Pertanika J. Sci. & Technol. 34 (S3): 55 - 69 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Detecting Fake Reviews on Online Products with Machine Learning

P. Dhanalakshmi[1]*, Ummadi Janardhan Reddy[2], Ravikanth Garladinne[3] , S. Ramachandra[4], E. R. Aruna[5], and Sharath Chandra Parashara[6]

[1]Department of Artificial Intelligence and Machine Learning, School of Computing, Mohan Babu University, Sree Sainath Nagar, 517102 Tirupati, Andhra Pradesh, India
[2]Department of Computer Science and Engineering, Faculty of Engineering and Technology, Alliance University, Chikkahagade Cross, Chandapura–Anekal Main Road, Anekal, 562106 Bengaluru, Karnataka, India
[3]Department of Computer Science and Engineering, KLEF College of Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, 522302 Guntur, Andhra Pradesh, India
[4]Department of Mechanical Engineering, Mother Theresa Institute of Engineering and Technology, Melumoi Post, 517408 Palamaner, Chittoor, Andhra Pradesh, India
[5]Department of Computer Science and Engineering, Vardhaman College of Engineering, Kacharam, Shamshabad, 501218 Hyderabad, Telangana, India
[6]Fenix Commerce Inc., 5434 Castle Manor Drive, San Jose, 95129 California, United States

ABSTRACT

Reviews play a crucial role in potential buyers' decision-making and enable companies to showcase the quality of their products. A growing number of deceitful reviews, published either intentionally or automatically as computer-generated, pose a significant threat to online review platforms. For consumers and companies, it is thus crucial to distinguish between reliable and deceptive online reviews. For this reason, the automatic classification of online reviews as being fake or genuine is a highly relevant text classification task. In this work, we present a framework that relies on the feature extraction from textual data, using Term Frequency–Inverse Document Frequency (TF-IDF) in combination with the vectorisation techniques Count Vectoriser and Hashing Vectoriser. The extracted feature vectors are then fed into three classification models: Support Vector Machine (SVM), Random Forest, and Logistic Regression. The Amazon review dataset was used for the experiments. The performance of all classifiers was measured using accuracy, precision, recall, and F1-score. Our results show that SVM performs best among the tested approaches, with TF-IDF in combination with the Count Vectoriser yielding strong results. By combining vectorisation with standard machine learning, our approach offers an efficient, easy-to-understand framework that is less complex

ARTICLE INFO

Article history:
Received: 03 August 2026
Accepted: 06 October 2026
Published: 20 November 2026

DOI: https://doi.org/10.47836/pjst.34.S3.03

E-mail addresses:
mallidhana5@gmail.com (P. Dhanalakshmi)
ummadi.janardhan@gmail.com (Ummadi Janardhan Reddy)
garladinne.ravikanth@gmail.com (Ravikanth Garladinne)
mail2sramachandra@gmail.com (S. Ramachandra)
er.aruna@vardhaman.org (E. R. Aruna)
hellosharathchandra@gmail.com (Sharath Chandra Parashara)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

than Deep Learning or Transformer-based approaches. Thus, classical machine learning methods can serve as a good starting point for automatically detecting and avoiding the display of fake online reviews across various applications, while also efficiently processing text with limited computational resources.

Keywords: Count Vectoriser, fake review detection, Hashing Vectoriser, logistic regression, machine learning, natural language processing, random forest, support vector machine, TF-IDF

INTRODUCTION

Online shopping has dramatically changed the nature of information gathering for products and services, and buyers now rely heavily on previous buyers' experiences with a product or service. Hence, online reviews contain substantial information regarding product characteristics, perceived quality, usability, and customer satisfaction. Review systems assist customers when purchasing products online.

However, in most cases, customers rely on authentic reviews. Furthermore, online reviews can be deceptive, as they may be intentionally written to promote a product positively, make a product more popular, or even be fake. There are even programs that automatically generate reviews, which makes manual checking of online reviews very time-consuming and nearly impossible when many reviews are uploaded to an online platform. Deceptive content in online reviews influences both buyers and sellers.

Reviewing buyers may base their purchasing decisions on false information. In addition, sellers may be adversely affected by competitors' fake and fraudulent reviews. Therefore, methods to detect false online reviews have become a highly relevant research topic. We treat fake-review detection as a binary classification problem between deceptive and legitimate reviews. In a binary classification problem, a machine learning model decides between two categories, e.g., whether a review is deceptive or legitimate.

