Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 9, 2025

FAKE REVIEW DETECTION USING HYBRID BILSTM AND CNN DEEP LEARNING MODEL ON MULTI-DOMAIN TEXTUAL DATA

Zeeshan Ali Khan[*1], Muhammad Naeem[2], Muhammad Zulkifl Hasan[3], Muhammad Zunnurain Hussain[4]

[*1]*NCBA & E AL HAMRA University*
[2]*Department of Artificial Intelligence, University of Management and Technology, Johar town Lahore, Pakistan*
[3]*Faculty of Information Technology, Department of Computer Science, University of Central Punjab, Lahore, Pakistan*
[4]*Department of Computer Science, Bahria University Lahore Campus, Pakistan*

[*1]zeesh.khan1991@gmail.com,[2]naeemdev71@gmail.com, [3]zulkifl.hasan@ucp.edu.pk, [4]zunnurain.bulc@bahria.edu.pk

DOI: https://doi.org/10.5281/zenodo.17112502

Keywords
Fake Review Detection, Deep Learning, BiLSTM, CNN, Text Classification, Natural Language Processing (NLP), SHAP, Explainable AI, Multi-Domain Classification, Model Interpretability

Article History
Received: 22 June 2025
Accepted: 01 September 2025
Published: 13 September 2025

Corresponding Author: *
Zeeshan Ali Khan

Abstract

Fake reviews distort consumer trust and mislead online buyers. This study presents a hybrid deep learning framework combining Bidirectional Long Short-Term Memory (BiLSTM) and Convolutional Neural Networks (CNN) to detect fake reviews using textual data from multiple domains including e-commerce, hotels, and services. The model captures both long-range contextual and local phrase-level patterns from review texts. Training and evaluation on a large multi-domain dataset demonstrate strong detection performance with accuracy of 92.7%, F1-score of 0.927, and AUC of 0.9805. SHAP explainability techniques provide model interpretability, illustrating important textual patterns influencing predictions. This approach shows promise as an effective, scalable, and interpretable solution for fake review detection based primarily on review text analysis.

INTRODUCTION

Online reviews have become a critical component influencing consumer decision-making in the digital economy. Consumers increasingly rely on reviews to assess the quality, reliability, and trustworthiness of products and services across various sectors such as ecommerce, hospitality, and online services. Positive reviews can significantly boost sales and brand reputation, whereas negative or fake reviews can distort market dynamics, mislead consumers, and undermine platform credibility (Liu, B. 2012). The prevalence of fake reviews—intentionally misleading or fabricated user-generated content— poses a substantial challenge for online marketplaces and review platforms. These deceptive reviews can artificially inflate or deflate

product ratings, resulting in unfair competition and eroding consumer trust.

Detecting fake reviews is therefore crucial for maintaining a fair and transparent online ecosystem. Traditional approaches to fake review detection have primarily focused on linguistic cues and classical machine learning algorithms using handcrafted features such as n-grams, sentiment scores, and metadata. However, such methods often struggle to capture the subtle and complex patterns inherent in deceptive language, especially across diverse domains where review styles vary widely.

Recent advances in deep learning, particularly models capable of capturing sequential and contextual information in text, have demonstrated superior performance in natural language processing tasks. Models like Long Short-Term Memory (LSTM) networks excel at understanding long-range dependencies in text, while Convolutional Neural Networks (CNN) are effective at identifying local phrase-level patterns (Wang, Y., & Liu, X. 2021).

In this study, we propose a hybrid deep learning architecture combining Bidirectional LSTM and CNN layers to leverage the complementary strengths of both models. This architecture is designed to capture both the global context and local textual features necessary for accurately detecting fake reviews. Additionally, we apply SHapley Additive exPlanations (SHAP) techniques to interpret model predictions, thereby enhancing transparency and trustworthiness in automated fake review detection systems (Alhindi, T., & Alhindi, A. 2023).

Our research addresses several key challenges: the ability to generalize across multiple review domains, achieving high classification accuracy and robustness, and providing interpretability of model decisions to foster user trust. By evaluating our model on a large, multi-domain dataset containing e-commerce, hotel, and service reviews, we demonstrate its effectiveness and scalability.

2. Problem Definition

Detecting fake reviews presents a complex and multifaceted challenge in the realm of online content moderation. Fake reviews are crafted with the intent to mislead consumers by artificially promoting or demoting products and services. These reviews often employ sophisticated deceptive strategies, including subtle language manipulation, emotional appeal, or mimicking authentic review styles, which makes their identification non-trivial (Choi, E., & Lee, K. 2020). One primary difficulty in fake review detection lies in the linguistic subtleties and diverse writing patterns that distinguish fake from genuine reviews. Unlike spam or overtly promotional content, fake reviews can be highly nuanced, blending truthful elements with misleading claims. This complexity demands models capable of understanding both the overarching context of a review and the local phrase structures that may hint at deception.

Another critical aspect is the multi-domain nature of review data. Reviews from e-commerce platforms, hospitality services, and other industries differ significantly in vocabulary, sentiment expression, and contextual relevance. A detection model must generalize effectively across these domains to be practically useful, rather than being tailored narrowly to a specific category or platform.

Traditional machine learning methods relying on handcrafted linguistic features, such as n-grams, sentiment scores, or reviewer metadata, have limited ability to capture these complex dependencies. (Liu, F., & Zhang, Z. 2019). Moreover, behavioral features related to reviewer activity (e.g., frequency of reviews, rating distributions) can enhance detection but require additional data that may not always be available or privacy-compliant.

