Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence
An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

Classification of Customer Reviews Using Machine Learning Algorithms

Behrooz Noori

To cite this article: Behrooz Noori (2021) Classification of Customer Reviews Using Machine Learning Algorithms, Applied Artificial Intelligence, 35:8, 567-588, DOI: 10.1080/08839514.2021.1922843

To link to this article: https://doi.org/10.1080/08839514.2021.1922843

Published online: 06 May 2021.

APPLIED ARTIFICIAL INTELLIGENCE
2021, VOL. 35, NO. 8, 567–588
https://doi.org/10.1080/08839514.2021.1922843

Taylor & Francis
Taylor & Francis Group

Classification of Customer Reviews Using Machine Learning Algorithms

Behrooz Noori

Department of Industrial Engineering, West Tehran Branch, Islamic Azad University, Tehran, Iran

ABSTRACT

The information resulting from the use of the organization's products and services is a valuable resource for business analytics. Therefore, it is necessary to have systems to analyze customer reviews. This article is about categorizing and predicting customer sentiments. In this article, a new framework for categorizing and predicting customer sentiments was proposed. The customer reviews were collected from an international hotel. In the next step, the customer reviews processed, and then entered into various machine learning algorithms. The algorithms used in this paper were support vector machine (SVM), artificial neural network (ANN), naive bayes (NB), decision tree (DT), C4.5 and k-nearest neighbor (K-NN). Among these algorithms, the DT provided better results. In addition, the most important factors influencing the great customer experience were extracted with the help of the DT. Finally, very interesting results were observed in terms of the effect of the number of features on the performance of machine learning algorithms.

Introduction

Customer reviews have been commonly recognized as valuable sources for marketing intelligence and sentiment analysis (Dickinger and Mazanec 2015). Sentiment analysis seeks to build a system for analyzing and evaluating customer reviews reflected on websites, blogs, Twitter, or Instagram. In recent years, with the expansion of online systems, customer reviews have a powerful impact on business development and attracting potential customers. Therefore, review categorization becomes the key technology to organize textual data. Review categorization is defined as assigning new documents to a set of pre-defined categories based on the classification patterns (Uğuz 2011). In fact, customer sentiment is very important for the hospitality industry and plays an important role in providing better quality services (e.g., more adaptation to customer requirements and customization of services), better customer relationship

CONTACT Behrooz Noori noori.b@wtiau.ac.ir, bnoori@gmail.com Department of Industrial Engineering, West Tehran Branch, Islamic Azad University, Hassan Azari Avenue, Ponak Square, Ashrafi Esfehani Highway, Tehran, Iran 1468763785

and customer preference management. Other benefits of sentiment analysis include shaping company marketing strategies, effective marketing campaigns, and anticipating customer satisfaction.

Accordingly, a great deal of information is currently available in hotels through in-house and structured systems, such as CRM system, or external and non-structured systems, such as social networks and websites. With the growth of the availability of unstructured data through websites and social networks, managing this huge information and discovering unknown patterns in a large dataset is absolutely essential. An analysis of this volume of data requires the organization to use powerful tools in Big Data technologies which have been used well in other areas such as health care, or finance (Talón-Ballestero et al. 2018). As a result, review mining and sentiment analysis can be developed as an important tool in this connection.

Machine learning techniques are often used to analyze and predict sentiments. The sentiment analysis is done at three levels: the level of the document, the sentence level, and the feature level. The analysis of sentiments at the document level examines whether the document is positive, negative or neutral (Tripathy, Agrawal, and Rath 2016). In this study, sentiment analysis at the document level and feature level has been considered.

Machine learning methods categorize reviews. In this paper, six methods of machine learning were used to classify sentiments. The accuracy of these methods has been studied to evaluate their performance. In addition, the voice of customer has been analyzed using feature sentiment analysis.

This article is organized as follows: The next section examines the research related to this work. Section 3 describes the model created for doing the analysis. Section 4 presents the results, and Section 5 provides the final result.

Related Works

Sentiment Analytics

Online user-generated content in various social media and websites, such as consumer experiences, user feedback, and product reviews, has increasingly become the primary information source for both consumers and businesses (Duan et al. 2016). Basically, customer reviews demonstrate customer experience in relation to the organization, which is very important in understanding customer thoughts. These reviews have a major impact on other customers' decisions and are the basis for business improvement. The number of reviews has increased over the past few years, and attention to hidden features in these reviews will definitely increase the performance of the hotels. In other words, while customers use these reviews in their decision-making process, companies use this information to grow products.