A variety of aspects need to be addressed when building a classifier, such as sufficient labelled data, proper data preprocessing, suitable feature representation, a proper classification model, and an appropriate way to measure performance. Reviews are typically text data that cannot be used directly by most conventional algorithms. Thus, the first step is converting the reviews into numerical feature vectors.

This will result in possibly very different feature spaces depending on the method used to transform the text. There are two main methods: Count Vectoriser (CV) and Hashing Vectoriser (HV). To improve feature extraction using CV, we can use TF-IDF weighting to assign greater importance to the few informative terms in a review. We employ three standard classifiers (SVM, Random Forest, and Logistic Regression) and analyse their behaviour in conjunction with the methods for representing text as feature vectors.

All experiments have been done using the same dataset of labelled Amazon reviews and the same evaluation protocol. The purpose of this research was to investigate the capability of relatively simple machine learning methods to correctly distinguish between fake and genuine online reviews. The research made use of a number of methods to transform the reviewed text into numerical vectors (features) that can be used by machine learning methods for classification, and compared the performance of Support Vector Machine, Random Forest and Logistic Regression classifiers when used in conjunction with Term Frequency-Inverse Document Frequency (TF-IDF) transformed reviews, generated using the Count Vectoriser and Hashing Vectoriser features.

The study, therefore, provides a controlled experimental analysis of six classifier-Vectoriser configurations. Accuracy, precision, recall, F1-score, and confusion matrices are considered to obtain a broader view of classification performance. Rather than attempting to establish superiority over modern transformer or large-language-model systems, the work focuses on understanding the behaviour of conventional machine-learning techniques under a common experimental setup.

LITERATURE SURVEY

Research on deceptive online reviews has grown dramatically with the advent of machine learning, deep learning, and natural language processing. Most studies investigated different types of information, including text characteristics, sentiment, user behaviour, relationships between users and products, and multimodal information.

Mutemi and Bação (2024) presented and evaluated machine-learning approaches for e-commerce fraud detection. We expand on this work and discuss the growing arsenal of automated detection methods for the vast array of online fraud emerging on the e-commerce scene.

Supervised learning approaches for detecting fake reviews have also been extensively studied. Mir et al. (2023) examined the application of supervised learning with BERT to identify fake reviews. They used supervised learning to classify deceptive reviews, using contextual language representations rather than simple frequency features. These contextual language features are created by a transformer-based language model such as BERT.

Related to the earlier discussion of misleading online content, Akhtar et al. (2023) have developed an approach using Artificial Intelligence (AI) to identify other types of on-line deception. In this paper, they presented an overview and comparison of current state-of-the-art approaches to detecting and preventing fake news and disinformation, using AI and machine learning techniques to analyse text-based content.

Hajek et al. (2023) investigated the use of fake reviews for aspect-based sentiment analysis. This is an important line of work, as it can reveal more about the relationship between a reviewer's expressed opinions and the relevant aspects of the product under review.

In related work, Zhang et al. (2023) explore the use of sentiment intensity for effective fake-review detection, employing Positive-Unlabeled learning to address the challenge of obtaining reliable labels for fraudulent reviews in practical settings.

Using such methods for learning with limited or even no positive examples for fraudulent reviews is an interesting topic for future research. Deceptive review classification using deep learning approaches has also been investigated. Mohawesh et al. (2023) introduced an explainable multi-view deep-learning ensemble for detecting fake reviews. The paper offers in-depth insight into the approach, highlighting its ability to combine multiple representations and to provide explanations for its predictions.

Additional work has examined the discrepancy between review scores and corresponding text. Work here has identified instances of deceptive reviews in which review scores and corresponding text are "inconsistent" (Sharifpour & Lahmiri, 2023).

We can gain insight not found by examining reviews one by chance. Zabeen et al. (2023) presented an uncertainty-aware framework that utilised LSTM and BERT to develop sequential and contextual representations for training and assessing a deceptive-review classification model.

Several recent studies have taken advantage of the wealth of additional information beyond the review text. Multimodal approaches, as well as behavioural studies, can leverage this data, which is often not available in plain text reviews. Li and Chen (2023), for example, propose a deep neural network that incorporates review text with other multimodal features and performs feature fusion to merge them into a single set of inputs.

