Computers, Materials & Continua
DOI:10.32604/cmc.2022.019625
Article

Tech Science Press

Data Analytics for the Identification of Fake Reviews Using Supervised Learning

Saleh Nagi Alsubari[1], Sachin N. Deshmukh[1], Ahmed Abdullah Alqarni[2], Nizar Alsharif[3], Theyazn H. H. Aldhyani[4,*], Fawaz Waselallah Alsaade[5] and Osamah I. Khalaf[6]

[1]Department of Computer Science & Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India
[2]Department of Computer Sciences and Information Technology, Albaha University, Saudi Arabia
[3]Department of Computer Engineering and Science, Albaha University, Saudi Arabia
[4]Community College of Abqaiq, King Faisal University, Al-Ahsa, Saudi Arabia
[5]College of Computer Science and Information Technology, King Faisal University, Al-Ahsa, Saudi Arabia
[6]Al–Nahrain University, Bagdad, Iraq
[*]Corresponding Author: Theyazn H. H. Aldhyani. Email: taldhyani@kfu.edu.sa

Received: 20 April 2021; Accepted: 15 June 2021

Abstract: Fake reviews, also known as deceptive opinions, are used to mislead people and have gained more importance recently. This is due to the rapid increase in online marketing transactions, such as selling and purchasing. E-commerce provides a facility for customers to post reviews and comment about the product or service when purchased. New customers usually go through the posted reviews or comments on the website before making a purchase decision. However, the current challenge is how new individuals can distinguish truthful reviews from fake ones, which later deceive customers, inflict losses, and tarnish the reputation of companies. The present paper attempts to develop an intelligent system that can detect fake reviews on e-commerce platforms using n-grams of the review text and sentiment scores given by the reviewer. The proposed methodology adopted in this study used a standard fake hotel review dataset for experimenting and data preprocessing methods and a term frequency-Inverse document frequency (TF-IDF) approach for extracting features and their representation. For detection and classification, n-grams of review texts were inputted into the constructed models to be classified as fake or truthful. However, the experiments were carried out using four different supervised machine-learning techniques and were trained and tested on a dataset collected from the Trip Advisor website. The classification results of these experiments showed that naïve Bayes (NB), support vector machine (SVM), adaptive boosting (AB), and random forest (RF) received 88%, 93%, 94%, and 95%, respectively, based on testing accuracy and the F1-score. The obtained results were compared with existing works that used the same dataset, and the proposed methods outperformed the comparable methods in terms of accuracy.

Keywords: E-commerce; fake reviews detection; methodologies; machine learning; hotel reviews

1 Introduction

With the rapid development of e-commerce that enables the purchasing and selling of products and services online, customers are increasingly using this online marketing website for purchasing to meet their needs. After purchasing, customers write reviews about their personal experiences, feelings, and emotions concerning the products and services [1]. These online reviews can play a significant role in enhancing a new customer's shopping experience. Largely positive reviews entice more customers to purchase specific products or brands. Positive opinions provide considerable financial gain, whereas negative opinions often cause sales losses in e-commerce [2,3]. Therefore, most merchants depend primarily on public opinion to reshape their business plans by improving the quality of products [4,5]. Typically, opinions are the key to any online blog, post, or review. Spam content can be defined as meaningless or unsolicited data that are merged into opinions and are used for advertising, promoting, disseminating information, and financial profit purposes [6]. Consumers make online purchasing decisions to obtain products or services, and for that purpose, they go through online product reviews that are available on e-commerce websites before purchasing. The fake reviews detection system is a subfield of natural language processing. It aims to analyze, detect, and filter the reviewer's comments, particularly on e-commerce websites, into fake or truthful reviews [7]. Fake opinion refers to the false or inaccurate information in reviews to misguide consumers into making the wrong purchase decision and affecting the revenue of products [8]. Spam opinions can be divided into three types: 1) Untruthful (fake) reviews that have been written intentionally to mislead readers or systems of opinion mining. They include unworthy positive evaluations of specifically targeted products to promote the product or service. Furthermore, they include negative reviews of worthy products to defame them. These are named type 1 spam content. 2) Reviews on brands only are characterized as subjective opinions targeted at brands instead of the products themselves. Such reviews are known as type 2 spam opinions. 3) Non-reviews have two sub-types: (a) advertisements and (b) unrelated reviews that contain no opinion, such as questions, answers, or unspecific text [9]. Some important characteristics of spam/fake reviews are as follows [10]:

(1) Few details on the reviewer: Users who have fewer social relationships and do not have profile data are normally identified as fraudulent or spammers.
(2) Review content resemblance: Spammers often write duplicate or near-to-duplicate reviews and post them on the web. Such reviews may indicate fake reviews.
(3) Short review text: Fraudsters are always concerned with creating rapid returns; therefore, they have a tendency to post a very short review, which consists of grammatical errors and capital words. They further focus on the trademark of the products.
(4) Unexpected uploading and posting of reviews in the same timestamp: The greatest mechanism to identify spam reviews is to look at the time the review was posted. If a group of reviews is typed and uploaded at the same time, then these reviews may be spam.
(5) Extravagant use of negative and positive words: Spammers frequently utilize excessive positive and negative emotions in the review content, which may not be necessary for the review's context.

According to the literature on the opinion spam detection domain, there are no specific features to differentiate between truthful and fake content. Thus, this research work aimed to improve the accuracy of the fake review identification system by supervised learning algorithms. For this purpose, extracting features from the review text was an important and meaningful task. Such features are sentiment score, strong positive words, strong negative words, and four grams, as well as the number of verbs, nouns, and adjectives. More details about these features can be found in Section 3.1.

