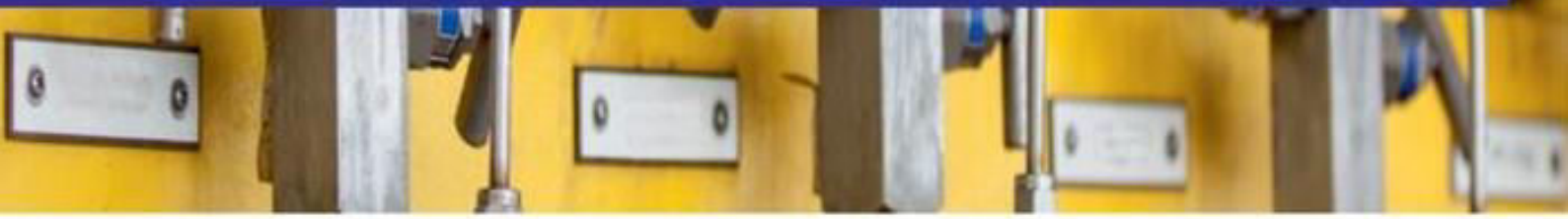
IJAREEIE

e-ISSN: 2278-8875
p-ISSN: 2320-3765

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

Volume 15, Issue 10, October 2026



INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 9.867

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering (IJAREEIE)

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 9.867 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 10, October 2026 ||

| DOI:10.15662/IJAREEIE.2026.1510013 |

ReviewLens: An Explainable NLP Framework for Suspicious Online Review Detection Using BiLSTM and Token-Level Evidence

Rashi Manjrekar, Chandana Vanage, Faiz Rangari, Dr Nayana Mahajan

Dept. of Electronics and Computer Science, Vidyalankar Institute of Technology, Mumbai, India

ABSTRACT: Online product reviews are an important information source for consumers, businesses, and e-commerce platforms, but fabricated, promotional, misleading, and spam reviews reduce the reliability of review-based decision-making. This paper presents ReviewLens, an explainable Natural Language Processing (NLP) framework for detecting suspicious online product reviews. Unlike systems that provide only a binary fake/genuine prediction, ReviewLens integrates URL-based review acquisition, text preprocessing, sequential deep learning, probabilistic classification, and explainable artificial intelligence. The primary classifier is a Bidirectional Long Short-Term Memory (BiLSTM) network trained on 270,727 labeled reviews divided, using a reviewer-disjoint strategy, into 189,189 training, 40,762 validation, and 40,776 held-out test reviews. The decision threshold (T = 0.62) is selected exclusively on validation data to prevent test-set leakage. On the held-out test set the model achieved 89.41% accuracy, 74.81% precision, 75.01% recall, 74.91% F1-score, 92.32% ROC-AUC, and 80.89% PR-AUC. Token/span occlusion estimates how textual components influence the suspiciousness probability, and a human-readable explanation layer highlights influential phrases and linguistic indicators such as promotional language, exaggerated claims, urgency, absolute statements, and excessive punctuation. These indicators serve as supporting evidence rather than proof of deception.

KEYWORDS: Fake Review Detection, Natural Language Processing, BiLSTM, Explainable AI, Token Attribution, Suspiciousness Detection.

I. INTRODUCTION

The rapid growth of e-commerce has made online customer reviews an essential source of information for consumer purchasing decisions and business analysis. Consumers rely on reviews to evaluate product quality, compare alternatives, and assess the experiences of previous buyers, while businesses use customer feedback to understand preferences and improve products and services. However, fabricated, misleading, promotional, and spam reviews can undermine the reliability of online feedback and reduce consumer trust. Jindal and Liu [1] established opinion spam as a distinct research problem, Ott et al. [2] demonstrated that deceptive reviews may exhibit distinguishable linguistic patterns, and Xu and Zhao [3] investigated deep linguistic features for identifying deceptive opinions.

Existing research on suspicious review detection has explored linguistic feature analysis, reviewer behavior modeling, deep learning, and contextual language representations. Mukherjee et al. [4] investigated the behavioral footprints of opinion spammers, and Fang et al. [5] reviewed developments in fake-review detection. Deep learning architectures such as CNN-BiLSTM, attention-based LSTM, and transformer-enhanced models have also been studied for their ability to learn textual features and contextual patterns from review data [6]–[8]. Many such systems, however, return only a label or a probability without showing which parts of the text influenced the result.