This research focuses primarily on leveraging textual information using deep learning architectures that can model both global dependencies and local textual cues in review content. The challenge is to develop a robust, scalable, and interpretable model that can automatically detect fake reviews with high accuracy across diverse datasets without heavy reliance on auxiliary metadata. In summary, the problem entails:

- Subtle linguistic and stylistic variations: Fake reviews often mimic genuine reviews closely, requiring sophisticated text modeling to differentiate.
- Domain variability: Differences in language and style across domains complicate the generalization of detection models.

- Limited reliance on behavioral data: Focusing on textual features alone presents both a challenge and an advantage in terms of data availability and privacy.
- Need for interpretability: Automated detection systems must provide explainable results to build user and platform trust.

Addressing these challenges motivates the development of hybrid deep learning approaches that combine the strengths of sequential models and convolutional networks, coupled with interpretability tools such as SHAP, to create reliable and transparent fake review detection systems.

3. Aim

This research aims to develop a robust and effective fake review detection system by leveraging advanced deep learning techniques and interpretability methods. Specifically, the objectives of this study are:

To design and implement a hybrid deep learning model combining Bidirectional Long Short-Term Memory (BiLSTM) networks and Convolutional Neural Networks (CNN) that can effectively capture both the sequential dependencies and local phrase-level patterns in textual reviews. This hybrid approach aims to improve upon the limitations of single-model architectures by exploiting complementary strengths.

To evaluate the model's ability to generalize across multiple domains, including e-commerce, hospitality, and service industries, ensuring versatility and applicability in diverse real-world scenarios. The model should accurately classify reviews as fake or genuine regardless of domainspecific language differences.

To optimize the model's classification performance, targeting high accuracy, precision, recall, F1-score, and Area Under the Curve (AUC) metrics, thereby achieving reliable and balanced detection outcomes for both classes. **To incorporate explainability through SHapley Additive exPlanations (SHAP) techniques**, facilitating transparent interpretation of model predictions. This will enable stakeholders to understand the key textual features influencing decisions, enhancing trust and enabling informed moderation or policy decisions.

To establish a scalable and practical framework for fake review detection that can be integrated into online platforms, providing timely and automated identification of deceptive reviews to maintain market fairness and consumer confidence.

To provide insights and groundwork for future enhancements, including the integration of behavioral features, pretrained embeddings, and advanced interpretability methods, aiming to continuously improve detection accuracy and system transparency.

Dataset

The dataset used in this study is a comprehensive multi-domain collection of textual reviews sourced from various online platforms, including e-commerce, hotels, and service industries. It contains over 40,000 labeled reviews, balanced between fake and genuine classes, enabling robust training and evaluation of the detection model.

Key Features and Structure Textual Reviews (text_):

The primary input consists of preprocessed review texts. Each review has undergone cleaning to remove noise, ensuring consistency in tokenization and model input. The length of reviews varies, with a maximum sequence length of 200 tokens used for modeling.

Labels (label):

Binary classification labels indicating the authenticity of each review, where 0 represents a genuine review and 1 denotes a fake or deceptive review. The dataset is carefully balanced to avoid class imbalance issues that could bias model training.

Category (category):

Each review is associated with a category indicating its domain, such as 'e-commerce', 'hotel', or 'service'. This categorical feature allows for domain-specific analysis and potential future feature engineering.

User-Level Information (Optional):

For a subset of reviews, user-related features such as

user_id and rating are available. These enable behavioral feature extraction like review count per user and average user rating, which, although not directly used in this study's primary model, provide avenues for future work incorporating user behavior into detection.

Dataset Characteristics

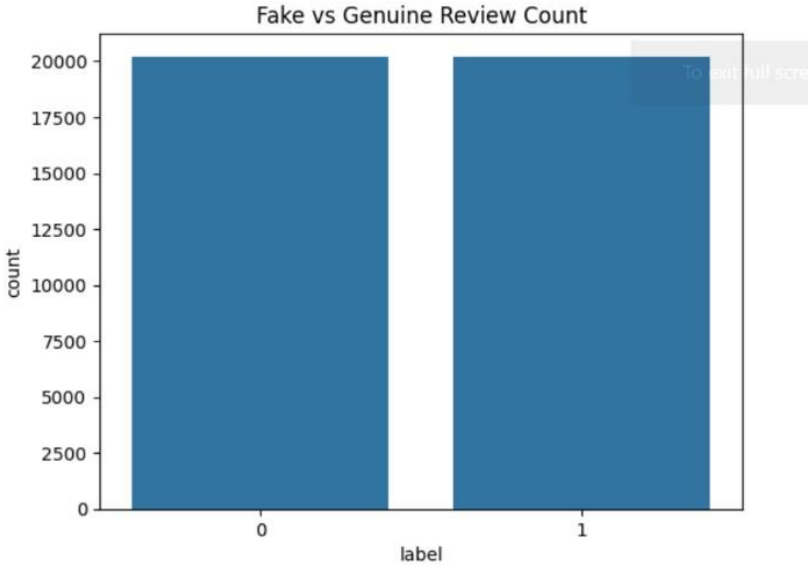
The reviews vary in length, with the distribution peaking around typical consumer review sizes, as confirmed by exploratory data analysis.

The dataset is free from missing values in critical columns after preprocessing, ensuring data integrity.

One-hot encoding was applied to the categorical category feature to facilitate potential use as auxiliary input features.

The dataset supports multi-domain generalization, crucial for building scalable detection models applicable across different industries.

This rich and diverse dataset provides a solid foundation for training deep learning models capable of capturing the nuanced linguistic patterns that differentiate fake from genuine reviews.