2 Related Work

2.1 Fake Review Detection

Fake review detection is a subfield of natural language processing. It aims to analyze, detect, and classify product reviews on online e-commerce domains into fake or truthful reviews. In the last two decades, fake review analysis has become a popular research topic. Many researchers have performed studies on fake/spam review identification due to its significant effect on customers and e-commerce businesses. The process of construction or extraction of important features from text data is called feature engineering. Fake review detection studies have two orientations: a behavioral approach (spammer reviewer features) and a linguistic features approach, which relies on review-centric features and relates only to the single review's content. Both features are mentioned and described below.

Stylometric-based features: These features are important and helpful for identifying the writing style of reviewers and detecting deception. They consist of two types of features. The first type is lexical characters, such as the number of characters in each word in the given text (N), the percentage of numeric to (N), the proportion of letters to N, the proportion of uppercase letters to N, the rate of spaces to N, the ratio of tabs to N, occurrences in the alphabet (A–Z), and the frequency of special characters (e.g., < > □ % I { } [l \ / # + - -;- * & @ $). The second type is a lexical word-based feature, which consists of the word count (T), a ratio of words presented in the sentence, proportion token length, rate of characters in words to N, the ratio of short words (1–3 characters), and the ratio of word length. Stylometric is the syntactic-based feature, which exemplifies the marks style of the reviewer when writing the review comment. Syntactic or grammar structures include the frequency of punctuation marks (. ? ! : ; , ") [11].

The maximum number of reviews per day: It was observed that about 70% of fraudsters type more than five reviews per day, whereas 90% of normal or truthful users can write only one review when purchasing a product or service. This considers the number of reviews that can be written by users to help identify spammers [12,13].

Proportion of positive reviews: Around 80% of fraudulent reviewers write down 85% of their reviews like positive reviews; for that matter, an increasing proportion of positive reviews might indicate a deceptive reviewer [14].

Length of the review text: As the review content consists of a group of words, from the work presented in [15], 75% of spammers cannot write more than 136 words per review. More than 90% of truthful reviewers write 200 words per review.

Reviewer deviation: Fraudster's ratings tend to be different from the normal average rating of truthful reviewers. Therefore, identifying users' rating variation might help in the process of spam detection [16].

Linguistic inquiry and word count (LIWC): LIWC is an analysis tool utilized by users to construct their dictionaries according to their interests. LIWC in fake review identification was used in the study presented by [17]; they accomplished better results by incorporating parts of speech (POS) features. LIWC is considered a deep linguistic feature. Examples of LIWC output features are self-reference (I, my, me), positive emotions (love, nice, sweet), negative emotions (Hurt, ugly, nasty), social words (talk, mate, they, child), big words (>6 letters), overall cognitive words (cause, know, ought), articles (a, an, the), and LIWC summary variables scores of authenticity, clout, cognitive words, and social words.

2.2 Datasets and Techniques Used for Fake/Spam Reviews and Spammer Detection

Goswami et al. [18] proposed a study on the impacts of reviewers' social interactions on fraud detection in online consumer reviews. In their experiment, Yelp's review dataset (135,413 reviews, of which 103,020 were recommended reviews and 32,393 were not-recommended reviews) was collected and preprocessed. Then, the behavioral and social interaction features of users were extracted, and the backpropagation neural network algorithm was implemented for the classification of reviews into genuine and fraud.

Jindal et al. [19] reported the first research for spam/fake review detection. The authors recognized three types of spam reviews, which were untruthful, reviews on brand only, and unrelated reviews. They used the supervised machine learning technique (logistic regression) to classify duplicates and near duplicates of Amazon product reviews into spam or non-spam, and the obtained result was 78% in terms of the area under the curve (AUC).

Mukherjee et al. [20] proposed an SVM model to detect fake product reviews. In terms of the dataset, they used real-life Yelp product reviews, which consisted of 5,678 reviews and 5,124 reviewers from hotels, in addition to 58,517 reviews and 35,593 reviewers from restaurants. In their experiment, two types of features were extracted: linguistic features, including n-grams, parts of speech, and LIWC, and behavioral features of the reviewer. To determine the difference between two distributions of fake and non-fake (truthful) review words, they applied Kullback–Leibler divergence (KL). The results of their methodology were 84% accuracy using the linguistic feature and 86% accuracy using reviewer features.

Ahmed et al. [21] proposed a linear support vector machine technique for fake review detection based on N-gram features. The authors evaluated the standard fake hotel review dataset that was collected from Tripadvisor.com. For feature extraction, the TF-IDF method was used, resulting in 90% accuracy.

Li et al. [22] attempted to define a general rule for identifying deceptive reviews. In their methodology, cross-domain datasets included 800 hotels' reviews collected from Amazon Mechanical Turk, in addition to 400 deceptive doctor reviews from domain experts. In terms of features, unigram, LIWC, and POS were used in their method. They used the multiclass classification method by using the Sparse Additive Generative Model (SAGE), which consists of a combination of generalized additive and topic models. Furthermore, SVM was applied to the same dataset and features. The achieved results of this experiment were 81% and 78% in terms of accuracy, respectively.

Savage et al. [23] proposed analyzing and detecting opinion spammers based on the rating deviation of the products. The authors concentrated on the dissimilarity between deceptive ratings and many opinions of truthful reviewers, and then they calculated the spasticity and honesty of each reviewer. Finally, they applied the binomial regression technique to detect reviewers having an abnormal attribution rating that deviated from public opinion.

Feng et al. [24] focused on the distribution footprints of reviewers and presented a correlation between spreading abnormalities and fraudulent activities. They evaluated the deceptive hotel reviews dataset that was collected from the Trip Advisor website. Fitzpatrick et al. [25] suggested that a deceptive review analysis is a particular implementation of a common recognition issue, where it is possible to use written and verbal clues.

