Hindawi
Complexity
Volume 2020, Article ID 8857570, 10 pages
https://doi.org/10.1155/2020/8857570

Research Article

Ensemble Machine Learning Model for Classification of Spam Product Reviews

Muhammad Fayaz,[1] **Atif Khan**,[2] **Javid Ur Rahman**,[3] **Abdullah Alharbi**,[4] **M. Irfan Uddin**,[5] **and Bader Alouffi**[6]

[1]*Center of Information Technology, University of Peshawar, Peshawar, Khyber Pakhtunkhwa, Pakistan*
[2]*Department of Computer Science, Islamia College Peshawar, Peshawar, Khyber Pakhtunkhwa, Pakistan*
[3]*Department of Computer Science, Govt Degree College, Dir, Khyber Pakhtunkhwa, Pakistan*
[4]*Department of Information Technology, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia*
[5]*Institute of Computing, Kohat University of Science and Technology, Kohat, Pakistan*
[6]*Department of Computer Science, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia*

Correspondence should be addressed to Muhammad Fayaz; fayaz@uop.edu.pk

Received 13 August 2020; Accepted 3 December 2020; Published 18 December 2020

Academic Editor: Dan Selisteanu

Nowadays, online product reviews have been at the heart of the product assessment process for a company and its customers. They give feedback to a company on improving product quality, planning, and monitoring its business schemes in order to increase sale and gain more profit. They are also helpful for customers to select the right products in less effort and time. Most companies make spam reviews of products in order to increase the products sales and gain more profit. Detecting spam product reviews is a challenging issue in NLP (natural language processing). Numerous machine learning approaches have attempted to detect and classify the product reviews as spam or nonspam. However, in order to improve the classification accuracy, this study has introduced an ensemble machine learning model that combines predictions from multilayer perceptron (MLP), K-Nearest Neighbour (KNN), and Random Forest (RF) and predicts the outcome of the review as spam or real (nonspam), based on the majority vote of the contributing models. In order to accomplish the task of spam review classification, the proposed ensemble and other benchmark boosting approaches are tested with 25 statistical features extracted from mobile application reviews of Yelp Dataset. Then, three different selection techniques are exploited to diminish the feature space and filter out the top 10 optimal features. The effectiveness of the proposed ensemble, the individual models, and other benchmark boosting approaches is again evaluated with 10 optimal features in terms of classification accuracy. Experimental outcomes illustrate that the proposed ensemble model outperformed the individual classifiers (MLP, KNN, and RF) and state-of-the-art boosting approaches like Generalized Boost Regression Model (GBM), Extreme Gradient Boost (XGBoost), and AdaBoost Regression Model in terms of classification accuracy.

1. Introduction

Product reviews refer to individual's feelings or views about certain products/services delivered to particular firms companies. In the modern technological world, online product reviews occupy a central place in the product evaluation process for a company and its customers. These reviews serve as feedback for a company to improve product quality, plan, and monitor its business ventures that result in boosting its productivity and profit. These also help customers in the right selection of products in less time and effort. Sometimes companies make spam reviews on products to enhance their sales. Detection of these reviews is a challenging task in NLP. An automatic approach is

therefore needed to detect spam reviews on products and will allow users to quickly see spam and nonspam product reviews.

Opinions like online product reviews provide the key source of information for consumers to guide them in purchasing products of interests. Customers post reviews about such products that serve feedback by mentioning their good and bad experiences [1]. These experiences have also great impact for businesses in the near future. As expected, these also provide the ways and means to modify decision-making of customers by posting misleading and false reviews. This unethical practice is termed opinion spamming, where spammers put down false reviews/opinions to either attract more customers or harm the reputation of a business or merchandise [1].

Bogus reviews/opinions can be classified into three groups: first, false reviews whose objective is to place bogus and untrusted details about a product to either harm or enrich its reputation; second, reviews that focused on products without mentioning any experience with other products; and third, nonreviews and advertisements comprised of text only and indirectly linked to the product [1]. The first group is very technical and problematic in identification, while the other two require less effort. The writer of these reviews/opinions is either a single spammer that works for a business or a cluster of spammers that exert effort in collaboration to get the end-results.

Google also pointed out concerns of fake reviews in an official report and clearly directed the innovators and users to not purchase and receive payments from firms that make available false reviews [2]. In certain countries, authorities have initiated actions against firms using false reviews to exaggerate their products. For example, a Canadian telecommunication company was penalised $1.25 million for posting fake encouraging reviews about its products. In addition, reputation of the CNN application was badly damaged by numerous adverse fake reviews that declined its rating and status in the Apple App Store [2]. In past, numerous efforts have been attempted to detect opinion spams. The researchers in [3] applied a machine learning methodology by using logistic regression to spot opinion spams. The authors in [4–6] adopted a supervised machine learning approach to notice spams. Other approaches [7,8] proposed a mix approach by combining supervised and semi-supervised learning to identify spam opinions. Despite these efforts, numerous shortcomings were noticed in the machine learning approaches. For instance, the use of numerous features is computationally expensive and provides poor flexibility and less accurate results.