This paper proposes ReviewLens, an explainable NLP framework for suspicious online product review detection. A user submits a product or review-page URL; the system extracts publicly accessible review text, preprocesses it, and obtains a suspiciousness score from a trained BiLSTM classifier. The score is converted to a class using a validation-selected threshold. Token/span occlusion then identifies the text that influenced the score, and a human-readable explanation layer presents that evidence in plain language. The main contributions are: (i) a BiLSTM-based suspicious-review classifier evaluated with a reviewer-disjoint train/validation/test protocol; (ii) validation-only threshold selection that keeps the test set untouched; (iii) instance-level evidence through token/span occlusion with cautious, human-readable explanations; and (iv) an end-to-end pipeline from URL-based review collection to a user interface. The output is a model classification, not proof of fraud.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 9.867 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 10, October 2026 ||

| DOI:10.15662/IJAREEIE.2026.1510013 |

II. LITERATURE REVIEW

Traditional machine-learning approaches rely on manually engineered textual, linguistic, and behavioral features [5]. Deep-learning architectures, including CNN-BiLSTM, attention-based LSTM, and hybrid recurrent models [6]–[8], learn complex textual patterns directly from review text. Graph-based methods [9], [10] capture relationships among reviewers and products, while Transformer architectures [12] and BERT [11] improve contextual language representation. However, these approaches may require additional metadata, computational resources, or explanation mechanisms.

Despite these developments, many review-detection systems provide only a classification label or probability without clearly identifying the textual evidence behind the prediction. This limits interpretability, particularly because suspicious language does not necessarily prove deception. Behavioral and graph-based approaches provide valuable additional evidence but depend on reviewer metadata and structured relationships that may not be available when analyzing an individual product URL. Evaluation is further complicated by noisy or proxy-generated labels, class imbalance, and data leakage when reviews from the same reviewer appear in different partitions.

The research gap addressed by ReviewLens is the integration of sequential deep-learning-based suspicious-review classification with instance-level textual evidence and human-readable explanations, supported by a reviewer-disjoint evaluation protocol and validation-only threshold selection. Unlike a system that reports only a binary label, ReviewLens aims to make the prediction process more transparent while explicitly recognizing that model-influential phrases and linguistic indicators do not constitute proof of fraud.

III. SYSTEM OVERVIEW AND DATA COLLECTION

A. System Overview

ReviewLens estimates whether an online review resembles the suspicious class represented in a labeled research dataset. It accepts a product or review URL when review text is publicly accessible, extracts the available reviews, preprocesses each text, obtains a suspiciousness score from the trained BiLSTM, and presents the prediction together with influential text spans and a human-readable explanation. The end-to-end workflow is shown in Fig. 1.

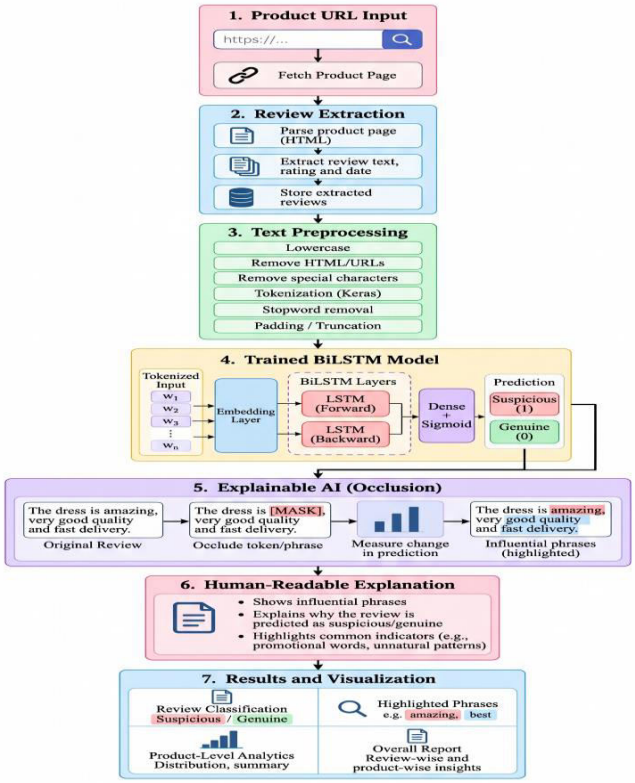


Fig. 1. ReviewLens end-to-end architecture, from URL input to results and visualization.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 9.867 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 10, October 2026 ||