Banerjee et al. [26] investigated details of the review text, such as comprehensibility, review length, writing style, and determinants of cognition features.

Zhang et al. [27] examined the influence of both textual and behavioral features using the restaurant and hotel review dataset. They compared these two features for fake review detection and found that behavioral features provided good results in fake review detection. Using a dataset collected from the restaurant domain, Luca et al. [28] explored some interesting findings and concluded that restaurant owners were more likely to post fake reviews once they had a weaker repute with a few customers and poor scorings. Wang et al. [29] developed a review graph to represent the associations between reviews, reviewers, and brands rated by reviewers. They used the same graph to recognize suspicious reviewers based on the iterative model.

Akoglu et al. [30] examined the network effects for spammer detection from two aspects: reviewer and review scoring for spammer discovery and grouping. They focused on the temporal features of the review that were concerned with the explosion of reviews and this effect on e-companies. Li et al. [31] presented a method that included behavioral and textual features for fake review detection.

Barbado et al. [32] presented a fake features structure for fake review identification based on an online Yelp product review dataset. The authors implemented different supervised machine learning techniques on the same dataset, which had reviewer features (personal, social, review activity, trust) and review-centric features (sentiment score). Their experimental results revealed that the AdaBoost algorithm obtained 82% accuracy.

Hajek et al. [33] applied two neural network models, deep feed-forward neural network and convolution neural network, based on the Amazon product review dataset. Then, they extracted feature sets, such as word emotions and N-grams. Their methodology results were 82% and 81% accuracy for DFFN and CNN methods, respectively.

3 Methodology

Fig. 1 presents the framework of the proposed methodology. It consists of seven steps.

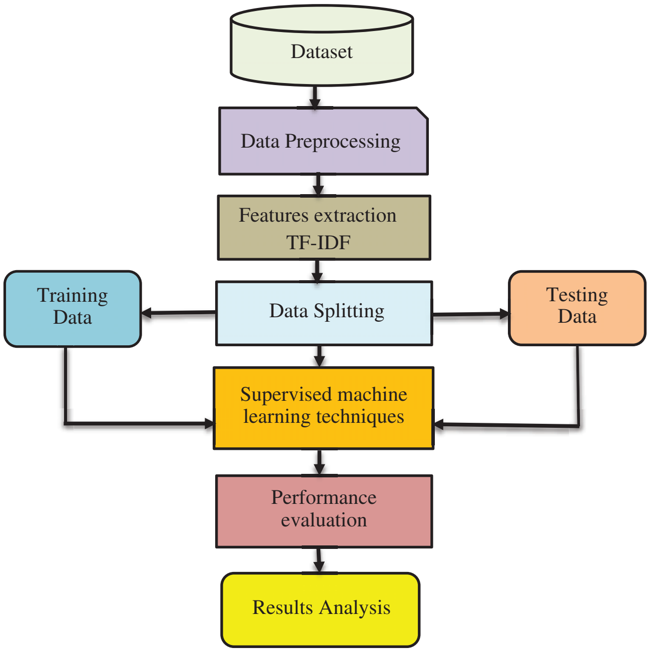

Figure 1: Framework for the proposed methodology

3.1 Dataset

In this experiment, the gold standard dataset was developed by Ott et al. [34] was used. It contains 1,600 hotel reviews collected from one popular hotel booking website, Trip Advisor. The authors of this dataset refined all 5- and 3-star rated reviews from 20 hotels in Chicago. The dataset was preprocessed by adding features such as review length, four-grams, sentiment score, and POS. Every review had the following features:

3.1.1 Name of the Hotel

This feature gives the details of the hotel name and the city where the reviews were selected.

3.1.2 Review Text

The user in text format wrote review contents. Based on this feature, the main analysis tasks were used for attaining a textual feature, such as sentiment analysis and linguistic features.

3.1.3 Sentiment Score

The sentiment score is a procedure for calculating and finding the polarity score (positive, negative, or neutral) of the given text. Negative fraudsters are usually accustomed to including negative words in their reviews more than positive ones when they reveal significant negative sentiment. In the same way, positive fraudsters are always accustomed to writing more affirmative (positive) words; therefore, the sentiment score should be calculated for every review text. The following formula was used for finding sentiment scores for each review text in the dataset:

$$S\left(r\right)=\frac{P\left(W\right)-N\left(W\right)}{T\left(W\right)} \tag{1}$$

where $S\left(r\right)$ indicates the sentiment (**S**) of the review; $P(W)$ refers to the number of positive words; $N(W)$ indicates the number of negative words; and $T(W)$ indicates the total number of positive and negative words in the review text.

3.1.4 Review Length

In this section, POS features were employed. POS tagging is the process of attachment of each word in the text's content with a POS tag based on its location and context in the sentence. Based on this method, the number of adjectives, nouns, prepositions, coordinating conjunctions, determiners pronouns, verbs, predetermines, and adverbs were extracted from the review text. The conclusion of this section was that truthful reviews had more nouns and adjectives, whereas fake reviews had more verbs and adverbs. An example of POS features is shown below.

Review 1: The hotel was so comfortable and nice; POS as demonstrated in the following Tab. 1 can represent this review.

3.1.5 N-Grams

The process of selecting N neighboring words from the text's contents as the features are called N-gram features [2]. When N = 1 (one word) at a time is assigned, it is known as a unigram. If two neighboring words (N = 2) at a time were chosen, thereafter, it is known as a bigram, and in the same way, when four neighboring words are assigned (N = 4) at the same time, it is known as four-grams.