This study attempts to use ensemble machine learning approach that combines the prediction from three classifiers, namely, Random Forest (RF), multilayer perception (MLP), and K-Nearest Neighbour (KNN), to improve the classification accuracy of spam product reviews. The selection of the three classifiers is made based on empirical analysis. The proposed ensemble machine learning model functions in the following way: At first, we extract 25 features from mobile application reviews of Yelp Dataset and represent the product reviews as feature vector. Too many features affect the performance of the model and not all features have the same contribution in predictive model, so nonvaluable features need to be filtered out. Therefore, we employ three different feature selection methods: Chi-square, Univariate, and Information Gain, to select the ten best features. Finally, for the task of spam review classification, the proposed ensemble model is used to classify the reviews as spam or real (nonspam). The effectiveness of the ensemble model is compared with the boosting approaches (XGBoost, GBM AdaBoost, and GBM Gaussian) that are benchmark techniques for this study. The contributions of this study are as follows:

(1) To propose an ensemble learning model for classification of spam product reviews

(2) To evaluate the ensemble classification model with all features extracted from spam product reviews in terms of classification accuracy

(3) To evaluate the effectiveness of the ensemble classification model with best features obtained using three feature selection techniques

The paper is planned as follows: Detailed literature is given in Section 2. Suggested ensemble mode is highlighted in Section 3. Numerous experimental results and a full discussion are given in Section 4. At the end, conclusion and future directions are presented in Section 5.

2. Related Work

Product reviews about user feedback and users satisfaction about particular products are mostly considered by customers to assess a particular product or service. These reviews help consumers in decision-making process of purchasing products. They also serve as guidelines for businesses to assess their future customers. Customer reviews related to mobile applications (apps) in the Google Play store use star ratings to indicate the quality of apps to other consumers. In addition, they serve as a tool for producers or manufacturers of such applications to upgrade their products/services to attract more customers [9]. In fact, product reviews have advantages for both the consumer and the producer; their role has still been reduced by spam reviews. The existence of such unethical reviews is a stumbling block in the way of online businesses, either illegally boosting their business profits or destroying their product reputation. The identification and classification of spam reviews is the need of the day to safeguard the trust of customers. Various studies [9,10] have presented a few techniques to identify spam reviews.

Technological advances have made it possible for online companies to sell mobile applications through playstores. Mobile users not only purchase products/applications from these playstores but also have the facility to post reviews that express their views on these products. Some of these product reviews/opinions are spams posted by fake users for their vested interests. The authors in [11] concentrated on fictitious opinions on crowdsourced fake review dataset by using n-gram-based classifiers to identify spam reviews. The study

in [12] also focused on detecting fictitious opinions and compared the behavioral features of spam reviews with factual-life Yelp Dataset through Support Vector Machine (SVM). Compared to simple *n*-gram-based approach, the results of approach relying on behavioral features improved the accuracy rate of spam classification. Both approaches are tested with generated datasets that is, fictional opinions that are powerless to depict spams finding in real-life product reviews. The authors in [13] used ranking-based approach to depict real-life product reviews and discovered the reviews burstiness in spotting opinion spammers. In order to single out spammers and legitimate users, they used Markov Random Field (MRF) and Loopy Belief Propagation (LBP) techniques to study the data.

The authors in [14] presented a good spam review detection method by using rating deviation of the review, reviewer's liveliness, and content-based information by adopting time series. The main shortcomings of this approach were high computational processing time and less competence in translating the semantic senses and information that are incorporated in the text of reviews. The study in [15] presented a model called the Fraud Eagle, which identified the network and graph associations of fake product reviews using [15] the iterative propagation-based classification approach. It is noted that the Fraud Eagle framework magnificently detected fake product reviews on online review site. The research work in [16] introduced the SPEAGLE framework that works on information collected from the review metadata like timestamps, ratings, texts, and network review information to identify spam product reviews.

In order to detect spam reviews in Chinese language review site, DianPing, the study in [17] attempted two approaches: the KNN and generic graph-based approaches. The results of these approaches have shown that the behavior of reviewers also helps to detect spams reviews. To detect fake and unknown reviews in the same research area, the study introduced another methodology entitled Positive-Unlabeled (PU) to identify false opinions by using supervised learning approach. Unknown reviews may be fake or genuine but fake reviews are always pretended. To extend these studies, researchers in [7] proposed another approach called mixing population and individual property method by using a novel semisupervised model that is functional to the PU learning (MPIPUL) method. It is noted that the PU learning operated well in balanced datasets but was not confirmed on imbalance datasets. The main problem that arises in these approaches is that the reviews are language-specific, that is, in Chinese language.

The current methodologies to detect spam product reviews/opinions undergo imbalance datasets; that is why the results may not be fully trusted. Therefore, the authors in [14] proposed another spam review detection method by using rating deviation of the review, reviewer's liveliness, and content-based information via time series. This was a good approach [14] for spam detection but it suffers from high computation processing time and is poor in interpreting the semantic values and information included in the text of reviews. Therefore, the study in [6] proposed a neural network approach that combines the recurrent neural network and convolutional neural network to know the unbroken document level depiction of the reviews. Compared to the discrete models, results of this work revealed that it has superior generalization ability.