On the other word, due to the increasing growth of social networks and websites, the obtained reviews are useful resources for analyzing and improving business and developing products and services. Often these reviews are not structured. With the help of machine learning methods, managers will be provided with information for future use.

The most important task of sentiment analytics is analyzing the polarity of sentiments and classifying texts into positive and negative emotional categories. To successfully discover, interpret, and communicate extracted opinions and detected polarity, sentiment analytics relies on multidisciplinary efforts including natural language processing and machine learning (Fu et al. 2018).

Existing approaches to sentiment analytics can be classified into two broad categories: semantic orientation approaches and machine learning approaches (Fu et al. 2018). Semantic orientation approaches hold that text is classified into affect categories on the basis of the presence of fairly unambiguous affect words, such as "happy," "sad," "afraid," and "bored." Semantic orientation approaches are popular thanks to their accessibility and economy. However, the weaknesses of these approaches include poor affect recognition given complex linguistic rules, and heavy dependence on the depth and breadth of the employed lexicon resources (Fu et al. 2018). For a domain lacking of such resources, machine learning approaches can mitigate the above limitations. By feeding a machine learning algorithm a training corpus of affectively annotated texts, machine learning approaches can not only learn the affective polarity of affect keywords but can also consider the polarity of other arbitrary keywords and word co-occurrence frequencies (Cambria 2017). However, machine learning approaches rely on statistical models that are meaningful when given a sufficiently large text input; therefore, the approaches can achieve better performance on the document or paragraph level compared to smaller text units, such as sentences or clauses (Fu et al. 2018).

With the proliferation of big data, there is an increasing trend for hotel industry researchers to adopt computational methods in their studies (Fu et al. 2018). In particular, sentiment analytics, which works as an effective method to automatically extract public opinions and analyze sentiment polarity from massive textual data (Cambria 2016), has thus piqued researchers' interest (Fu et al. 2018).

A large amount of studies by different authors were conducted where machine learning procedures were undertaken on hotel industry and hospitality data (Moro, Rita, and Coelho 2017). Ye, Zhang, and Law (2009) studied sentiment classification methods in online reviews from travel blogs, comparing them with three machine learning techniques (Ye, Zhang, and Law 2009). Cao, Duan, and Gan (2011) studied the impact of online review features hidden in the reviews on the number of helpful votes by applying text mining for extracting the review's characteristics. In order to meet the requirement of

customized services, sentiment classification of online reviews has been applied to study the unstructured reviews so as to identify users' opinions on certain products (Cao, Duan, and Gan 2011). Wang et al. (2013) focused on selecting text features more effectively and efficiently, in order to improve sentiment classification of Chinese online reviews (Wang et al. 2013). Schuckert et al. (2015) reviewed and analyzed articles related to online reviews in hotel industry and hospitality area. Hu and Chen (2016) studied hotel review helpfulness (Hu and Chen 2016). Sánchez-Franco, Navarro-García, and Rondán-Cataluña (2019) described a supervised classification approach to identify the most relevant terms and their influence on hotel ratings. They applied the NB approach to solve the problem of dealing with huge numbers of service reviews (Sánchez-Franco, Navarro-García, and Rondán-Cataluña 2019). Ma, Cheng, and Hsiao (2018) provided a critical review of the origin, development and process of sentiment analysis and a demonstration for hospitality researchers and students on how to perform sentiment analysis using a sample study (Ma, Cheng, and Hsiao 2018). Al-Smadi et al. (2018) implemented and compared neural network and SVM for aspect-based sentiment analysis of Arabic hotels' reviews (Al-Smadi et al. 2018). Tran, Ba, and Huynh (2019) proposed a framework to summarize the customer's reviews by latent dirichlet allocation (LDA) model. They focused on aspect-based sentiment analysis (Tran, Ba, and Huynh 2019). Furthermore, Bi et al. (2019) proposed a method for modeling customer satisfaction from online reviews (Bi et al. 2019).

So far, however, there has been little discussion about the predictive accuracy of sentiment analytics, and hence issues regarding how to properly select machine learning method are seldom addressed in the hotel industry literature (Fu et al. 2018). In addition, there are scare studies about feature sentiment extraction. Therefore, in this study, we will investigate them.

