International Journal of Research and Review
[Indexed, Peer-Reviewed Journal]
Volume 13; Issue: 3; March 2026
Website: www.ijrrjournal.com
Research Paper *E-ISSN: 2349-9788; P-ISSN: 2454-2237*

Multi-Modal System for Fake Review Detection on E-Commerce Platforms

Putti Venkata Siva Teja[1], P. Purna Chandrika[1], N. Lavanya Sai[2], M. Sushma[3], Md. Allabakshu[4], T. Mahesh[5]

[1,2,3,4,5]Department of Information Technology,
Dhanekula Institute of Engineering & Technology, Vijayawada, Andhra Pradesh, India.

Corresponding Author: Putti Venkata Siva Teja

DOI: https://doi.org/10.52403/ijrr.20260338

ABSTRACT

The expansion of e-commerce has made online reviews a primary factor in consumer purchasing decisions. However, the rise of fake reviews reduces trust in these platforms and creates confusion for buyers. This project develops a multi-modal detection system that analyzes both text and image-based reviews to identify fraudulent content. For textual data, we use Natural Language Processing (NLP) with the NLTK library and TF-IDF feature extraction to find patterns in deceptive writing. For images, Perceptual Hashing (pHash) is used to flag duplicate photos often used in fake reviews. Our system categorizes reviews using a Naive Bayes classifier and visualizes the results for specific products via bar graphs. Experimental results show high accuracy, helping consumers make more informed choices.

Keywords: Natural Language Processing, Machine Learning, Image Hashing, E-Commerce Systems, NL TK, TF-IDF.

INTRODUCTION

The rapid expansion of e-commerce platforms has fundamentally changed the way consumers purchase products and services in the modern era. In this digital marketplace, online reviews have become a cornerstone of the decision-making process, providing critical information regarding the quality, performance, and overall satisfaction of a product before a transaction occurs. Many shoppers depend heavily on these peer reviews to gain a realistic understanding of a product based on the actual experiences of previous users. However, the integrity of these platforms is currently threatened by the proliferation of deceptive online reviews. These fake reviews are often strategically posted to either artificially inflate the rating of a specific product or to maliciously damage the reputation of competing items. Such "opinion spam" is frequently composed to mimic the tone and style of genuine customer feedback, making it difficult for the average user to distinguish between authentic experiences and manufactured praise. Furthermore, the scale of this problem is amplified by the use of automated tools and coordinated groups capable of generating massive volumes of fraudulent content. Because of these factors, the detection of deceptive reviews has become an essential area of study within machine learning and data mining. Traditional research in this field primarily focused on Natural Language Processing (NLP) to analyze writing styles, sentiment, and specific word patterns. However, modern e-commerce sites now allow users to supplement their text with images to verify their purchases. Relying solely on

text analysis is no longer sufficient to achieve high accuracy in these multi-modal environments. To address these limitations, this study proposes a comprehensive system that analyzes both text and image data. By integrating NLP techniques like tokenization and stop-word removal with image processing methods such as Perceptual Hashing (pHash), the system can identify both linguistic inconsistencies and suspicious visual duplicates. This multi-modal approach aims to increase the accuracy of fake review detection, ultimately helping consumers make more reliable and informed purchasing decisions.

LITERATURE REVIEW

The rapid expansion of e-commerce platforms has fundamentally altered consumer behavior regarding the procurement of products and services. In this digital landscape, online reviews have emerged as a primary information source, offering critical insights into product quality and performance benchmarks. These user-generated evaluations are essential for mitigating risk before a purchase decision is finalized. Many consumers rely on these collective experiences to bridge the information gap between the digital listing and the physical product. However, the integrity of these platforms is currently challenged by the proliferation of deceptive online reviews. Such fraudulent content can mislead consumers and significantly undermine the systemic credibility of e-commerce marketplaces [1, 2]. The primary motivation behind posting fake reviews is typically to artificially inflate the rating of a specific product or to systematically disparage the reputation of competing items. In most cases, these deceptive entries are carefully crafted to mimic the linguistic style and sentiment of genuine customer feedback. Furthermore, large volumes of such reviews are often generated through automated scripts or coordinated groups. Consequently, the detection of deceptive reviews has become a critical research objective within the fields of machine learning and data mining [3, 4]. Earlier investigations in this domain primarily focused on analyzing textual content using Natural Language Processing (NLP) techniques. These methods examine linguistic features, including writing style and sentiment polarity, to differentiate between authentic and fabricated reviews. However, modern e-commerce websites now frequently allow users to supplement text with images. In these multi-modal environments, relying solely on text-based analysis is often insufficient for achieving high detection accuracy [5, 6]. To improve the reliability of detection, contemporary methods have begun incorporating the analysis of both text and visual data. For textual analysis, techniques such as tokenization, stop-word removal, and stemming are applied to isolate significant features. These features are then converted into a numerical format using the Term Frequency-Inverse Document Frequency (TF-IDF) method. For visual analysis, Perceptual Hashing (pHash) is utilized to identify duplicate or near-duplicate images that may indicate fraudulent activity [7, 8]. This research proposes a multi-modal fake review detection system that processes both textual and image-based data. The system utilizes the NLTK library for text preprocessing and pHash for image similarity checks. By allowing users to input specific Product IDs, the system generates a visual representation of the findings via bar graphs, distinguishing between genuine and fake content to assist consumers in informed decision-making [9, 10].