Table 1: Representation of review by POS

Word in the review	POS tagging
The	DT (Determiner)
Hotel	NN (Noun singular)
Was	VBD (Verb, past tense)
So	RB (adverb)
Comfortable	JJ (Adjective)
And	CC (Coordinating conjunction)
Nice	JJ (Adjective)

3.2 *Preprocessing Steps*

Before performing the feature extraction step, the data needs to be exposed to certain cleanings, such as punctuation removal to strip punctuation marks from the review's text (? ! : ; ," .), stop word removal to clean the review sentences from article words ('the,' 'a,' 'an,' 'in'), stripping unnecessary words and characters from the whole dataset, and data tokenization to split each sentence of review contents into separated words, keywords, phrases, and pieces of information.

3.3 *Feature Extraction (TF-IDF)*

TF-IDF refers to the term frequency-inverse document frequency [35]; it is considered one of the feature extraction and representation methods used in text classification systems. It is used in natural language understanding and information retrieval. Furthermore, TF-IDF is a statistical approach used to scale how significant a term or word is to a document in the dataset. It has two parts: term frequency, which is used to calculate the frequency of particular words in documents to discover the resemblance between documents. The formula for TF is as follows:

$$TF(w)_d = \frac{\mathrm{n_w(d)}}{|\mathrm{d}|} \tag{2}$$

Set D indicates a group of documents, and d acts as a document. $\mathrm{d} \in \mathrm{D}$ is a document as a collection of sentences and words, w. $\mathrm{n_w}$ (d) represents the numbers of frequent words, w, presented in document d. Therefore, the volume of document d can be computed as follows:

$$|\mathrm{d}| = \sum_{\mathrm{w} \in \mathrm{d}} \mathrm{n_w}\,(\mathrm{d}) \tag{3}$$

The number of times that word appeared in the document was calculated in the above formula. The second part is inverse document frequency (IDF), which was used for calculating the number of documents in the corpus divided by the number of documents where that specific word appeared. The formula for computing IDF is as follows:

$$IDF(w)_d = 1 + log\left(\frac{|D|}{|\{d: D|w \in d\}|}\right) \tag{4}$$

Therefore, computing the TF-IDF for word w related to document d and corpus D can be performed by the following formula:

$$TF.IDF = TF(w)_d \times IDF\,(w)_D \tag{5}$$

TF-IDF shows the classifier that words are more or less repeated in the document.

3.4 Supervised Machine Learning Techniques

This subsection deals with the different supervised algorithms used for classifying the review text as fake or truthful. After converting the dataset into the form of TF-IDF features and before starting to train the machine learning classifier, the dataset was divided into 80% training set and 20% test set. In this experiment, four different supervised classifiers were applied, namely a support vector machine (SVM), a naive Bayes (NB), a random forest (RF), and adaptive boosting (AB).

3.4.1 Support Vector Machine

An SVM is a popular supervised probabilistic algorithm that can be utilized to divide the data sequentially and non-sequentially [36]. SVM is used for text categorization and provides good efficiency in high-dimensional vector space. Moreover, it represents the data training samples in space maps. The data points of the various classes are discriminated by a maximum margin inside the hyper-plane. Its decision boundary is the extreme margin for resolving training samples. The radial basis function (RBF) was applied in this method, and its equation was as follows:

$$K\left(X, X'\right) = exp\left(-\frac{\|\boldsymbol{X}-\boldsymbol{X'}\|^{\boldsymbol{2}}}{\boldsymbol{2\sigma^2}}\right) \tag{6}$$

where X, X' is the data training set, and it indicates the feature vectors of the dataset. A $\|\boldsymbol{X}-\boldsymbol{X'}\|^{\boldsymbol{2}}$ represents the squared Euclidean difference between the two feature inputs, and σ is a free parameter. The kernel function applied in this model is the RBF for detecting fake reviews.

3.4.2 Naive Bayes

Naive Bayes (NB) is a supervised machine learning method used for classification. It can be used to calculate the likelihood of an event given the likelihood of another event that previously occurred. It is based on the conditional probability theorem [37]. Through the text classification tasks, data contains a high dimension, meaning that each word represents one feature in the data. However, this model predicts the probability of each word in a text sentence and considers it a feature of any one of the dataset classes. Eq. (7) is expressed as a conditional probability for the NB algorithm as follows:

$$P(A \mid B) = \frac{P(B \mid A)(A)}{P(B)} \tag{7}$$

where A is the class label that is fake or truthful and B is the piece of text. In this formula, the probability of A given that B is true equals the likelihood of B given that A is true times the likelihood of A, divided by the likelihood of B. The kernel function used in this model is the linear function.

3.4.3 Random Forest

Random Forest (RF) is a widely used method in machine learning techniques [38]. RF, as the name suggests, is a forest of trees. It consists of several decision trees that can help in making a decision. Each tree in the random forest is made by the same strategy of making a single decision tree. By making a decision, votes of a small decision tree are taken and a class will be decided by majority vote. RF is known as the divide and conquer approach. It uses a few weak learners to generate a strong linear relationship. Every single tree in the classifier has a root node that is constructed of N data points or samples. Each node t in the tree also consists of N_t data points located at a split S_t for creating two sub-nodes that are t_L (left node) and t_R (right node). For calculating and determining the best split of the data points that had the highest information, an impurity measure is computed using the Gini index function as expressed in Eq. (8).

$$Gini = 1 - \sum_{i=1}^{C} (P_i)^2 \tag{8}$$

where $\boldsymbol{P_i}$ is the probability of a piece of text that is presented and observed in the dataset, and C is the number of classes.

