Published online on the journal's webpage: **http://journals.telkomuniversity.ac.id/index.php/dinda**

Journal of Dinda

Data Science, Information Technology, and Data Analytics

Vol. 6 No. 2 (2026) 79 - 84 E-ISSN: 2809-8064

Bridging Technology and Law: Fake Review Detection Using Naive Bayes Classification and Its Regulatory Implications in E-Commerce

Niamatullah zubair[1*], Somantri[2]

[1*]Informatics Engineering , Faculty of Engineering, Computer and Design, Nusa Putra University

[2] Informatics Engineering , Faculty of Engineering, Computer and Design, Nusa Putra University

[1*]niamatullah.zubair_ti23@nusaputra.ac.id, [2]Somantri@nusaputra.ac.id

Abstract

With the rapid growth of e-commerce, online reviews have become an essential information source for consumers. However, it has also become an emerging trend in posting fake reviews, which is misleading. Such fake reviews not only misguide the consumers in making decisions but also create an unfair market for competing businesses and harm e-businesses that rely on them as an information source. This paper presents a study on the detection of fake online reviews using machine learning techniques and the applicability of such techniques in making them statutory for digital e-businesses. The study used an Amazon review dataset with benchmark verified fake and genuine labels, pre-processed and transformed using the TF-IDF technique. Three machine learning classifiers-Naive Bayes, Logistic Regression, and Random Forest-were applied and their performances compared. The results obtained indicated that Logistic Regression gave the best classification accuracy of 88.26%, followed by Naive Bayes with 85.68% and Random Forest with 84.99%. The Naive Bayes model showed 84.64% precision, 87.21% recall, and 85.91% F1-score, proving competitive performance with modest computational requirements. In a general sense, these are results that point to the fact that machine learning has quite a substantial potential in assisting to find and validate online reviews. However, there is further regulatory oversight and enforcement that is needed for greater consumer protection as well as transparency and trust in digital marketplaces.

Keywords: *Artificial intelligence; consumer protection; fake online reviews; machine learning; Naive Bayes classification*

1. Introduction

In today's digital world, online reviews play a significant role in helping consumers make decisions about products/services when shopping through digital platforms. They can assist consumers in making informed buying decisions before they purchase a product or service by offering them the experience of other users. Thus, online reviews help establish consumer confidence in their purchases and an influencer for consumer behaviors and market competition.

The proliferation of fake reviews has led to a significant 'trust deficit' in digital marketplaces, where consumers are increasingly skeptical of even genuine feedback [1] From a legal perspective, this erosion of trust is a direct challenge to the principles of fair competition and consumer autonomy. The legal implications encompass broader issues of digital platform liability and the adequacy of existing consumer protection statutes in the face of automated misinformation [2] Recent studies emphasize that understanding detection mechanisms is inseparable from analyzing the regulatory environment that governs these technologies [3].

As more and more businesses use fake and misleading reviews, the reliability of online reviews continues to be called into question. Fake reviews can be created to change the way a consumer thinks about a particular product, and these fake reviews can also negatively impact the reputation of other, competing businesses. As such, the creation and use of fake online reviews not only presents consumers with inaccurate information about a given business' products and services; they also create an uneven playing field for companies that rely on honest customer feedback.[4]

AI technology has evolved to the point where they can now create realistic and believable written material, making it challenging to tell the difference between real feedback and false feedback. Hence, people who relied upon manual identification methods to detect this "fake" content no longer have these methods as adequate, as they now have to utilize automated detection methods [5].

Received: 05-06-2026 | Revised: 15-97-2026 | Published: 21-08-2026

Aside from these challenges, fake reviews can also create legal challenges. For instance, a company that posts fake reviews may be perceived as deceiving and therefore could be violating consumer protection laws. A company that engages in this type of conduct may have a competitive advantage because consumers will have false or inaccurate information about that business. Given this reality, many now believe that regulatory frameworks should be put in place to ensure proper transparency and accountability within all types of digital media [6].

Text analysis through machine learning offered an empirical method of discovery within text and has developed based on textual features of language like the occurrence, frequency and style of words used; how often they are used throughout an entire text; and their tone to classify a text as a true or false review. Many basic algorithms (Naive Bayes) have provided a well-documented success rate for classifying texts correctly based on these attributes.

