JEAS
Journal of Economics and Administrative Sciences
E-ISSN: 2148-1792 P- ISSN: 1302-2024
Volume 5, Supplement Issue 2
www.jeasweb.org

TF-IDF and MRC in Mining Opinions of e-Commerce Customers: Examination and Analysis

Shamim Fashi
Master of Information Technology Management; Islamic Azad University, Science and Research Branch, Tehran, Iran.

ABSTRACT

Today, the Internet has become a common platform for an extensive range of business interactions. As such, the purpose of this study was to analyze the reviews of online customers on products purchased in online stores. The present research is an applied study that employs descriptive methods for its purposes. The statistical population of this research consisted of users of Amazon online retail store with verified purchases. The data required for analysis was collected using Machine Reading Comprehension (MRC) and term frequency-inverse document frequency (TF-IDF) methods and were hence applied to the mentioned data set, as a result of which the customers' reviews were classified into three categories, namely positive, neutral and negative. The proposed method employs the random forest algorithm to train a set of decision trees following the initial processing of users' reviews and the subsequent extraction of features, for which the MATLAB environment has been utilized. To evaluate the method of this research, the criteria of accuracy, recall, and F-score were used. The research sought in their experiments to determine the optimal values for the parameters effective in the random forest algorithm, the output of which was compared with single and ensemble methods as well the so-called idea mining methods recently introduced to the virtual stores. The results indicated that the random forests algorithm exhibits superior performance in terms of data mining evaluation criteria. After implementing the proposed method in MATLAB software, the results were compared with previous methods. The results show that the proposed method is able to achieve an accuracy of 87%, a precision of 89%, the F-score of 88%, and recall of 91%.
Keywords: Opinion Analysis, Virtual Stores, TF-IDF, MRC.

Introduction

Today, e-shopping has garnered a great share of attention worldwide, such that 53% of people of the countries party to the Organization for Economic Co-operation and Development (OECD) had a minimum of one online purchase in 2016 (OECD, 2018). Given the many advantages it offers, such as the quick access to product information, the ability to compare different features, and the reduced cost of information search for customers, the inclination to use online shopping tools have been significantly on the rise (Jane et al., 2017; Singh and Matsui, 2017).

The success of online stores seems correlated with the intention of consumers to adopt the corresponding technologies, hence leading to the emergence of the notion of consumer behavior (Olah Khan et al., 2017;

Lim et al., 2016; and Rodriguez and Trujillo, 2013). In the meanwhile, authentic analysis of reviews remains of paramount significance to the ever-rising expansion of developments. The task of mining and analysis of reviews and sentiment is dedicated to the extraction of users' opinions, recognizing their term polarity, and identifying users in various fields (Kantardzic et al., 2011).

Shahidi et al. (2017) performed a study to study big data and unstructured information, in which they proposed a lexicon-based method for summarizing and extracting hidden and valuable information. The results show that the performance accuracy of this algorithm and method is lower than machine learning methods. Vasavi (2018) studied the extraction of hidden patterns within road accident data using machine learning techniques. Keshwani et al. (2018) sought in their study to predict the market movement of gold, silver, and crude oil using sentiment analysis. Kang et al. (2018) proposed a new method in their article for opinion mining using ensemble text hidden Markov models in text classification. The results clearly indicated that the proposed method was able to handle texts classification tasks with improved precision. The results also indicate that the appropriate selection of test data sets is effective in optimizing the method.

Kouloumpis et al. (2017) performed a study to examine Twitter sentiment analysis through hashtags. Features extracted in this article included n-gram, Lexicon, Part-of-Speech, and Micro-Blogging, which categorized tweets into three categories, namely, positive, negative, and neutral. Chen et al. (2017) conducted a study to examine social opinion mining for supporting buyers' complex decision making, in which a supervised method called CRF was used for idea mining. The accuracy of the method was tested on two purchased products. Positive and negative comments were each scored between 0 and 5.

Bertola and Patti (2016) examined a plethora of ontology-based effective models to organize artworks in the social semantic web using the title of information source and tags as textual works that visitors employ to exhibit artworks in a social-operating system. Nakov and Rosenthal (2016) performed a study to examine the sentiment analysis of Twitter and other social media texts using a proposed semantic evaluation task.

Jorge et al. (2015) employed the data heterogeneity method in the proposed model for examining opinion mining and information fusion. Chaovalit et al. (2015) proposed semantic-based idea mining in their model. The proposed method exhibited an accuracy of 85.54%, a better performance compared to other semantic-based methods. Chen et al. (2014) dedicated their academic endeavors on thematic aspect of classification methods, in which ensemble aspects have employed extract semantic classes. Zhiqiang et al. (2014) used PLSA and LDA semantic modelling approaches in their proposed method for text-based sentiment analysis in the text. Using the aforementioned two methods, words are separated and analyzed within the text. Lasota et al. (2012) showed that one of the important factors in the efficiency of sentiment analysis methods is the ensemble algorithm used to integrate scores.