3.4.4 Adaptive Boost

Adaptive boost is one of the supervised machine learning techniques that relates to a specific method to learn a boosted classifier [39]. It is a classification method used to construct strong learners from a linear combination of weak learners. In the adaptive boost model, each training sample utilizes a weight to decide the probability of being selected for a training set, and the final vote for classification is performed based on the weighted votes of weak learners. The formula for adaptive boost is as follows:

$$f_T(x) = \sum_{i=1}^{C} a_t h_t(x) \tag{9}$$

where f_T is a weak linear relationship that takes x as the input and gives a value representing a class. $a_t h_t(x)$ as the set of weak learners that are considered the last classifier.

3.5 Performance Evaluation

This subsection presents an evaluation of how proficiently the proposed models can classify and differentiate between fake and truthful review texts in terms of false-positive and false-negative rates. The performance analysis of the classifiers used was obtained from the confusion matrices, which are 2 * 2 tables for the calculation of five different evaluation metrics. In the confusion matrices depicted in Figs. 2–5, on the Y-axis, zero (0) indicates the truthful review class and one (1) indicates the fake review class. All five performance evaluation metric formulas are presented and discussed in the following section.

$$Accuracy = \frac{TP + TN}{FP + FN + TP + TN} \times 100 \tag{10}$$

$$Precision = \frac{TP}{TP + FP} \times 100 \tag{11}$$

$$Sensitivity = \frac{TP}{TP + FN} \times 100 \tag{12}$$

$$specificity = \frac{TN}{TN + FP} \times 100 \tag{13}$$

$$F1 - score = 2 * \frac{precision \times Sensitivity}{precision + Sensitivity} \times 100 \tag{14}$$

where true negative (TN) represents the total number of samples that were effectively predicted as truthful reviews by the classifier. False negative (FN) represents the total number of samples that were incorrectly classified as fake reviews. True Positive (TP) denotes the total number of samples that were successfully classified as fake reviews. False Positive (FP) is the sum of samples that were incorrectly categorized as truthful reviews.