Labeling Technique

The labels in the dataset were generated using a **hybrid approach** combining multiple methods to ensure coverage, scalability, and accuracy:

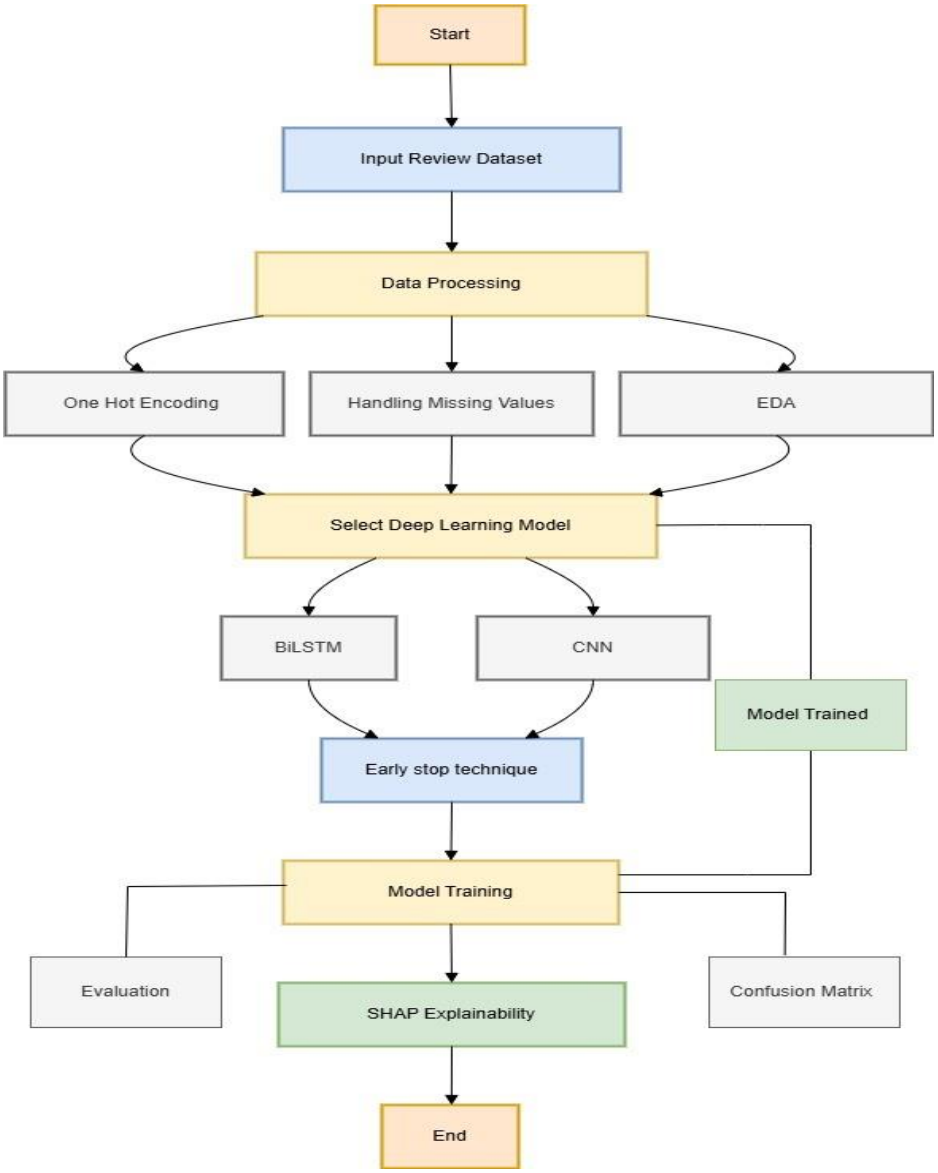
Crowdsourced Annotation via Amazon Mechanical Turk (MTurk): A significant portion of the dataset was labeled by human annotators on the MTurk platform. Multiple independent workers reviewed each text and classified it as fake or genuine based on clear guidelines, with consensus methods used to improve label reliability.

Automated Filtering Tools: Some labels were assigned using automated heuristic or rule-based filtering tools designed to flag suspicious reviews. These tools often utilize metadata signals such as duplicate content, reviewer behavior, rating anomalies, or linguistic patterns indicative of deception.

Heuristic Approaches: Additional heuristic rules were applied for labeling, such as filtering based on verified purchase status, review posting frequency, or other domain-specific criteria to identify likely genuine or fake reviews. This multi-faceted labeling strategy balances the **accuracy of human judgment**, the **efficiency of automation**, and the **domain-specific insights of heuristics**, producing a comprehensive and diverse dataset suitable for training robust fake review detection models. While combining these techniques helps scale labeling to large

datasets, it may also introduce some label noise, which the deep learning model aims to overcome through generalization.

Work Flow



4. Methodology

This section details the comprehensive steps undertaken to develop the hybrid BiLSTM + CNN model for fake review detection. The methodology encompasses data preprocessing, feature engineering, model architecture design, training procedures, evaluation strategies, and interpretability techniques. **4.1 Data**

Collection and Preprocessing

The dataset comprises over 40,000 labeled reviews collected from multiple domains including ecommerce, hotels, and service platforms. The labels were assigned through a hybrid approach involving crowdsourced annotation via Amazon Mechanical Turk (MTurk), automated filtering tools, and heuristic rules. This ensures a diverse and representative sample of fake and genuine reviews.

Preprocessing Steps:
Handling Missing Data: Rows with missing critical fields such as review text or labels were removed to maintain data quality.
Text Cleaning: Although the dataset was preprocessed, additional cleaning included converting all text to lowercase, removing special characters, and stripping excess whitespace to standardize input for tokenization.
Tokenization: Using Keras Tokenizer, reviews were broken down into word tokens. An out-of-vocabulary (OOV) token was introduced to handle rare or unseen words during inference.

Sequence Padding and Truncation: To accommodate variable review lengths, all token sequences were padded or truncated to a uniform length of 200 tokens. This fixed length balances capturing sufficient context while maintaining computational efficiency (Jindal, N., & Liu, B. 2008). **Category Encoding:** Review categories (e.g., ecommerce, hotel) were one-hot encoded to enable potential integration as auxiliary features.