Classification

Classification is one of the most commonly used methods in machine learning. It is a process of finding a set of models that allows data classes to be identified and distinguished. The aim of classification is to determine the class of future data objects by using past information. In classification, a training set is usually used to learn the model, and the learned information is then tested on the test set. Many classification algorithms have been developed in the literature so far since there is no perfect algorithm for all data sets (Gulsoy and Kulluk 2019).

Proposed Sentiment Classification Framework

It is a challenging task to study the predictive accuracy of sentiment analytics for hotel industry. Such task is more methodological (e.g., choosing design

factors) than technical (e.g., improving a new classification algorithm) (Fu et al. 2018). Hence, we need an integrative effort to examine how different methods in sentiment analytics influence predictive accuracy, and how to ensure predictive accuracy of semantic analytics via a systematic approach. Specifically, this article attempts to address the following research questions:

(1) What are the key steps of sentiment analytics for hotel industry?
(2) What are the key design factors of feature engineering?
(3) How do these design factors influence the predictive accuracy of sentiment analytics for hotel industry?
(4) How can machine learning methods be systematically incorporated to improve the predictive accuracy of sentiment analytics?

The parts of proposed system structure are shown in Figure 1. These parts are explained in the following subsections:

Data Collection

The review data were collected from the TripAdvisor.com. TripAdvisor.com is one of the most famous and largest travel website. A corpus or data collection can be defined as a set of text documents that can be classified under many subsets (Hu and Chen 2016). The corpus contains 400 documents of different lengths. In this data collection, each document was saved in a separate database.

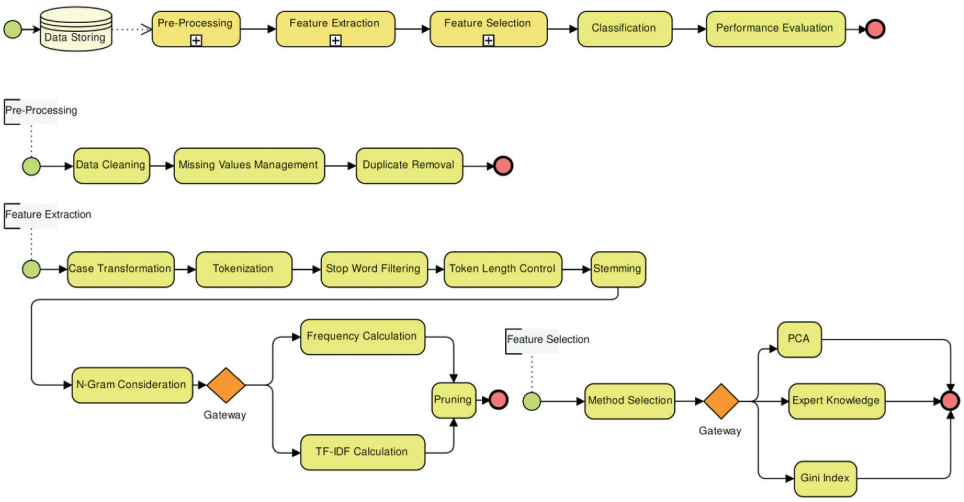

Figure 1. System structure.

Pre-processing and Feature Extraction

Text pre-processing is an important step in the text classification process. This step can reduce the errors and enhance the accuracy of classification (Bahassine et al. 2018; Uysal and Gunal 2014). The main objective of this endeavor is to get rid of noisy and nonmeaningful words (Bahassine et al. 2018).

Each review was subject to the following procedure:

Data Cleaning

- duplicate removal, delete digits, punctuation marks, and numbers.

- delete stop-words and non-useful words like: pronouns, articles and propositions. (Bahassine et al. 2018)

Transformation and Tokenization

Tokenization is the process of splitting reviews into pieces called tokens.

Removing of Stop-words

Words such as conjunctions and pronouns that are not related to the concept of the text are called stop-words. This process involves removing certain common words such as ‘a’, ‘an’, the’, etc., that occur commonly in all documents. It is important to removing these high-frequency words because they may misclassify the documents (Uğuz 2011).

Stemming

Stemming is a process of reducing inflected words into one form (stem or root) by removing prefixes, suffixes and infixes (Bahassine et al. 2018). The stemming process leaves out the root forms of the words. Thereby, terms sharing the same root that seem like different words due to their affixes can be determined. For example, “computer”, “computing”, “computation”, and “computes” all have the same comput root (Uğuz 2011).