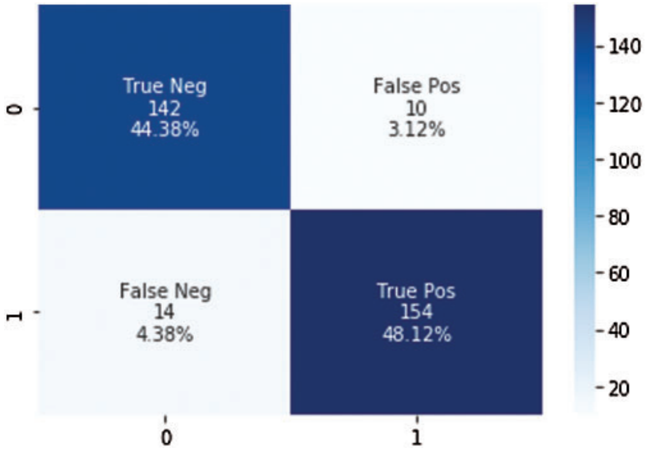

Figure 2: Confusion matrix for SVM

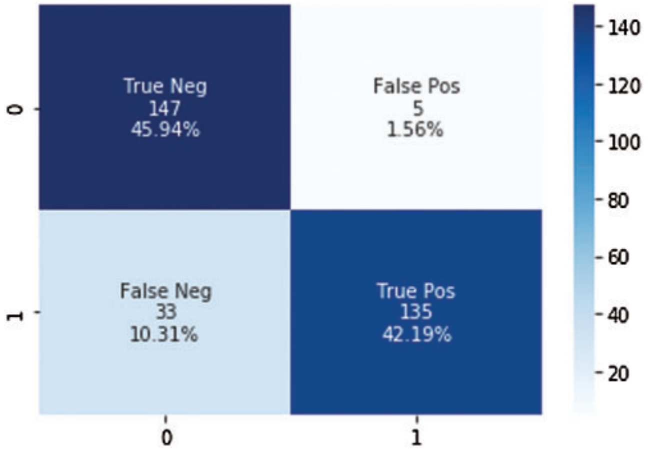

Figure 3: Confusion matrix for NB

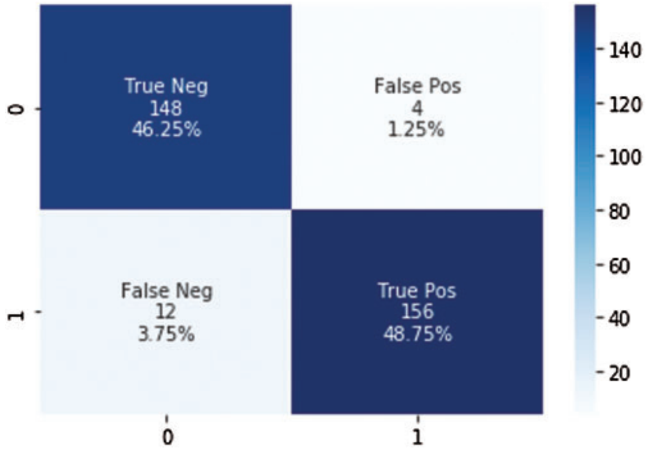

Figure 4: Confusion matrix for RF

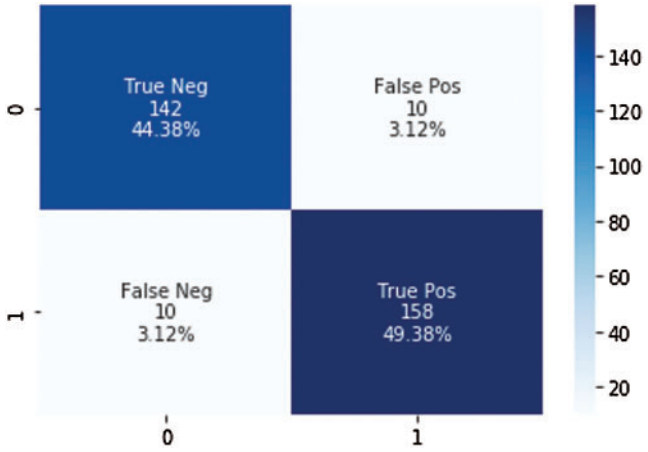

Figure 5: Confusion matrix for Adaboost

3.6 Experimental Results and Discussion

This subsection presents the obtained results of experiments that were executed to estimate the efficiency of four different supervised classifiers based on the standard fake hotel review dataset. In terms of features, POS and four-grams, as well as sentiment scores, were used for training and testing the proposed classifiers, which were NB, RF, Ada Boost, and SVM. The main task of the proposed classifiers was employed to detect and classify the review text into a fake or truthful review. The classification results are visualized and depicted in Fig. 6.

By comparing the classification results of the proposed classifiers, the RF classifier provided the best performance at detecting fake reviews and outperformed other classifiers with a 95% accuracy and F1-score metric. The sample classifications through RF were based on the majority voting of multi-decision trees. The Adaboost classifier provided equal numbers of positive and negative samples and had better results than SVM and NB classifiers with a 94% sensitivity metric. The naïve Bayes classifier had the highest misclassification, yielding an 88% accuracy and F1-score metric.

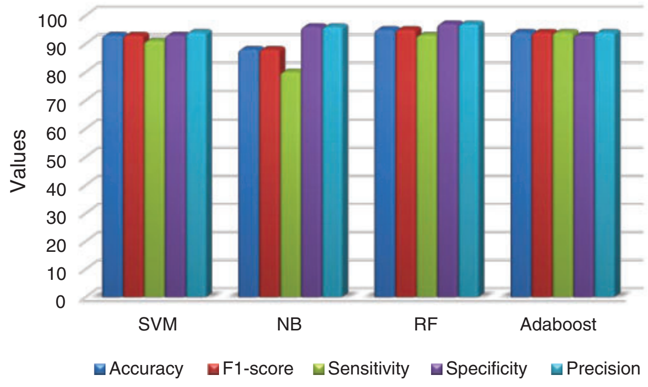

Figure 6: Visualization of the classification results

3.7 Wordcloud

Wordcloud is a technique used to visualize the most important and frequently used words in a given text. Here, wordcloud was applied to visualize the repeated words in truthful and deceptive review texts.

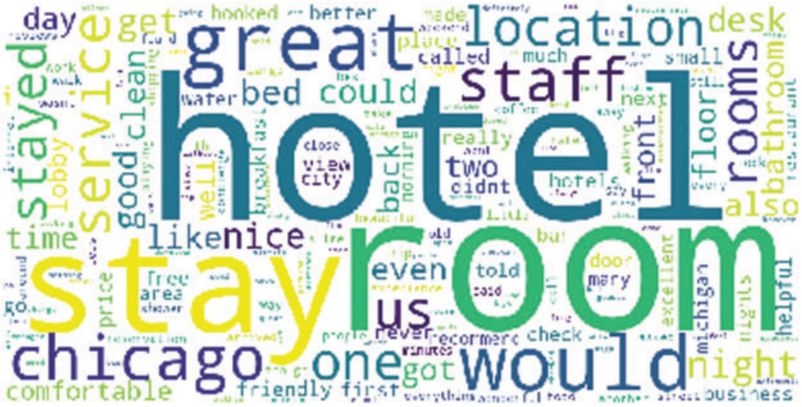

Figure 7: Wordcloud for truthful reviews

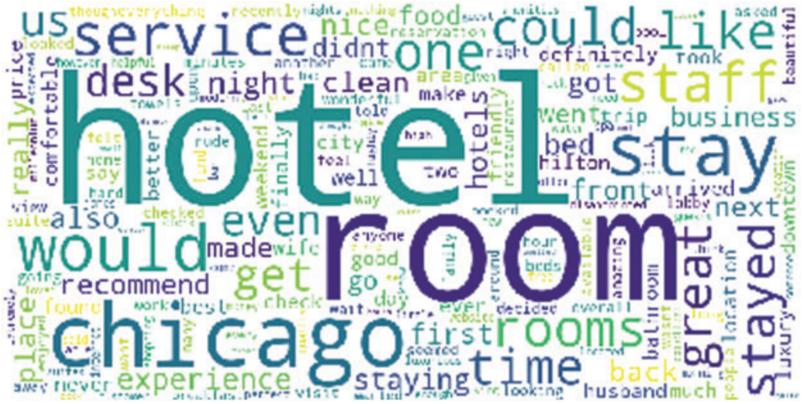

Figure 8: Wordcloud for deceptive reviews

4 Comparative Analysis

While developing a novel system for spam/fake review detection, it is necessary to understand what techniques and datasets have already been utilized in previous studies. In this section, the results obtained by the proposed models were analyzed and compared with existing approaches based on the same dataset and accuracy metrics. Tab. 2 shows the comparative analysis of the results of the proposed models with existing works.

