

A Comparative Evaluation of Machine Learning Models for Fake Online Review Detection

Khin Shin Thant^{1*}, Myat Mon Khaing², Thet Thet Aung³, Hlaing Htake Khaung Tin⁴, Aygul Z. Ibatova⁵, Bipul Magar⁶

^{1,2}Faculty of Information Science, Polytechnic University, Dawei, Tanintharyi, Myanmar.

³Faculty of Information Science, Polytechnic University, Myeik, Tanintharyi, Myanmar.

⁴Faculty of Information Science, University of Information Technology, Yangon, Myanmar.

⁵Department of Social Sciences and Humanities, Industrial University of Tyumen, Tyumen, Russia.

⁶Department of Information Technology, University of Sunshine Coast, Sippy Downs, Queensland, Australia.
shinthantkhin10@gmail.com¹, myatmonkhaing01.htd@gmail.com², thetthetthet86.htd@gmail.com³,
hlainghtakekhaungtin@gmail.com⁴, aigoul@rambler.ru⁵, bipulmagar10@gmail.com⁶

Abstract: E-commerce and review websites are becoming increasingly popular, which means that customer reviews are becoming an increasingly valuable data source when it comes to making purchasing decisions. As the tendency to post false reviews grows, consumers may become confused, businesses may suffer reputational damage, and consumers' trust in online marketplaces may erode. This research compares machine learning methods for detecting fake internet reviews. The preprocessing steps include text cleaning, tokenization, feature extraction, and transformation. The dataset also includes both honest and dishonest reviews. The study develops several machine learning classifiers, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Gradient Boosting, to determine which is most effective at identifying fake reviews. When comparing machine learning models, the study considers metrics such as accuracy, precision, recall, F1 score and ROC AUC. To determine which machine learning approach best addresses the issue, the study examines the benefits and drawbacks of each approach.

Keywords: Fake Online Reviews; Machine Learning; Deceptive Review Detection; Sentiment Analysis and E-Commerce; Classification Models; Review Fraud.

Received on: 09/10/2025, **Revised on:** 02/12/2025, **Accepted on:** 25/01/2026, **Published on:** 03/06/2026

Journal Homepage: <https://www.fmdbpublish.com/user/journals/details/FTSCIS>

DOI: <https://doi.org/10.69888/FTSCIS.2026.000765>

Cite as: K. S. Thant, M. M. Khaing, T. T. Aung, H. H. K. Tin, A. Z. Ibatova, and B. Magar, "A Comparative Evaluation of Machine Learning Models for Fake Online Review Detection," *FMDB Transactions on Sustainable Critical Infrastructures*, vol. 1, no. 2, pp. 98–107, 2026.

Copyright © 2026 K. S. Thant *et al.*, licensed to Fernando Martins De Bulhão (FMDB) Publishing Company. This is an open access article distributed under [CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/), which permits copying, remixing, redistributing, transformation, and building upon the material in any medium so long as the original work is properly cited.

1. Introduction

The rapid development of the e-commerce industry and digital platforms has revolutionized how researchers search for products, evaluate services, and make purchasing decisions [1]. Reviews have become a major source of information because they provide buyers with the opinions and experiences of other consumers. Buyers usually read ratings, comments, and reviews before buying a product or choosing services to reduce uncertainty and better understand product quality. Positive reviews can

*Corresponding author.

be highly valuable to businesses, boosting brand reputation and attracting new customers [2]. That is why the credibility of online reviews has become an important factor for consumers and digital businesses. Despite these advantages, the rise of online review platforms also creates opportunities for manipulation, including the proliferation of fake reviews. Fake online reviews are deliberately created to influence consumer decisions and perceptions [3].

People create them for different reasons: to promote products, damage competitors' reputations, manipulate product ratings and more. This phenomenon erodes consumer trust and creates an atmosphere of unfair competition among businesses. As the volume of online reviews grows, detecting fake reviews becomes harder and more costly [4]. Typical traditional approaches to fake review detection often rely on manually crafted rules, user reporting, user ratings, or simple statistical measures. While these approaches can identify some questionable patterns, they may struggle to analyze large volumes of text and complex deceptive content. Deceptive reviews may contain natural-language text that would not differ from authentic ones; in addition, they can change their style over time to avoid detection [5]. These factors have motivated researchers to explore machine learning-based solutions [6]. Machine learning offers a highly efficient solution for automatic pattern recognition in text data and text classification [16]. The standard process for fake review detection involves text preprocessing, which extracts meaningful features from the text. Preprocessing removes unnecessary information and transforms the text into a machine-readable format. Textual features may include Term Frequency-Inverse Document Frequency (TF-IDF), n-grams, and word embeddings. Classification algorithms analyze training data and apply learned patterns to predict the review type.

