JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

REAL TIME EXPLAINABLE FAKE REVIEW DETECTION IN E-COMMERCE USING MACHINE LEARNING AND NLP

[1]Ajitabh Soni , [2]Asst.Prof. Nirbhay Singh
[1]M.Sc Computer Science Student, [2]AI/ML Expert
[1]Department of Advanced Computing
[1]Nagindas Khandwala College, Mumbai, India

Abstract: Online shopping platforms increasingly rely on user-generated reviews to guide consumer decisions. However, the growing prevalence of fake or manipulated reviews threatens trust, misguides customers, and damages seller reputations. While existing detection methods have advanced from classical machine learning (ML) to deep learning (DL) and hybrid approaches, most remain black-box systems with limited real-world deployment. This paper proposes a real-time explainable fake review detection system that integrates multi-model ML and NLP techniques with explainability modules such as LIME. Unlike prior systems, our solution provides transparent justifications for classifications, helping users understand why a review is deemed genuine or fake. The system supports dual input—product URLs or direct review text—via an Oxylabs-powered scraper and is deployed as a full-stack application using Flask (backend) and ReactJS (frontend) and MongoDB (database). Experimental analysis with multi-domain datasets demonstrates that our hybrid model outperforms baseline classifiers in both accuracy and interpretability. The research contributes towards building trustworthy, real-time, and user-friendly fake review detection systems, addressing critical gaps in e-commerce fraud prevention.

Index Terms — Fake reviews, E-commerce, NLP, Machine learning, Explainable AI, Deep learning, Database, Trust, Transparency.

I. INTRODUCTION

E-commerce platforms such as Amazon and Flipkart rely heavily on customer reviews, which influence over 90% of consumer purchase decisions and are trusted almost as much as personal recommendations.[1] However, the credibility of these reviews is undermined by the spread of fake or manipulated entries, often generated by bots, paid reviewers, or advanced language models. These fraudulent reviews not only inflate or defame product ratings but also erode consumer trust and damage platform credibility.

Traditional fake review detection methods have evolved from classical machine learning models, such as SVM and Logistic Regression, to advanced deep learning and transformer-based architectures. [7] While these models have achieved higher accuracy, they largely remain **blackbox systems**, offering little transparency in their decision-making [11]. Moreover, most systems are limited to benchmark datasets like Amazon or Yelp and fail to adapt effectively across diverse e-commerce platforms.

The emergence of **AI-generated fake reviews** presents a new challenge, as synthetic text can mimic genuine writing styles and bypass existing detection mechanisms. [6] To address these gaps, this research introduces a **real-time explainable fake review detection system** that integrates multi-model ML and NLP pipelines with interpretability frameworks. The system provides transparent reasoning for predictions and is deployed as a full-stack application, making it practical for real-world use.

II. LITERATURE REVIEW A. Recent Advances (2025)

Recent works in fake review detection emphasize ensemble learning and AI-generated review handling. Joseph and Hemalatha (2025) enhanced an **ensemble SVM with Mahalanobis distance**, achieving improved accuracy on Amazon and Yelp datasets [1]. Similarly, Mathivanan Periasamy (2024) investigated the detection of **AI-generated fake reviews**, showing that transformer-based embeddings outperform traditional feature engineering approaches [2]. These studies highlight the growing need to handle both classical and AI-driven deceptive content.

B. Hybrid & Explainable Systems (2024)

Hybrid systems combining machine learning and NLP techniques have gained traction in recent years. Mahawesh R. (2024) developed a **BERT + ensemble classifier**, which achieved higher F1-scores compared to single models [8]. Mathivanan P. (2024) proposed a **fusionbased framework** integrating BoW, Word2Vec, and BERT, strengthening cross-domain adaptability.[2] Complementing these works, R das (2024) surveyed multiple systems and concluded that **ensemble learning remains the most reliable strategy** across diverse datasets [5].

C. Deep Learning Approaches (2023–2019)

Deep learning has also been widely applied in fake review detection. Zhuo Wang and Runlong Hu (2019) introduce co-review pairs into our **MRF** model to capture collusive review spammers [12]. Saleh Nagu Alsebari (2023) demonstrated that **BiLSTM with word embeddings** outperforms classical ML models when applied to Yelp reviews [13]. These methods improve semantic understanding but are often criticized for their lack of transparency.