Proposed System

The objective of the proposed research is to identify the existence of fake reviews on online shopping websites by examining the reviews posted on the websites as well as the images accompanying the reviews. This is done by following these steps: collecting the reviews from available sources. A dataset is created that contains the review text along with the corresponding images. After gathering the data, the dataset is

arranged and prepared so that it can be used for further processing. The next step is data preprocessing. Data preprocessing improves the quality of collected data. Natural Language Processing techniques are applied for textual data using the Natural Language Toolkit library for processing the data. In this stage, the text data is tokenized, unnecessary words are removed, and the review data are cleaned. All these operations are done to make the data more suitable for analysis. Meanwhile, image data is also processed in a way that useful information from images can be obtained in later stages. In the next stage, feature extraction is done on the preprocessed data. In this stage, data is transformed into a numerical format that is suitable for analysis. For textual data, the TF-IDF algorithm is applied to obtain useful words in reviews and transform the data into a numerical format. For image data, perceptual hashing (pHash) is used to measure the similarity between images. In case images are found similar in multiple reviews, it could be a potential indicator of suspicious activity. After extracting the features, the dataset is split into training and testing subsets for model development and evaluation. While the training data set can be used to learn how to differentiate between real and fake reviews, the testing data set can be used to determine how well the system is performing. Finally, the performance of the system can be determined through the evaluation measures such as accuracy, precision, recall, and F1 score. The results of the analysis are represented through a bar graph that displays the number of real and fake reviews according to a given product ID. This method could help in the detection of fake reviews and could also help users in making more informed purchasing decisions when shopping online.

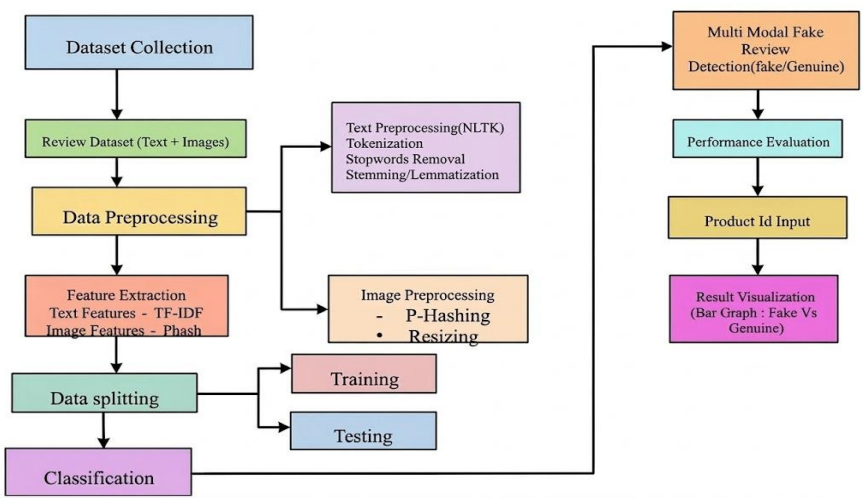

Figure 1: Architecture of the Multi-Modal Detection System

As illustrated in Figure 1, the system follows a sequential pipeline starting from data acquisition to the final visualization. The text and images associated with a Product ID are extracted and passed through separate processing streams. The textual stream uses the NLTK library for cleaning and TF-IDF for vectorization, while the image stream uses pHash to identify recycled visual content. These features are then merged and processed by the Naive Bayes classifier to determine the review's authenticity.