The process of spam detection includes spotting the users' accounts from where spamming activities are performed for malicious objectives. The various detection approaches like *n*-gram, and linguistic, and pattern-based are unable to detect well-equipped spammers who write their reviews in a manner that seems real. Therefore, the study in [18] introduced an approach based on heterogeneous graphs to catch and associate the connection existing among the reviewers and the reviews. This methodology is exempted from the use of any textual content from the reviews and can increase the chances to identify opinion spammers in a better way. The research work in [19] concentrated on network footprints and presented a two-step approach to detect spammer's products and groups. This approach comprises two main modules, that is, Network Footprint Score (NFS) and GroupStrainer. The results showed that this methodology surpassed those approaches that had studied iTunes and Amazon datasets in spam detection with a high accuracy rate.

The authors in [20] used Naïve Bayes, max entropy, support vector model (SVM), and RF techniques for the iPhone mobile review dataset collected from Kaggle. Part of speech (POS) tagging and vectoring features are exploited to detect spam reviews. The best accuracy was given by RF. The authors in [9] used sentiment analysis as a feature for movies reviews dataset to detect spam reviews. Naïve Bayes provided improved performance than other machine learning classifiers. The authors in [21] used Naïve Bayes (NB), SVM, KNN, and Decision Tree (DT) for classification of movies products reviews via sentiments analysis stop words or without stop words used as features vector space or feature vector. The authors in [22] used Count Vectorizer and TF-IDF features using the SVM classifier on MTurk and Yelp Amazon Dataset of different product reviews. The study in [23] used logistic regression, Naïve Bayes, RF, SVM, and deep neural network classifier on the dataset of Amazon product reviews using TF-IDF features and found that deep neural network performs better than other machine learning classifiers. The authors in [24] developed an automatic system to identify rumours in online business reviews by classifying them as rumours and nonrumors using several machine learning classifiers.

In recent research studies, researchers also exploit the capabilities of supervised boosting approaches based on statistical features to achieve good and result-oriented accuracy in detecting fake product reviews. To make better the uncovering of opinion spams in cellphone application playstores, the authors in [25] proposed a methodology based on statistical features that are modeled via supervised boosting techniques like the GBM and XGBoost and to appraise two polyglot datasets of English language and Malay language. The appraisal of this study highlighted that XGBoost is utmost appropriate for spotting spam opinion in the English language dataset; on the other hand, the GBM Gaussian is appropriate for the Malay dataset, and, compared to other approaches, statistical-based features had attained better correctness rate for both datasets. We

propose an ensemble supervising machine learning approach for classification of spam products reviews. The detailed framework/structure of this methodology is given below.

3. Proposed Methodology

This section presents methodology of the suggested ensemble model for spam product reviews classification, as shown in Figure 1. The methodology consists of three phases, preprocessing, feature extraction, feature selection, as well as classification model for spam product reviews.

3.1. Preprocessing. In computational linguistics, data preprocessing is a vital step in cleaning the unwanted data, so that the cleaned data can be efficiently used before any further processing or providing it as input to the system. This phase comprises sentence segmentation, tokenization, stop word removal, and words stemming, which are discussed below.

3.1.1. Sentence Segmentation. It is used to detect text boundary and split the text into sentences. Commonly, exclamation (!), interrogation (?), and full stop (.) signs are used as indicators to segment the text.

For instance, we have the following text: "I purchased this product. It is the finest product available in this marketplace." We will get the two following sentences after segmentation:

Input product review text:

"I purchased this product. It is the finest product available in this marketplace."

Output:

Segment 1: "I purchased this product."
Segment 2: "It is the finest product available in this marketplace."

3.1.2. Tokenization. At this step, the sentences are divided into distinct words by dividing them at whitespaces like tabs, blanks, and punctuation signs, that is, dot (.), comma (,), semicolon (;), colon (:), and so forth. These are the main indications for dividing the text into tokens.

3.1.3. Stop Words Removal. Words that repeatedly occurred in a sentence are called stop words. These consist of prepositions (in, on, at, etc.), conjunctions (and, also, thus, etc.), articles (a, an, and the), and so forth. These words have little meaning in the text documents, and removing them from the text will help to improve the system performance.

3.1.4. Word Stemming. Word stemming plays a vital role in preprocessing. In order to capture the related concept, this step changes the derived words to their base or stem words. The renowned stemming algorithm, Porter's stemming (Porter, 1980), is adopted to remove suffixes like -ing, -es, and -ers from the text words. For example, the words "rising" and "rises" will be changed to the base form "rise" after stemming.