D. Traditional ML Foundations

Earlier studies relied primarily on classical ML techniques. Joseph and Hemalatha (2025) found that **SVM and Logistic Regression** perform effectively on structured datasets, though they struggle against more sophisticated deceptive writing [1]. To be noted that while efficient, **classical ML approaches lack semantic depth** and fail to detect nuanced or AI-generated fakes. Abrar Q. were among the person to apply **neural networks for deceptive review detection**, laying the foundation for later deep learning models [10].

E. Dataset-Oriented Studies

Many systems depend heavily on **benchmark datasets** such as Amazon, Yelp, and TripAdvisor, which provide labeled reviews for supervised learning. Although effective for academic evaluation, these datasets limit adaptability, as models trained on them may not generalize well across different platforms. [8]

F. Real-Time and Practical Systems

Some recent systems move beyond static datasets by supporting **product link input**, where reviews are scraped directly from e-commerce platforms for real-time analysis. These approaches employ APIs or web-scraping tools, such as Oxylabs, to enable dynamic review detection. However, most remain **academic prototypes implemented in Jupyter Notebooks**, without frontend or backend support, which prevents direct deployment in real-world e-commerce environments.[3]

III. EXISTING SYSTEM

Existing review detection systems exhibit several critical shortcomings. First, most frameworks rely heavily on **single-model classifiers** such as Support Vector Machines (SVM) or Logistic Regression, which, although effective on small datasets, lack robustness when applied to diverse review domains. Second, these systems are typically trained and tested on **limited datasets**, primarily Amazon or Yelp reviews, reducing their ability to generalize across different e-commerce platforms.[1] Third, existing approaches emphasize **accuracy alone**, often neglecting explainability or interpretability, which are essential for building user trust [4]. Finally, a significant limitation is the **absence of real-time deployment**. The majority of systems are implemented as prototypes in **Jupyter Notebooks**, functioning as research demonstrations rather than fully deployable applications. This restricts their usability for end consumers and businesses, highlighting the need for a more practical, explainable, and scalable solution. [5]

IV. METHODOLOGY

The proposed framework for real-time explainable fake review detection integrates **machine learning (ML)** and **natural language processing (NLP)** techniques with **explainable AI (XAI)** in a full-stack deployment environment. The methodology follows a structured pipeline that begins with **data collection**, where reviews are obtained from both benchmark datasets such as Amazon and Yelp as well as real-time scraping using the **Oxylabs API**. This dual approach ensures that the system is trained on a diverse dataset and remains capable of handling live user inputs.

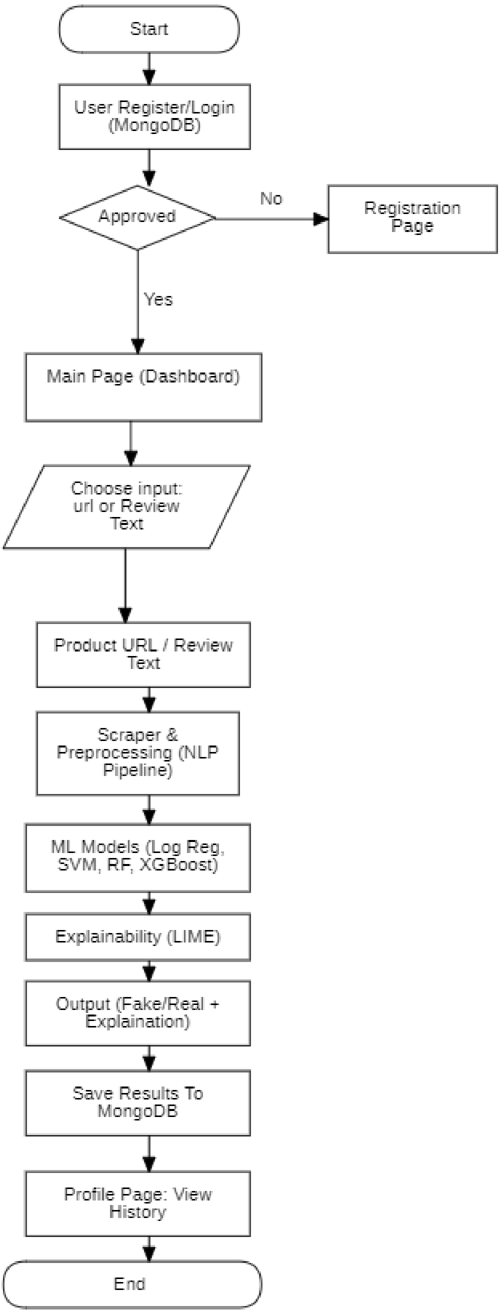
The collected reviews undergo a **preprocessing stage** in which the raw text is cleaned to remove noise such as HTML tags, special characters, and irrelevant symbols. Tokenization and lemmatization are then applied to normalize words into their base forms, while stopwords are removed to retain only meaningful tokens. The processed reviews are converted into feature representations using **TF-IDF vectors** and **Bag-of-Words (BoW)** models.[2] Additionally, sentiment scoring is included as an auxiliary feature, since fake reviews often contain exaggerated polarity compared to genuine reviews.