To our knowledge, a comprehensive method is yet to be proposed for examining the opinions and sentiments of Internet users together, based on which it would be able to identify and provide appropriate services to customers. As such, proposing a method that can use these analyzes to increase the quality of services and, consequently, increase the satisfaction of Internet customers is of paramount significance. Accordingly, the purpose of this study was to analyze the opinions and sentiments of customers of virtual stores using TF-IDF and MRC criteria.

Research methods

The current research is an applied study that employs descriptive methods for its purposes. The statistical population of the research consisted the customers of Amazon website with verified purchase. Valid scientific articles from reputable journals corresponding to the field of research and library archives were used to collect the data required for the purposes of the study.

The method presented in the present study is consisted of two main parts. In the first part, the available data set is processed and the desired properties are extracted therefrom. In the second part, a learning model is developed and evaluated using a random forest algorithm. It should be noted that both steps of the proposed method are implemented in MATLAB environment. In this study, it is assumed that the training data set is available offline. The data set used includes the opinions of users on YouTube, each of which is attributed a positive, negative or neutral sentiment. Finally, the present study focuses solely on English-language content.

This research has employed a hybrid algorithm consisted of MRC and TFIDF to analyze the opinion of users. As such, the current study seeks to offer an improvement for classification methods on data mining with the aim of sensing the sentiment of users, per their reviews and comments, towards a product or service. These sentiments can be either positive, negative or neutral. The proposed method is divided into two main parts. In the first part, the existing data set is pre-processed and the necessary properties are hence extracted therefrom. It should be nonetheless stated that user comments are perceived as raw unstructured textual data that are of low analytical value without prior pre-processing. There are to two major stages in the classification of comments. In the first stage, users' comments are processed as their actual form, that is, text and conversational sentences. The purpose of this stage was to develop a new data set that can be prospectively used to train the learning model

The output of this stage is the very same user feedbacks, albeit in a new feature space, one that can be readily employed to train the learning model. The second stage pertains to training the proposed learning model using ensemble classification. Finally, the learning model is used to classify incoming ideas following the completion of the training stage.

1. Data pre-processing

For the pre-processing phase, first all the terms within the existing data are converted to lower case. Then, URLs and other web-based addresses are omitted from users' comments. All punctuation and extra characters, such as spaces and numbers, are also removed from the comments. These tasks are the first step in the preprocessing phase.

Next, each sentence is broken down into its constituent words, a process called tokenizing. The output of this step is comments, each of which is converted into a vector of words. This process is often considered a sub-task of parsing input. Repetitive words such as suffixes or auxiliary verbs, which are often repeated in everyday sentences, and are hence of low analytical value, are omitted. As such, all the important words in all the reviews in the dataset are extracted thus far. Two approaches are considered here for performing the main task in data processing, which is the extraction of meaningful features for the learning methods.

In the statistical approach to feature extraction, each n-gram, or shingle, is a contiguous sequence of n items from a given sample of text or speech, the attributable values for which are 1, 2 or 3 in this study. For example, in the phrase “In the name of God, Most Gracious, Most Merciful”, after the deletion of the punctuation and preposition “in” and assuming n = 3, the extracted features of the phrase include the set {“name of God”, “God Most Gracious”, “Gracious Most Merciful”}. As such, all possible n-grams are extracted from the comments in the dataset. Then, to quantify these features, a statistical method known as Term Frequency-Inverse Document Frequency, or TF.IDF for short, is used in the form of Equation (1).

In this method, each word (phrase) is given a weight based on its frequency in the document. The document here is an “opinion”

$$TF.IDF_{t,i} = tf_{t,i} * \log [N / df_t]$$

In which, t represents a word (phrase) while i represents an opinion (document). $TF.IDF_{t,i}$ is hence the weight calculated for the word t in opinion i. $tf_{t,i}$ refers to the frequency of the phrase t in document I, while df_t represents the frequency of features in which t is observed. N is the total number of available opinions.

Each opinion here is restructured as a vector, the elements of which are shingles in the feature space whose value is the weight of TF-IDF calculated for that elements. Elements that are non-existent in an opinion are attributed a weight of zero.