As e-commerce continues to develop rapidly, with many consumers now depending heavily on user-generated content, researchers have noticed a growing need to establish methods for detecting fake online reviews. Thus far, researchers have primarily investigated different computational techniques for this task using machine learning techniques and natural language processing to identify fraudulent reviews. Much like typical text categorization tasks performed with traditional machine learning algorithms (Naive Bayes, Support Vector Machines [SVM], and Decision Trees), Naive Bayes has been successfully applied as a tool for classifying fraudulent reviews. However, Naive Bayes is easier to use than other methods (e.g., SVMs or Decision Trees), works well even for small sample sizes (i.e., less than 500 reviews), and is less computer-intensive than other methods [7]. The aforementioned techniques mainly rely upon textual features such as word counts, sentence length, and sentiment analysis.

On the regulatory side, the Federal Trade Commission (FTC) in 2024 announced a landmark final rule explicitly banning the sale and purchase of fake reviews [8]. Similarly, the UK's Competition and Markets Authority has provided updated guidance for businesses to comply with consumer protection laws regarding review integrity [9]. These ethical and legal frameworks are essential for tackling deceptive digital marketing practices [10].

Recent scholarship has emphasized the shift toward more resilient detection models using behavior and textual information [11]. For instance, combining convolutional neural networks (CNN) with optimization algorithms has shown promise in unmasking deceptive patterns [12]. Furthermore, comprehensive surveys of machine learning techniques highlight the ongoing battle between deceptive tactics and evolving detection models [1].

In several studies, the characteristics of deceptive reviews have been identified. For instance, [4] found that the use of inflated language, multiple repetitions of a common phrase, and excessively positive promotional language are the major identifying features of fake reviews. The use of these patterns can be captured by machine learning models and help in classification. More recently, researchers have been looking at ways to improve detection accuracy beyond just using text to include such things as reviewer behavior, rating patterns, and temporal activity. Evidence has shown that using both text and behavioral features collectively in the design of a fake review detecting system provides a better overall understanding of review authenticity [13].

AI-generated content presents new problems. More advanced language models can create extremely believable and accurate reviews that look very similar to reviews that a real person wrote. Traditional ways to find fake reviews will probably not be able to find these more sophisticated types of fakes [6] With regard to the law, fake reviews are being treated as a type of deceptive and unhealthy business practice; they can cause consumers to be misled, prevent them from being able to accurately evaluate their choices as customers and give certain businesses an unfair advantage over their competition. Businesses that post fake reviews may be violating consumer protection laws and other regulations regarding misleading advertising.

As discussed by [5], if we want to maintain our confidence in and use of digital platforms, we will need to combine technological solutions with other laws to provide an effective legal framework for the companies using these platforms. Businesses must take responsibility for their actions through regulation and be held accountable to ensure online platforms will monitor and take down deceptive content. All research indicates, despite machine learning being a very effective method of identifying fake reviews, that to solve the fake review problem requires a cooperative approach between technology to identify the deception and enforcing legal remedies to address the issue. Both perspectives must be taken to produce fairness and transparency in and protect consumers from deceptive content available online through e-commerce.

Most past research has either focused on enhancing machine learning algorithms for fake review detection or discussing the legal ramifications of deceptive online practices. Few studies have combined both aspects while employing benchmark datasets with verified labels. Besides, most studies only consider a single classification model, which makes it hard to ascertain the efficacy of the various machine learning approaches for fake review detection in e-commerce.

Based on these issues, this study intends to use machine learning as a means of detecting fraudulent reviews posted on the internet. The study will also consider the potential legal repercussions relating to the practice of posting fraudulent reviews.

Additionally, the focus of the study is to analyze Amazon reviews and assess the success of a Naive Bayes statistical approach for identifying false reviews and corresponding legal implications.

This paper presents an analysis that fuses technological and legal aspects of the detection of fake online reviews. Previous research has focused on improving detection algorithms or exploring the challenges of regulation separately; this research combines both in a single framework. It provides a comparative analysis of Naive Bayes, Logistic Regression, and Random Forest methods using a benchmark dataset of Amazon reviews with verified labels. At the same time, it looks into the legal and regulatory aspects of deceptive reviews in electronic marketplaces. By combining computational analysis with consumer protection and governance of platforms, this paper helps to improve our understanding of the detection of fake reviews and, consequently, of building trustworthy and responsible e-commerce environments.