3.2. Features Extraction. Features play a significant role in text classification problems. The purpose of this step is to mine features from review datasets for product reviews classification problem. In this study, we extracted 25 features from mobile application reviews of Yelp Dataset. Almost all of these features are statistical and can be calculated directly from text. The proposed ensemble and other benchmark boosting approaches are tested with all 25 features for the task of spam reviews classification, as shown in Table 1. The description of all these features is presented in Table 2.

3.3. Features Selection. It is generally not a good idea to use all 25 features (in our case) to classify product reviews as spam or nonspam, since all features do not have the same relevance in constructing a reliable and accurate predictive model. Some features are valuable and contribute more to model prediction, while others are less valuable and have a serious impact on the effectiveness of the model. Moreover, the relevant and valuable features avoid overfitting, enhance accuracy, and lessen the training time of the predictive model. In order to address this issue, we exploited three feature filtering techniques, that is, Chi-square, Information Gain, and Univariate, to diminish the features space size in order to obtain optimal features as discussed in Section 3.5. Using these feature selection techniques, ten important and relevant statistical features are selected out of twenty-five from mobile application reviews of Yelp Dataset.

To get rid of nonvaluable and extra features, three features selection techniques (Chi-square, Univariate, and Information Gain) are adopted. Table 3 demonstrates that Chi-square technique selected the ten most salient features for the Yelp Dataset. Univariate selection technique chose the ten best features from the same dataset, as illustrated in Table 4. Finally, Table 5 depicts the ten best features selected using Information Gain. The next section highlights the fact that all classifiers with selected optimal features have performed well in comparison to all features. More specifically, given the reduced optimal features, the proposed ensemble model outperformed all the classifiers, as discussed in Section 4.2.

3.4. Classification Model for Spam Product Reviews. The focus of this phase is to categorize the product reviews as spam or real (nonspam), using ensemble learning model. Ensemble learning assists in enhancing the results (outcomes) of machine learning by integrating numerous models. This approach allows the creation of an improved predictive model compared to a single model.

In this study, Simple Majority Voting Ensemble or Voting Classifier has been employed to combine the predictions from multiple machine learning algorithms (MLP, Random Forest, and KNN) in order to get an improved combined result. Once the Voting Classifier has been trained, it can be used to predict the label of new instance

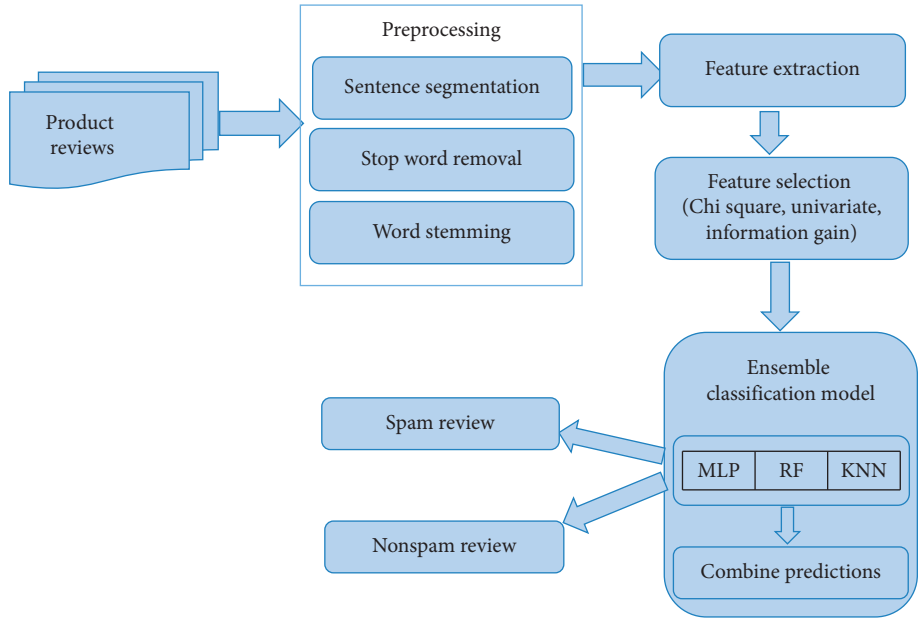

Figure 1: Proposed ensemble model for spam product reviews classification.

based on majority vote of contributing models. In order to evaluate the effectiveness of the individual models and ensemble model, initially, we train and test the individual models on the product review dataset using 10-fold cross validations. Then, we trained our proposed ensemble classifier on the same review dataset using the 10-fold cross validation.

MLP, RF, and KNN are state-of-the-art algorithms and have been proven to be very effective in addressing text classification problems. RF is normally used as baseline in text classification problems by researchers. It is an ensemble learning approach for the classification job and operates by creating a number of decision trees at training time and predicts the most frequent class decided by the contributing decision trees. The KNN algorithm works by calculating the distance (given in equations (1)–(3)) between a query and all examples in the data, picking the specified number of examples (K) that are nearest to the query [9].

KNN distance formula is

$$\text{Euclidean} = \sqrt{\sum_{i=1}^{k} (xi - yi)^2}, \tag{1}$$

$$\text{Manhattan} = \sum_{i=1}^{k} |Xi - Yi|, \tag{2}$$

$$\text{Minkowski} = \sum_{i=1}^{k} \left(|xi - yi|^q\right)^{1.q}. \tag{3}$$

Table 1: Classification results using all features on Yelp Dataset.

