Smart Detection of Review Fraud on E-Commerce Sites

Ayush Rathore[1], Sourabh Pandey[2], Yashdeep Sahu[3], and Kavyashree G. M[4]

[1,2,3] *Students, Department of Computer Science and Engineering, Sir M Visvesvaraya Institute of Technology*

[4] *Assistant Professor, Department of Computer Science and Engineering, Sir M Visvesvaraya Institute of Technology*

***Abstract*—The surge of e-commerce has led to an increase in user-generated content, particularly reviews, many of which are fabricated by companies to boost product visibility and sales. These reviews, often posted by bots or paid individuals, undermine trust in digital marketplaces. This research introduces a comprehensive approach that integrates machine learning (ML) and natural language processing (NLP) to detect such deceptive reviews. Utilizing extensive preprocessing and advanced feature extraction techniques, our system effectively identifies subtle textual patterns and sentiments that differentiate genuine content from fraudulent ones. Both supervised and unsupervised ML models, including ensemble methods, were evaluated. Real-world datasets demonstrated the framework's ability to achieve high detection accuracy while minimizing false positives, thus contributing to the integrity and reliability of online shopping platforms.**

***Keywords*—Machine Learning, NLP, XGBoost, Logistic Regression, Decision Tree, Review Authenticity, Random Forest**

I. INTRODUCTION

In today's digital commerce landscape, e-commerce platforms have reshaped how consumers engage with brands and products. Online reviews significantly influence purchasing decisions, acting as digital endorsements. Unfortunately, this system is plagued by deceptive reviews intended to manipulate customer perception and behavior. The proliferation of fake reviews not only skews public opinion but also diminishes trust and distorts fair market practices.

This study proposes a robust solution combining natural language processing and machine learning to identify and mitigate the impact of fake reviews. Online reviews carry linguistic intricacies and behavioral patterns that can be exploited to differentiate real reviews from fabricated ones. Our methodology focuses on extracting relevant linguistic features and applying diverse ML models to automate the detection of suspicious content.

We adopt both supervised learning techniques to leverage labeled data and unsupervised methods like clustering and anomaly detection to discover hidden patterns. These tools collectively enable our system to adaptively detect and flag deceptive reviews, ultimately fostering transparency and credibility in online marketplaces.

II. LITERATURE SURVEY

Jindal & Liu (2008) - Opinion Spam and Analysis:

- Distinct from traditional spam: Opinion spam differs from web/email spam, requiring novel detection techniques due to its focus on fabricated reviews and ratings.
- Types of opinion spam: Identifies three categories - duplicate reviews, non-reviews (ads/irrelevant content), and untruthful reviews (biased/fake opinions).
- Review-centric features: Proposes using linguistic and behavioral features (e.g., review length, rating deviation) rather than link-based metrics.
- Duplicate detection: Highlights that duplicate reviews are a common but limited subset of spam, often posted by the same user/IP.

Ott et al. (2013) - Negative Deceptive Opinion Spam

- Dataset creation: Introduces the first gold-standard dataset of 400 negative deceptive hotel reviews via Amazon Mechanical Turk.
- Automated detection: Achieves 86% accuracy using n-gram text classifiers, outperforming human judges who scored near chance.
- Linguistic patterns: Identifies features like reduced spatial detail and exaggerated language in deceptive reviews.
- Sentiment-deception interaction: Compares negative spam with Ott et al.'s prior positive

spam dataset, noting consistent psycholinguistic markers.

Mukherjee et al. (2013) - What Yelp Fake Review Filter Might Be Doing?

- KL-divergence approach: Uses information-theoretic methods to distinguish "forced" fake reviews (paid) from "natural" ones (organic)56.
- Data sources: Combines crowdsourced pseudo-reviews (Mechanical Turk) with labelled Yelp.com data for analysis56.
- Behavioural insights: Reveals forced fake reviewers use more verbs and fewer nouns, while natural fakes mimic genuine reviews5.
- Classification performance: Achieves up to 89% accuracy using SVM classifiers with linguistic and syntactic features5.

Rayana & Akoglu (2015) - Collective Opinion Spam Detection

- SpEagle Framework: Introduces SpEagle, a unified model that combines metadata (e.g., text, timestamps, ratings) with relational data (user-review-product networks) to collectively detect spam users, fake reviews, and targeted products.
- Semi-Supervised Extension (SpEagle+): Enhances SpEagle by allowing the incorporation of a small set of labeled data (e.g., known spam users) without retraining, Improving detection accuracy while maintaining efficiency.
- Behavioral and Linguistic Features: Utilizes features such as review burstiness (multiple reviews in a short time), rating deviations, and linguistic patterns to compute spam scores for users, reviews, and products, informing prior probabilities in the network model.
- Scalability with SpLite: Proposes SpLite, a lightweight version of SpEagle that uses only key review features (e.g., review length, rating variance) to reduce computational overhead while maintaining approximately 90% of the original method's accuracy.

III. METHODOLOGY

A. Data Collection and Preprocessing

We used the publicly available OSFHOME dataset, which contains 40,000 customer reviews—split evenly into 20,000 genuine and 20,000 fake entries. Reviews marked as 'OR' (Original Reviews) were considered real, and those tagged 'CG' (Computer Generated) were treated as fake.