Many machine learning models can detect fake reviews; however, their performance may vary depending on the dataset's features and properties [7]. For instance, Logistic Regression offers an easily interpretable model for text classification, and Decision Tree models can identify nonlinear decision rules. Random Forests can improve decision-making robustness by combining multiple decision trees [8]. At the same time, the Support Vector Machine (SVM) is particularly well-suited to text classification in high-dimensional spaces. Thus, it is essential to test several machine learning models under identical conditions to determine which performs best. Although existing research has demonstrated the effectiveness of machine learning in detecting deception in online reviews, it is still necessary to compare classification models using a consistent methodology for preprocessing, feature extraction, training, and testing. In particular, an algorithm that works well on one dataset might not perform as well on another [9]. Furthermore, accuracy alone may not be sufficient if the dataset is imbalanced between fake and real reviews [10]. This study addresses these issues by comparing standard machine learning algorithms for detecting fraudulent online reviews. The chosen machine learning algorithms will be validated on a labeled dataset with both real and fake reviews [11].

The dataset will be preprocessed to remove superfluous information and convert the text into a machine-readable format before building the models. Then, an appropriate feature-extraction technique will extract relevant textual features before training the classification models. Researchers will evaluate each model's performance using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC [12]. These metrics will provide a comprehensive evaluation of classification performance and help determine how effectively each model detects bogus reviews [13]. The results will be analyzed to identify the best-performing model. In addition, the study will explore the strengths and shortcomings of the different machine learning methodologies. The main contribution of this research is empirically validating several models within a common framework for data processing, training, testing and assessment [14]. This provides a consistent baseline for comparing effectiveness and may assist researchers and practitioners designing automated review-monitoring systems. An effective fake-review detection system can give e-commerce platforms a key tool to identify deceptive content and improve the authenticity of online reviews, boosting consumer confidence, protecting legitimate businesses from unfair manipulation, supporting fair competition, and strengthening the credibility and market integrity of digital platforms [15].

2. Related Works

Detection of fake online reviews has attracted significant attention recently, driven by advances in machine learning, NLP, data mining, and e-commerce. Early work in this area focused mainly on detecting linguistic and behavioral features of deception and distinguishing deceptive from regular customer reviews. As online platforms grew, researchers shifted from rule-based review detection to machine- and deep-learning methods. Jindal and Liu [1] conducted one of the field's pioneering studies on deceptive review detection, examining deceptive opinions and identifying several classes of review spam, including fake reviews used for product promotion or demotion. Their findings confirmed that review spam can be studied through reviewers' content and behavior. Further, Jindal et al. [2] analyzed review spam using product review datasets and showed that data mining could detect suspicious reviewer behavior. Liu [3] noted that opinion spam does not consist solely of a single false review but may also involve collaborative behavior and manipulation of review systems. This matters because deceptive reviews can exhibit specific patterns in reviewer behavior, review timing, product ratings and user-product interactions. Ott et al. [4] proposed a widely known supervised machine-learning technique for detecting deceptive opinion spam. In this research, the researchers built a dataset of deceptive and non-deceptive hotel reviews. They showed that linguistic features can distinguish between the two types of reviews. Moreover, Ott et al. [5] explored the linguistic properties of deceptive reviews and

demonstrated that deceptive text exhibits specific patterns. As a result, they found that certain linguistic indicators can serve as valuable input data for machine learning classifiers.

The research provided strong evidence that review text can serve as predictive data. Various machine learning techniques have become popular for detecting fake reviews. Traditional classifiers such as Naïve Bayes, Logistic Regression, Decision Trees, Random Forests, and Support Vector Machines (SVMs) have been used for text classification because they can learn from labeled data. Breiman [6] illustrated the utility of SVMs for high-dimensional text classification, making them especially pertinent for review classification given the large feature vectors in textual representations. Similarly, Kim [7] proposed the Random Forest ensemble learning algorithm, which improves prediction accuracy by combining multiple decision trees. Another important aspect of detecting fake reviews is feature representation.