Classifiers	Accuracy (%)
RF	85.75
KNN	85.02
MLP	84.50
Ensemble classifier (MLP, KNN, and RF)	**88.13**
GBM Gaussian	84.74
XGBoost	85.59
GBM AdaBoost	85.87

Maximum accuracy achieved is highlighted in bold.

In classification problems, different K values in KNN algorithm will yield different classification results; however, a good value of K is determined by conducting the experiment several times with different values of K and then choosing the one that gives good classification results.

RF works by developing a number of decision trees at training time and predicting the most frequent class decided by the contributing decision trees. RF employs Gini Index and Entropy for classification purposes, which are given in the two following equations:

$$\text{Gini Impurity} = \sum_{i=1}^{c} fi(1 - fi), \tag{4}$$

$$\text{Entropy} = \sum_{i=1}^{c} -f \log(fi). \tag{5}$$

MLP, colloquially, is often referred to as neural networks known as "vanilla," particularly when having only one hidden layer [10]. As mentioned earlier in this section, this

Table 2: All features in Yelp Dataset.

S. no	Reviews	Description
1	avg_consine_simiarity_text	Average cosine similarity of reviews text
2	Polarity_text	Polarity of review text
3	stdev_rating_avg_revrating_app	Standard deviation review text and average review rating
4	rev_pos_ascend	Part of speech in review text in ascending order
5	avg_consine_simialriy_title	Average cosine similarity in review title
6	rev_pos_descend	Part of speech in reviews text in descending order.
7	avg_levenshtein_dist_text	Average Levenshtein distance in review text
8	avg_levenshtein_dist_title	Average Levenshtein distance in review title
9	rev_body_len	Length of review body text
10	rev_rating	Review rating
11	authomated_readability_index_text	Automated readability index for review text
12	avg_num_letters_per_word	Average number of letters per word in review text
13	num_unique_words_text	Number of unique words in review text
14	stdev_revApp_rating	Standard deviation for review on application and review rating
15	avg_words_freq_text	Average words in frequent review text
16	app_score	Application score in review text
17	stdev_num_words_title_text	Standard deviation for number of words and review text title
18	numeric_text_ratio	Numeric review text ratio
19	only_rev	Only review on application
20	num_unique_words_title	Number of unique words in review title
21	first_rev	First review on product
22	brand_names_in_title	Brand names in review title
23	avg_words_freq_title	Average words used frequently in title
24	unique_words_to_words_text_ratio	Unique words to words ratio in review text
25	unique_words_to_words_title_ratio	Unique words to words title ratio in review text

Table 3: Ten best features chosen by Chi-square for Yelp Dataset.

S. no	Features	Description
1	app_score	Application review scores
2	rev_body_len	Reviews body length
3	rev_pos_ascend	Reviews part of speech in ascending order
4	rev_pos_descend	Reviews part of speech in descending order
5	avg_cosine_similarity_text	Average cosine similarity for review text
6	rev_rating	Reviews rating
7	stdev_revApp_rating	Standard deviation of reviews application and review rating
8	num_unique_words_text	Numbers of unique words in review text
9	stdev_revrating_avgrevratingapp	Standard deviation for review rating and average review rating of mobile application
10	brand_names_in_text	Brand names in reviews text

Table 4: Ten best features selected by Univariate for Yelp Dataset.

S. no	Features	Description
1	app_score	Application review scores
2	rev_body_len	Review body length
3	rev_pos_descend	Reviews part of speech in descending order
4	avg_cosine_similarity_text	Average cosine similarity in review text
5	avg_levenshtein_dist_text	Average Levenshtein distance between reviews texts
6	rev_rating	Reviews rating
7	stdev_revApp_rating	Standard deviation of review application rating
8	num_unique_words_text	Number of unique words in review text
9	stdev_revrating_avgrevratingapp	Standard deviation of review rating and average review rating application
10	brand_names_in_text	Brand names in review text

Table 5: Ten best features selected by Information Gain for Yelp Dataset.

S. no	Features	Description
1	rev_rating	Reviews rating
2	stdev_revApp_rating	Standard deviation of review rating and rating application
3	stdev_revrating_avgrevratingapp	Standard deviation of review rating and average review rating application
4	avg_cosine_similarity_text	Average cosine similarity in review text
5	polarity_text	Polarity of review text
6	rev_pos_ascend	Reviews part of speech in ascending order
7	rev_pos_descend	Reviews part of speech in descending order
8	avg_levenshtein_dist_text	Average Levenshtein distance between reviews text
9	automated_readability_index_text	Automated readability index (ARI) of review body
10	avg_num_letters_per_word	Average number of letters per word in review body

study has proposed an ensemble learning model that integrates the effective ML algorithms, namely, RF, KNN, and MLP, and utilizes the statistical features (extracted from product reviews) for the task of categorizing product reviews as spam or nonspam.