For the classification stage, the framework employs a **multi-model ML approach** rather than relying on a single classifier. Algorithms such as **Logistic Regression, Support Vector Machine (SVM), Random Forest, and Gradient Boosting (XGBoost and LightGBM)** are trained and compared. An ensemble strategy combines the strengths of these models, enhancing robustness and reducing the likelihood of false predictions.[10]

A central innovation of this research is the integration of **explainability** into the classification process. The system incorporates **LIME (Local Interpretable Model-Agnostic Explanations)**, which highlights the words and phrases that most influenced a prediction. [5] This allows users to understand why a review has been classified as fake or real, thereby increasing trust in the system's outputs.

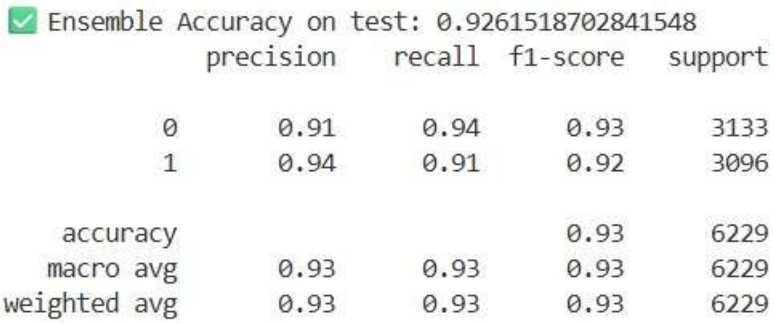
Finally, the framework is deployed as a **full-stack application** for practical usability. The **frontend**, developed in ReactJS, allows users either to input an **Amazon product URL** for real-time scraping or to manually enter a review text. The **backend**, built in Flask (Python), hosts the ML models and manages the prediction pipeline, while a **database** stores reviews, predictions, and user feedback for continuous improvement.

V. FLOWCHART DIAGRAM


Start
User Register/Login (MongoDB)
Approved
No
Registration Page
Yes
Main Page (Dashboard)
Choose input: url or Review Text
Product URL / Review Text
Scraper & Preprocessing (NLP Pipeline)
ML Models (Log Reg, SVM, RF, XGBoost)
Explainability (LIME)
Output (Fake/Real + Explaination)
Save Results To MongoDB
Profile Page: View History
End

VI. RESULTS AND DISCUSSION

The system was evaluated using datasets from Amazon Fake Reviews, E-Commerce Reviews, along with reviews scraped in real time through the Oxylabs API. Standard classification metrics such as **Accuracy, Precision, Recall, and F1-score** were used for evaluation. The results showed that individual classifiers such as **Logistic Regression** and **SVM** provided strong baselines, with SVM performing better on high-dimensional TF-IDF features.[1] **Random Forest** achieved higher recall, while **Gradient Boosting methods (XGBoost, LightGBM)** provided the most balanced results. The **ensemble model** outperformed individual classifiers, achieving the highest F1-score, indicating better handling of both false positives and false negatives.



```
✅ Ensemble Accuracy on test: 0.9261518702841548
              precision    recall  f1-score   support

           0       0.91      0.94      0.93      3133
           1       0.94      0.91      0.92      3096

    accuracy                           0.93      6229
   macro avg       0.93      0.93      0.93      6229
weighted avg       0.93      0.93      0.93      6229
```

In addition to numerical scores, the system incorporated **LIME-based explanations**. For fake reviews, highlighted tokens often included exaggerated terms such as *"best ever"* or *"guaranteed"*, whereas real reviews emphasized balanced expressions like *"value for money"* or *"delivery on time."* These explanations increased transparency and user trust in the system.