Data Collection

The amazon fine food reviews dataset which can be accessed through the Kaggle platform has been used for this research the dataset used for this research contains the customer reviews of the food items

available on amazon this dataset contains thousands of data related to the reviews of the food items with various attributes such as productid UserId, ProfileName, Helpfulness, Numerator helpfulness, Denominator score time summary and text these attributes contain the information related to the product as well as the customer who has written review text is analyzes to understand customer opinions while the productid helps associate each review with its corresponding product after obtaining the dataset the analysis of the collected reviews are processed using natural language processing methods implemented through the TF-IDF method will be used to convert the text data into numerical form.

Data Pre-Processing

Before the development of the model, the text of the review, along with the images associated with the review, is processed based on the requirement for the analysis. For text data, Natural Language Processing is performed using the NLTK library. First, the text being reviewed is preprocessed by removing the punctuation, symbols, and other unwanted characters. Then, the separation of the words from each other is done using the tokenization method. Common words that do not contribute to the analysis process are removed. This is done by applying the stop word filtering method. After the completion of the preprocessing, the TF-IDF method is applied for the transformation of the text into numerical feature values, which can be used for the analysis. Similarly, the images associated with the reviews are prepared before the analysis, such that the useful information can be obtained for the detection of suspicious/duplicated content.

Confusion Matrix

Confusion matrix is a performance evaluation metric for the fake review detection system. This works on the principle of comparing the labels that the model has predicted with the labels that are available in the dataset. By doing this, it becomes possible to identify the correct as well as the incorrectly classified reviews. Using the information that is available through the confusion matrix, it becomes possible to calculate the accuracy, precision, recall, F1 score, etc. This will provide a clear idea about how effectively the model can differentiate between the real reviews and the fake reviews.

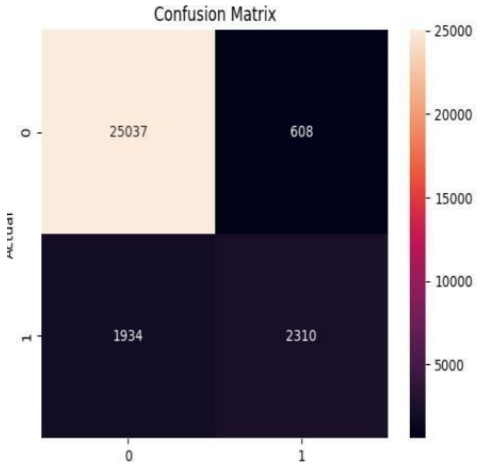

Figure 2: Confusion Matrix

TRAINING

In this work, the Naive Bayes classifier is applied to develop the fake review detection model. Once the preprocessing stage is complete, the review dataset with the textual information and the imagerelated data is ready for training. The textual reviews are transformed into numerical representations using the TF-IDF method, while image information is analyzed using perceptual hashing (pHash). The processed feature set, together with the corresponding review labels indicating whether a review is genuine or fake, is provided to the Naive Bayes algorithm. During this stage, the model learns the relationships between the extracted features and their associated classes, enabling it to identify patterns that distinguish authentic reviews from fraudulent ones.

Testing

After the training phase, the dataset is divided into training and testing sets in a 70:30 ratio. Finally, the Naive Bayes model,

which has already been trained, is used on the test data to check whether the review belongs to the original category or the fake category. The output values are compared with the actual values to check the performance of the model. Various performance metrics are calculated to check the effectiveness of the model to detect the fake reviews posted on various e-commerce sites. In the project, the Naive Bayes algorithm has been used as a classifier to classify the reviews as genuine or fake reviews. Naive Bayes is a probabilistic approach to machine learning and text classification because of the ability to process huge amounts of data with high efficiency. It is derived from Bayes' theorem and estimates the probability that a review belongs to a particular class based on the extracted features from the data.

Classification

After the reviews are preprocessed, the features are obtained using the TF-IDF algorithm for text data and perceptual hashing algorithm for image data. These obtained features are then used as input to the Naive Bayes classifier algorithm. It learns from the obtained features from the given data set and then classifies the reviews in the test data set into different categories. After that, the efficiency of the classifier is measured using accuracy, precision, recall, and F1 score to measure the efficiency of the classifier in detecting fake reviews.