| DOI:10.15662/IJAREEIE.2026.1510013 |

B. Offline Labeled Dataset

The model-development dataset was derived from a public Amazon fake-review research dataset. It contains 270,727 labeled review records with fields including review_id, review_text, title, rating, label, reviewer_id, product_id, review_date, and source. The prepared counts are 214,509 class-0 and 56,218 class-1 reviews; class 0 is treated as genuine/non-suspicious and class 1 as suspicious. Labels are research/proxy labels rather than perfect ground truth: the preparation process uses available reviewer and product indicators to construct review-level labels and removes exact duplicate review texts. Reported performance therefore measures agreement with these prepared labels and does not establish that every predicted review is objectively fraudulent.

C. Runtime Review Collection from a URL

At runtime, the user enters a product or review-page URL. The backend requests the publicly accessible page and searches exposed structured data and page markup (JSON-LD, embedded JSON, or review-related HTML elements) for review objects and review text, retaining titles, ratings, identifiers, and other metadata when exposed. Extracted records are validated: an item must contain review text rather than only a rating or product description; whitespace is normalized and markup artifacts are removed; empty records are discarded and identical texts are deduplicated where appropriate; and the source URL and available metadata are retained for traceability. Newly collected reviews are treated as inference examples and are never silently added to the labeled training or test sets.

A URL may yield no reviews because some sites render content only through JavaScript, require authentication, rate-limit requests, or block automated access. ReviewLens does not bypass CAPTCHA, login, or access controls; if review text is not publicly available, the correct result is a clear extraction error rather than a fabricated review.

D. Reviewer-Disjoint Dataset Partitioning

The prepared dataset is divided using a reviewer-disjoint strategy, which prevents reviews from the same reviewer from appearing in more than one partition. This reduces the chance that the model recognizes reviewer-specific patterns instead of generalizing to unseen reviewers. The split (Table I, Fig. 2) is approximately 70% training, 15% validation, and 15% test; it is a three-way split rather than a strict 70/30 train/test split.

TABLE I. DATASET PARTITIONS

Partition	Records	Purpose
Training	189,189	Fit model parameters
Validation	40,762	Model selection and threshold selection
Held-out test	40,776	Final performance evaluation
Total	270,727	Prepared dataset

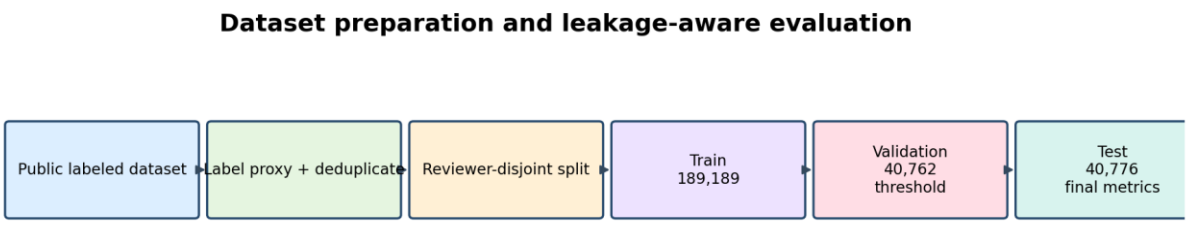


Fig. 2. Dataset preparation, reviewer-disjoint splitting, and final evaluation.

IV. NLP PREPROCESSING AND TEXT REPRESENTATION

A. Normalization and Tokenization

A review is represented as a token sequence R = [w1, w2, …, wn]. The model does not assign separate labels to individual sentences; it classifies the review as a sequence, and the explanation component later examines tokens or short spans to estimate which text influenced the review-level prediction. Raw text is normalized (repeated spaces and line breaks standardized, HTML tags and decoding artifacts removed), tokenized, and converted to integer indices using the saved tokenizer. A vocabulary is constructed from the training data only, so that vocabulary-dependent preprocessing cannot

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 9.867 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 10, October 2026 ||

| DOI:10.15662/IJAREEIE.2026.1510013 |

leak information from the validation or test partitions. Shorter sequences are padded and longer ones truncated to a fixed model length so that reviews can be processed in batches.

Training and inference preprocessing must match exactly. Steps such as lowercasing, stopword removal, stemming, or lemmatization are applied at inference only if they were part of the training pipeline, because applying them only at inference changes token IDs and model behavior. The saved tokenizer and model are always loaded together.