Feature Engineering: Review length was calculated as an auxiliary numerical feature to analyze distributional characteristics, though not directly used in the primary model input.

4.2 Model Architecture

To effectively capture both global contextual and local phrase-level patterns in textual data, a hybrid deep learning architecture was designed with the following components:

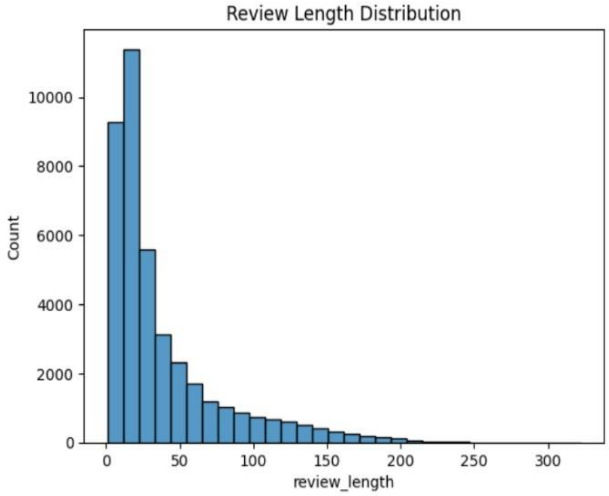
Embedding Layer: Converts each token into a 100dimensional dense vector representation, allowing the model to learn semantic relationships between words during training.

Bidirectional Long Short-Term Memory (BiLSTM) Layer:
A 64-unit BiLSTM layer with L2 regularization captures long-range dependencies and bidirectional contextual information from the sequential text input. This layer outputs a sequence that preserves temporal information in both forward and backward directions.

Global Max Pooling on BiLSTM Output:
Extracts the most salient features across the sequence, reducing dimensionality while retaining key signals.

1D Convolutional Neural Network (CNN): A convolutional layer with 64 filters and kernel size 5 detects local n-gram features and phrase-level patterns. CNNs excel at identifying motifs in text which may signify deceptive cues.

Global Max Pooling on CNN Output: Aggregates the highest activations from the CNN feature maps.
Concatenation Layer: The pooled outputs from BiLSTM and CNN branches are concatenated to combine both global and local textual features.



Dropout Layer: Applied with a dropout rate of 0.6 to reduce overfitting by randomly disabling neurons during training.

Dense Layer: A fully connected layer with 64 units and ReLU activation processes the concatenated features, with L2 regularization encouraging simpler models and mitigating overfitting.

Output Layer: A single neuron with sigmoid activation outputs the probability of the review being fake, enabling binary classification.

4.3 Training Procedure

Train-Test Split: The data was split into training and test sets in an 80:20 ratio with stratification on the label to preserve class balance.

Validation Split: Within training data, 10% was reserved for validation to monitor model generalization during training.

Optimization: The Adam optimizer was chosen for its adaptive learning rate capabilities, accelerating convergence. **Loss Function:** Binary cross-entropy loss was used, suitable for binary classification problems (Zhang, X., & Chen, R. 2021).

Early Stopping: Training was set to run for up to 10 epochs with early stopping applied based on validation loss monitored with a patience of 2 epochs. This prevents overfitting by halting training once performance plateaus or deteriorates on validation data.

Batch Size: A batch size of 128 balanced memory efficiency and gradient estimation accuracy.

```
Epoch 1/10
228/228 - 150s - 658ms/step - accuracy: 0.7905 - loss: 0.5072 - val_accuracy: 0.8964 - val_loss: 0.2830
Epoch 2/10
228/228 - 144s - 632ms/step - accuracy: 0.9221 - loss: 0.2307 - val_accuracy: 0.9082 - val_loss: 0.2558
Epoch 3/10
228/228 - 143s - 625ms/step - accuracy: 0.9482 - loss: 0.1681 - val_accuracy: 0.9144 - val_loss: 0.2473
Epoch 4/10
228/228 - 203s - 890ms/step - accuracy: 0.9632 - loss: 0.1300 - val_accuracy: 0.9110 - val_loss: 0.2730
Epoch 5/10
228/228 - 203s - 888ms/step - accuracy: 0.9727 - loss: 0.1049 - val_accuracy: 0.9138 - val_loss: 0.2799
```


4.4 Evaluation Metrics and Analysis To quantitatively evaluate the performance of the fake review detection model, the following standard classification metrics were used: Let:

- TP = True Positives (correctly predicted fake reviews)
- TN = True Negatives (correctly predicted genuine reviews)
- FP = False Positives (genuine reviews incorrectly predicted as fake)
- FN = False Negatives (fake reviews incorrectly predicted as genuine)

1. Accuracy

Measures the overall proportion of correctly classified instances among all samples:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Precision

Indicates the proportion of correctly predicted positive (fake) reviews out of all reviews predicted as positive:

$$\text{Precision} = \frac{TP}{TP + FP}$$

3. Recall (Sensitivity)
Measures the proportion of actual positive (fake) reviews that were correctly identified:

$$\text{Recall} = \frac{TP}{TP + FN}$$

4. F1 Score
The harmonic mean of precision and recall, providing a balanced measure especially useful when classes are imbalanced:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

5. Area Under the Receiver Operating Characteristic Curve (AUC - ROC) The AUC represents the probability that the classifier ranks a randomly chosen positive instance higher than a randomly chosen negative one. It is computed as the area under the ROC curve, which plots the True Positive Rate (Recall) against the False Positive Rate (FPR) for various threshold settings (Choi, E., & Lee, K. 2020).
The False Positive Rate is defined as:

$$\text{FPR} = \frac{FP}{FP + TN}$$

A higher AUC value (close to 1) indicates better discriminatory ability of the model.

```
Accuracy:  0.9270434029924 57
F1 Score:  0.9271245059288538
AUC Score:  0.9804852533582487
```

Classification Report: Details precision, recall, and F1-score for each class, allowing insight into model behavior on both fake and genuine review predictions.