2. Research Methods

The systematic steps of the research process, ranging from initial data collection to final legal analysis, are detailed in Figure 1 below.

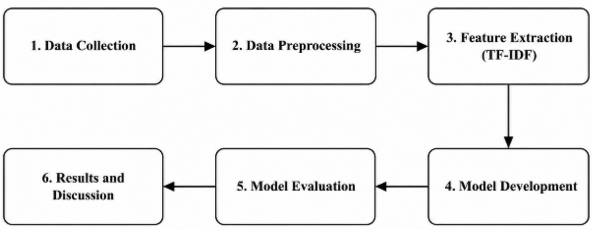

Figure 1. Methodology

2.1. Research Design

Our research uses quantitative methods of statistical analysis combined with machine learning technology to determine whether or not Internet reviews are fraudulent or authentic. The intent of this approach is to assess the efficacy of computational models in identifying deception by identifying the patterns found in language and measuring them through objective measures of word frequency and grammatical structure.

2.2. Dataset Description

This study used the Fake Reviews Dataset from Kaggle [14]. It contains reviews that were previously classified into two categories: Computer Generated (CG) and Original Reviews (OR). The use of a benchmark dataset that contains verified labels removes the need for manual or rule-based labelling and makes the classification process more reliable. After cleaning the data and removing duplicates, the dataset ended up pretty balanced — it had roughly the same number of reviews in each class. Fake reviews are those that have been identified and labelled within the benchmark dataset as deceptive or computer-generated content. The benchmark labels are treated as the ground truth for model training and evaluation.

2.3. Data Pre-processing

The reviews collected were put through a few pre-processing steps before the model could be trained. All text was first converted to lowercase, and then everything apart from alphabets was removed. Punctuation marks, special characters, URLs, and unnecessary spaces were removed in the second step. Reviews that were duplicates were removed so that repeated information did not affect the model's performance. The last step was to convert the cleaned textual data into a numerical representation using the Term Frequency-Inverse Document Frequency (TF-IDF) technique.

2.4. Model Implementation

The Naive Bayes Classifier (NBC) is a probabilistic model grounded in Bayes' Theorem, which calculates the probability of a review belonging to a class (C) given its observed textual features $(x_1, \dots, x_n)$.

$$P(C|x_1, \dots, x_n) = \frac{\left(P(C) * \prod P(x_i|C)\right)}{P(x_1,\dots,x_n)} \tag{1}$$

The model is characterized by the conditional independence assumption, which treats each word as an independent feature. By multiplying the individual likelihoods ($\prod P(xi|C)$) and the class prior ($P(C)$), the classifier predicts the most likely category for a review based on its word distribution. This approach is particularly effective for high-dimensional text classification tasks where computational efficiency is required [1]. Furthermore, recent research confirms that the probabilistic nature of Naive Bayes makes it a robust baseline for processing large-scale datasets in the context of e-commerce fraud detection [15].

A classification task using the (NBC) algorithm is very popular for text classification because it is easy to implement, computationally efficient, and performs well in high-dimensional feature spaces. In this study, the primary model is the Naive Bayes classifier as it provides an effective probabilistic baseline for fake review detection, especially when word frequency patterns are involved. The model's assumption of conditional independence allows it to perform remarkably well even with smaller datasets, making it a highly interpretable and resource-efficient choice for e-commerce fraud detection [11].

To extend the comparison beyond this baseline, Logistic Regression and Random Forest were also used. This multi-model approach allows a more comprehensive assessment of how different algorithmic structures, ranging from probabilistic models to ensemble learners, will capture the nuances of deceptive online reviews.

The dataset was split into training and testing sets with an 80:20 stratified split. This ensured that 80% of the data was used to develop the model and the remaining 20% was used for independent evaluation. Stratified sampling was employed, using a fixed random seed of 42 to keep the original class distribution (Fake vs. Genuine) intact in both subsets, thus avoiding any sampling bias and making the experimental results reproducible.