Recently, many hybrid approaches have been introduced to exploit both the sequential information in reviews and the contextual information in words. For this purpose, Mohawesh et al. (2024) enhanced an LSTM network by integrating it with the powerful pre-trained language model RoBERTa. Reviews are also associated with reviewers' behaviours.

We can therefore combine review text with information from reviewers' behaviour in a detection system. Liu et al. (2024) use the RoBERTa model to generate text representations and, on top of this, add several types of reviewer behaviours to their framework. Often, suspicious activity leaves a 'fingerprint' with the behaviour of reviewers on a website rather than in the linguistic content of individual reviews.

To model the relationships among reviewers, products, and corresponding reviews, graph-based methods have been extended in a strong manner so far. Specifically, Yao et al. (2024) presented a graph contrastive learning framework that incorporates hierarchical relationships and label consistency for review detection.

Recently, transformer-based models have become popular for detecting deceptive content. In this line of work, Hyder et al. (2024) proposed a method, DeceptiveBERT, to identify deceptive reviews using BERT representations. Since these models are equipped with contextual language models, they can capture inter-word relationships, which are challenging for frequency-based methods to capture.

More recent work has used large language models and advanced deep learning techniques. In this category of work, Fan et al. (2025) introduced a deep learning method that uses characteristics derived from chaos theory to classify sophisticated deceptive online reviews. Thilini Jayasinghe and Dassanayaka (2025) compared the performance of various deep neural networks in detecting fraudulent Amazon reviews.

The area of fake-review detection has moved on from simple text classification using features to more complex deep learning models, including transformers, behavioural models such as graph-based methods, and, more recently, multimodal and large language models. All the studies reviewed here are influenced by both the content representation of the review and its classification. Importantly, no single source can solve the problem of fake reviews for all scenarios. The useful signals are hidden in the review textual patterns as well as in the reviewer's sentiment, in the reviewer's ratings, in the reviewer's behaviour, as well as in the entities' inter-relations.

RESEARCH GAP

The state of the art for the detection of fake reviews is presented in increasing order of technological complexity, i.e., from conventional machine-learning methods to deep learning methods, to transformer-based methods, to behavioural models, to graph models, to multimodal methods, and finally to large language models (Fan et al., 2025; Hyder et al., 2024; Liu et al., 2024; Mir et al., 2023; Mohawesh et al., 2023; Mohawesh et al., 2024; Thilini Jayasinghe & Dassanayaka, 2025; Yao et al., 2024). Even if the presented methods can successfully detect fake reviews, most are very complex and require substantial computational power.

Most studies also have several limitations. The performance of existing models is often reported across different data sets, with different types of labelling, class distributions, and preprocessing steps, feature representations, and evaluation measures. Thus, even if a model is reported to outperform other models on a particular benchmark, this does not necessarily mean it will also outperform them on other, similar benchmarks. In order to provide a simple and understandable framework for the comparison of various text representations and machine learning approaches within a fixed setup of experiments, the use of two simple text representations (Count Vectoriser and Hashing Vectoriser) in combination with TF-IDF, and three standard classifiers (Support Vector Machine, Random Forest, and Logistic Regression) for the detection of fake reviews appears to be promising. These methods are easy to evaluate and to interpret.

Position of the Present Study

This study investigates the capability of TF-IDF-based representations, built with the Count Vectoriser and the Hashing Vectoriser, to evaluate the performance of SVM, Random

Forest, and Logistic Regression classifiers in detecting fake reviews. The same experimental framework is then used to analyse the impact of vectorisation and classification on performance. The goal of this study is to provide a simple, interpretable tool for detecting fake reviews and to determine the best combination of text representation and a traditional classifier on a given dataset of Amazon reviews. The results will not be compared against the state of the art, which comprises more recent transformer- and LL_miss-based approaches, but they will provide a concrete, and thus practical, evaluation of fake-review detection using lightweight methods.

EXISTING METHODOLOGY

Existing GPT-2-Based Approach

Previous research has investigated the use of generative language models to create and identify deceptive online reviews. GPT-2 can be used to generate synthetic review content by learning linguistic patterns from existing review data. Such generated reviews can be used as an experimental source for studying the characteristics of artificially produced review text. In the existing approach considered in this study, GPT-2 was used to generate synthetic reviews with different review lengths and rating characteristics.