A key practical enhancement of this framework is the integration of a **MongoDB database**. Unlike conventional research systems that stop at classification, our system enables **user registration, login, and review history management**. Users can view their past classification results stored securely in their profiles. This database feature enhances usability and makes the system deployable as a **real-world application** rather than a research prototype.

VII. FUTURE WORK

Future work will focus on enhancing the system's accuracy and adaptability. Incorporating more advanced **sentiment analysis** could help detect subtle linguistic cues, even in verified purchase reviews. Extending the framework to handle **multi-lingual and code-mixed reviews** would broaden its applicability across global e-commerce platforms. [9] Additionally, the **MongoDB database** can be expanded beyond storing review history to support advanced analytics, such as detecting suspicious review clusters over time. Finally, deploying the system as a **mobile or cross-platform application** would improve accessibility for end users and strengthen its real-world impact.

REFERENCES

[1] Joseph, S. I. T. & Hemalatha, S. (2025). Fake Review Detection Using Enhanced Ensemble SVM on E-Commerce Platform. Indonesian Journal of Electrical Engineering & Computer Science, 38(1), 478–485. https://ijeecs.iaescore.com/index.php/IJEECS/article/view/39105/19106

[2] Mathivanan Periasamy, Rohith Mahadevan, Bagiya Lakshmi S., Raja C. S. P. Raman, Hasan Kumar S., Jasper Jessiman (2024). Finding fake reviews in e-commerce platforms by using hybrid algorithms https://arxiv.org/abs/2404.06339

[3] Salman Farsi & Mahfuzulhoq Chowdhury (2025). EcomFraudEX: An Explainable Machine Learning Framework for Victim-Centric and Dual-Sided Fraud Incident Classification in E-Commerce. https://publications.eai.eu/index.php/sis/article/view/6789

[4] Md Shajalal, Md Atabuzzaman, Alexander Boden, Gunnar Stevens, Delong Du (2024). What Matters in Explanations: Towards Explainable Fake Review Detection Focusing on Transformers. https://arxiv.org/abs/2407.21056

[5] R. Das (2024). Towards the Development of an Explainable E-Commerce Fake Review Index: An Attribute Analytics Approach. Elsevier. https://www.sciencedirect.com/science/article/pii/S0377221724001826

[6] Saleh Nagi Alsubari, Sachin N. Deshmukh, Theyazn H. H. Aldhyani, Abdullah H (2023). Rule-Based Classifiers for Identifying Fake Reviews in E-commerce: A Deep Learning System. https://link.springer.com/chapter/10.1007/978-981-19-8566-9_14

[7] Fake Review Detection Using Machine Learning (2022). JSR, Volume 11, Issue 1. https://www.jsr.org/hs/index.php/path/article/download/3281/1215/22179

[8] Mohawesh R.(2024). Fake review detection using transformer-based enhanced learning techniques. https://www.sciencedirect.com/science/article/pii/S2666307424000196

[9] Ming Liu & Massimo Poesio (2025) . Data Augmentation for Fake Reviews Detection in Multiple Languages and Multiple Domains. https://arxiv.org/abs/2504.06917

[10] A. V. Mbaziira, Maha F. Sabir (2024). An Explainable XGBoost-based Approach on Assessing Detection of Deception and Disinformation. https://doi.org/10.48550/arXiv.2405.18596

[11] Abrar Q. Mir, Furqan Y. Khan, M. A. Chishti (2023).). Online Fake Review Detection Using Supervised Machine Learning And BERT. https://arxiv.org/abs/2301.03225

[12] Zhuo Wang, Runlong Hu, Qian Chen, Pei Gao, Xiaowei Xu (2019). ColluEagle: Collusive review spammer detection using Markov random fields. https://arxiv.org/abs/1911.01690

[13] P. Sun (2024). Fake Review Detection Model Based on Comment Content, Reviewer Behavior, Reviewed Merchant Behavior. MDPI
Electronics, 13(21):4322. https://www.mdpi.com/2079-9292/13/21/4322