Term Weighting

After the words are transformed into terms, the presentation form of the document, which means the expression thereof, terms have to be determined. This process is called term weighting. Thereby, each document could be written in a vector form depending on the terms they contained. To obtain the weight vector, term frequency–inverse document frequency (TF-IDF) feature weighting algorithm is used as its weight scheme. N-gram refers to a sequence of n tokens based on words.

Pruning of the Words

The pruning process basically filters less frequent features in a document collection. The term vector is very high-dimensional and sparse. Also, it is

seen that a number of elements in the term vector is "0". Therefore, we prune the words that appear less than two times in the documents. This process decreases the term vector dimension further (Uğuz 2011).

Feature Selection

Feature selection is a process that selects a subset from the original feature set according to some criteria of feature importance (Uğuz 2011). A major problem of sentiment categorization is the high dimensionality of the feature space due to a large number of terms. This problem may cause the computational complexity of machine learning methods used for sentiment categorization to be increased and may bring about inefficiency and results of low accuracy due to irrelevant terms in the feature space. For a solution to this problem, two techniques are used in this study: feature ranking and feature selection (Uğuz 2011).

Sentiment mining has become a heated research in recent years. One of the important means of sentiment mining is sentiment categorization. For many problems of sentiment categorization, a good feature selection method can not only reduce the computational complexity but also increase the categorization performance. Feature selection is a process that selects a set of new features from the original features and forms a distinct feature space. Apart from this, feature selection is also perceived as a prerequisite for text categorization, so its significance and importance can be imagined (Wang et al. 2015).

On the other hand, feature extraction produces a large feature set and creates a high-dimensional vector space, which will ultimately lower the efficiency and the effectiveness of sentiment classification. As a result, it is critical to select features with significant sentiment distinguishing ability and reduce the dimension of vector space (Wang et al. 2013). Features selection is effective in the reduction of large data in text classification. It can enhance the classification process. Feature selection deletes irrelevant and noisy data and chooses a representative subset of all data to minimize the complexity of the classification process (Dadgar, Araghi, and Farahani 2016).

Numerous techniques of features selection can be detected in the literature such as: Chi-square (Bahassine et al. 2018) and Gini index (Manek et al. 2017). The present research tried to introduce a modified version of Gini feature selection method which will be presented hereafter (Bahassine et al. 2018). A Gini Index-based feature selection method solves the problem mentioned above. The experiments showed that the weight by Gini Index method has better classification performance (Manek et al. 2017). At the end of the feature selection step, terms of high importance in documents are acquired through the Gini method.

In the current study, feature selection and feature sentiment extraction are used to manage the high dimensionality of a feature space composed of a large

number of terms, remove redundant and irrelevant features from the feature space and thereby decrease the computational complexity of the machine learning algorithms used in the text categorization and increase performances thereof (Uğuz 2011).

In the first stage, each term in the text is ranked depending on their importance for the classification in decreasing order using the Gini method. Therefore, terms of high importance are assigned to the first ranks and terms of less importance are assigned to the following ranks. In the second stage, the PCA method selected for feature sentiment extraction.

Principal Component Analysis

Principal component analysis (PCA) is a popular method for reducing the dimension of data while preserving most of their variations. In a few words, PCA consists in extracting the main modes of variation of the data around their mean via the computation of new synthetic variables named main components (Cardot and Degras 2018).

It has previously been observed that the PCA showed promising results in the feature selection process. The PCA is based on the following steps: 1) Convert training and test datasets into numerical form; 2) Find covariance matrix of datasets; 3) Calculate Eigen values and Eileen vector of the covariance matrix; 4) Sort Eigen non-increasing Eigen values; 5) Keep the top k vectors; and 6) Train, test and evaluate the reduced datasets (Zainuddin, Selamat, and Ibrahim 2018).

Classification Methods

In this study, six separate machine learning methods are used in text categorization; the SVM, ANN, NB, DT, C4.5 and kNN methods are used due to their usability and accuracy in text categorization. The reason for using a classifier is to compare the performances of the six methods in the text categorization. Brief descriptions of these methods are given, as follows.

SVM

SVM is a supervised learning method used for categorization. It is a useful methodology that finds the best possible surface to separate the positive samples from the negative samples. The basic goal of SVM, behind the training process, is to find a maximum margin hyperplane to solve the feature review's classification task. There are unlimited possible boundaries to separate the two classes. To select the best class, it is important to choose a decision boundary that has a maximum margin between any points from both classes. The decision boundary with a maximum margin would be less likely to make prediction errors, which is close to the boundaries of one of the classes (Ali, Kwak, and Kim 2016). The dot kernel type was selected because of its better

performance against radial and polynomial kernel types. The C parameter set 2.