B. Data Preprocessing

As the dataset was already balanced, further sampling was unnecessary. Preprocessing involved several cleaning steps:

- Removal of stop words, punctuation, emojis, and URLs
- Conversion to lowercase
- Null value elimination
- Text normalization using stemming (preferred over lemmatization for efficiency)

Processed data was then split into training (70%) and testing (30%) subsets. Subsequently, feature vectors were generated using Count Vectorizer, TF-IDF, and pre-trained Word2Vec embeddings.

C. Feature Extraction Techniques

Bag of Words (BoW): Converts text into a vector based on word occurrence.

TF-IDF: Highlights the importance of uncommon words in the corpus.

Word2Vec: Embeds semantic relationships in vector space using CBOW and Skip-gram

D. Language Mode

BERT (Bidirectional Encoder Representations from Transformers): A deep learning model capable of learning bidirectional context from large corpora. Fine-tuning BERT enables robust classification performance on text classification tasks like fake review detection.

E. Classifiers and Models

- Logistic Regression
- Naive Bayes (Multinomial and Complement)
- Support Vector Machine (SVM)
- Decision Tree (DT)
- Random Forest (RF)
- XGBoost and AdaBoost
- Stochastic Gradient Descent (SGD)
- K-Nearest Neighbors (KNN)
- Ensemble Voting: Combines RF, XGBoost, and LR for improved performance.

F. Evaluation Metrics

To evaluate each model, we used the following standard metrics:

- Accuracy: (TP + TN) / (TP + TN + FP + FN)
- Precision: TP / (TP + FP)
- Recall: TP / (TP + FN)
- F1-Score: Harmonic mean of precision and recall

These metrics help quantify the performance, especially the balance between false positives and false negatives.

IV. RESULTS AND DISCUSSIONS

A) Model Selection and Evaluation Summary

To ensure optimal performance and generalization, multiple classification algorithms were evaluated during model development. Each model underwent structured hyperparameter tuning and was tested for accuracy on the test set. The table below summarizes the algorithms used, their tuning strategies, evaluation status, and final usage decision:

Algorithm	Library Used	Hyperparameter Tuning	Accuracy
Logistic Regression	sklearn.linear_model	C: [2, 6, 10]	0.89
Random Forest	sklearn.ensemble	n_estimators, max_depth	0.76
Naive Bayes	sklearn.naive_bayes	var_smoothing	0.81
MLP Classifier	sklearn.neural_network	Static (5,2) layer config	0.79
BERT (Fine-tuned)	transformers	Pretrained + Trainer setup	0.77

V. CONCLUSION

This study presents a comprehensive and data-driven approach for detecting fraudulent reviews on e-commerce platforms by integrating natural language processing with various machine learning and deep learning techniques. The framework was tested on a balanced real-world dataset comprising genuine and computer-generated reviews, ensuring the model's applicability to practical scenarios.

The methodology encompassed robust preprocessing steps, multiple feature extraction techniques including Bag of Words, TF-IDF, and word embeddings such as Word2Vec and GloVe, and a wide spectrum of classification algorithms. The study compared traditional classifiers like Logistic Regression, Naive Bayes, and Random Forest with advanced ensemble and deep learning models including BiLSTM and BERT. Among these, the Logistic Regression model achieved the highest accuracy of 89%, underscoring its effectiveness in contextual understanding and text classification tasks.

This work not only addresses the technical aspects of review classification but also reinforces the importance of trust and transparency in digital commerce. By reducing the influence of fake reviews, such systems can safeguard consumer interests, promote fair competition among sellers, and preserve the credibility of online marketplaces.

For future work, we propose extending this framework to support multilingual datasets and dynamic model retraining to adapt to evolving fake review patterns. Real-time deployment as a browser extension or API for e-commerce platforms could make this research highly impactful in the ongoing battle against digital misinformation.

REFERENCES

[1]. Jindal, N., & Liu, B. (2008). Opinion spam and analysis. Proceedings of the International Conference on Web Search and Data Mining.

[2]. Ott, M., Cardie, C., & Hancock, J. T. (2013). Negative deceptive opinion spam. Proceedings of NAACL-HLT.

[3]. Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. (2013). What Yelp Fake Review Filter Might Be Doing? Proceedings of ICWSM.

[4]. Rayana, S., & Akoglu, L. (2015). Collective Opinion Spam Detection: Bridging Review Networks and Metadata. Proceedings of KDD.

[5]. Anonymous, "A Unified Framework for Detecting Author Spamicity," Expert Systems with Applications, vol. 114, pp. 393–404, 2018.

[6]. J. Salminen, et al., "Creating and Detecting Fake Reviews of Online Products," Information & Management, vol. 59, no. 2, 2022.

[7]. Anonymous, "Fake Review Detection Using XGBoost and SGD Classifiers," Revue d'Intelligence Artificielle, vol. 37, no. 5, pp. 751–758, 2024.

[8]. Anonymous, "Ontology-Based Sentiment Analysis for Fake Review Detection," Expert Systems with Applications, vol. 210, 2022.

[9]. P. Hajek, et al., "Detecting Fake Reviews Through Topic Modelling," Journal of Business Research, vol. 156, Jan. 2023.