2.5. Model Evaluation

The machine learning models were assessed on their performance using metrics such as accuracy, precision, recall, F1-score, and confusion matrix analysis. Accuracy is the measure of how correct the overall classification process is, while precision and recall measure the model's ability to correctly identify reviews that belong to each category. The F1-score was used as it gives an equal balance between precision and recall. This is in addition to a five-fold cross-validation used to check for stability and reliability of the classification models.

2.6. Implementation Environment

All experiments were conducted using Google Collab[16], a cloud-based platform that supports Python programming and machine learning libraries. This environment enables efficient execution of the model without requiring high local computational resources. The specific configuration used for this research included a Python 3 runtime type, a CPU hardware accelerator, and the Latest (recommended) runtime version.

2.7. Methodology Overview

The overall methodology of this study follows a structured process, beginning with dataset collection and ending with result analysis. The process includes data selection, pre-processing, feature extraction, model training, and evaluation.

3. Results and Discussion

The Figure 2 Distribution of reviews in the benchmark dataset. The dataset contains a relatively balanced number of Computer Generated and Original Reviews, which minimizes the chances for classification bias and allows a more dependable evaluation of model performance.

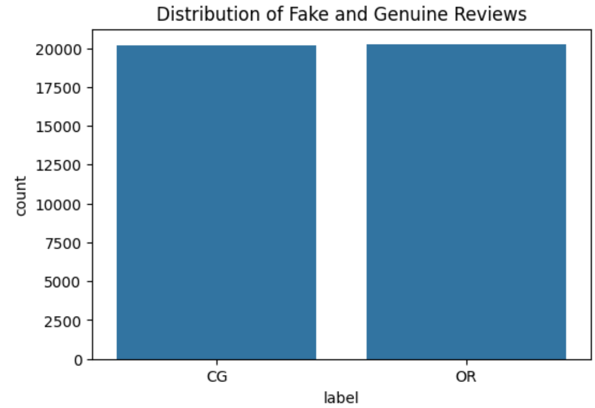

Figure 2. distribution of fake and real reviews

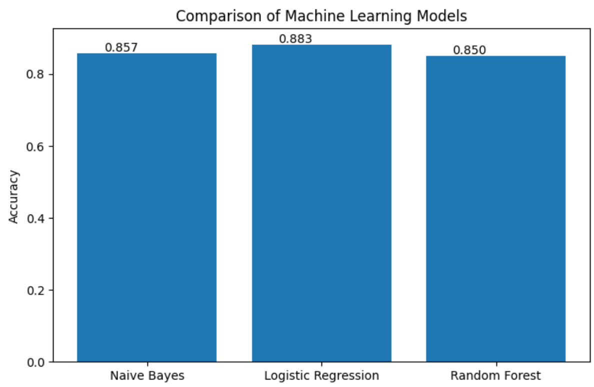

Figure 3. comparison of ML Models

To assess the performance of various machine learning methods for fake review detection, three classification algorithms—Naive Bayes, Logistic Regression, and Random Forest—were implemented and their outcomes compared. The classification results are on Table 1.

Table 1. Classification Accuracy of Machine Learning Models

Model	Accuracy	Precision	Recall	F1-score
Naive Bayes	85.68%	86.00%	86.00%	86.00%
Logistic Regression	88.26%	88.00%	88.00%	88.00%
Random Forest	84.99%	85.00%	85.00%	85.00%

Table 1 presents the results of the classification accuracy. Logistic Regression achieved the highest accuracy, 88.26%, followed by Naive Bayes with 85.68%, and Random Forest with 84.99%. The superior performance of Logistic Regression is probably due to its ability to model well the high-dimensional textual data from TF-IDF vectorization. Though not the best in terms of predictive performance, Naive Bayes returned results that could be considered competitive while keeping computational simplicity and interpretability. Random Forest, on the other hand, had the lowest accuracy among the three, which can be ascribed to the sparsity of textual features from TF-IDF.

To enhance the analysis, extra performance metrics were computed for the Naive Bayes model. The results are given in Table 2.