Table 2: Comparative analysis with existing works

Paper	Dataset used	Features used	Technique	Result (%)
Shojaee et al. [11]	1600 hotel reviews from mechanical Turk and Trip Advisor websites.	Review's features	Naive Bayes SAGE	81 84
Fei et al. [13]	1600 hotel reviews from mechanical Turk and Trip Advisor websites.	Features of review and reviewer	Markov model	72
Ahmed et al. [21]	1600 hotel reviews from mechanical Turk and Trip Advisor websites.	Review's features	LSVM	90
Ott et al. [34]	1600 hotel reviews from Mechanical Truk and trip advisor websites.	Review's features	Naive Bayes SVM	89 88
Narayan et al. [35]	1600 hotels reviews from (AMT) Amazon.com	Review's features	Logistic regression	86
The proposed work.	1600 hotel reviews from Mechanical Turk and Trip Advisor websites	Review's features	Naïve Bayes SVM Ada boost RF	88 **93** **94** **95**

5 Conclusions

Fake reviews affect both customers and the e-commerce domain. Thus, fake review identification has gained significant interest in the domains of academic research and business. Based on fake hotel reviews, four supervised machine-learning techniques, namely naïve Bayes, support vector machine, random forest, and adaptive boost, were studied and implemented for fake review identification. For feature extraction, the TF-IDF method was used. By comparing the classification results of experiments, the random forest classifier provided better performance in detecting fake reviews and outperformed other classifiers, achieving a 95% accuracy and F1-score metric. The Adaboost classifier attained a higher (94%) sensitivity metric. A comparative analysis of methodologies was performed for fake review detection, which included features extraction methods, as well as the dataset used. According to the presented studies, most of the introduced studies have used the same feature extraction methods. After reviewing the literature, no large labeled fake review dataset was found. Many researchers have used a small-size dataset that was created by Narayan et al. [35]. However, the experimental results revealed that the proposed models outperform the compared methods.

6 Suggestions

Some opinion mining tasks, as follows, are suggested for supporting a business and merchants in collecting and analyzing large numbers of customer reviews:

a. Classification of sentiment that defines whether an opinion is negative, positive, or neutral.
b. Discovering the features of a reviewed entity and earning the opinion of the reviewer about a particular object.
c. Comparative sentences and relationship discovery that can be performed between one object with one or more similar objects.
d. Supervised machine-learning methods were able to classify the opinions of customers into fake and truthful with the highest accuracy and outperformed the human judgment for differentiating between fake and truthful opinions.
e. Fake opinions have two effects on consumers: 1) they drive the consumer to make bad decisions when purchasing a product or service; and 2) they alter the consumer's trust in e-commerce product reviews.

7 Limitations and Future Work

The dataset used in the experiments was limited to the hotel domain. The features used for training the proposed models were less. Future research should combine a large-scale dataset with several textual and behavioral features for detecting fake reviews on different e-commerce domains.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] S. N. Alsubari, S. N. Deshmukh, M. H. Al-Adhaileh, F. W. Alsaade and T. H. Aldhyani, "Development of integrated neural network model for identification of fake reviews in e-commerce using multidomain datasets," *Applied Bionics and Biomechanics*, vol. 2021, pp. 1–11, 2021.

[2] Y. Li, X. Feng and S. Zhang, "Detecting fake reviews utilizing semantic and emotion model," in *2016 3rd Int. Conf. on Information Science and Control Engineering*, Beijing, China, pp. 317–320, 2016.

[3] X. Hu, J. Tang, H. Gao and H. Liu, "Social spammer detection with sentiment information," in *2014 IEEE Int. Conf. on Data Mining*, Shenzhen, China, pp. 180–189, 2014.

[4] F. Long, K. Zhou and W. Ou, "Sentiment analysis of text based on bidirectional LSTM with multi-head attention," *IEEE Access*, vol. 7, pp. 141960–141969, 2019.

[5] V. W. Feng and G. Hirst, "Detecting deceptive opinions with profile compatibility," in *Proc. of the 6th Int. Joint Conf. on Natural Language Processing*, Nagoya, Japan, pp. 14–18, 2013.

[6] S. J. Delany, M. Buckley and D. Greene, "SMS spam filtering: Methods and data," *Expert Systems with Applications*, vol. 39, no. 10, pp. 9899–9908, 2012.

[7] L. Li, B. Qin, B. W. Ren and T. Liu, "Document representation and feature combination for deceptive spam review detection," *Neurocomputing*, vol. 254, pp. 33–41, 2016.

[8] S. Sarika, M. S. Nalawade1 and S. S. Pawar, "A survey on detection of shill reviews by measuring its linguistic features," *Int. J. Emerg. Trends Technol. Comput. Sci.*, vol. 3, no. 6, pp. 269–272, 2014.

[9] Q. Peng, "Store review spammer detection based on review relationship," in *Advances in Conceptual Modeling*. Berlin, Heidelberg, Germany: Springer, pp. 287–298, 2014.

[10] N. Hussain, H. T. Mirza, I. Hussain, F. Iqbal and I. Memon, "Spam review detection using the linguistic and spammer behavioral methods," *IEEE Access*, vol. 8, pp. 53801–53816, 2020.

[11] S. Shojaee, M. Murad, A. B. Azman, N. M. Sharef and S. Nadali, “Detecting deceptive reviews using lexical and syntactic features,” in *2013 13th Int. Conf. on Intelligent Systems Design and Applications*, Salangor, Malaysia, pp. 53–58, 2013.

[12] A. Heydari, M. A. Tavakoli, M. N. Salim and Z. Heydari, “Detection of review spam: A survey,” *Expert Systems with Applications*, vol. 42, no. 7, pp. 3634–3642, 2015.

[13] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos *et al.,* “Exploiting burstiness in reviews for review spammer detection,” in *Proc. of the Int. AAAI Conf. on Web and Social*, Media, Massachusetts, USA, pp. 175–184, 2013.

[14] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter and H. Al Najada, “Survey of review spam detection using machine learning techniques,” *Journal of Big Data*, vol. 2, no. 1, pp. 1–24, 2015.