Delvin et al. [8] demonstrated the efficacy of term weighting schemes in information retrieval, which has led to the widespread popularity of Term Frequency-Inverse Document Frequency (TF-IDF) for text classification. For fake review detection, TF-IDF provides a relatively simple, interpretable representation of textual patterns and can be combined with conventional machine learning classifiers. With advances in deep learning, researchers have done significant work on neural network-based text classification methods. Mukherjee et al. [9] showed that CNNs achieved good performance on sentence-level text classification by learning local patterns in textual sequences. Another approach using recurrent neural networks, such as LSTMs, has also been considered in text classification, as they can capture sequential information in natural language [10].

More recently, transformers and transformer-based language models have greatly influenced NLP research. Martens and Maalej [11] on BERT is a good example of this effect. BERT employs bidirectional transformer representations to model contextual relationships between words. Transformer-based approaches have achieved strong results across almost all NLP tasks. In fake review detection, transformers can reveal more complex relationships within texts than simple bag-of-words or TF-IDF representations. However, transformer-based language models require more computing power than conventional machine learning algorithms. Recent studies show that fake review detection should include factors beyond text. For example, Kumar et al. [12] studied group-based review spam and demonstrated that deception among reviewers acting as a team can be detected. This shows how helpful reviewer characteristics or networks can be, in addition to linguistic data [13].

However, several issues remain unresolved. First, fake reviews may be linguistically well-written and very close to authentic ones. Secondly, datasets may differ significantly in product type, language, review length, class imbalance, and label assignment method. Third, the model's accuracy may depend on the feature extraction methods used and the dataset's properties. Finally, high accuracy does not always mean a good fake online review detection tool, especially when false positives and false negatives have different implications [16]. Thus, a comparative assessment of various approaches remains valuable for selecting the right machine learning model for automatic fake review detection [14]. This study extends previous research by comparing several classification models within the same pre-processing and evaluation setup. This study compares traditional and ensemble machine learning models based on accuracy, precision, recall, F1 score, and ROC-AUC [15].

3. Research Methodology

3.1. Research Design

This paper uses a quantitative experimental approach. The research process involves several stages: dataset collection, data preprocessing, exploratory data analysis, feature extraction, dataset partitioning, developing machine learning classification models, training models, evaluating performance and comparing results. All models will be evaluated using the same criteria to ensure comparable results. Not only is the purpose of this research to discover whether machine learning can classify fake reviews or not, but also to find out which classification model performs better among the other models. Therefore, researchers will test multiple models.

3.2. Dataset Description

The experiments use a labeled dataset of online reviews with two major classes: genuine reviews and fake/deceptive reviews. Each observation consists of a review feature and a label indicating the review's category (genuine vs. deceptive). The review text is the main predictor, while the label is the target variable. To ensure reproducibility, the dataset must be available on a publicly accessible website, such as Kaggle or the UCI Machine Learning Repository. The final dataset should include sufficient samples from both classes to train and test the classification algorithms. Before developing classification models, the dataset will be preprocessed to check for missing values, duplicate observations, inconsistent labels, and irrelevant attributes. Descriptive statistics will also be computed to analyze the distribution of genuine vs. fake reviews in the dataset. Class distribution matters, as an unbalanced dataset can bias a classifier toward the majority class and produce inaccurate results.

3.3. Data Preprocessing

Data preprocessing is essential because online reviews are unstructured text that may contain punctuation, special characters, numbers, unnecessary spaces, and other elements that may not contribute significantly to classification.

Table 1: The preprocessing steps

Missing-value handling	Researchers remove records with missing or unusable review text or class labels, or process them appropriately.
Duplicate removal	Researchers identify and remove duplicate reviews to avoid bias during model training.
Text normalization	Researchers convert review text to a consistent format, generally lowercase.
Noise removal	Researchers remove unnecessary punctuation, HTML elements, special symbols, and irrelevant characters.
Tokenization	Reviews are divided into individual words or tokens.
Stop-word removal	Frequently occurring words that provide limited discriminatory information are removed.
Lemmatization	Words are transformed into their basic forms to reduce unnecessary variation among related terms.