RESULTS

The performance of the proposed multimodal fake review detection system was evaluated using several key metrics: accuracy, precision, recall, and the F1-score. These metrics are derived from a confusion matrix, which summarizes the model's predictions by comparing predicted labels against the actual labels in the dataset.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$\text{Precision} = TP / (TP + FP) \quad (2)$$

$$\text{Recall} = TP / (TP + FN) \quad (3)$$

$$\text{F1-Score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

In the above equations, TP stands for True Positive, TN stands for True Negative, FP stands for False Positive, and FN stands for False Negative.

The results obtained using the proposed system clearly indicate the efficiency of the proposed system in detecting fraudulent reviews.

Table 1: Performance Metrics

Performance Metrics	Naive Bayes Model
Accuracy	0.91
Precision	0.91
Recall	0.91
F1-Score	0.91

Visualization Of Results

The system provides a clear identification of correct and incorrect classifications through a confusion matrix, which compares predicted labels against the actual labels in the dataset. As shown in the Confusion Matrix shown in figure 2, the model successfully identified a high volume of reviews, allowing for the precise calculation of the performance metrics listed in the table 1. The user interface of the Fake Review Detection System allows for real-time analysis by entering a Product ID, manual text, or uploading an image. The final output is presented to the user through a clear visual indicator. For instance, when a suspicious image is detected via perceptual hashing, the system flags it as a "Fake Review". Finally, the system can visually present the distribution of results for a specific product ID in the form of a bar graph, representing the total count of real versus fake online reviews to help consumers make informed purchasing decisions as shown in Figure 3.