In the semantic approach to feature extraction, each word is considered a stand-alone feature. That is, every 1-gram is a feature. This approach employs a lexicon-based method instead of TF-IDF method or other statistical approach to quantify features. One of the purposes of the research is to compare these two approaches with each other, for which, a tool known as MRC Machine Usable Dictionary is used. It is is a web-based dictionary database that contains 150837 words and has more 26 lexico-semantic features per word and is used in artificial intelligence applications, especially natural language processing.

In this research, three features of familiarity, concreteness and imageability were used. The values extracted from the MRC for each of these features range from 100 to 700. These features are optionally selected. After extracting all the feature from the available opinions, each word is inputted to the online MRC tool, and the values of the aforementioned features are returned as output. The average output values are then stored as the final value for that particular term. In this method, as in the previous method, each comment is structured as a vector, again, the, elements of which are the extracted features.

Finally, after performing each of the above approaches, the output of the preprocessing stage is matrix, the rows of which represent ideas, while the columns are features in the feature space.

2. Classification of opinions

This study sought the contribution of ensemble methods to analyze the opinions. The random forest model is hence trained and evaluated using the data generated in the previous step. The purpose of making this model is to analyze the opinions of users in virtual stores in order to identify the sentiment of the author of the comment and categorize it as either positive, negative or neutral.

3. Estimation of prediction error

In random forests, a separate test suite is not required to estimate bias-free error, i.e. the precision of the final model. Errors can be estimated at runtime. At the time of generation of each tree, about one third of the training data belonging to that tree is discarded as out-of-bag (OOB) errors and are thus excluded from the tree generation process.

4. Extractable information from the proposed model

A random forest tree is best used when facing with numerous variables and/or complex space, for which all the features are employed. After determining the most important features, the forest can be re-generated using only the important, or the desired, features. If the criterion for determining the best failure feature in each node is the Gini index, when the failure of a node occurs on the m feature The Gini impurity value for the sub-nodes generated will be lower than that of the parent node. By adding up this reduction in impurities for each m variable across the forest, it provides an instant insight to that variable, which is consistent with the previous method.

Findings

1. Test data set

The data used in this study is extracted from the opinion pool of YouTube users. This dataset contains 1256 comments. In this dataset, each comment is written by a user and has a specific tag that indicates the author's sentiment. This tag can be either “positive”, “negative” or “neutral”. The existing data are presented in the form of XML files which, after pre-processing and loading into a format usable for processing, are entered as input to the first phase of the proposed algorithm. An example of the comments is shown in the

XML files in Figure 1. In these files, each comment is presented in the form of XML tags. Each sentence is placed in a *sentence* tag. The sentiment of the user (also known as polarity) in the aspectTerm tag, along with the aspect of the product that the user is commenting on and the start and end position in the sentence. The comments and the sentiment of the author are extracted by processing XML files with the help of MATLAB libraries and hence inputted in the proposed method.

```
<sentence id="3">
    <text>My overall experience with this monitor was very poor .</text>
    <aspectTerms>
        <aspectTerm from="32" to="38" polarity="negative" term="monitor" pos="nn"/>
    </aspectTerms>
</sentence>
```

Figure 1: An example of user comments in the form of XML tags

2. Examining various parameters in the model

To evaluate the random forest algorithm, OOB error estimation is used, which is a method for evaluating this algorithm, and the graphs are drawn according to this criterion. Some criteria for evaluating classification algorithms were used in data mining such as accuracy, recall (sensitivity), precision and F-score.

2.1. Number of trees

The following experiment was performed to examine the effect of the number of trees in the forest on the efficiency of the algorithm and to select the appropriate number of trees in the forest.

In case the statistical approach is deemed optimal, the possible values for n in the extraction of n-grams will be 1, 2 and 3. For each individual n, the resulting feature space is different. For n = 1, the size of the feature space is 2520, for n = 2, the feature space would have 5989 features, while for n = 3, a feature space with a dimension of 5308 is generated. This experiment is reiterated for each feature space. Moreover, in case of adopting semantic approach for feature extraction, the feature space will include separate words in the opinions. Like the case in the statistical approach for n = 1, in which every 1-gram is a feature. As previously stated, the feature space will have 2520 dimensions.

The number of trees varies from 10 to 150, with increments of 10. In all experiments, according to the main random forest construction algorithm, one-third of the training data were excluded as OOB errors and other evaluation criteria were used as experimental data.