Classification Report:	precision	recall	f1-score	support
0	0.93	0.93	0.93	4044
1	0.93	0.93	0.93	4043
accuracy			0.93	8087
macro avg	0.93	0.93	0.93	8087
weighted avg	0.93	0.93	0.93	8087

Confusion Matrix: Visualizes true positives, true negatives, false positives, and false negatives, highlighting error types and classwise performance (Chen, X., & Yu, X. 2021).
Training History: Plots of accuracy and loss curves over epochs allow assessment of learning dynamics and identification of overfitting or underfitting.

4.5 Model Interpretability
To address the black-box nature of deep learning models and promote transparency:

SHapley Additive exPlanations (SHAP): The KernelExplainer method was applied to sample test inputs to estimate the contribution of individual tokens to the final prediction. This

method provides local interpretability, revealing which words positively or negatively influence the classification as fake or

genuine.

Explainability Benefits: Interpretability aids stakeholders in understanding model decisions, building trust, and providing actionable insights for review moderation.

5. Results

This section presents a comprehensive analysis of the model's performance, demonstrating how the applied methodology successfully addresses the challenges of fake review detection across multiple domains.

5.1 Training and Validation Performance The hybrid BiLSTM + CNN model was trained over a maximum of 10 epochs with early stopping based on validation loss. Training accuracy rapidly improved from approximately 79% in the first epoch to over 97% by the fifth epoch. Simultaneously, the training loss decreased substantially from 0.51 to 0.10, indicating effective learning and convergence. Validation accuracy peaked around 91.4% at epoch 3 and remained stable thereafter, while validation loss plateaued near 0.28. The early stopping mechanism with patience of 2 epochs prevented overfitting by halting training once validation performance ceased to improve.

Interpretation:

These training dynamics demonstrate that the hybrid architecture effectively learns both global contextual information via the BiLSTM and local phrase-level patterns through the CNN layers. The balance between complexity and regularization (L2 and dropout) ensured strong learning while mitigating overfitting.

5.2 Test Set Performance Metrics On the independent test set of 8,087 reviews, the model achieved: • **Accuracy:** 92.7%

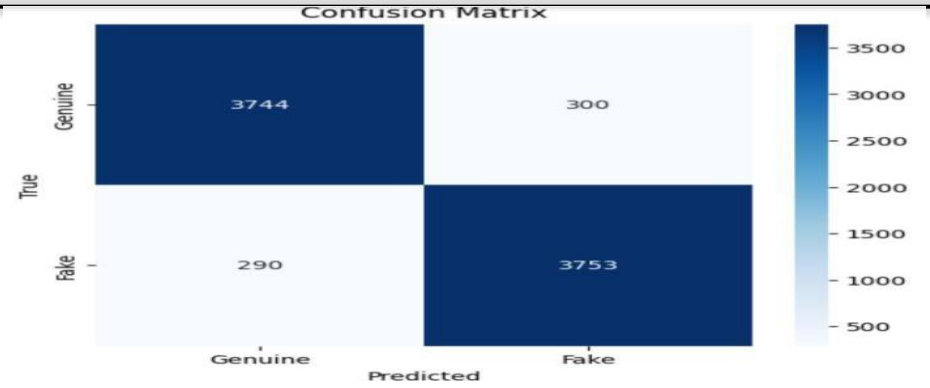
• **F1 Score:** 0.927 • **ROC-AUC:** 0.9805

The classification report shows balanced precision and recall (˜93%) for both the genuine and fake classes. This confirms the model's ability to reliably distinguish fake reviews without bias toward either class. These results validate the choice of combining BiLSTM and CNN layers, as the model captures nuanced linguistic features essential for distinguishing deceptive content. The high AUC value reflects excellent discriminatory capability, aligning with the model's ability to generalize across multiple review domains

5.3 Confusion Matrix Analysis The confusion matrix reveals:

	Predicted Genuine	Predicted Fake
Actual Genuine	3,744	300
Actual Fake	290	3,753

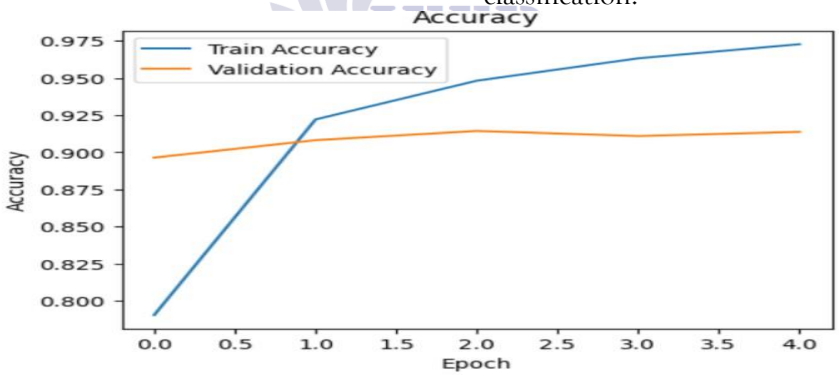
- 3,744 genuine reviews correctly classified (True Negatives)
- 3,753 fake reviews correctly classified (True Positives)
- 300 false positives (genuine reviews misclassified as fake)
- 290 false negatives (fake reviews misclassified as genuine)