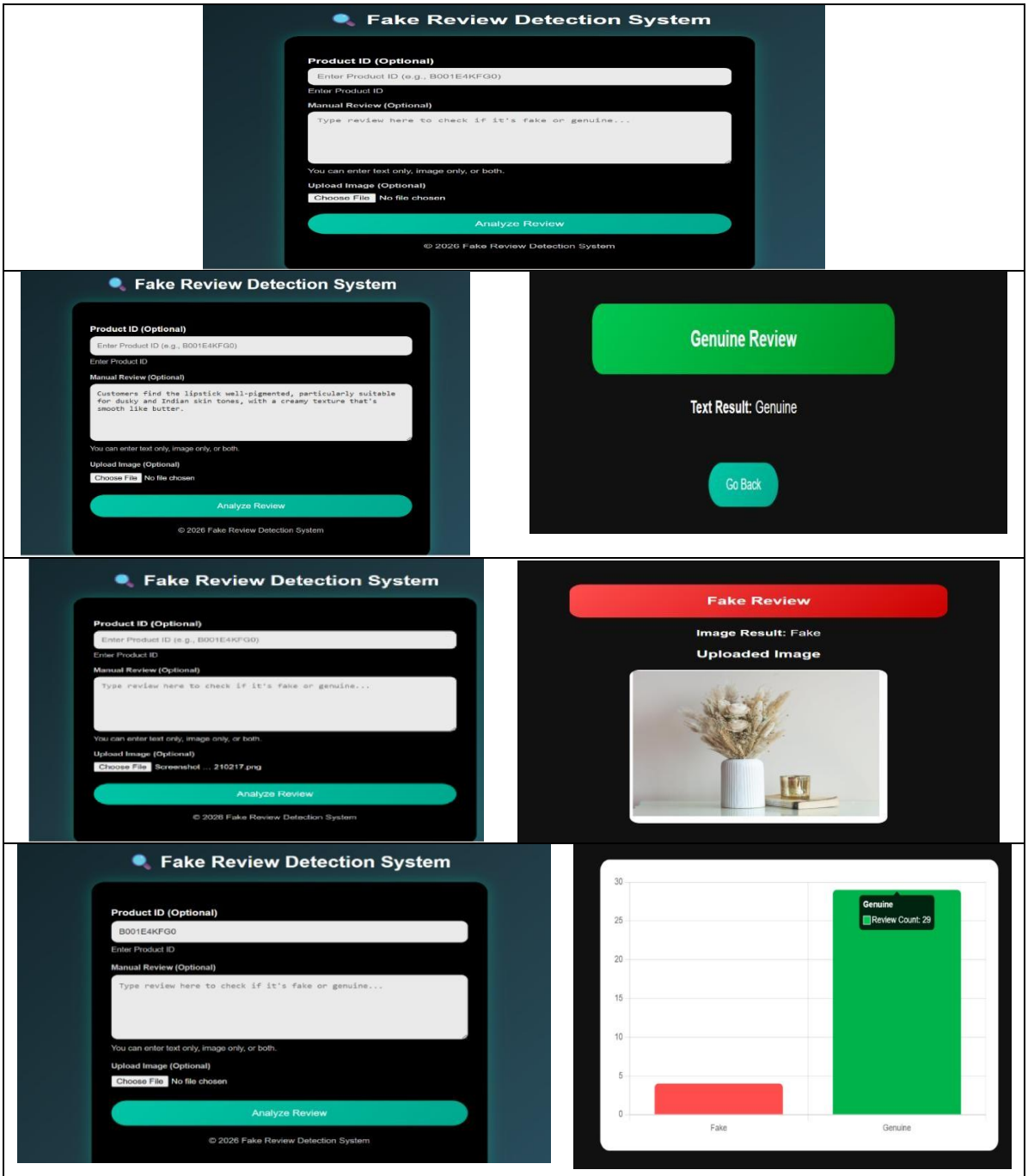

Figure 3: System User Interface

CONCLUSION

The proposed multimodal fake review detection system successfully addresses the growing challenge of deceptive content on e-commerce platforms by integrating both textual and visual analysis. By utilizing Natural Language Processing (NLP) techniques and the NLTK library for preprocessing, alongside TF-IDF for feature extraction, the system identifies significant linguistic patterns within review text. Simultaneously, the implementation of perceptual hashing (pHash) provides a robust mechanism for detecting duplicate or near-duplicate images, which often indicate fraudulent activities. The application of a Naive Bayes classifier facilitates the efficient categorization of reviews into genuine or fake classes, achieving a high-performance metric of 0.91 across accuracy, precision, and recall. Ultimately, this research demonstrates that combining

multimodal features significantly enhances detection accuracy over traditional text-only methods, providing consumers with a reliable tool for making informed purchasing decisions and restoring trust in online review ecosystems.

Declaration by Authors

Acknowledgement: None
Source of Funding: None
Conflict of Interest: No conflicts of interest declared.

REFERENCES

1. Jindal N, Liu B. Opinion spam and analysis. Proceedings of the 2008 International Conference on Web Search and Data Mining. 2008 Feb 11-12:219-230. doi: 10.1145/1341531.1341560.
2. Ott M, Choi Y, Cardie C, Hancock JT. Finding deceptive opinion spam by any stretch of the imagination. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011 Jun 19-24; 1:309-319.
3. Rayana S, Akoglu L. Collective opinion spam detection: Bridging review networks and metadata. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015 Aug 10-13:985-994. doi: 10.1145/2783258.2783370.
4. Akoglu L, Chandy R, Faloutsos C. Opinion fraud detection in online reviews by network effects. Proceedings of the International AAAI Conference on Web and Social Media. 2013 Jul 8;7(1):2-11.
5. Asaad WH, Allami R, Ali YH. Fake Review Detection Using Machine Learning. Revue d'Intelligence Artificielle. 2023 Oct 31;37(5):1111-1120. doi: 10.18280/ria.370505.
6. Gupta R, Jindal V, Kashyap I. Recent state-of-the-art of fake review detection: a comprehensive review. The Knowledge Engineering Review. 2024 Jan 15;39: e8. doi: 10.1017/S026988892300029X.
7. Hajek P, Hikkerova L, Sahut JM. Fake review detection in e-Commerce platforms using aspect-based sentiment analysis. Journal of Business Research. 2023 Nov 1; 167:114143. doi: 10.1016/j.jbusres.2023.114143.
8. Veluru SR, Erukude ST, Marella VC. Multimodal Detection of Fake Reviews using BERT and ResNet-50. 2025 4th International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). 2025 Sep 11-13:877-882. doi: 10.1109/ICIMIA64123.2025.1070012.
9. Tufchi S, Yadav A, Ahmed T. A comprehensive survey of multimodal fake news detection techniques: advances, challenges, and opportunities. International Journal of Multimedia Information Retrieval. 2023 Jun 1;12(2):28. doi: 10.1007/s13735-023-00296-3.
10. Liu C, He X, Yi L. Determinants of multimodal fake review generation in China's e-commerce platforms. Scientific Reports. 2024 Apr 11;14(1):8524. doi: 10.1038/s41598-024-59024-4. PMID: 38605094.

How to cite this article: Putti Venkata Siva Teja, P. Purna Chandrika, N. Lavanya Sai, M. Sushma, Md. Allabakshu, T. Mahesh. Multi-Modal system for fake review detection on e-commerce platforms. *International Journal of Research and Review*. 2026; 13(3): 327-333. DOI: *https://doi.org/10.52403/ijrr.20260338*

******