According to Figures 2 and 3, as the number of trees increases OOB errors exhibit a decreasing trend. That is, as the number of trees increases, so does the classification ability of the forest

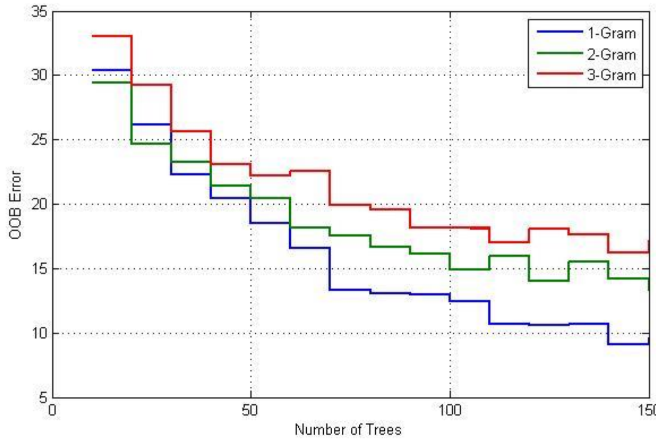


Figure 2: Estimation of OOB errors in testing the effect of number of trees, based on statistical approach of feature extraction

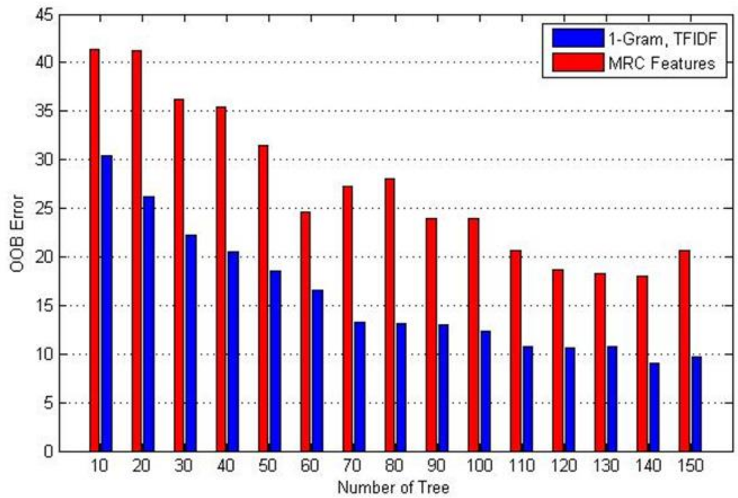


Figure 3: Estimation of OOB errors in testing the effect of the number of trees, based on the semantic approach of feature extraction

2.2. Parameter m

Another parameter that can be adjusted in the random forest model is known as the m parameter. This parameter represents the number of features in each node of the tree randomly extracted from the set of features, which is in turn used to determine the best feature for prospective splitting. During the generation of random forest, this parameter is considered for all fixed trees, and the typical values selected are Sqrt (nVariable) or Log (nVariable).

In this experiment, both feature extraction approaches were used to extract and quantify the desired features, from which an experimental random forest is generated. Table 1 presents the results of algorithm evaluation in these three different modes and based on both feature extraction approaches. Figure 4 plots the results on comparing different values of the m parameter based on the OOB error.

Table 1: Evaluation of a random forest model using different values of the m parameter

Parameter m	Feature space	OOB error	Accuracy	Recall	Precision	F-score
nVariable	1-gram	9.6%	0.89	0.91	0.87	0.88
	2-gram	13.3%	0.87	0.89	0.85	0.86
	3-gram	17.16%	0.81	0.85	0.82	0.83
	MRC Fea.	20.63%	0.77	0.8	0.76	0.77
Sqrt(nVariable)	1-gram	12.1%	0.85	0.89	0.83	0.85
	2-gram	15%	0.82	0.87	0.82	0.85
	3-gram	20.6%	0.73	0.78	0.79	0.78
	MRC Fea.	26%	0.67	0.76	0.74	0.74
Log(nVariable)	1-gram	12%	0.85	0.9	0.81	0.85
	2-gram	14.3%	0.83	0.87	0.81	0.83
	3-gram	19.1%	0.74	0.77	0.79	0.77
	MRC Fea.	24%	0.69	0.79	0.75	0.76

It is clear from the results that reducing the number of properties to sqrt (nVariable) and log (nVariable) values would not lead to promising results, particularly considering that not only the OOB error has increased, but also evaluation criteria have decreased.