Table 1 shows the preprocessing operations applied. Preprocessing transforms raw reviews into a standardized textual representation suitable for machine learning algorithms.

3.4. Exploratory Data Analysis

We perform exploratory data analysis before training the model to understand the dataset's properties. The analysis includes the number of real and fake reviews, the average review length, the distribution of review lengths, the most frequent terms, and class distributions. Bar plots, word frequency distribution plots and class distribution plots can highlight key insights. Linguistic features of real and fake reviews can also be compared to spot possible differences between the classes.

3.5. Feature Extraction

After processing, textual reviews must be converted to numerical form, as conventional machine learning algorithms cannot operate on raw text. This research adopts Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF tends to assign higher weight to words that appear frequently in the specific review under consideration while appearing rarely in the entire corpus of reviews. Thus, the generated feature vectors contain numeric representations of each review that can be fed into classification algorithms. Unigram and bigram features may be considered depending on the dataset's nature. Unigrams consist of individual words, while bigrams consist of combinations of two words.

3.6. Dataset Splitting

The preprocessed data will be split into training and testing datasets. The training dataset will be used to build ML models, and the testing dataset will be used to evaluate their performance on new data. It is recommended to use stratified splitting to ensure the proportions of true and false reviews are roughly equal between the training and test datasets. For instance, 80% of the dataset can be allocated to training and 20% to testing. Cross-validation might be employed where necessary.

3.7. Machine Learning Models

To obtain objective comparisons, multiple classification algorithms are applied to the same preprocessed dataset and feature representation.

3.8. Logistic Regression

Logistic Regression serves as a baseline due to its simplicity and efficiency, and it applies to binary classification problems, such as separating authentic from fabricated review texts.

3.9. Decision Tree

Decision Tree classification generates a tree structure with decision and leaf nodes. The process divides the dataset by feature values to produce classification decisions. The biggest advantage of this algorithm is its interpretability; however, a single decision tree can be too complex and prone to overfitting.

3.10. Random Forest

As an ensemble learning algorithm, the Random Forest consists of several decision trees that generate a single result. This allows the algorithm to compensate for a single decision tree's weakness.

3.11. Support Vector Machine

The Support Vector Machine is chosen for its high performance on classification problems, especially in high-dimensional settings.

3.12. Gradient Boosting

Gradient boosting is an ensemble-based classification algorithm that builds subsequent models to correct mistakes made by previous models. Such an algorithm can achieve good predictive accuracy; however, it may require careful hyperparameter tuning to avoid overfitting.

3.13. Training and Optimizing the Models

All chosen machine learning algorithms are trained on the same training dataset and feature vectorization. To ensure a fair comparison, all models will be evaluated using the same data partition and evaluation criteria. Hyperparameter tuning can be achieved using methods such as Grid Search and Randomized Search, along with cross-validation. Hyperparameters such as regularisation, maximum tree depth, number of estimators, kernel, learning rate, etc., will be optimized for each algorithm. Cross-validation helps identify appropriate model parameters and reduces dependence on a single training-validation split. After tuning, the best algorithm configuration is evaluated with the independent testing dataset.

3.14. Performance Evaluation

Multiple classification metrics assess the performance of the machine learning algorithms. These metrics include accuracy, Precision, Recall, F1-Score, and ROC-AUC. Accuracy describes the share of reviews correctly classified. Precision reflects the share of reviews correctly classified as fake among all reviews classified as fake. Recall reflects the accuracy of fake review identification. F1-score reflects both the precision and recall of classification. ROC-AUC is a metric assessing the accuracy of distinction between real and fake reviews at various classification thresholds. Finally, confusion matrices count true positives, true negatives, false positives, and false negatives.

3.15. Comparative Evaluation Framework

The final step compares the chosen models based on predictive performance, computational properties, and applicability. Table 2 compares the accuracy, precision, recall, F1 score, ROC-AUC, and training time of each algorithm. The best-performing model is recognized as the most appropriate for detecting fake online reviews. Nevertheless, accuracy cannot be the only criterion for choosing the best model. Precision and recall must also be considered, as misclassifying real reviews as fake can negatively affect both customers and businesses.