Artificial Neural Networks

An artificial neural network (ANN), usually called neural network (NN), is a mathematical model or computational model that is inspired by the structure and functional aspects of biological neural networks. A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation (the central connectionist principle is that mental phenomena can be described by interconnected networks of simple and often uniform units). In most cases, an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Modern neural networks are usually used to model complex relationships between inputs and outputs or to find patterns in data. A feed-forward neural network is an artificial neural network where connections between the units do not form a directed cycle. In this network, the information moves in only one direction, forward, from the input nodes, through the hidden nodes (if any) to the output nodes. There are no cycles or loops in the network.

Back propagation algorithm is a supervised learning method which can be divided into two phases: propagation and weight update. The two phases are repeated until the performance of the network is good enough. In back propagation algorithms, the output values are compared with the correct answer to compute the value of some predefined error-function. By various techniques, the error is then fed back through the network. Using this information, the algorithm adjusts the weights of each connection in order to reduce the value of the error function by some small amount. After repeating this process for a sufficiently large number of training cycles, the network will usually converge to some state where the error of the calculations is small. In this case, one would say that the network has learned a certain target function.

One of the most distinctive techniques is the use of ANN in sentimental analysis. The increasing growth of ANN as a machine learning tool is mainly due to the development of hardware and learning algorithms that enable the implementation of networks with different layers called deep learning. Regardless of the technique used, the key issue is that there is a data set that can be used to feed machine learning algorithms. This article used a feed-forward neural network trained by a back propagation algorithm (multi-layer perceptron).

Naive Bayes

NB Tree is a supervised classifier that combines Bayesian rule and decision tree. This algorithm uses the Bayes rule to calculate the likelihood of each given class of instances, assuming that the properties are the conditional

independent of the label given (Gulsoy and Kulluk 2019). A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. In simple terms, a Naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class (i.e. attribute) is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 4 inches in diameter. Even if these features depend on each other or upon the existence of the other features, a Naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple.

The advantage of the Naive Bayes classifier is that it only requires a small amount of training data to estimate the means and variances of the variables necessary for classification. Because independent variables are assumed, only the variances of the variables for each label need to be determined and not the entire covariance matrix.

Decision Trees

A decision tree is a tree-like graph or model. It is more like an inverted tree because it has its root at the top and it grows downwards. This representation of the data has the advantage compared with other approaches of being meaningful and easy to interpret. Each interior node of tree corresponds to one of the input attributes. The number of edges of a nominal interior node is equal to the number of possible values of the corresponding input attribute. Outgoing edges of numerical attributes are labeled with disjoint ranges. Each leaf node represents a value of the *label* attribute given the values of the input attributes represented by the path from the root to the leaf. Decision trees in the Simple Cart algorithm are constructed by dividing each decision node into two different branches using various separation criteria (Gulsoy and Kulluk 2019).

C4.5 Decision Tree Classifier

The decision tree is a well-known machine learning approach to automate the induction of classification trees based on training data. In a typical decision tree training algorithm, there are usually two phases. The first phase is tree growing where a tree is built by greedily splitting each tree node. Because the tree can overfit the training data, in the second phase, the overfitted branches of the tree are removed. C4.5 is a univariate decision tree algorithm. At each node, only one attribute of the instances are used for decision-making (Uğuz 2011). J48 algorithm is the name of C4.5 algorithm in Weka data mining software and it is one of the best-known decision tree-based algorithms. The algorithm reaches a decision result from the nodes formed by dividing the data over the attribute with the highest information gain (Gulsoy and Kulluk 2019).

In our application, by using C4.5 decision tree algorithm, in the pruning phase, the post-pruning method is used to decide when to stop expanding a decision tree (Uğuz 2011). The confidence factor is used for pruning the tree. In our study, the confidence factor is assigned as 0.25. The pruned trees consist of 2 instances per leaf.

KNN

The KNN algorithm is a well-known instance-based approach that has been widely applied to text categorization due to its simplicity and accuracy. To categorize an unknown document, the KNN classifier ranks the document's neighbors among the training documents and uses the class labels of the k most similar neighbors. Similarity between two reviews may be measured by the Euclidean distance, cosine measure, etc. The similarity score of each nearest neighbor document to the test document is used as the weight of the classes of the neighbor document. If a specific category is shared by more than one of the k-nearest neighbors, then the sum of the similarity scores of those neighbors is obtained from the weight of that particular shared category (Uğuz 2011).