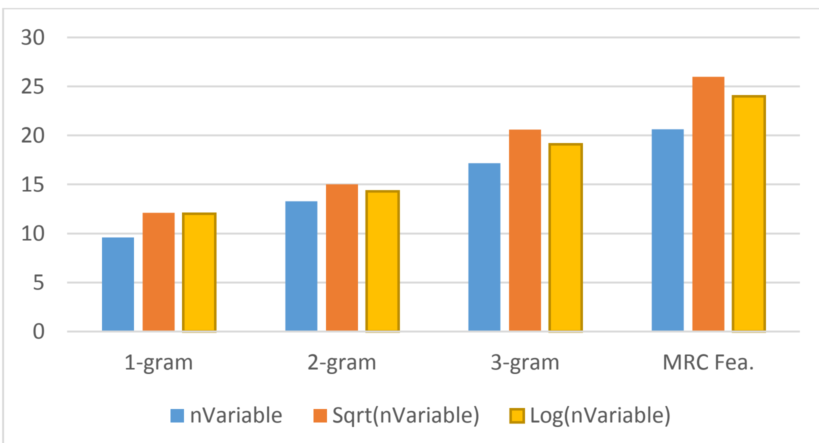


Figure 4: OOB error for different values of the m parameter

2.3. Splitting method

In this experiment, a random forest algorithm was implemented with the following conditions.

1. Number of trees: 150

2. nVariable: Given the previous experiment and the fact that reducing the value of m may have an adverse effect on the performance of the algorithm, its value was fixed as the number of variables. i.e., the features. In this experiment, as in the previous experiment, the desired features are extracted and quantified using the two aforementioned approaches, following which a training forest model is generated. That is,

this experiment is performed for each feature extraction approach. The size of the feature space in both approaches, as well as for different values of n, will be the same as in the previous experiment.

3. Split selection criteria: methods of gain ratio, information gain and Gini index are employed to compare between various splitting methods.

In Table 2, the results for the aforementioned splitting methods using both feature extraction approaches are shown. In Figure 5, the classification criteria are compared based on the OOB error.

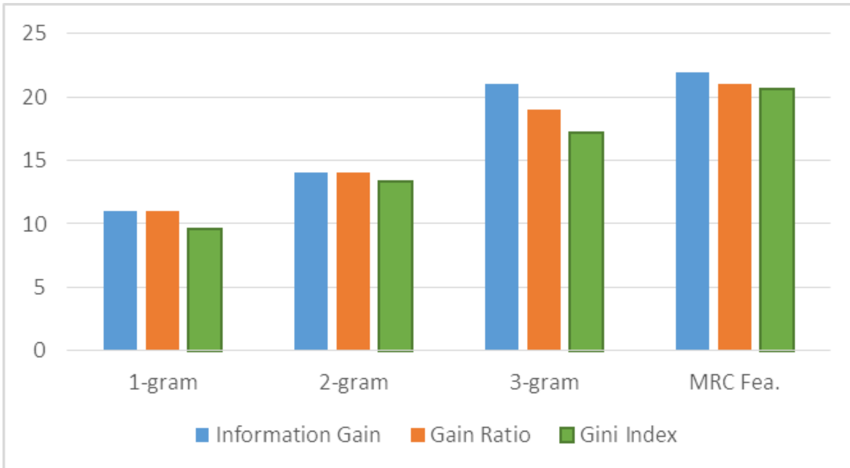


Figure 5: OOB errors for different splitting methods

Table 2: Evaluation of random forest method based on different parameters for splitting method

Splitting method	Feature space	Error	Accuracy	Recall	Precision	F-score
Information Gain	1-gram	11%	0.87	0.89	0.87	0.87
	2-gram	14%	0.84	0.85	0.82	0.83
	3-gram	21%	0.79	0.8	0.82	0.8
	MRC Fea.	22%	0.77	0.77	0.77	0.77
Gain Ratio	1-gram	11%	0.87	0.89	0.87	0.87
	2-gram	14%	0.84	0.85	0.82	0.83
	3-gram	19%	0.78	0.84	0.81	0.82
	MRC Fea.	21%	0.75	0.77	0.77	0.77
Gini Index	1-gram	9.6%	0.89	0.91	0.87	0.88
	2-gram	13.3%	0.87	0.89	0.85	0.86
	3-gram	17.16%	0.81	0.85	0.82	0.83
	MRC Fea.	20.63%	0.77	0.8	0.76	0.77

The results indicate the Gini impurity index outperforms other splitting methods. After determining the best values for the parameters of the random forest, this method was evaluated with other methods. For this purpose, the proposed method was evaluated with single algorithms, ensembles, as well as the most optimal methods proposed for data mining. First, to evaluate the proposed method with single learning methods, support vector machines, Naïve Bayes (NB) methods, and decision trees methods, all of which are among the more widely used data mining methods, were utilized.

Furthermore, there is a particular difference between this experiment and the previous ones, as the experiment is implemented only using the statistical approach of feature extraction with n = 1. The logic behind this decision is that, in the previous experiments, features based on 1-grams outperformed those based on the 2-gram and 3-gram, and it also outperformed the second approach, that is, the semantic approach. The results of this evaluation are shown in Table 3.