B. Embedding Representation

The tokenizer maps the review to integer IDs, R → [i1, i2, …, in]. An embedding layer maps each ID to a dense vector ei ∈ ℝ^d, and the model receives E = [e1, e2, …, en] rather than raw text. Unlike simple word-frequency representations, embeddings provide a distributed representation that allows the model to learn useful textual patterns during training.

V. BiLSTM CLASSIFIER AND DECISION THRESHOLD

A. BiLSTM Model

ReviewLens employs a Bidirectional Long Short-Term Memory (BiLSTM) network [13] to learn contextual and sequential relationships within review text. An LSTM uses memory cells and gating mechanisms to retain relevant information while reducing the difficulty of learning long-term dependencies in conventional recurrent networks. A BiLSTM consists of a forward LSTM that reads the sequence from left to right and a backward LSTM that reads it from right to left, so that the classifier can use context from both sides of a token. The hidden states are

$$ht(f) = LSTM(f)(et, ht{-}1(f)), \quad ht(b) = LSTM(b)(et, ht{+}1(b)), \quad ht = [ht(f) ; ht(b)] \tag{1}$$

The combined sequence representation h is passed to a dense layer that produces a logit z, and a sigmoid activation yields the suspicious-class probability:

$$p = P(y = 1 \mid R) = \sigma(z) = 1 / (1 + e^{(-z)}) \tag{2}$$

where y = 1 denotes the suspicious class and y = 0 the genuine/non-suspicious class. The model is trained by minimizing the binary cross-entropy loss over N training reviews:

$$L = -(1/N)\ \Sigma i\ [\ yi \log(pi) + (1 - yi) \log(1 - pi)\ ] \tag{3}$$

The training partition fits the model parameters, the validation partition supports model selection and threshold selection, and the held-out test partition is reserved for final evaluation.

B. Threshold and Class Decision

The model outputs a score p between 0 and 1. The evaluation records a threshold T = 0.62, selected on validation data only; the test set was not used for threshold selection. The decision rule is

$$\hat{y} = 1 \text{ if } p \geq T, \text{ otherwise } \hat{y} = 0 \tag{4}$$

A review with p ≥ 0.62 is labeled “Suspicious”, and a review with p < 0.62 is labeled “Likely genuine”. The score is an estimate from the learned model, so the interface uses wording such as “classified as suspicious” or “likely genuine” rather than asserting that a review has been proved fake.

VI. EXPLAINABLE AI AND HUMAN-READABLE EXPLANATION

A. Token/Span Occlusion

ReviewLens estimates text influence through token/span occlusion. The original review is first evaluated to obtain p_original. A token or short span is then masked or removed, and the same trained model evaluates the perturbed review to obtain p_masked. The contribution estimate of span i is

$$Ci = p_original - p_masked,i \tag{5}$$

A positive Ci means that masking the span reduced the suspiciousness score, so the original span supported the suspicious prediction under this perturbation; a negative value means that masking increased the score. This is a local, model-dependent influence estimate, not proof that the text is deceptive. The interface ranks influential tokens/spans and highlights the strongest evidence (Fig. 3). Explanations are based on measured model changes rather than hard-coded phrases that would be shown regardless of the prediction.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 9.867 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 10, October 2026 ||