Table 2. Performance Metrics of NBC

Metric	Value
Accuracy	85.68%
Precision	84.64%
Recall	87.21%
F1-score	85.91%

The Naive Bayes classifier attained a precision of 84.64%, meaning that about 85% of the predicted classifications were right. The recall value, at 87.21%, shows that the model can correctly identify reviews that belong to each category. Additionally, the F1-score (85.91%) shows a balanced trade-off between precision and recall, indicating that the model performs consistently across both classes.

The confusion matrix depicted in Figure 4 gives an elaborate view of the classification results that the Naive Bayes model achieved.

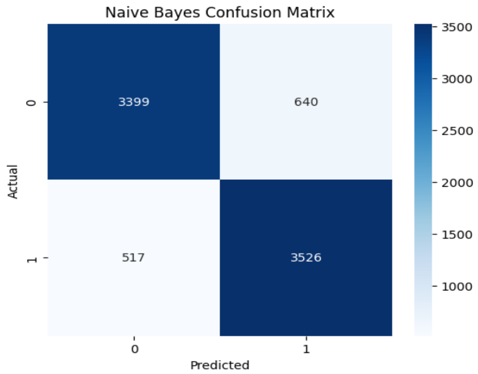

Figure 4. Confusion Matrix of Naive Bayes

The model correctly classified 3,399 Computer Generated reviews and 3,526 Original Reviews. However, 640 reviews were misclassified, 517 of which were in the opposite category. The relatively small difference between false positives and false negatives indicates that the classifier is not strongly biased toward one category and maintains relatively balanced performance in detecting both classes of reviews.

To evaluate the stability and reliability of the proposed approach, five-fold cross-validation was conducted. The results are presented in Table 3.

Table 3. Cross-Validation Results

Fold	F1-score
1	0.8356
2	0.7313
3	0.7606
4	0.8271
5	0.8193
Mean	0.7948

An average cross-validation F1-score of 79.48% shows that the model performs fairly consistently over different subsets of the dataset. There was some variation between the folds, but it can be said that the proposed approach has acceptable generalization capabilities for effective classification of deceptive online reviews.

The results of this study show that machine learning models can be used successfully to identify fake online reviews in e-commerce. Such a system, from a practical perspective, would help e-commerce platforms reduce misinformation, increase transparency in their marketplaces, and shield consumers from manipulative practices. From a legal perspective, fake reviews can be considered deceitful commercial practices that contravene consumer protection laws and disturb fair market competition. The more advanced artificial intelligence gets at creating credible fake reviews, the more necessary it is to have automated detection systems backed up by a stronger regulatory framework. However, automated detection systems will never be 100% accurate and therefore should be seen as supporting tools in the decision-making process and not as self-sufficient legal enforcement instruments.

4. Conclusions

This study aimed to apply machine-learning methodologies in the detection of fake review listings and assess the possible legal implications for e-commerce platforms. Results show that the use of the Naive Bayes classification algorithm is a good approach to predict the authenticity of reviews, having an accuracy of 85.68%, precision of 84.64%, recall of 87.21%, and F1-score of 85.91%. Logistic Regression gave slightly higher classification accuracy at 88.26% but the Naive Bayes model showed competitive performance with simplicity, computational efficiency, and interpretability such that it could be considered as a practical baseline approach for fake review detection.

In general, these findings suggest that simple machine-learning methods may be able to capture the general pattern of text data and not accurately identify all types of deceptive reviews. Similar limitations have been observed in other studies; performance is negatively affected by various factors such as model complexity, fake content, and linguistic diversity [6]. Although a benchmark dataset with verified labels was used in this study, it is still limited because only textual features were considered and no advanced deep learning approaches were investigated. Thus, future research should focus on larger and more diverse datasets, incorporate behavioral and metadata features, and explore advanced machine-learning and deep-learning models to achieve better detection results.

Fake online reviews have proven very problematic in court because they can mislead consumers and tilt the playing field among businesses. Fake review technology can also violate consumer protection laws. The problem of fake reviews can only be solved through better technology for detecting review fraud and more regulations that create transparency and accountability for digital platforms. Machine-learning systems are not perfect, though, so they should be used as decision support rather than completely automated enforcement. Perhaps in the long run, however, machine learning might be effectively deployed in the fight against fake reviews to restore trust in digital marketplaces.