However, the GPT-2-based synthetic review generation approach is considered an existing methodology and is not used as the dataset for the proposed machine-learning experiments presented in Section 4. The proposed experiments are conducted independently on the Amazon review dataset, which contains labelled computer-generated and human-written reviews. This distinction is made to clearly separate the previously investigated synthetic review generation approach from the dataset used to train and evaluate the proposed classification models.

The existing methodology provides a basis for understanding how artificially generated review content can be produced and how such content may differ from human-written reviews. However, the present study focuses on the subsequent classification task of distinguishing fake reviews from genuine reviews using conventional machine-learning algorithms and alternative text-vectorisation techniques.

Proposed Methodology

The proposed methodology focuses on detecting fake and genuine online product reviews using text-based feature representations and supervised machine learning classification as shown in Figure 1. The approach consists of data preparation, text preprocessing, feature extraction, TF-IDF-based feature weighting, classification, and performance evaluation. Count Vectoriser and Hashing Vectoriser are considered alternative text vectorisation techniques, while Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR) are used as classification algorithms.

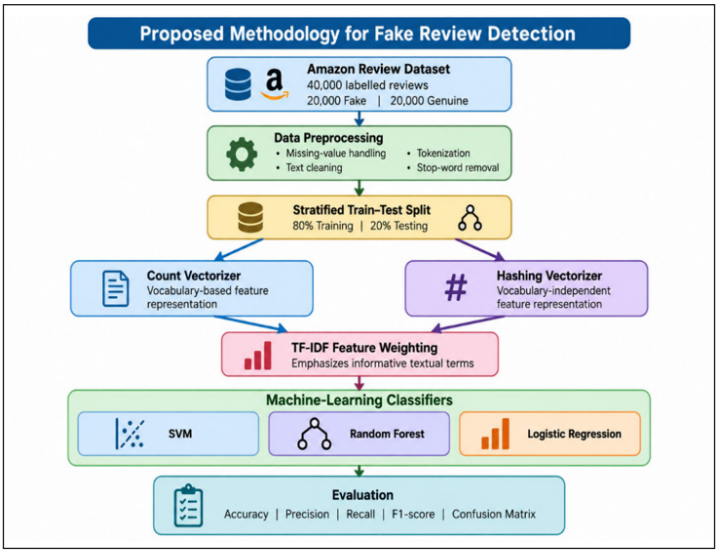


Figure 1. Proposed methodology for fake review detection using text vectorisation and machine learning

The proposed methodology is designed to provide a consistent comparison of different combinations of text representations and machine-learning classifiers. The overall process begins with the labelled Amazon review dataset and proceeds through preprocessing and numerical feature representation before classification. The trained models are finally evaluated using standard classification metrics to assess their effectiveness at distinguishing fake reviews from genuine ones.

Dataset Description

The proposed model uses an Amazon product review dataset containing 40,000 labelled reviews. The dataset consists of 20,000 reviews labelled as computer-generated (fake) and 20,000 labelled as human-written (genuine), resulting in an equal class distribution of 50% fake and 50% genuine. The balanced class distribution allows the classification models to be evaluated without introducing a class-frequency bias toward either category, as shown in Table 1.

Table 1
Distribution of fake and genuine reviews in the Amazon review dataset

Review Class	Number of Reviews	Percentage
Fake / Computer-generated	20,000	50%
Genuine / Human-written	20,000	50%
Total	40,000	100%

In contrast to the GPT-2-based synthetic review generation approach discussed in Section 3, the Amazon review dataset described in this section is the dataset used for the proposed machine-learning experiments. Each review is assigned a predefined class label, enabling supervised machine-learning algorithms to learn the textual characteristics of fake and genuine reviews.

Data Preprocessing

The review text is first pre-processed to remove unnecessary information and prepare the data for feature extraction. The preprocessing procedure includes removing irrelevant characters, handling missing or incomplete text, converting text into a consistent format, and preparing the reviews for numerical representation. These preprocessing steps reduce noise in the review text and improve the data's suitability for subsequent feature extraction and classification.