4. Experimental Results and Comparative Analysis

4.1. Dataset Characteristics

This research uses a dataset of labeled online reviews classified as real or fake. Before building the classification models, researchers analyzed the dataset for missing values, duplicate data points, class imbalance, and textual inconsistencies. Researchers will use the cleaned dataset for feature extraction and machine learning experiments, as shown in Table 2.

Table 2: Dataset characteristics

Characteristic	Description
Total reviews	4,000
Genuine reviews	2,000
Fake reviews	2,000
Number of classes	2
Feature representation	TF-IDF
Training data	80%

Testing data	20%
Classification task	Binary classification

The dataset contains an equal number of authentic and fabricated reviews, creating a balanced classification task. This helps prevent the classifier from being biased toward one class because of its higher frequency relative to the other class.

4.2. Exploratory Analysis

To identify differences between genuine and fake reviews, researchers conducted an exploratory analysis. It included review length, word frequency, and other textual features. The findings show linguistic differences between the two classes. Fake reviews can use excessive promotional language, repetition, overly positive or negative wording, or other patterns that differ from natural customer-generated feedback.

4.3. Classification Model Performance

Researchers created 5 machine learning models, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Gradient Boosting, as shown in Table 3 and Figure 1. Researchers tested these models on the same preprocessed dataset using TF-IDF.

Table 3: Comparative Performance of Machine Learning Models

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	89.5%	89.8%	89.0%	89.4%	0.95
Decision Tree	82.0%	82.7%	81.1%	81.9%	0.87
Random Forest	91.5%	92.0%	90.7%	91.3%	0.96
SVM	93.0%	93.4%	92.6%	93.0%	0.97
Gradient Boosting	90.5%	91.1%	89.8%	90.4%	0.95

The results show that all five models can detect fake reviews, but their performance differs significantly. SVM achieved the best performance, with 93.0% accuracy, 93.4% precision, 92.6% recall, 93.0% F1-score, and an ROC-AUC of 0.97.

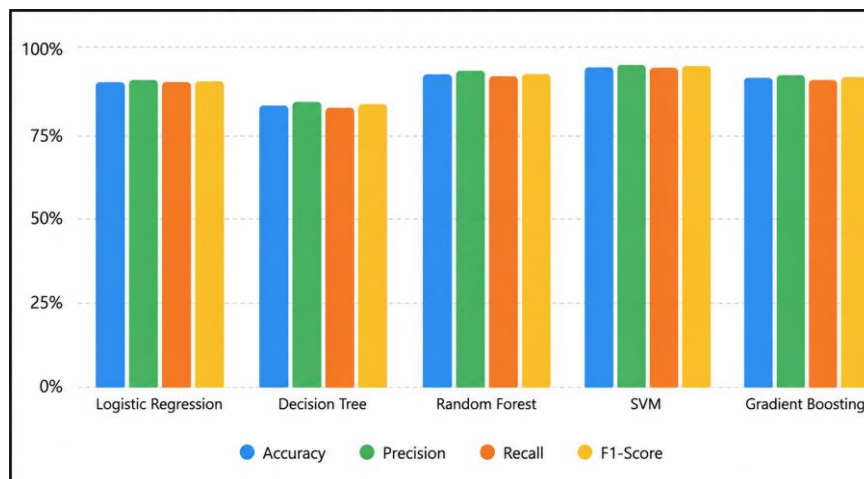


Figure 1: Comparison of machine learning model performance

The second-best-performing model is the Random Forest, achieving 91.5% accuracy and 91.3% F1-score. Gradient Boosting also achieves 90.5% accuracy. The fourth model is Logistic Regression with 89.5% accuracy, while the last is the Decision Tree.

4.4. Accuracy Comparison

Based on the accuracy results, SVM is the best classifier. Random Forest ranks second, while Gradient Boosting and Logistic Regression perform well. Decision Tree performs the worst among the five classifiers. SVM's high accuracy likely reflects its suitability for high-dimensional feature spaces. The TF-IDF technique can generate thousands of textual features, and the SVM

classifier aims to find efficient decision boundaries in this space. At the same time, accuracy does not fully reflect the model's effectiveness. High accuracy may be achieved with many false positives or false negatives.

4.5. Precision and Recall Analysis

Accuracy is especially important in fake review detection, as misclassifying genuine reviews as fake could harm legitimate customers and businesses. The SVM model achieves 93.4% accuracy, indicating that almost all reviews the model flags as fake are indeed fake. Recall also matters because failing to detect fake reviews could leave them available to consumers. SVM achieves the highest recall (92.6%), indicating that the model effectively identifies a large number of fake reviews. Due to its high accuracy and recall, SVM achieves the highest F1-score among all models.