4. Experimental Settings

4.1. Datasets for Evaluation. The proposed ensemble model is assessed on Yelp Dataset [14], which is a publicly available dataset that contains English reviews/opinions from several hotels and restaurants. This dataset is widely used in spam reviews detection problem. The dataset contains a total of 2526 opinions/reviews taken from Yelp's hotel reviews. It includes 389 spam and 2136 normal opinions.

For the task of spam reviews classification, we also evaluated the proposed ensemble classifier with the benchmark models [11] in terms of performance metric, that is, classification accuracy. The benchmarks model used boosting techniques such as XGBoost, GBM AdaBoost, and GBM Gaussian classifiers and the proposed ensemble model combined the predictions from machine learning classifiers such as KNN, RF, and MLP.

Table 6: Classification results using top 10 features selected by Chi-square.

Classifiers	Accuracy (%)
RF	85.81
KNN	84.75
MLP	84.90
Ensemble classifier (MLP, KNN, and RF)	**89.26**
GBM Gaussian	84.74
XGBoost	85.03
GBM AdaBoost	85.59

Maximum accuracy achieved is highlighted in bold.

Table 7: Classification results using top 10 features selected by Univariate.

Classifiers	Accuracy (%)
RF	85.72
KNN	84.46
MLP	84.50
Ensemble classifier(MLP, KNN, and RF)	**88.70**
GBM Gaussian	84.74
XGBoost	85.31
GBM AdaBoost	85.59

Maximum accuracy achieved is highlighted in bold.

4.2. Evaluation Results and Discussion. First of all, the preprocessing methods are applied over the given dataset to split the review text into sentences, tokenize the sentences into words, and to remove the stop words. Word stemming is then performed on the rest of words. Initially, we extracted all 25 statistical features from Yelp Dataset for the task of spam review classification. We evaluated the effectiveness of the proposed ensemble approach, the individual classifiers (MLP, RF, and KNN), and other benchmark boosting approaches for the spam review classification task. We know that all features do not have the same significance in constructing a reliable and accurate predictive model. Some features are valuable and contribute more to model prediction, while others are less valuable and adversely affect the model performance. In order to get rid of nonvaluable and extra features, three features selection techniques (Chi-square, Univariate, and Information Gain) are adopted to extract the top 10 features from review dataset. In order to investigate the impact of reduced optimal features set on classification accuracy, the proposed ensemble learning model, the individual models, and boosting approaches are tested again using the top 10 best features obtained using mentioned feature selection techniques.

To achieve the spam reviews classification job, the RF, KNN, and MLP classifiers are adopted to classify the reviews as spam or nonspam. The training and testing of all the classifiers including RF, KNN, and MLP, the proposed ensemble model, and the boosting approaches (GBM Gaussian, XGBoost, and GBM AdaBoost) are performed using stratified 10-fold cross validation (SCV). In stratified 10-fold cross validation (SCV), the folds are picked in such a way that each fold contains roughly the same number of class labels.

Classification results of all classifiers adopted in this work, with all features and features chosen by means of 3 selection techniques, are presented in this section. The individual models such as RF, KNN, and MLP, ensemble model (RF, KNN, and MLP), GBM Gaussian, XGBoost, and GBM AdaBoost are used in this study. At first phase, the performances of individual models such as RF, KNN, and MLP, ensemble model, and boosting models for all 25 features of Yelp Dataset are compared. The results in Table 1

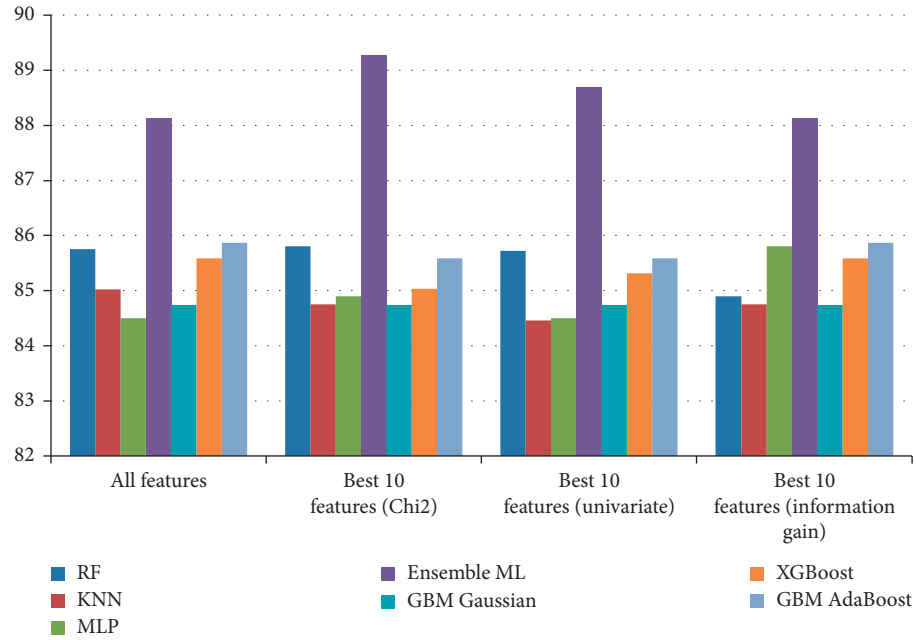

Figure 2: Classifiers accuracy on top 10 features set obtained using three feature selection techniques (Chi-square, Univariate, and Information Gain) for Yelp Dataset.

show that the proposed ensemble model has the highest accuracy of 88.13% as compared to other classifiers.