Table 3: Evaluation of the proposed method using single learning methods

Method	Accuracy	Recall	Precision	F-score	Time required for generating the model (s)
The proposed model	0.89	0.91	0.87	0.88	1519
SVMs	0.66	0.7	0.68	0.68	14
NB	0.72	0.74	0.69	0.71	5

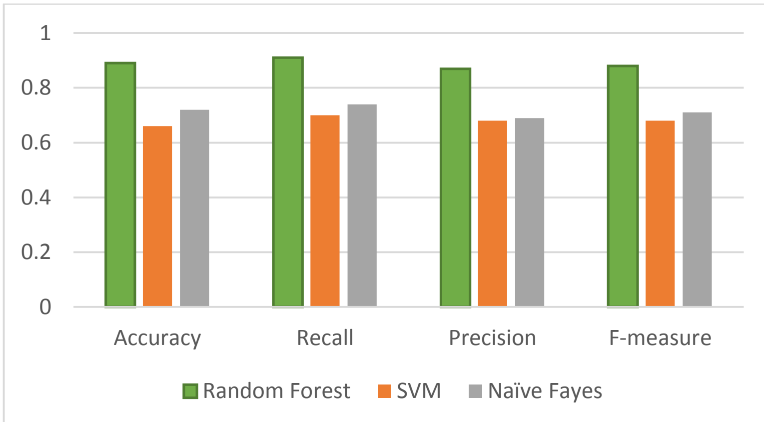


Figure 6: Comparison of random forest algorithms using single learning algorithms

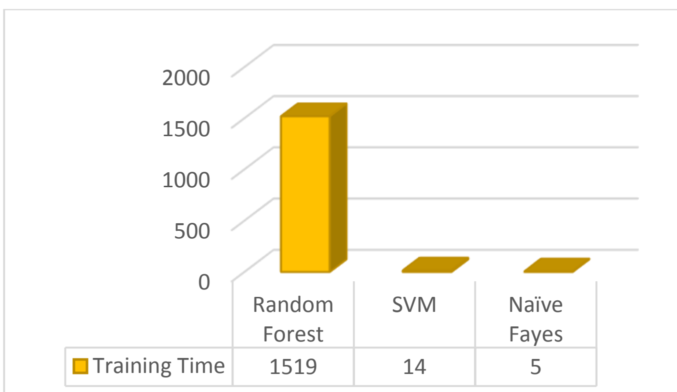


Figure 7: Comparison of random forest algorithms using single learning algorithms in terms of model construction time (training)

Table 3 indicates that the ensemble random foresting algorithm greatly outperforms single algorithms, because in a feature space with very high dimensions, single models are less capable in learning the whole problem space.

Also, to compare the random forest algorithm with other ensemble methods, a set of NB and SVM models were trained using the bootstrap aggregation (bagging method), the results of which are presented in Table 4. Figure 8 present the results obtained on comparison of random forest algorithms using ensemble learning algorithms. while Figure 9 shows the time required to generate training models.

Table 4: Evaluation of the proposed method using ensemble learning methods

Method	Accuracy	Recall	Precision	F-score	Time required for generating the model (s)
The proposed model	0.89	0.91	0.87	0.88	1519
SVMs	0.73	0.81	0.8	0.8	1645
NB	0.78	0.85	0.79	0.81	336

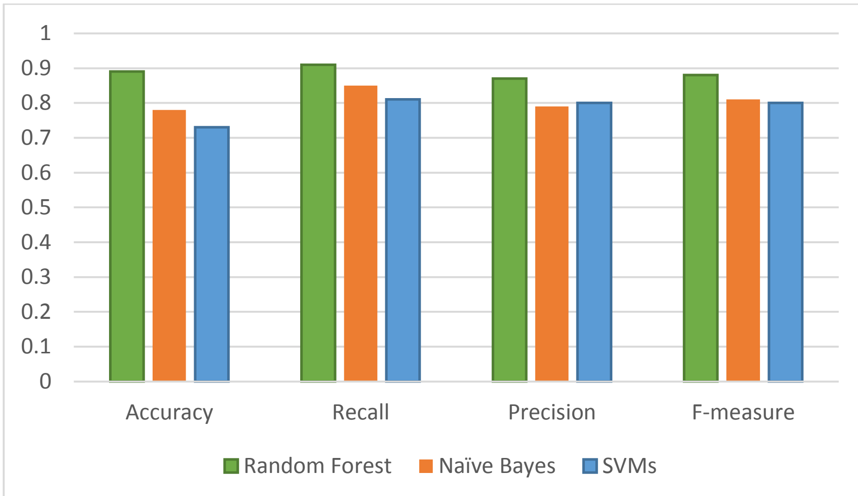


Figure 8: Comparison of random forest algorithms using ensemble learning algorithms

