal deception theory in a real-world, high-stakes environment," *J. Crim. Psychol.*, vol. 10, no. 3, pp. 185–199, Jan. 2020, doi: 10.1108/JCP-07-2019-0025.

[8] A. Asiri and F. Alotaibi, "Deceptive Reviews Dataset," vol. 2, Jul. 2025, doi: 10.17632/y2s4973hsg.2.

[9] N. Jindal and B. Liu, "Analyzing and Detecting Review Spam," in *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, Oct. 2007, pp. 547–552. doi: 10.1109/ICDM.2007.68.

[10] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding Deceptive Opinion Spam by Any Stretch of the Imagination," Jul. 22, 2011, *arXiv*: arXiv:1107.4557. doi: 10.48550/arXiv.1107.4557.

[11] R. Mohawesh *et al.*, "Fake Reviews Detection: A Survey," *IEEE Access*, vol. 9, pp. 65771–65802, 2021, doi: 10.1109/ACCESS.2021.3075573.

[12] J. Lu *et al.*, "BSTC: A Fake Review Detection Model Based on a Pre-Trained Language Model and Convolutional Neural Network," *Electronics*, vol. 12, no. 10, p. 2165, 2023, doi: 10.3390/electronics12102165.

[13] MafasRaheem and Yap Seng Chong, "E-Commerce Fake Reviews Detection Using LSTM with Word2Vec Embedding," *J. Comput. Inf. Technol.*, vol. 32, no. 2, pp. 65–80, Sep. 2024, doi: 10.20532/cit.2024.1005803.

[14] M. S. Javed, H. Majeed, H. Mujtaba, and M. O. Beg, "Fake reviews classification using deep learning ensemble of shallow convolutions," *J. Comput. Soc. Sci.*, vol. 4, no. 2, pp. 883–902, Nov. 2021, doi: 10.1007/s42001-021-00114-y.

[15] P. Hajek and J.-M. Sahut, "Mining behavioural and sentiment-dependent linguistic patterns from restaurant reviews for fake review detection," *Technol. Forecast. Soc. Change*, vol. 177, p. 121532, Apr. 2022, doi: 10.1016/j.techfore.2022.121532.

[16] R. Mohawesh, S. Xu, M. Springer, Y. Jararweh, M. Al-Hawawreh, and S. Maqsood, "An explainable ensemble of multi-view deep learning model for fake review detection," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 35, no. 8, p. 101644, Sep. 2023, doi: 10.1016/j.jksuci.2023.101644.

[17] M. S. Jacob and P. Selvi Rajendran, "Fuzzy artificial bee colony-based CNN-LSTM and semantic feature for fake product review classification," *Concurr. Comput. Pract. Exp.*, vol. 34, no. 1, p. e6539, 2022, doi: 10.1002/cpe.6539.

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[18] W. Zhang, Y. Du, T. Yoshida, and Q. Wang, "DRI-RCNN: An approach to deceptive review identification using recurrent convolutional neural network," *Inf. Process. Manag.*, vol. 54, no. 4, pp. 576–592, Jul. 2018, doi: 10.1016/j.ipm.2018.03.007.

[19] X. Wang, K. Liu, and J. Zhao, "Detecting Deceptive Review Spam via Attention-Based Neural Networks," in *Natural Language Processing and Chinese Computing*, X. Huang, J. Jiang, D. Zhao, Y. Feng, and Y. Hong, Eds., Cham: Springer International Publishing, 2018, pp. 866–876. doi: 10.1007/978-3-319-73618-1_76.

[20] W. Liu, W. Jing, and Y. Li, "Incorporating feature representation into BiLSTM for deceptive review detection," *Computing*, vol. 102, no. 3, pp. 701–715, Mar. 2020, doi: 10.1007/s00607-019-00763-y.

[21] H. M. Alawadh, A. Alabrah, T. Meraj, and H. T. Rauf, "Correction: Alawadh et al. Semantic Features-Based Discourse Analysis Using Deceptive and Real Text Reviews. Information 2023, 14, 34," *Information*, vol. 15, no. 12, p. 824, Dec. 2024, doi: 10.3390/info15120824.