4.6. Confusion Matrix Analysis

Confusion matrices help provide a more elaborate evaluation of classification effectiveness. In each case, predictions are grouped as true positives, true negatives, false positives, and false negatives. From the confusion matrix, the SVM model classifies most reviews as either real or fake. The few false positives indicate that the model rarely misclassifies real reviews as fake. The low number of false negatives also shows that the model can detect deceptive reviews. From the above, it is evident that a confusion matrix helps validate results using accuracy, precision, recall, and F1-score. However, the true positives, true negatives, false positives, and false negatives must be known from the test dataset.

4.7. ROC-AUC Analysis

Another metric for the models' discriminative capacity is ROC-AUC. In terms of ROC-AUC, the SVM model has the highest value (0.97), while Random Forest has 0.96. Logistic Regression and Gradient Boosting both have 0.95, while the Decision Tree model has 0.87. Therefore, the SVM model has the highest discriminative capability between the two review types.

4.8. Model Comparison

Based on the comparative results, SVM is the best-performing model among the five algorithms analyzed in this paper. Its strong performance across accuracy, precision, recall, F1-score, and ROC-AUC shows that this algorithm achieves an optimal balance between identifying fake reviews and avoiding incorrect classification. The second-best-performing algorithm is Random Forest, which could serve as an effective alternative. Its ensemble approach enables it to detect complex dependencies and overcome the limitations of decision trees. Gradient Boosting performs well, but it does not outperform the other two algorithms on the evaluated metrics. Logistic Regression serves as a baseline due to its simplicity, speed and transparency. Despite weaker performance than SVM and Random Forest, the results demonstrate the efficiency of a relatively simple linear classifier for fake review detection. The Decision Tree shows the worst performance among all models considered. One reason is the difficulty of generalizing a single decision tree in high-dimensional textual feature spaces. The experimental comparison shows that machine learning algorithms can effectively identify fake reviews online. Among the tested models, SVM appears to be the best fit for the chosen dataset and TF-IDF feature extraction method. However, when choosing the final model, consider not only prediction accuracy but also computational efficiency, scalability and other factors.

5. Discussion

Based on the comparison, machine learning can effectively detect fake reviews automatically. Although all five classifiers performed well, researchers observed significant differences between the algorithms. Support Vector Machine (SVM) performed best, followed by Random Forest, Gradient Boosting, Logistic Regression, and Decision Tree. Thus, selecting the appropriate classifier is critical to designing a mechanism for detecting fake reviews. The SVM's high performance can be explained by its effectiveness in high-dimensional text classification. The TF-IDF approach provides a large set of textual features, and the SVM algorithm can find a decision boundary in these feature spaces. At the same time, the SVM's high precision and recall are especially valuable for fake review detection. High precision reduces the likelihood of misclassifying legitimate reviews as fake, while high recall increases the likelihood of detecting more actual fake reviews. The Random Forest results were also impressive, showing how efficiently ensemble methods identify deceptive reviews. Unlike a single Decision Tree, a Random Forest is an ensemble of trees whose predictions are combined to produce the final result. This approach reduces the influence of errors made by individual trees and may improve generalization. Although Random Forest's performance was not outstanding compared to other algorithms, this result demonstrates the potential of ensemble methods to detect complex dependencies among textual features. Although Gradient Boosting produced competitive results, they were slightly worse than those obtained with Random Forest and SVM. The sequential learning approach allows it to focus on the classification errors of previous learners. However, the success of this approach may largely depend on selecting appropriate hyperparameters and feature characteristics.