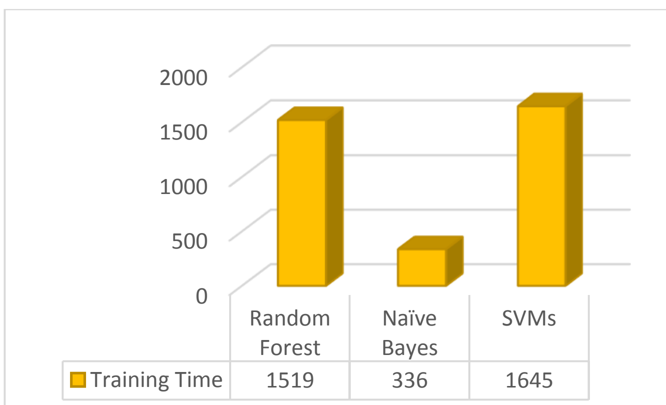
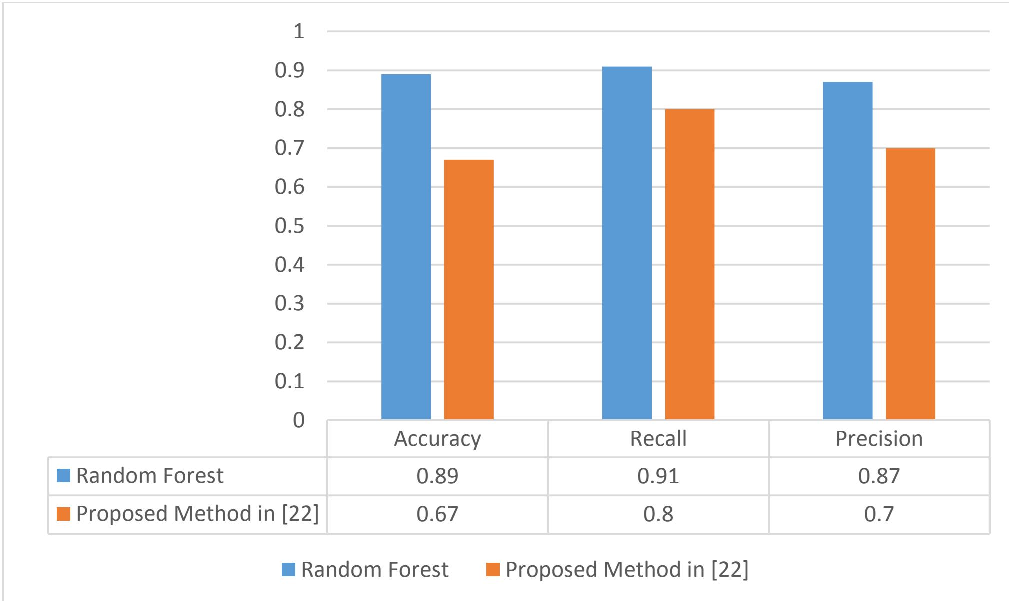


Figure 9: Comparison of random forest algorithms using ensemble learning algorithms in terms of model construction time (training)

Findings from Table 4 and Figure 8 indicate that the proposed algorithm outperforms other ensemble methods. Therefore, the optimal selection of methods constituting an ensemble method is of paramount importance to the efficiency of the learning model.

Moreover, Figure 9 reveals that ensemble methods made up of SVMs required more time for training than the other two methods.

In [26], a method for analyzing the opinions of users in virtual stores is presented, in which different aspects of a product are extracted to assign a sentiment, which is positive, negative, or neutral. The results presented in this study are compared with the proposed method of this research based on different performance criteria in Figure 10. In this evaluation, a random forest algorithm with 150 trees, m for the total number of features, and Gini index for splitting method is trained.



	Accuracy	Recall	Precision
Random Forest	0.89	0.91	0.87
Proposed Method in [22]	0.67	0.8	0.7

Figure 10: Comparison of random forest algorithm with the algorithm presented in [22]

The results from Figure 10 reveal that the proposed method outperforms other methods in all three evaluation criteria. It is noteworthy, however, that the method presented in [26] is lexicon-based method. The main drawback of this method is the lack of lexical resources in the field under study, effectively rendering it less efficient than machine learning-based approaches.