| DOI:10.15662/IJAREEIE.2026.1510013 |

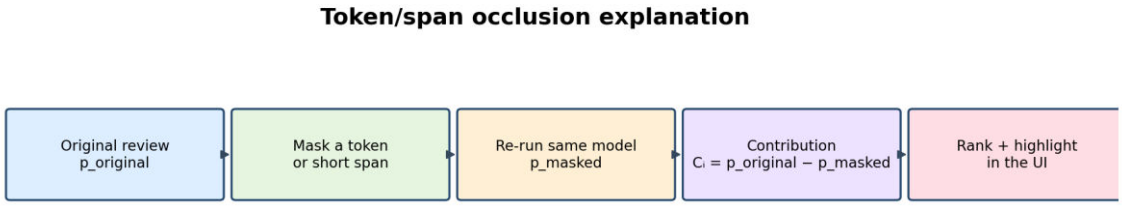


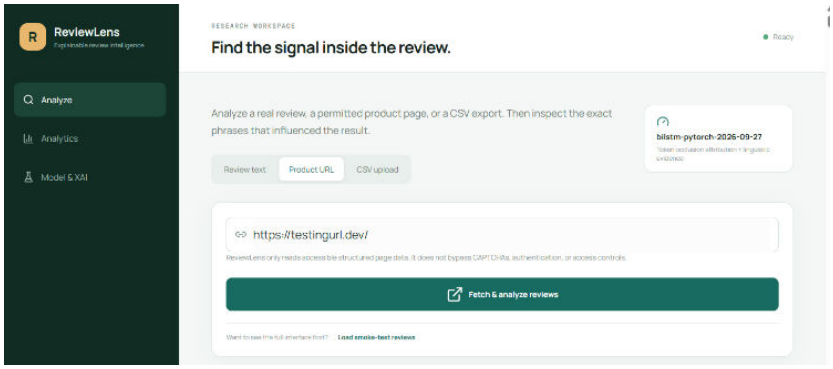
Fig. 3. Token/span occlusion and evidence ranking.

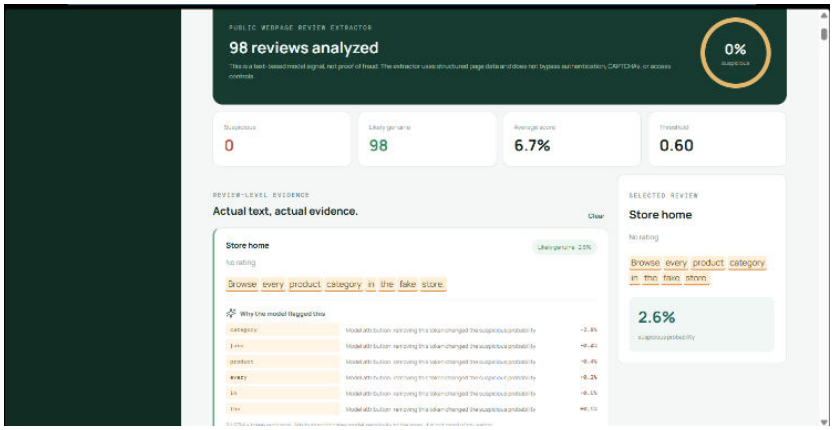
B. Human-Readable Explanation

Occlusion estimates which text affected the model; the natural-language explanation describes those spans in understandable terms. Optional supporting linguistic indicators include promotional imperatives (e.g., “MUST BUY”), exaggerated wording (“BEST EVER”), urgency (“BUY NOW”), absolute claims, and excessive punctuation. These indicators are not independent proof of fraud. For a suspicious prediction, a cautious explanation states that promotional or urgent wording contributed to the model’s suspiciousness score and that the highlighted phrase influenced the prediction, but that the wording alone does not prove the review is fake. For a low score, the explanation does not describe the review as “flagged”; it states that the review was classified as likely genuine under the current model and threshold and that highlighted phrases indicate model influence rather than suspicion.

VII. RUNTIME OUTPUT AND USER INTERFACE

When the user submits a URL, the backend attempts public review extraction. If review text is found, it is validated and processed with the saved tokenizer; the BiLSTM returns a suspiciousness score that is compared with T = 0.62, and the occlusion module evaluates perturbed versions of the review. The frontend then displays the review text, predicted class, score, highlighted spans, and human-readable explanation (Fig. 4). When data are available, it can also display the number of reviews analyzed, the predicted genuine and suspicious counts, the suspicious percentage, and an overall status. Labels such as Low, Moderate, or High suspicion are application indicators, not legal or factual determinations.





| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 9.867 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 10, October 2026 ||

| DOI:10.15662/IJAREEIE.2026.1510013 |

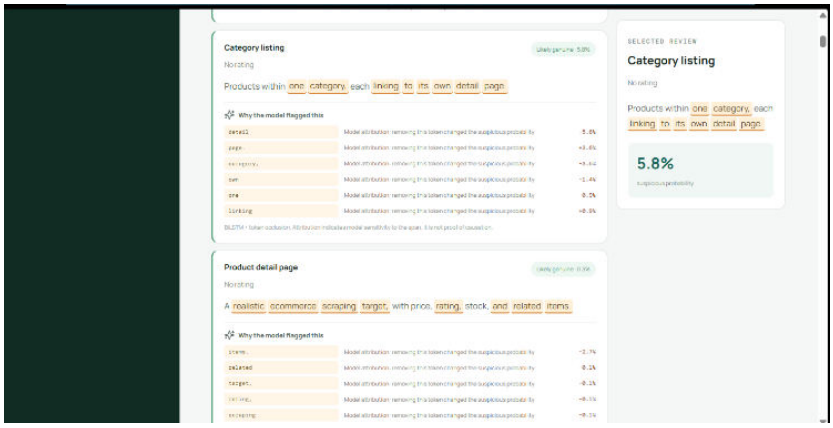


Fig. 4. ReviewLens user interface: (a) URL input and analysis page; (b) summary of analyzed reviews with review-level evidence; (c) selected review with highlighted influential phrases and explanation.