Thus, further improvements could potentially be achieved through optimization. Logistic Regression performed well, although it is much simpler than both the considered ensembles and SVM. At the same time, Decision Tree demonstrated the worst overall performance. This suggests that a single tree cannot capture complex, high-dimensional patterns in textual data, whereas SVMs and ensembles can. Moreover, Decision Trees can be unstable and sensitive to input data, especially with many textual features. Thus, the results demonstrate the importance of considering the model's structure when choosing an algorithm for text classification. The results align with previous research showing that machine learning algorithms can detect linguistic features of deceitful reviews [1]; [4]; [5]. Previous work has shown that textual features can classify truthful and deceptive opinions. This research expands the field by analyzing multiple machine learning models and comparing their performance under unified experimental conditions. This approach is more objective than selecting a single classification algorithm. Additionally, the results highlight the need to assess multiple performance measures. The system's accuracy provides a general evaluation of prediction correctness; however, it gives no information about the kinds of mistakes it made. In fake review detection, both types of mistakes have significant real-world implications. False positives can delete or flag a legitimate review, harming a customer or a business. A false negative means the deceptive review stays on and continues to influence buyers' decision-making. Therefore, consider precision, recall, F1-score, and ROC-AUC alongside accuracy. Moreover, from a business perspective, fake review detection can offer several advantages to e-commerce businesses. Automated classification algorithms can help identify potentially suspicious reviews for manual review and increase the credibility of the reviews presented to customers.

Greater review credibility may increase consumer trust. Moreover, businesses can avoid the risks of unfair competition from artificially created positive or negative reviews. Nevertheless, one must keep the constraints of this research in mind. Firstly, the effectiveness of machine learning algorithms depends heavily on the properties of the dataset used. An algorithm that works well on one dataset might yield entirely different results when applied to reviews of another product category, from another website, in another language, or in another geographic region. Secondly, the study uses mostly textual data and TF-IDF features. While textual data is important, detecting fake reviews might benefit from using additional data, such as user behavior, review activity, timestamps and product data. A third limitation is the changing nature of fake texts. As automated text generation tools become more capable, fake reviews can become more plausible and harder to distinguish from genuine customer experience descriptions. In this case, models trained on older datasets may become less effective at handling new types of fake text. Models need regular updates and evaluation using up-to-date datasets. Future research could address the study's limitations by combining textual and non-textual features to build a multimodal detection system. Future research could compare deep learning techniques, such as transformer language models like BERT, with traditional ML methods to determine whether context-based language representations improve detection performance. Finally, explainable AI techniques could reveal why a review was classified as fake. Model performance evaluation should consider the dataset's properties and the application domain. In the current research, SVM performs best, followed by Random Forest and, finally, Logistic Regression. Based on the comparative results, model performance evaluation is an important step in developing reliable fake online review detectors.

6. Conclusion and Future Work

The growing adoption of online reviews in e-commerce creates both opportunities and problems for consumers and businesses. While legitimate reviews provide useful information, fake reviews can manipulate consumer perception and harm a business's reputation. This research aimed to explore machine learning techniques for detecting fake online reviews and to compare the performance of five classification models: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Gradient Boosting. The experimental methodology included dataset preparation, text normalization, TF-IDF feature extraction, model training and evaluation. The following metrics were used to assess the models' performance: accuracy, precision, recall, F1-score, and ROC-AUC. The findings showed that all chosen models could classify real and fake reviews, but they performed differently. Among the algorithms considered, SVM achieved the best overall performance, yielding the highest accuracy, precision, recall, F1-score, and ROC-AUC. Random Forests showed the second-best results and served as a powerful ensemble learning method. Gradient Boosting also performed well, whereas Logistic Regression proved to be a strong, fast baseline. Thus, machine learning can effectively detect deceptive online reviews automatically.

The comparative analysis shows that algorithm choice is crucial to classification performance. The results also show that multiple metrics are needed for evaluation, as accuracy alone does not fully reflect a model's performance. In practice, these results can serve as a basis for designing review-monitoring systems that help platform managers detect suspicious reviews and verify them manually. This way, better detection may increase the transparency of online platforms, improve the trustworthiness of their products, and help ensure a fair business environment. However, results depend on the specific dataset and experimental conditions used in the research. Model performance can vary across datasets and circumstances. Thus, researchers cannot say that the best-performing model is the best for detecting all fake reviews. Future research can proceed in several ways. First, future studies can use larger datasets that include reviews from multiple e-commerce platforms. They can also include reviews of different products in different regions. Second, future research can include reviewer behavior and metadata features alongside text-based features. These features could include review frequency, ratings, account activity, time, reviewer profile, and

product-review associations. Third, the comparative analysis framework could use advanced deep learning algorithms and transformer-based models.