[22] J. Blake, A. S. M. Miah, K. Kredens, and J. Shin, "Detection of AI-Generated Texts: A Bi-LSTM and Attention-Based Approach," *IEEE Access*, vol. 13, pp. 71563–71576, 2025, doi: 10.1109/ACCESS.2025.3562750.

[23] H. Liao, Y. Liang, S. Chen, L. Xiang, Z. Chang, and Y. Xiao, "STSG: A Short Text Semantic Graph Model for Similarity Computing Based on Dependency Parsing and Pre-trained Language Models," *Appl. Artif. Intell.*, vol. 38, no. 1, p. 2321552, Dec. 2024, doi: 10.1080/08839514.2024.2321552.

[24] Y. Hao *et al.*, "An End-to-End Model for Question Answering over Knowledge Base with Cross-Attention Combining Global Knowledge," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, R. Barzilay and M.-Y. Kan, Eds., Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 221–231. doi: 10.18653/v1/P17-1021.

[25] G. Wang, G. Shang, P. Pu, X. Li, H. Peng, and C.-H. Wu, "Fake Review Identification Methods Based on Multidimensional Feature Engineering," *Mob. Inf. Syst.*, vol. 2022, Jan. 2022, doi: 10.1155/2022/5229277.

[26] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," May 19, 2016, *arXiv*: arXiv:1409.0473. doi: 10.48550/arXiv.1409.0473.

[27] G. M. Shahariar, Md. T. R. Shawon, F. M. Shah, M. S. Alam, and Md. S. Mahbub, "Bengali fake reviews: A benchmark dataset and detection system," *Neurocomputing*, vol. 592, p. 127732, Aug. 2024, doi: 10.1016/j.neucom.2024.127732.

[28] M. Bayer, M.-A. Kaufhold, and C. Reuter, "A Survey on Data Augmentation for Text Classification," *ACM Comput Surv*, vol. 55, no. 7, p. 146:1-146:39, Dec. 2022, doi: 10.1145/3544558.

[29] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Glob. Transit. Proc.*, vol. 3, no. 1, pp. 91–99, Jun. 2022, doi: 10.1016/j.gltp.2022.04.020.

[30] J. f. George, A. m. Mills, G. Giordano, M. Gupta, V. m. Tennant, and C. c. Lewis, "Toward a Greater Understanding of the Use of Nonverbal Cues To Deception in Computer-Mediated Communication," *IEEE Trans. Prof. Commun. Prof. Commun. IEEE Trans. IEEE Trans Profess Commun*, vol. 66, no. 2, pp. 131–149, Jun. 2023, doi: 10.1109/TPC.2023.3263378.

[31] T. R. Levine, "Truth-default theory and the psychology of lying and deception detection," *Curr. Opin. Psychol.*, vol. 47, p. 101380, Oct. 2022, doi: 10.1016/j.copsyc.2022.101380.

[32] P. Faulkner, "Lying and Deception: Theory and Practice," *Ethics*, vol. 121, no. 4, p. 799, Jul. 2011, doi: 10.1086/661116.

[33] M.-T. Luong, H. Pham, and C. D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," Sep. 20, 2015, *arXiv*: arXiv:1508.04025. doi: 10.48550/arXiv.1508.04025.

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[34] "Fuzzy artificial bee colony-based CNN-LSTM and semantic feature for fake product review classification", doi: 10.1002/cpe.6539.
[35] A. Alghaligah, A. Alotaibi, Q. Abbas, and S. Alhumoud, "Optimized Hybrid Deep Learning for Enhanced Spam Review Detection in E-Commerce Platforms," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 1, 2025, doi: 10.14569/IJACSA.2025.0160134.
[36] J. Li, Y. Fu, D. Liu, and R. Xu, "Improving Fake Product Detection with Aspect-Based Sentiment Analysis," in *Cognitive Computing – ICCC 2020*, Springer, Cham, 2020, pp. 39–49. doi: 10.1007/978-3-030-59585-2_4.

Received: August 06, 2025