[15] S. Noekhah, E. Fouladfar, N. Salim, S. H. Ghorashi and A. A. Hozhabri, “A novel approach for opinion spam detection in e-commerce,” in *Proc. of the 8th IEEE Int. Conf. on E-Commerce with focus on E-Trus*, Mashhad, Iran, pp. 1–8, 2014.

[16] M. Ott, C. Cardie and J. T. Hancock, “Negative deceptive opinion spam,” in *Proc. of the 2013 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, pp. 497–501, 2013.

[17] S. N. Alsubari, M. B. Shelke and S. N. Deshmukh, “Fake reviews identification based on deep computational linguistic features,” *International Journal of Advanced Science and Technology*, vol. 29, no. 8s, pp. 3846–3856, 2020.

[18] K. Goswami, Y. Park and C. Song, “Impact of reviewer social interaction on online consumer review fraud detection,” *Journal of Big Data*, vol. 4, no. 1, pp. 1–19, 2017.

[19] N. Jindal and B. Liu, “Opinion spam and analysis,” in *Proc. of the 2008 Int. Conf. on Web Search and Data Mining*, Palo Alto, California, USA, pp. 219–230, 2008.

[20] A. Mukherjee, V. Venkataraman, B. Liu and N. Glance, “What yelp fake review filter might be doing,” in *Proc. of the Int. AAAI Conf. on Weblogs and Social*, Media, Massachusetts, USA, pp. 409–418, 2013.

[21] H. Ahmed, I. Traore and S. Saad, “Detecting opinion spams and fake news using text classification,” *Security and Privacy*, vol. 1, no. 1, pp. 1–15, 2018.

[22] J. Li, M. Ott, C. Cardie and E. Hovy, “Towards a general rule for identifying deceptive opinion spam,” in *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, Maryland, pp. 1566–1576, 2014.

[23] D. Savage, X. Zhang, X. Yu, P. Chou and Q. Wang, “Detection of opinion spam based on anomalous rating deviation,” *Expert Systems with Applications*, vol. 42, no. 22, pp. 8650–8657, 2015.

[24] S. Feng, L. Xing, A. Gogar and Y. Choi, “Distributional footprints of deceptive product reviews,” in *Proc. of the Sixth Int. AAAI Conf. on Weblogs and Social*, Media, Dublin, Ireland, pp. 98–105, 2012.

[25] E. Fitzpatrick, J. Bachenko and T. Fornaciari, “Automatic detection of verbal deception,” *Computational Linguistics*, vol. 43, no. 1, pp. 269–271, 2015.

[26] S. Banerjee, A. Y. Chua and J. J. Kim, “Using supervised learning to classify authentic and fake online reviews,” in *Proc. of the 9th Int. Conf. on Ubiquitous Information Management and Communication*, Bali Indonesia, pp. 1–7, 2015.

[27] D. Zhang, L. Zhou, J. L. Kehoe and I. Y. Kilic, “What online reviewer behaviors really matter? Effects of verbal and nonverbal behaviors on detection of fake online reviews,” *Journal of Management Information Systems*, vol. 33, no. 2, pp. 456–481, 2016.

[28] M. Luca and G. Zervas, “Fake it till you make it: Reputation, competition, and Yelp review fraud,” *Management Science*, vol. 62, no. 12, pp. 3412–3427, 2016.

[29] G. Wang, S. Xie, B. Liu and S. Y. Philip, “Review graph based online store review spammer detection,” in *2011 IEEE 11th Int. Conf. on Data Mining*, Vancouver, BC, Canada, pp. 1242–1247, 2011.

[30] L. Akoglu, R. Chandy and C. Faloutsos, “Opinion fraud detection in online reviews by network effects,” in *ICWSM*, Massachusetts, USA, pp. 2–11, 2013.

[31] F. H. Li, M. Huang, Y. Yang and X. Zhu, “Learning to identify review spam,” in *Twenty-Second Int. Joint Conf. on Artificial Intelligence*, Barcelona, Catalonia, Spain, pp. 2488–2493, 2011.

[32] R. Barbado, O. Araque and C. A. Iglesias, “A framework for fake review detection in online consumer electronics retailers,” *Information Processing & Management*, vol. 56, no. 4, pp. 1234–1244, 2019.

[33] P. Hajek, A. Barushka and M. Munk, “Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining,” *Neural Computing and Applications*, vol. 32, no. 23, pp. 17259–17274, 2020.

[34] M. Ott, Y. Choi, C. Cardie and J. T. Hancock, “Finding deceptive opinion spam by any stretch of the imagination,” in *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oregon, pp. 309–319, 2011.

[35] R. Narayan, J. K. Rout and S. K. Jena, “Review spam detection using opinion mining,” in *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*. Singapore: Springer, pp. 273–279, 2018.

[36] V. S. Shirsat, R. S. Jagdale and S. N. Deshmukh, “Sentence level sentiment identification and calculation from news articles using machine learning techniques,” in *Computing, Communication and Signal Processing*. Singapore: Springer, pp. 371–376, 2019.

[37] W. Etaiwi and A. Awajan, “The effects of feature selection methods on spam review detection performance,” in *2017 Int. Conf. on New Trends in Computing Sciences*, Amman, Jordan, pp. 116–120, 2017.

[38] G. Louppe, L. Wehenkel, A. Sutera and P. Geurts, “Understanding variable importances in forests of randomized trees,” in *Advances in Neural Information Processing Systems*, Tahoe, USA, vol. 26, pp. 431–439, 2013.

[39] P. S. Toke, R. Mutha, O. Naidu and J. Kulkarni, “Enhancing text mining using side information,” *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 5, no. 5, pp. 793–797, 2016.