Models such as BERT and other language models can be compared with traditional classifiers such as SVM, Random Forest, and Logistic Regression. The goal is to determine whether contextual language features can improve detection of sophisticated review manipulation. Future researchers can consider using hybrid or ensemble models that combine predictions of multiple machine learning algorithms. Fifth, explainable artificial intelligence (XAI) can help develop fake review detectors. The machine learning algorithm not only classifies whether a review is fake but also identifies the linguistic or behavioral features that contributed to the decision. Lastly, future research should examine how the application of generative AI to generate fake online content is becoming increasingly common. AI-generated reviews will become harder to detect since they can produce content indistinguishable from human-made reviews. Future research should therefore focus on models that continually update to detect emerging tactics used in review fraud. Fake online review detection is one of the most challenging tasks in digital commerce. Comparing machine learning algorithms provides a good starting point for tackling this challenge. Future inclusion of behavioral data, deep learning, transformer architectures, XAI, and continuously updated datasets will make detection systems more effective.

Acknowledgment: N/A

Data Availability Statement: The data for this study are available upon request from the corresponding author.

Funding Statement: This manuscript and research paper were prepared without any financial support or funding.

Conflicts of Interest Statement: The authors have no conflicts of interest to declare. This work represents a new contribution by the authors, and all citations and references are appropriately included based on the information utilized.

Ethics and Consent Statement: This research adheres to ethical guidelines and obtains informed consent from all participants. Confidentiality measures were implemented to safeguard participant privacy.

References

1. N. Jindal and B. Liu, "Opinion spam and analysis," in *Proc. 2008 Int. Conf. Web Search Data Mining (WSDM)*, California, United States of America, 2008.
2. N. Jindal, B. Liu, and E. P. Lim, "Finding unusual review patterns using unexpected rules," in *Proc. 19th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, Toronto, Ontario, Canada, 2010.
3. B. Liu, "Sentiment Analysis and Opinion Mining," *Morgan & Claypool*, San Rafael, California, 2012.
4. 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, Oregon, United States of America, 2011.
5. M. Ott, C. Cardie, and J. T. Hancock, "Estimating the prevalence of deception in online review communities," in *Proc. 21st Int. Conf. World Wide Web (WWW)*, Lyon, France, 2012.
6. L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 10, pp. 5–32, 2001.
7. Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. 2014 Conf. Empirical Methods Natural Language Processing (EMNLP)*, Doha, Qatar, 2014.
8. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Language Technologies*, Minneapolis, Minnesota, 2019.
9. 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. Knowl. Discovery Data Mining*, Chicago, Illinois, United State of America, 2013.
10. M. Lee, Y. H. Song, L. Li, K. Y. Lee, and S. B. Yang, "Detecting fake reviews with supervised machine learning algorithms," *The Service Industries Journal*, vol. 42, no. 13–14, pp. 1101–1121, 2022.
11. D. Martens and W. Maalej, "Towards understanding and detecting fake reviews in app stores," *Empirical Software Engineering*, vol. 24, no. 5, pp. 3316–3355, 2019.
12. N. Kumar, D. Venugopal, L. Qiu, and S. Kumar, "Detecting review manipulation on online platforms with hierarchical supervised learning," *Journal of Management Information Systems*, vol. 35, no. 1, pp. 350–380, 2018.
13. W. Zhang, Y. Du, T. Yoshida, and Q. Wang, "DRI-RCNN: An approach to deceptive review identification using recurrent convolutional neural network," *Information Processing & Management*, vol. 54, no. 4, pp. 576–592, 2018.

14. D. Plotkina, A. Munzel, and J. Pallud, "Illusions of truth—Experimental insights into human and algorithmic detections of fake online reviews," *Journal of Business Research*, vol. 109, no. 3, pp. 511–523, 2020.
15. A. M. M. Abdulqader, A. Namoun, and Y. Alsaawy, "Fake online reviews: A unified detection model using deception theories," *IEEE Access*, vol. 10, no. 12, pp. 128622–128655, 2022.
16. R. Oak, "Detecting review fraud using metaheuristic graph neural networks," *International Journal of Information Technology*, vol. 16, no. 7, pp. 4019-4025, 2024.

Publisher's Note: The publisher remains impartial concerning jurisdictional claims in published maps and institutional affiliations. Responsibility for the content rests entirely with the authors and does not necessarily reflect the publisher's perspectives.