Data Partitioning and Validation

The labelled Amazon review dataset is divided into training and testing subsets using an 80:20 stratified split. The training subset is used for model development, while the testing subset is kept separate and is used only for final performance evaluation. Stratified partitioning preserves the proportion of fake and genuine reviews in both subsets.

The classification models are trained using the training data and subsequently evaluated using the unseen test data. Accuracy, precision, recall, F1-score, and confusion matrices are considered to assess the classification performance. The test subset is not used during model training, thereby providing an independent evaluation of the trained models.

Feature Representation

After preprocessing, the textual reviews are converted into numerical representations because machine-learning classifiers cannot directly process raw text. In this study, the Count Vectoriser and the Hashing Vectoriser are investigated as alternative approaches to converting review text into numerical feature representations. TF-IDF weighting is subsequently used to represent the importance of textual terms.

Tokenisation

Tokenisation divides the review text into individual words or tokens. During preprocessing, unnecessary stop words and irrelevant textual elements are removed to obtain a cleaner representation of the review content. The resulting tokens are then used to represent numerical features.

Count Vectoriser

Count Vectoriser converts the pre-processed review text into a numerical document-term matrix. Each row represents a review, and each column represents a term in the vocabulary, while the corresponding matrix value represents the occurrence of that term in the review.

In this study, Count Vectoriser provides an explicit vocabulary-based representation of the review text before TF-IDF weighting is applied. The purpose of using Count Vectoriser is to preserve word-frequency information, allowing the classifiers to identify differences in word usage between fake and genuine reviews. The resulting representation is subsequently combined with TF-IDF weighting to emphasise informative textual features.

Hashing Vectoriser

Hashing Vectoriser provides an alternative text representation to Count Vectoriser. Instead of explicitly constructing and storing a vocabulary, it applies a hashing function to map terms directly into a predefined feature space. By not creating an explicit vocabulary, such an approach can avoid consuming a lot of memory during feature generation.

We compare the performance of Hashing Vectoriser to that of the Count Vectoriser in the experiments proposed in this work, in order to verify whether or not a vocabulary-independent representation can deliver results as good as those obtained by using a vocabulary-dependent representation when combined with the aforementioned classification methods (i.e., SVM, Random Forest and Logistic Regression).

Count Vectoriser and Hashing Vectoriser are feature-representation techniques. They transform text into numerical features which can then be used with classification algorithms such as SVM, Random Forest or Logistic Regression.

TF-IDF Feature Weighting

The Term Frequency-Inverse Document Frequency (TF-IDF) score is used to weight the importance of features in a review. The feature's Term Frequency (TF) within a review is combined with the feature's Inverse Document Frequency (IDF) across the entire review collection. The Inverse Document Frequency score decreases for features that occur in a large proportion of documents.

High weight will be given to terms that are very frequent in a single review but rare elsewhere in the dataset. To enhance the numerical representations derived from the review text, TF-IDF is used in this study. Two alternative Vectorisers, Count Vectoriser and Hashing Vectoriser, are also investigated. The use of TF-IDF weighting is furthermore compared to the weighting of no features at all.

Model Training and Classification

Three supervised learning methods in machine learning, SVM, RF, and LR, are implemented for the binary classification between fake and genuine reviews in this study.

The classifiers are trained individually using the feature vectors from the Count Vectoriser and Hashing Vectoriser with TF-IDF weighting. In the training phase, the classifiers are developed to recognise two review types, and in the testing phase, the individual classifiers are used to predict whether reviews are fake or real on previously unseen test data.

The same data processing and feature representation are applied to all classifiers to allow for a fair comparison of their classification performance.

Model Evaluation

We evaluate the performance of our proposed classification approaches by using standard measures such as accuracy, precision, recall, F1-score and a confusion matrix. Accuracy is the proportion of correctly classified reviews, as shown in Table 2. Precision measures the proportion of identified positive samples that are positive. Recall measures the ability to identify samples of interest. The F1-score is the harmonic mean of both precision and recall.

A confusion matrix has been used to look at the number of correctly/incorrectly classified instances of fake and genuine reviews. Using multiple measures to assess a classification model's performance yields more accurate results than relying on accuracy alone.