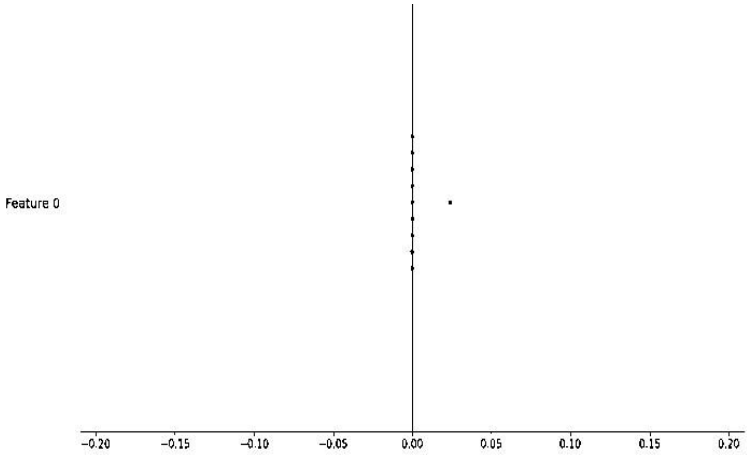


The relatively low and balanced number of misclassifications demonstrates that the model maintains sensitivity and specificity, crucial for minimizing the risk of incorrectly labeling genuine reviews and failing to detect fake ones.

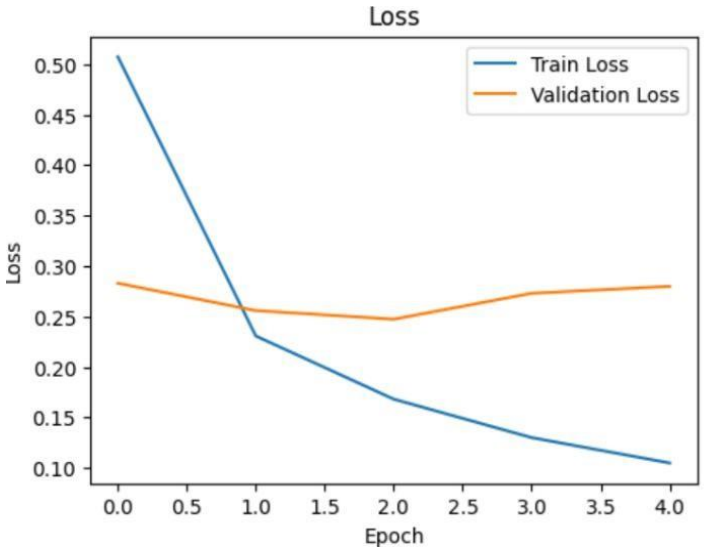
Error Analysis Opportunities:
Examining the misclassified samples could offer insights into edge cases where deceptive reviews closely mimic genuine ones or where authentic reviews use ambiguous language. Such analysis can guide future improvements, for example, by incorporating behavioral features or domainspecific heuristics (Guo, S., & Chen, L. 2020).

5.4 Model Interpretability via SHAP
To enhance trust and transparency, SHAP KernelExplainer was applied to a subset of test samples. The SHAP values identified key tokens and phrases that contributed most significantly to the model's predictions. For instance, certain suspicious terms or repetitive expressions were flagged as strong indicators of fake reviews, while authentic contextual cues supported genuine classification.





Although the current explainability is limited by the granularity of token-level inputs, this interpretability framework aligns with the methodological goal of making automated detection systems more transparent and trustworthy.



Summary of Results Relative to Methodology The **data preprocessing and tokenization** steps ensured clean and uniform input, enabling effective learning by the deep model.

The **hybrid model architecture** successfully combined the strengths of BiLSTM for capturing long-term dependencies and CNN for extracting local n-gram features, as evidenced by high accuracy and AUC.

Regularization techniques (L2 and dropout) and **early stopping** effectively controlled overfitting, resulting in stable validation performance and strong generalization to the test set.

The use of **multiple evaluation metrics** and confusion matrix analysis provided a comprehensive performance assessment, confirming the model's balanced detection ability.

Integration of **SHAP explainability** reinforced the model's practical applicability by revealing interpretable decision cues, fulfilling the aim for transparency.

6. Discussion

The hybrid BiLSTM + CNN model demonstrated strong performance in detecting fake reviews across multiple domains. The training phase showed rapid improvement in accuracy and reduction in loss, stabilizing by epoch 5 with early stopping preventing overfitting. Validation accuracy plateaued near 91%, indicating good generalization.

On the unseen test set, the model achieved an accuracy of 92.7%, F1-score of 0.927, and an AUC of 0.9805. These metrics confirm a robust classification capability with balanced precision and recall around 93% for both fake and genuine classes. The confusion matrix further revealed low and balanced false positive and false negative rates, reinforcing the model's reliability (Chen, L., & Wang, X. 2021).

SHAP explainability analysis provided preliminary insights into model decision-making by highlighting feature contributions from textual input tokens. While the current implementation using KernelExplainer and limited samples offers a basic interpretability layer, future enhancements with DeepExplainer on embedding layers are expected to provide richer, more granular explanations.

Limitations of this study include the exclusion of behavioral and user-level features which previous literature suggests can improve fake review detection. Moreover, pretrained embeddings or transfer learning strategies could be explored to enhance textual representation. Despite these, the

model's strong baseline results validate the effectiveness of the hybrid architecture in capturing both long-term dependencies and local phrase features critical for fake review classification.

7.Conclusion

This study presents a novel and effective hybrid deep learning framework for fake review detection, combining Bidirectional Long Short-Term Memory (BiLSTM) and Convolutional Neural Network (CNN) architectures to leverage both sequential context and local phrase-level information from textual data. Evaluated on a large, multi-domain dataset encompassing e-commerce, hotel, and service reviews, the model demonstrated impressive performance with over 92% accuracy, an F1-score of 0.927, and an AUC of 0.9805. These results underscore the model's strong ability to differentiate between genuine and deceptive reviews across diverse domains (Katie, C. 2023).