References

[1] M. J. Abd and M. H. Hussein, "Fake reviews detection in e-commerce using machine learning techniques: a comparative survey," in *BIO Web of Conferences*, EDP Sciences, Apr. 2024. doi: 10.1051/bioconf/20249700099.

[2] J. M. Martínez Otero, "Fake reviews on online platforms: perspectives from the US, UK and EU legislations," *SN Social Sciences*, vol. 1, no. 7, p. 181, 2021, doi: 10.1007/s43545-021-00193-8.

[3] K. Waitz, "A SHIFT IN THE TIDES? THE WELCOMED PROPOSAL OF HARSHENED FTC GUIDELINES FOR SOCIAL MEDIA REVIEWS AND ADVERTISING." [Online]. Available: https://www.ftc.gov/about-ftc

[4] W. H. Asaad, R. Allami, and Y. H. Ali, "Fake Review Detection Using Machine Learning," *Revue d'Intelligence Artificielle*, vol. 37, no. 5, pp. 1159–1166, 2023, doi: 10.18280/ria.370507.

[5] A. M. Santos and N. Antonio, "Improving trust in online reviews: a machine learning approach to detecting artificial intelligence-generated reviews," *Information Technology and Tourism*, vol. 27, no. 3, pp. 739–766, Sep. 2025, doi: 10.1007/s40558-025-00329-z.

[6] R. Gupta, V. Jindal, and I. Kashyap, "Recent state-of-the-art of fake review detection: a comprehensive review," Nov. 21, 2024, *Cambridge University Press*. doi: 10.1017/S0269888924000067.

[7] A. H. Alshehri, "An Online Fake Review Detection Approach Using Famous Machine Learning Algorithms," *Computers, Materials and Continua*, vol. 78, no. 2, pp. 2767–2786, 2024, doi: 10.32604/cmc.2023.046838.

[8] Federal Trade Commission, "Federal Trade Commission Announces Final Rule Banning Fake Reviews and Testimonials," 2024. [Online]. Available: https://www.ftc.gov/news-events/news/press-releases/2024/08/federal-trade-commission-announces-final-rule-banning-fake-reviews-testimonials

[9] Competition and M. Authority, "Short guide for businesses: publishing consumer reviews and complying with consumer protection law," 2025. [Online]. Available: https://www.gov.uk/government/publications/fake-reviews-cma208/short-guide-for-businesses-publishing-consumer-reviews-and-complying-with-consumer-protection-law

[10] Miracle Eze, "The ethics of digital marketing: Tackling fake reviews and deceptive practices," *International Journal of Management & Entrepreneurship Research*, vol. 7, no. 1, pp. 16–42, Jan. 2025, doi: 10.51594/ijmer.v7i1.1796.

[11] D. Zhang, W. Li, B. Niu, and C. Wu, "A deep learning approach for detecting fake reviewers: Exploiting reviewing behavior and textual information," *Decis. Support Syst.*, vol. 166, Mar. 2023, doi: 10.1016/j.dss.2022.113911.

[12] N. Deshai and B. Bhaskara Rao, "Unmasking deception: a CNN and adaptive PSO approach to detecting fake online reviews," *Soft comput.*, vol. 27, no. 16, pp. 11357–11378, Aug. 2023, doi: 10.1007/s00500-023-08507-z.

[13] P. Sun *et al.*, "Fake Review Detection Model Based on Comment Content and Review Behavior," *Electronics (Switzerland)*, vol. 13, no. 21, Nov. 2024, doi: 10.3390/electronics13214322.

[14] J. Salminen, "Amazon reviews." Accessed: Jun. 05, 2026. [Online]. Available: https://www.kaggle.com/datasets/mexwell/fake-reviews-dataset?resource=download

[15] I. Journal, "FAKE REVIEW DETECTION USING DEEP LEARNING," *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, vol. 08, no. 01, pp. 1–11, Jan. 2024, doi: 10.55041/ijsrem27893.

[16] Niamatullah zubair, "Fake Review Detection Model Implementation." Accessed: Jun. 05, 2026. [Online]. Available: https://colab.research.google.com/drive/1Lhayw_FBNaHh7kIdRoLfyWRC4wp5O-c0?usp=sharing