Table 2
Experimental configurations used for fake-review classification

Vectoriser	Feature Weighting	Classifier	Accuracy
Count Vectoriser	TF-IDF	Support Vector Machine (SVM)	91%
Count Vectoriser	TF-IDF	Random Forest (RF)	87%
Count Vectoriser	TF-IDF	Logistic Regression (LR)	86%
Hashing Vectoriser	TF-IDF	Support Vector Machine (SVM)	88%
Hashing Vectoriser	TF-IDF	Random Forest (RF)	85%
Hashing Vectoriser	TF-IDF	Logistic Regression (LR)	82%

RESULTS AND DISCUSSIONS

We conduct an experimental investigation of 6 combinations of 2 text vectorisation methods and 3 supervised classification methods, all using the same set of labelled reviews. The goal of the comparison is to test whether the text representation of reviews influences classification performance. Therefore, Count Vectoriser and Hashing Vectoriser are tested separately from the TF-IDF feature-weighting tool.

Comparative Performance

Experimental results revealed that the choice of Vectoriser greatly affected the performance of the machine learning methods, as shown in Figure 2. Using a combination of TF-IDF and Count Vectoriser, the SVM method achieved the highest accuracy of 91%, while the Random Forest method achieved 87% and the Logistic Regression method achieved 86%.

For TF-IDF with Hashing Vectoriser, the SVM achieved an accuracy of 88%, the Random Forest an accuracy of 85%, and the Logistic Regression an accuracy of 82%, as shown in Figure 3.

Across all tested setups, the highest accuracy was achieved by SVM combined with TF-IDF and Count Vectoriser, at 91%. The results indicate that the vocabulary-based Count Vectoriser representation provides better classification performance than the Hashing Vectoriser representation for all three classifiers under the experimental conditions considered in this study shown in Table 3.

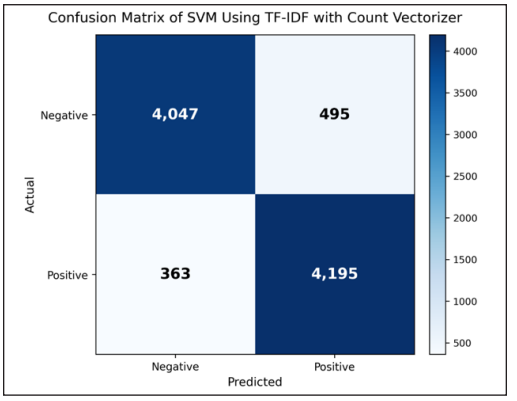


Figure 2. Confusion matrix of SVM using TF-IDF with Count Vectoriser

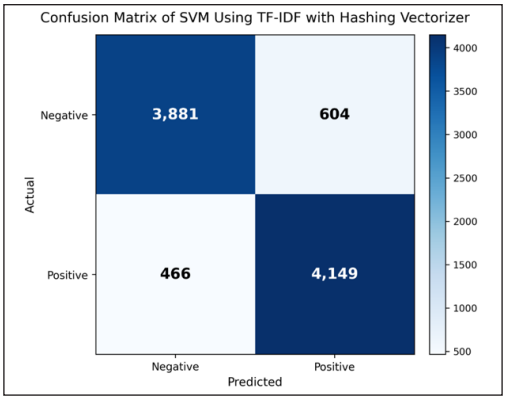


Figure 3. Confusion matrix of SVM using TF-IDF with Hashing Vectoriser

Table 3
Comparative accuracy of the evaluated machine-learning models

Vectoriser	Feature Weighting	Classifier	Accuracy
Count Vectoriser	TF-IDF	Support Vector Machine (SVM)	91%
Count Vectoriser	TF-IDF	Random Forest (RF)	87%
Count Vectoriser	TF-IDF	Logistic Regression (LR)	86%
Hashing Vectoriser	TF-IDF	Support Vector Machine (SVM)	88%
Hashing Vectoriser	TF-IDF	Random Forest (RF)	85%
Hashing Vectoriser	TF-IDF	Logistic Regression (LR)	82%

DISCUSSION

The results demonstrate that both the feature representation and the classification algorithm influence the performance of fake-review detection shown in Figure 4. Count Vectoriser combined with TF-IDF provides higher accuracy than Hashing Vectoriser combined with TF-IDF for SVM, Random Forest, and Logistic Regression.