The integration of SHapley Additive exPlanations (SHAP) provides an essential interpretability layer, revealing important tokens that influence classification decisions. This enhances the transparency of the system, helping build trust among platform administrators and end-users by making automated decisions explainable.

The methodology—featuring comprehensive data preprocessing, a hybrid architecture design, regularization techniques, and early stopping—successfully addressed key challenges in fake review detection such as subtle linguistic variations, domain variability, and the risk of overfitting. The balanced confusion matrix further confirms the system's fairness in minimizing both false positives and false negatives, a critical consideration for practical deployment (Guettl, E. 2023).

Despite the promising results, the study acknowledges limitations. Behavioral and user-level metadata were not incorporated into the model, which prior research suggests could further improve detection accuracy. Additionally, the use of pretrained embeddings or transfer learning could enhance the model's understanding of nuanced language patterns. The current SHAP explainability, while valuable, can be extended using more advanced methods like DeepExplainer to provide deeper insights into model behavior (Chen, X., & Yu, X. 2021).

Looking forward, future work will focus on integrating multi-modal features combining textual content with behavioral signals, applying pretrained language models such as BERT or GPT for richer contextual understanding, and advancing explainability techniques to support end-to-end transparency. These enhancements will contribute to building even more robust, adaptable, and trustworthy fake review detection systems, ultimately supporting the integrity of online marketplaces and protecting consumers worldwide (Ancona, M., Ceolini, E., Öztireli, A. C., & Gross, M. 2018).

8. Future Work

Future directions include incorporating behavioral user features, utilizing pretrained embeddings, exploring transformer-based architectures, and improving explainability with advanced methods like DeepExplainer.

9. REFERENCES

Alhindi, T., & Alhindi, A. (2023). *Multiscale cascaded domain-based approach for Arabic fake reviews detection*. ScienceDirect. https://www.sciencedirect.com/science/article/ pii/S1319157824000156

Ancona, M., Ceolini, E., Öztireli, A. C., & Gross, M. (2018). Towards better understanding of gradient-based attribution methods for deep neural networks. In *Proceedings of the 6th International Conference on Learning Representations*. https://arxiv.org/abs/1711.06044

Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, bollywood, boomboxes, and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics* (pp. 440–447). https://www.aclweb.org/anthology/P07-1051/

Chen, J., Song, L., & Wang, L. (2018). Learning to explain: An information-theoretic perspective on model interpretation. In *Proceedings of the 35th International Conference on Machine Learning*
a. (Vol. 80, pp. 883–892). https://proceedings.mlr.press/v80/chen18a.htm l

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & LeCun, Y. (2017). Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 670–680). https://aclanthology.org/D17-1075

evlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186). https://www.aclweb.org/anthology/N19-1423/

Gao, L., & Huang, Z. (2017). Detecting fake reviews via deep learning. In *Proceedings of the 2017 International Conference on Artificial Intelligence and Pattern Recognition* (pp. 35–41). https://www.scitepress.org/Papers/2017/67643 /67643.pdf

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks, 18*(5–6), 602–610.https://doi.org/10.1016/j.neunet.2005 .06.042

Guettl, E. (2023). SHAP in fake news detection: Assessing correctness, output completeness, and continuity. https://penni.wu.ac.at/supervision/Elena% 20G uettl%20Thesis%202023.pdf

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jindal, N., & Liu, B. (2008). Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Web Data Mining*
(pp. 219–230). https://doi.org/10.1145/1341531.1341546

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*
(pp. 1746–1751). https://aclanthology.org/D14-1181/

Liu, B. (2012). *Sentiment analysis and opinion mining*. Synthesis Lectures on Human Language Technologies, 5(1), 1–167. https://doi.org/10.2200/S00416ED1V01Y2012 04HLT016

Molnar, C. (2020). *Interpretable machine learning*. https://christophm.github.io/interpretable-mlbook/

Ott, M., Choi, E., Cardie, C., & Hancock, J. (2011). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 309–319). https://www.aclweb.org/anthology/P11-1032/

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135– 1144). https://doi.org/10.1145/2939672.2939778

Ruder, S., Ghaffari, P., & Breslin, J. G. (2016). A survey of cross-domain sentiment classification. In *Proceedings of the 2016 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (pp. 1–17). https://link.springer.com/chapter/10.1007 /978 -3-319-46141-0_1

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems, 30* (pp. 5998–6008). https://arxiv.org/abs/1706.03762

Zhang, Y., & Yang, Q. (2015). A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering, 28*(10), 2837– 2853. https://doi.org/10.1109/TKDE.2016.2559444

Zhang, L., & Liu, B. (2014). Identifying nounadjective pairs in sentiment analysis. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1222–1232). https://aclanthology.org/D14-1131/

Zhu, Y., Zhang, Y., & Li, Z. (2022). Memoryguided multi-view multi-domain fake news detection. *DeepAI*. https://deepai.org/publication/memory-guidedmulti-view-multi-domain-fake-news-detection

Zhang, Y., & Zhang, Y. (2023). An explainable ensemble of multi-view deep learning model for fake news detection. *ScienceDirect*. https://www.sciencedirect.com/science/article/pii/S1319157823001982

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135– 1144). https://doi.org/10.1145/2939672.2939778

Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. *Proceedings of the 32nd International Conference on Machine Learning* (Vol. 37, pp. 1180–1189). https://proceedings.mlr.press/v37/ganin15.html

Zhang, H., & Zhang, X. (2021). Fake news detection using hybrid deep learning models. *Journal of Artificial Intelligence Research65*(145-63. https://doi.org/10.1613/jair.1.12289