Model-evaluation analytics and live URL predictions are kept separate: held-out test metrics describe performance on a fixed labeled dataset, whereas live URL analysis describes predictions made for newly extracted reviews.

VIII. RESULTS AND DISCUSSION

The system was evaluated on the reviewer-disjoint held-out test set of 40,776 reviews, using threshold T = 0.62 selected on validation data. The reported metrics are accuracy = (TP + TN)/(TP + TN + FP + FN), the proportion of all predictions that are correct; precision = TP/(TP + FP), the fraction of predicted suspicious reviews that are suspicious in the labeled test set; recall = TP/(TP + FN), the fraction of labeled suspicious reviews that are detected; and F1-score = 2 × Precision × Recall/(Precision + Recall), the harmonic mean of precision and recall. ROC-AUC and PR-AUC summarize ranking quality across all thresholds.

TABLE II. HELD-OUT TEST PERFORMANCE

Metric	Held-out test result
Accuracy	89.41%
Precision	74.81%
Recall	75.01%
F1-score	74.91%
ROC-AUC	92.32%
PR-AUC	80.89%
Threshold	0.62

The confusion matrix on the test set is [[30,007, 2,171], [2,149, 6,449]] (Fig. 5), i.e., 30,007 true negatives, 2,171 false positives, 2,149 false negatives, and 6,449 true positives. The system achieves an overall accuracy of 89.41%, indicating that it correctly classifies a substantial proportion of the test reviews. Precision of 74.81% and recall of 75.01% are closely balanced, giving an F1-score of 74.91%. The ROC-AUC of 92.32% shows good separation between the two classes, while the lower PR-AUC of 80.89% reflects the class imbalance (about 21% of records are in the suspicious class), under which accuracy alone overstates performance on the minority class.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 9.867 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 10, October 2026 ||

| DOI:10.15662/IJAREEIE.2026.1510013 |

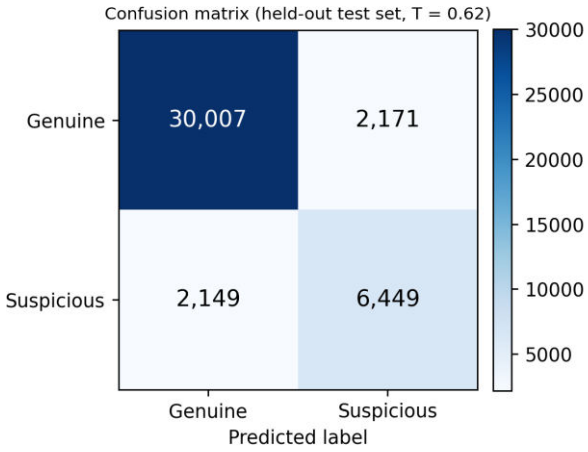


Fig. 5. Confusion matrix on the held-out test set (40,776 reviews, T = 0.62).

Beyond classification performance, a principal contribution of ReviewLens is its emphasis on explainability. Rather than limiting the output to the predicted class, the system shows which phrases influenced the score and describes them cautiously, helping users and researchers assess the reliability of an individual prediction. These results are measured against prepared research labels and should be interpreted with the label-quality limitation in mind.

IX. LIMITATIONS

Public review access differs by website, and browser visibility does not guarantee backend extraction. Research/proxy labels can contain noise. Reviewer-disjoint splitting reduces reviewer leakage but does not guarantee generalization to all sites and product categories. Enthusiastic positive wording can be genuine, so language indicators are not proof of deception. Occlusion results depend on the perturbation strategy and should be treated as model-influence estimates. Finally, the same tokenizer and preprocessing must be used for training and inference, and the test set must not be used to choose thresholds or repeatedly tune model settings.