At the second phase, feature space is reduced by using the Chi-square feature selection method and the top 10 best features are selected for Yelp Dataset as given in Table 3. Given the 10 best features, all the classifiers are applied on Yelp Dataset for the spam review classification task.

The classification outcomes, given in Table 6, demonstrate that the proposed ensemble model achieved accuracy of 89.26% and performed better than individual models and other boosting approaches on ten best features selected using Chi-square features selection technique.

Referring to the results given in Table 6, the proposed ensemble model achieved the highest accuracy of 89.26 percent, RF obtained the second highest accuracy of 85.72 percent, the accuracy of GBM AdaBoost was 85.59 percent, XGBoost accuracy was 85.03 percent, and GBM Gaussian got the lowest accuracy of 84.74 percent. Among the individual models, RF achieved the highest accuracy, which is even better than boosting approaches. However, MLP obtained the lowest accuracy of 84.50 percent.

At the third phase, Univariate features selection method is utilized to filter the feature space in order to choose the top 10 best features for Yelp Dataset as shown in Table 4. For the spam review classification task, and given the 10 best features, all the classifiers are applied to Yelp Dataset.

The classification outcomes, given in Table 7, reveal that the proposed ensemble model attained accuracy of 88.70% and performed better than individual models and other boosting approaches on ten best features selected using Univariate selection technique.

Similarly, at the fourth phase, Information Gain is used to select the 10 optimal features from all features of Yelp Dataset, as shown in Table 5.

For the spam review classification task, and given the 10 best features, all classifiers are applied to Yelp Dataset. The classification outcomes, for the task spam review classification using 10 best features, are given in Table 7, revealing that the proposed ensemble model achieved accuracy of 88.13% and performed better than individual models and other boosting approaches.

Figure 2 depicts the classification accuracy of the proposed ensemble model and other benchmark classifiers on all features and top 10 features obtained using Chi-square, Univariate, and Information Gain selection techniques.

From the results shown in Figure 2, we observed that accuracy of classifiers either improved, remained unaffected, or slightly downgraded with the best selected features compared to the accuracy of classifiers obtained using all features. It is also noteworthy that the proposed ensemble model (RF, KNN, and MLP) surpasses all the benchmark classifiers on top 10 best features obtained using aforementioned selection techniques. From the empirical results given in Tables 1 and 6–8, we concluded the following:

(1) The accuracy of the proposed ensemble model improved with best features obtained using Chi-square and Univariate selection techniques, while it remained constant with IG

(2) The accuracy of the GBM Gaussian remained constant on all feature selection techniques

(3) The accuracy of XGBoost and GBM AdaBoost either remained constant or slightly downgraded on best features

(4) The accuracy of RF and KNN classifiers either improved or slightly downgraded on best features,

Table 8: Classification results using top 10 features selected by Information Gain.

Classifiers	Accuracy (%)
RF	84.90
KNN	84.75
MLP	85.81
Ensemble classifier (MLP, KNN, and RF)	**88.13**
GBM Gaussian	84.74
XGBoost	85.59
GBM AdaBoost	85.87

Maximum accuracy achieved is highlighted in bold.

while the accuracy of MLP improved or remained constant on best features

(5) Overall, the classification accuracy of the proposed ensemble model is superior to those of all individual models as well as other boosting approaches

5. Conclusion and Future Work

tSpam product review classification is a difficult task in the area of opinion mining. Numerous research efforts have been attempted to address this issue. However, in this study, we present an ensemble model that combines the predictions from MLP, KNN, and RF to classify product reviews as spam or nonspam. For the task of spam review classification, we studied the impact of all 25 statistical features on the proposed ensemble model, the individual models, and other boosting approaches. We found from the empirical results that the proposed ensemble model outperformed all classifiers in terms of classification accuracy. In next step, we employed feature selection techniques (Chi-square, Univariate, and Information Gain) to extract the top 10 features from the reviews dataset. The performances of the proposed ensemble model and other classifiers are evaluated using 10 best features obtained using the three selection techniques; and we found from the experimental outcomes that the ensemble model surpassed all the classifiers in terms of accuracy for the task of spam review classification achieved on Yelp Dataset. Hence, it is verified from results that the proposed ensemble approach is superior to other algorithms such as boosting approaches like Extreme Gradient Boost (XGBoost), the Generalized Boosted Regression Model (GBM), and AdaBoost Regression Model. In the future, we want to explore the deep learning approach and longest short-term memory with weighted TF-IDF embedding for the task of spam review classification.