Conclusion

The purpose of this study was to analyze the reviews of virtual store users to optimize services using TF-IDF and MRC criteria. Moreover, this study sought to examine the semantic and lexical load of opinions and sentiments to complement its contribution to this perceived gap in literature. In this regard, a model with four evaluation criteria including accuracy, precision, recall and F-score was proposed and evaluated. The results of this study show that the proposed model has been able to exhibit improvements over the aforementioned criteria compared to other models. Given that this study was performed on the Amazon website, the proposes model has shown the capacity to improve the classification performance of the opinions, or “Reviews” of the consumers, hence providing them with improved efficiency. Evidence suggests that, owing to the volume of unstructured information in the sphere of user interactions and product purchases, previous models have been highly incapable of extracting valuable information in an efficient manner to meet all aspects required by the customer. Yet, the present study has been able to improve the accuracy by 16% compared to the Support vector machines and 11% by the Bayesian method. Also, regarding the recall criteria, it exhibited respective improvements of 7% and 8% compared to the aforementioned methods. For precision, it offered a 7% improvement over the SVM method and an 8% improvement over the Bayesian method. Last but not least, the proposed method offered an improvement of 8% over the SVM method and 7% over the Bayesian method, all of which indicate that the proposed method has been proficient in classifying user comments.

The results further revealed that the analysis of users' opinions and feedbacks can be highly effective in optimizing the services offered, thus enhancing customer satisfaction by offering products tailored to their needs and tastes. The semantic load of terms and lexicons were used to adopt efficient decisions regarding the users' opinions, much to the improvement of all four evaluation criteria. The TF-IDF and MRC were combined to propose an efficient method for analyzing users' opinions and sentiments, which shows a precision of 87%, accuracy 89% of classification, 88% F rating, and 91% readability. Using a combination of the semantic load of words and a random forest to analyze users' opinions increased the accuracy of the classification to 87%, which was a significant improvement over other methods.

According to the findings of the study, online shops are advised to develop efficient methods for analyzing users' opinions and sentiment to compensate for the deficiencies caused by the absence of face-to-face interaction for individuals, the ultimate purpose of which would be to satisfy customers. In addition, the store can directly receive the positive and negative feedback by developing a dedicated website with attractive design, a variety of products, optimized after-sales service, as well as creating forums on the site, improving the quality of customer service in the aftermath.

References

Bertola, F. and V. Patti, Ontology-based affective models to organize artworks in the social semantic web. Information Processing & Management, 2016. 52(1): p. 139-162.

Chaovalit, P. and L. Zhou. Movie review mining: A comparison between supervised and unsupervised classification approaches. in Proceedings of the 38th annual Hawaii international conference on system sciences. 2015. IEEE.

Chen, L. and L. Qi, Social opinion mining for supporting buyers' complex decision making: exploratory user study and algorithm comparison. Social Network Analysis and Mining, 2017. 1)4 :(p. 301-320.

Chen, Z. B.Liu, Mining topics in documents : standing on the shoulders of big data, in: Proceedings of the 20thACMSIGKDDInternational Conference on Knowledge Discovery and Data Mining,ACM ,pp.1116–1125, 2014.

Jorge A. Balazs, Juan D., "Opinion Mining and Information Fusion: A Survey", Preprint submitted to Information Fusion June 4, 2015.

Kang, J. Ahn, K. Lee, Opinion mining using ensemble text hidden Markov models for text classification, Expert Systems With Applications 94, 218–227, 2018.

Kantardzic, M., Data mining :concepts, models, methods, and algorithms. 2011: John Wiley & Sons.

Keshwani, K., P. Agarwal, and D. Kumar, Prediction of Market Movement of Gold, Silver and Crude Oil Using Sentiment Analysis, in Advances in Computer and Computational Sciences. 2018, Springer. p. 101-109.

Kouloumpis, E., T. Wilson, andJ.D. Moore, Twitter sentiment analysis: The good the bad and the omg! Icwsm, 2017. 11: p. 538-541.

Lasota, T., T. Łuczak, and B. Trawiński. Investigation of rotation forest method applied to property price prediction. in Artificial Intelligence and Soft Computing. 2012. Springer.

Nakov P., S. Rosenthal,. "Developing a Successful SemEval Task in Sentiment Analysis of Twitter and Other Social Media Texts", data mining, 2016.

Shahidi, S., Ahmadi Abkenari, F., Mir Roshandel, S. (2017). Pattern recognition in opinion mining Using lexicon-based methods and opinion polarity, *3rd International Conference on Pattern Recognition and Image Analysis in Iran*. Shahrekord University - Iranian Association of Machine Vision and Image Processing.

Vasavi, S., Extracting Hidden Patterns Within Road Accident Data Using Machine Learning Techniques, in Information and Communication Technology. 2018, Springer. p. 13-22.

Zhiqiang, T. W.Wenting, DLIREC: Aspect term extraction and term polarity classification system, in: Proceedingsofthe8thInternational WorkshoponSemanticEvaluation,2014, pp.235–240, 2014.