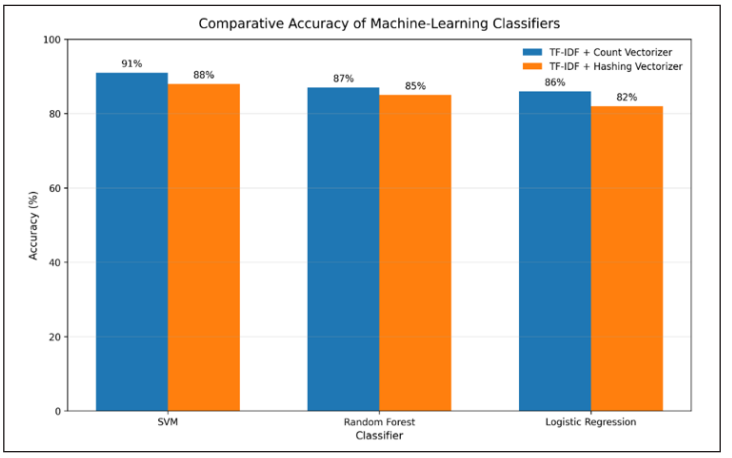


Figure 4. Comparative accuracy of machine-learning classifiers using Count and Hashing Vectorisers

SVM provides the best performance among the evaluated classifiers, achieving 91% accuracy with Count Vectoriser and 88% accuracy with Hashing Vectoriser. Random Forest achieves 87% and 85%, respectively, while Logistic Regression achieves 86% and 82%.

The comparatively higher performance of SVM indicates that it is effective in distinguishing the textual patterns associated with fake and genuine reviews in the selected dataset. However, the results are specific to the dataset and experimental configuration used in this study and should not be interpreted as evidence that SVM will outperform other approaches on all fake-review datasets.

Overall, the experiments demonstrate that conventional machine-learning algorithms, combined with appropriate textual feature representations, can achieve effective performance in fake-review classification. The best-performing configuration in the present study is SVM with TF-IDF and Count Vectoriser, which achieves an accuracy of 91%.

CONCLUSION

The purpose of this research is to investigate online product review classification (i.e. deceptive vs genuine) by using conventional machine learning approaches and text-based feature representations. We conducted experiments on a large dataset of 40,000 labelled Amazon reviews, comprising 20,000 computer-generated and 20,000 human-written reviews.

The text was transformed into numeric vectors by two different methods. We used both the Count Vectoriser and the Hashing Vectoriser. Both methods were combined with TF-IDF weighting and then used in three different supervised learning classifiers: Support Vector Machine, Random Forest and Logistic Regression. Our experiments achieved different accuracy values on the six different configurations. The best results were achieved by SVM (91%) with Count Vectoriser and TF-IDF, followed by Random Forest (87%) and Logistic Regression (86%). Using the Hashing Vectoriser with TF-IDF yielded accuracy values of 88%, 85% and 82% for SVM, Random Forest and Logistic Regression, respectively.

Our results show that, when classifying fake reviews with conventional machine-learning classifiers, the quality of the feature representation is crucial for achieving high accuracy. In our experiments, the best results were achieved using the Count Vectoriser and its default parameters across all methods. Our lightweight setup enables experimental investigations into deceptive-review classification without relying on large, compute-intensive transformer or large language model architectures.

It can also serve as a simple baseline in experimental setups or even in production with very limited computational resources. The findings reported here are based on a single dataset with a specific set of experimental parameters. A well-balanced, large dataset of labelled reviews for e-commerce systems, which may include many more features (linguistic, behavioural, temporal, etc.) than those used in this study, will likely need to be evaluated with other models and experimental settings. Thus, the results obtained here should not be used to assess the performance of the given models on other datasets.

Future work can explore additional features, such as reviewer-level behavioural features, temporal features, product-level features, and more linguistic features. We also plan to explore the use of Transformer-based architectures and large language models to investigate the added value of the contextual information encoded in these representations compared to the results obtained with the conventional methods used in this work.

ACKNOWLEDGEMENT

The authors would like to thank our universities for providing the facilities and support needed to carry out this research. We also extend our gratitude to our colleagues for their helpful insights and feedback during the preparation of this manuscript.