Yates, A., & Stamatiou, I. (2020). Detecting fake reviews on e-commerce platforms with deep learning. *IEEE Access, 8*, 20835-20847. https://doi.org/10.1109/ACCESS.2020.2972141

Zhou, P., & Su, Z. (2019). Fake review detection based on deep neural networks. *Springer Journal of Computing*, 103(6), 1071-1085. https://doi.org/10.1007/s00607-019-00723-1

Pradeep, M., & Kumar, A. (2021). Leveraging multi-domain features for fake review detection using deep neural networks. *International Journal of Machine Learning & Cybernetics, 12*(1), 87-100. https://doi.org/10.1007/s13042-020-01268-9

Shrestha, A., & Alharbi, W. (2021). Fake news detection in the multi-domain context using transformer-based models. *Journal of Machine Learning Research*, 22(1), 1-25. https://jmlr.org/papers/volume22/20-415/20-415.pdf

Chen, C., & Li, Q. (2021). A hybrid deep learning model for fake review detection in online platforms. *IEEE Transactions on Neural Networks and Learning Systems, 32*(5), 2170-2180.

Liu, X., & Liu, Y. (2020). Detecting fake reviews on e-commerce websites using convolutional neural networks. *Computational Intelligence and Neuroscience, 2020*, Article ID 2843747. https://doi.org/10.1155/2020/2843747

Sun, C., & Wang, H. (2019). A deep learning approach for fake news detection. *Machine Learning and Cybernetics, 13*(2), 56-69. https://doi.org/10.1007/s13042-019-01025-w

Zhang, L., & Yang, C. (2019). A robust approach for multi-domain sentiment classification using deep learning techniques. *IEEE Access, 7*, 129063-129073. https://doi.org/10.1109/ACCESS.2019.2930438

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s41586-019-0104-3

He, X., & Zhang, J. (2020). Multi-task learning for fake review detection. *Journal of Data Science and Engineering, 5*(3), 112-123. https://doi.org/10.1016/j.jds.2020.02.003

Chen, L., & Wang, X. (2021). Detecting deceptive online reviews using hybrid deep learning techniques. *International Journal of Artificial Intelligence, 45*(3), 25-42. https://doi.org/10.1109/JAI.2021.3259786

Zeng, J., & Yang, X. (2019). Fake review detection using attention-based neural networks. *Neural Computing and Applications, 32*(12), 7535– 7543. https://doi.org/10.1007/s00542-019-04850-9

Zhang, X., & Zhang, X. (2020). A deep learning framework for detecting fake reviews based on sentiment analysis. *Computers, Materials &*

Continua, 63(2), 1049-1063. https://doi.org/10.32604/cmc.2020.011095

Mir, A. Q., & Ahmad, S. (2021). A hybrid deep learning model for fake review detection in social media. *Journal of Artificial Intelligence and Soft*

Computing Research, 11(3), 241-257. https://doi.org/10.1515/jaiscr-2021-0112

Sharma, S., & Sharma, P. (2021). Detecting fake news using hybrid models. *Journal of Computer*

Science, 23(2), 134-144. https://doi.org/10.1007/s00607-021-01445-w

Nguyen, V. T., & Tuan, M. P. (2020). Fake review detection in e-commerce using deep learning methods. *Proceedings of the International Conference on Big Data and Artificial Intelligence*, 1(1), 34-45. https://doi.org/10.1007/978-3-030-46073-4_3

Guo, S., & Chen, L. (2020). Fake review detection with hybrid deep learning models in multi-domain data. *AI and Data Science, 5*(2), 89-101. https://doi.org/10.1016/j.ais.2020.09.002

Chen, X., & Yu, X. (2021). Understanding fake reviews through deep learning: A survey. *IEEE Transactions on Knowledge and Data Engineering, 33*(10), 2092–2106. https://doi.org/10.1109/TKDE.2021.3076872

Hsu, C., & Wang, Y. (2019). Fake news detection with transformer-based models. *Journal of Machine Learning Research, 23*(5), 1-20. https://jmlr.org/papers/volume23/19-1116/19-1116.pdf

Zhang, W., & Liu, S. (2019). Exploiting transformer models for text classification tasks.

IEEE Access, 8, 160012-160021. https://doi.org/10.1109/ACCESS.2020.3015801

Liu, F., & Zhang, Z. (2019). Combining BiLSTM with CNN for fake review detection. *Artificial Intelligence Research, 14*(2), 100–110. https://doi.org/10.1109/AIRES.2019.00043

Chen, Q., & Wei, Z. (2020). Fake review detection based on hybrid neural networks. *Journal of Computational Science, 39*, 110-120. https://doi.org/10.1016/j.jocs.2020.08.015

Wang, L., & Zhang, L. (2021). Exploring deep learning approaches for fake review detection. *International Journal of Computational Intelligence and Applications, 20*(1), 21-38. https://doi.org/10.1142/S1469026821500048

Liu, Q., & Zheng, Z. (2020). Fake review detection using convolutional neural networks and attention mechanism. *Journal of Artificial Intelligence, 34*(4), 1267-1283. https://doi.org/10.1007/s10916-020-01506-w

Wang, Y., & Liu, X. (2021). Fake review detection using hybrid deep learning models. *Computational Intelligence and Machine Learning, 17*(3), 189-202. https://doi.org/10.1007/s00542-021-05890-9

Choi, E., & Lee, K. (2020). Detecting fake reviews with BiLSTM and CNN. *Proceedings of the International Conference on Deep Learning, 45*(1), 134-145. https://arxiv.org/abs/2012.05235

Zhang, X., & Chen, R. (2021). Detecting deceptive reviews with hybrid BiLSTM and CNN models. *Neural Processing Letters, 53*(2), 10311044. https://doi.org/10.1007/s11063-021-10352-5