X. CONCLUSION

ReviewLens uses a BiLSTM classifier to model the sequential structure of review text and produces a suspiciousness probability that is converted into a binary decision using a validation-selected threshold. Through token/span occlusion, it identifies the textual components that influence the prediction and presents them as highlighted evidence, and a human-readable explanation layer describes linguistic patterns such as promotional imperatives, exaggeration, urgency, absolute claims, and excessive punctuation while explicitly avoiding the claim that these patterns prove fraud. On a reviewer-disjoint held-out test set of 40,776 reviews the system achieved 89.41% accuracy and a 74.91% F1-score. By combining URL-based collection, classification, probability estimation, evidence highlighting, and cautious explanation in one system, ReviewLens is more suitable for trustworthy review analysis than a conventional black-box binary classifier. Future work will focus on transformer architectures, behavioral and graph-based features, multimodal evidence, and stronger evaluation against AI-generated deceptive reviews.

REFERENCES

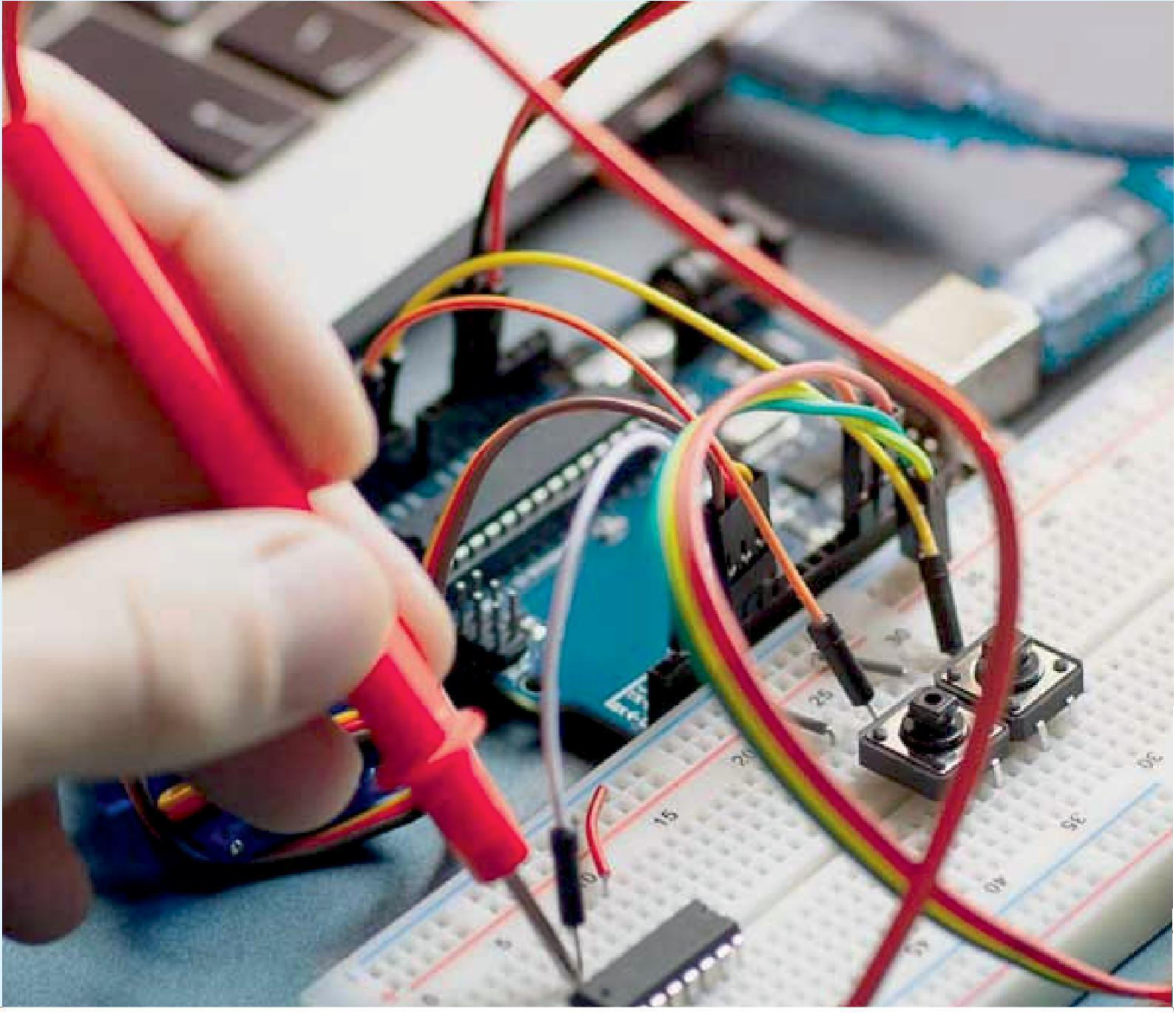
[1] N. Jindal and B. Liu, “Opinion spam and analysis,” in Proc. 2008 Int. Conf. Web Search and Data Mining (WSDM), Palo Alto, CA, USA, 2008, pp. 219–230, doi: 10.1145/1341531.1341560.
[2] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, “Finding deceptive opinion spam by any stretch of the imagination,” in Proc. 49th Annu. Meeting Assoc. Comput. Linguistics: Human Language Technologies, Portland, OR, USA, 2011, pp. 309–319.
[3] Q. Xu and H. Zhao, “Using deep linguistic features for finding deceptive opinion spam,” in Proc. COLING 2012, Mumbai, India, 2012, pp. 1341–1350.
[4] A. Mukherjee, A. Kumar, B. Liu, J. Wang, M. Hsu, M. Castellanos, and R. Ghosh, “Spotting opinion spammers using behavioral footprints,” in Proc. 19th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2013, pp. 632–640, doi: 10.1145/2487575.2487580.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 9.867 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 10, October 2026 ||