REFERENCES

Akhtar, P., Ghouri, A. M., Khan, H. U. R., Amin ul Haq, M., Awan, U., Zahoor, N., Khan, Z., & Ashraf, A. (2023). Detecting fake news and disinformation using artificial intelligence and machine learning to avoid supply chain disruptions. *Annals of Operations Research, 327*(2), 633-657. https://doi.org/10.1007/s10479-022-05015-5

Fan, Y., Fan, H., Zhang, F., & Wang, Z. (2025). Large language model meets chaos: A new deep learning model for fake review detection. *Electronic Commerce Research and Applications, 73*, Article 101521. https://doi.org/10.1016/j.elerap.2025.101521

Hajek, P., Hikkerova, L., & Sahut, J.-M. (2023). Fake review detection in e-commerce platforms using aspect-based sentiment analysis. *Journal of Business Research, 167*, Article 114143. https://doi.org/10.1016/j.jbusres.2023.114143

Hyder, S. B., Tariq, N., Moqurrab, S. A., Ashraf, M., Yoo, J., & Srivastava, G. (2024). BERT-based deceptive review detection in social media: Introducing DeceptiveBERT. *IEEE Transactions on Computational Social Systems, 11*(6), 7234-7243. https://doi.org/10.1109/tcss.2024.3403937

Li, X., & Chen, L. (2023). Fake review detection using deep neural networks with multimodal feature fusion method. In *2023 IEEE 29th International Conference on Parallel and Distributed Systems (ICPADS)* (pp. 2869-2872). IEEE. https://doi.org/10.1109/icpads60453.2023.00411

Liu, J., Quan, P., & Zhang, W. (2024). A study on fake review detection based on RoBERTa and behavioral features. *Procedia Computer Science, 242*, 1323-1330. https://doi.org/10.1016/j.procs.2024.08.131

Mir, A. Q., Khan, F. Y., & Chishti, M. A. (2023). *Online fake review detection using supervised machine learning and BERT model* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2301.03225

Mohawesh, R., Salameh, H. B., Jararweh, Y., Alkhalaileh, M., & Maqsood, S. (2024). Fake review detection using transformer-based enhanced LSTM and RoBERTa. *International Journal of Cognitive Computing in Engineering, 5*, 250-258. https://doi.org/10.1016/j.ijcce.2024.06.001

Mohawesh, R., Xu, S., Springer, M., Jararweh, Y., Al-Hawawreh, M., & Maqsood, S. (2023). An explainable ensemble of multi-view deep learning model for fake review detection. *Journal of King Saud University - Computer and Information Sciences, 35*(8), Article 101644. https://doi.org/10.1016/j.jksuci.2023.101644

Mutemi, A., & Bação, F. (2024). E-commerce fraud detection based on machine learning techniques: Systematic literature review. *Big Data Mining and Analytics, 7*(2), 419-444. https://doi.org/10.26599/bdma.2023.9020023

Sharifpour, K., & Lahmiri, S. (2023). Fake review detection using rating-sentiment inconsistency. In *2023 International Conference on Machine Learning and Applications (ICMLA)* (pp. 926-931). IEEE. https://doi.org/10.1109/icmla58977.2023.00137

Thilini Jayasinghe, J. M., & Dassanayaka, S. (2025). Detecting deception: Employing deep neural networks for fraudulent review detection on Amazon. *Neural Computing and Applications, 37*(26), 21715-21742. https://doi.org/10.1007/s00521-025-11485-y

Yao, J., Jiang, L., Shi, C., & Yan, S. (2024). Fake review detection with label-consistent and hierarchical-relation-aware graph contrastive learning. *Knowledge-Based Systems, 302*, Article 112385. https://doi.org/10.1016/j.knosys.2024.112385

Zabeen, S., Hasan, A., Islam, M. F., Hossain, M. S., & Rasel, A. A. (2023). Robust fake review detection using uncertainty-aware LSTM and BERT. In *2023 IEEE 15th International Conference on Computational*

Intelligence and Communication Networks (CICN) (pp. 786-791). IEEE. https://doi.org/10.1109/cicn59264.2023.10402342

Zhang, S., Zhu, A., Zhu, G., Wei, Z., & Li, K.-C. (2023). Building fake review detection model based on sentiment intensity and PU learning. *IEEE Transactions on Neural Networks and Learning Systems, 34*(10), 6926-6939. https://doi.org/10.1109/tnnls.2023.3234427