Data Availability

The data are publicly available at https://www.yelp.com/dataset.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This research was supported by Taif University Researchers Supporting Project number (TURSP-2020/231), Taif University, Taif, Saudi Arabia.

References

[1] H. Ahmed, I. Traore, and S. Saad, "Detecting opinion spams and fake news using text classification," *Security and Privacy*, vol. 1, no. 4, p. e9, 2018.

[2] D. Martens and W. Maalej, "Towards understanding and detecting fake reviews in app stores," *Empirical Software Engineering*, vol. 24, no. 6, pp. 3316–3355, 2019.

[3] N. Jindal and B. Liu, "Analyzing and detecting review spam," in *Proceedings of the Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pp. 547–552, Omaha, NE, USA, October 2007.

[4] J. Li, M. Ott, C. Cardie, and E. Hovy, "Towards a general rule for identifying deceptive opinion spam," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1566–1576, Baltimore, MD, USA, June 2014.

[5] Y. Lin, T. Zhu, H. Wu, J. Zhang, X. Wang, and A. Zhou, "Towards online anti-opinion spam: spotting fake reviews from the review sequence," in *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, pp. 261–264, Beijing, China, August 2014.

[6] Y. Ren and D. Ji, "Neural networks for deceptive opinion spam detection: an empirical study," *Information Sciences*, vol. 385-386, pp. 213–224, 2017.

[7] Y. Ren, D. Ji, and H. Zhang, "Positive unlabeled learning for deceptive reviews detection," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 488–498, Doha, Qatar, October 2014.

[8] H. Li, Z. Chen, B. Liu, X. Wei, and J. Shao, "Spotting fake reviews via collective positive-unlabeled learning," in *Proceedings of the 2014 IEEE International Conference on Data Mining*, pp. 899–904, Shenzhen, China, December 2014.

[9] A. Sharaff and A. Soni, "Analyzing sentiments of product reviews based on features," in *Proceedings of the 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 710–713, Tirunelveli, India, May 2018.

[10] A. Mikheev, "Document centered approach to text normalization," in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 136–143, Athens, Greece, July 2000.

[11] M. Hazim, N. B. Anuar, M. F. Ab Razak, and N. A. Abdullah, "Detecting opinion spams through supervised boosting approach," *PloS One*, vol. 13, no. 6, Article ID e0198884, 2018.

[12] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pp. 219–230, Palo Alto, CA, USA, February 2008.

[13] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," 2011, http://arxiv.org/abs/1107.4557.

[14] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "What yelp fake review filter might be doing?" in *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, Cambridge, MA, USA, July 2013.

[15] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Exploiting burstiness in reviews for review

spammer detection," in *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, Cambridge, MA, USA, July 2013.

[16] A. Heydari, M. Tavakoli, and N. Salim, "Detection of fake opinions using time series," *Expert Systems with Applications*, vol. 58, pp. 83–92, 2016.

[17] L. Akoglu, R. Chandy, and C. Faloutsos, "Opinion fraud detection in online reviews by network effects," in *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, Cambridge, MA, USA, July 2013.

[18] S. Rayana and L. Akoglu, "Collective opinion spam detection: bridging review networks and metadata," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 985–994, Sydney, Australia, August 2015.

[19] C. Xu, J. Zhang, K. Chang, and C. Long, "Uncovering collusive spammers in Chinese review websites," in *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pp. 979–988, San Francisco, CA, USA, October 2013.

[20] G. Wang, S. Xie, B. Liu, and S. Y. Philip, "Review graph based online store review spammer detection," in *Proceedings of the 2011 IEEE 11th International Conference on Data Mining, ICDM*, pp. 1242–1247, Vancouver, Canada, December 2011.

[21] J. Ye and L. Akoglu, "Discovering opinion spammer groups by network footprints," in *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 267–282, Porto, Portugal, September 2015.

[22] N. M. Danish, S. M. Tanzeel, N. Usama, A. Muhammad, A. Martinez-Enriquez, and A. Muhammad, "Intelligent interface for fake product review monitoring and removal," in *Proceedings of the 2019 16th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, pp. 1–6, Mexico City, Mexico, September 2019.

[23] E. Elmurngi and A. Gherbi, "An empirical study on detecting fake reviews using machine learning techniques," in *Proceedings of the 2017 Seventh International Conference on Innovative Computing Technology (INTECH)*, pp. 107–114, Luton, UK, August 2017.

[24] A. Habib, S. Akbar, M. Z. Asghar, A. M. Khattak, R. Ali, and U. Batool, "Rumor detection in business reviews using supervised machine learning," in *Proceedings of the 2018 5th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC)*, pp. 233–237, Kaohsiung, Taiwan, November 2018.

[25] A. V. Sandifer, C. Wilson, and A. Olmsted, "Detection of fake online hotel reviews," in *Proceedings of the 2017 12th International Conference for Internet Technology and Secured Transactions (ICITST)*, pp. 501-502, Cambridge, UK, December 2017.