| DOI:10.15662/IJAREEIE.2026.1510013 |

[5] F. F. A. et al., "Fake reviews detection: A survey," IEEE Access, vol. 9, pp. 65771–65802, 2021, doi: 10.1109/ACCESS.2021.3075573.
[6] "Detection of spam reviews through a hierarchical attention architecture with N-gram CNN and Bi-LSTM," Information Systems, vol. 101, 2021, Art. no. 101865, doi: 10.1016/j.is.2021.101865.
[7] "A hybridized approach for enhanced fake review detection," IEEE Transactions on Computational Social Systems, vol. 11, no. 6, pp. 7448–7466, 2024, doi: 10.1109/TCSS.2024.3411635.
[8] "Fake review detection using transformer-based enhanced LSTM and RoBERTa," International Journal of Cognitive Computing in Engineering, vol. 5, pp. 250–258, 2024, doi: 10.1016/j.ijcce.2024.06.001.
[9] R. A. Duma, Z. Niu, A. S. Nyamawe, and A. A. Manjotho, "An analysis of graph neural networks for fake review detection: A systematic literature review," Neurocomputing, Art. no. 129341, 2025, doi: 10.1016/j.neucom.2025.129341.
[10] J. Zhao, M. Shao, H. Tang, J. Liu, L. Du, and H. Wang, "RHGNN: Fake reviewer detection based on reinforced heterogeneous graph neural networks," Knowledge-Based Systems, 2023, Art. no. 111029, doi: 10.1016/j.knosys.2023.111029.
[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, Minneapolis, MN, USA, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, 2017.
[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
[14] M. Ott, "Linguistic models of deceptive opinion spam," in Proc. 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Baltimore, MD, USA, 2014, p. 31, doi: 10.3115/v1/W14-2606.
[15] "Research on false review detection methods: A state-of-the-art review," Journal of King Saud University—Computer and Information Sciences, vol. 34, no. 9, pp. 7530–7546, 2022, doi: 10.1016/j.jksuci.2021.07.021.
[16] "Machine learning approaches for fake reviews detection: A systematic literature review," Journal of Web Engineering, vol. 22, no. 5, pp. 821–848, 2023.
[17] "Fake reviews detection using deep learning: A survey," in Proc. Intelligent Methods, Systems, and Applications (IMSA), Giza, Egypt, 2024, doi: 10.1109/IMSA61967.2024.10652819.
[18] "Smart fake review detection with NLP, BERT & behaviour analysis," in Proc. 4th Int. Conf. Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2025, doi: 10.1109/ICAAIC64647.2025.11330838.
[19] "Fake review detection in e-commerce using machine learning and NLP technique," in Proc. 3rd Int. Conf. Inventive Computing and Informatics (ICICI), 2025, doi: 10.1109/ICICI65870.2025.11069636.
[20] A. Jakhar and A. Indian, "Explainable fake review detection: A hybrid deep learning model for E-commerce platforms to enhance customer trust," Journal of Retailing and Consumer Services, vol. 104796, 2026, doi: 10.1016/j.jretconser.2026.104796.



INNO SPACE
SJIF Scientific Journal Impact Factor

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

International Journal of Advanced Research

in Electrical, Electronics and Instrumentation Engineering

9940 572 462 6381 907 438 ijareeie@gmail.com

Scan to save the contact details