At the phase when classification is done by means of the KNN, the most important parameter affecting classification is k-nearest neighbor number. Usually, the optimal value of k is empirically determined. In our study, k value is determined so that it would give the better classification accuracy (k = 1 is determined). In addition, in the phase of finding the k-nearest neighborhood, mix Euclidean distance is used as the distance metric (Uğuz 2011).

Evaluation of the Performance

Various performance criteria can be calculated. Among these criteria, accuracy shows the proportion of correctly classified instances within the whole data. It helps assess the overall performance of the classifier. However, the accuracy criterion is not satisfactory for imbalanced datasets. In this situation recall and precision criterions can be used. Recall is the ratio of the number of correctly classified positive instances to the total number of positive instances, and precision is the ratio of the number of positively predicted true positive instances to the total number of instances that are predicted to be positive. F-measure combines precision and recall criteria by taking harmonic averages. While evaluating performance of the model, apart from these criteria, the ROC-area value can also be considered. The ROC curve is a graph drawn using true positive and false positive values. The ROC area value is also expected to be close to 1 (Gulsoy and Kulluk 2019). In this study, the accuracy, precision, recall, F-measure and ROC area are calculated for each feature set and method.

Results

Pre-processing and Feature Extraction

Pre-processing, dimension reduction, and classification processes are implemented by the Rapid Miner Studio, and Weka. A 10-fold cross-validation procedure is preferred for the performance evaluation stage. After data cleaning, the first step consisted of tokenizing and removing the stop words because they are useless for the classification. In the study, stop words are removed. After removing the stopwords, the dataset contains 2640 unique words. In the second step, the WordNet algorithm is used for stemming. We compared its output with Porter algorithm and WordNet was better for our application. In the third step, the document vectors are built with the TF-IDF weighting scheme. The total number of terms finally extracted is 1892. We call it third feature set. Thereby, a document-term matrix is acquired with a dimension of 400*1800 at the end of preprocessing.

Feature ranking was applied via the Gini index method to reduce the high dimension of the feature space. In this phase, the effects of the individual feature ranking operation by the Gini Index method on classifier performance are examined. Accordingly, features are ranked in decreasing order using the Gini index. After feature ranking, we used PCA to find feature sentiment. We selected 100 important features and called in second feature set. Finally, we selected top 25 features and called it first set. Initially, six classifiers were selected and applied on the whole of the document-term feature space.

Performance Comparison

The accuracy performance metric values obtained by machine learning methods on the sentiment data set are shown in Table 1. When Table 1 is examined, it can be seen that the NB is given the maximum accuracy (65.5%) in first dataset. Therefore, NB is very suitable for small feature set. It can be seen that the SVM is given the maximum accuracy (68.6%) in second dataset. Moreover, DT and C4.5 are given the maximum accuracy (98.9%) in third dataset. The

Table 1. Accuracy results of a 10-fold cross validation as a function of the number of selected features. Best results are in boldface.

Classifiers	Number of features		
	25	100	1800
Accuracy			
SVM	62.70%	**68.58%**	77.52%
ANN	63.5%	67.3%	55.1%
NB	**65.5%**	65.4%	70.6%
DT	56.6%	64.2%	**98.9%**
C 4.5	54.8%	63.8%	**98.9%**
K-NN	61.7%	65.9%	72.0%

obtained accuracy is pretty high for real world problems. In the third dataset, SVM algorithm is the second algorithm with 75.1% accuracy, and it is followed by K-NN, and NB, respectively. ANN algorithm is the worst algorithm with 55.1% accuracy on large feature set. When the obtained maximum number of features is taken into consideration, DT and C4.5 algorithms are the best algorithms. Therefore, the DT and C4.5 produced significant results in terms of accuracy and third feature set. The overall accuracies of the algorithms are indicated in Figure 2. The accuracies of the selected algorithms are indicated in Figure 3.

When the performances of the algorithms are evaluated in the first data set in terms of recall performance metric, ANN classifier can be seen to give the best result of 0.633 (Table 2). NB, K-NN, SVM, DT and C4.5 algorithms are following this algorithm, respectively for first feature set. Furthermore, SVM can be seen to give the best result of 0.675. ANN, K-NN, NB, C4.5 and DT algorithms are following this algorithm, respectively for second feature set. Recall values support the results obtained with accuracy values. Figure 4. shows recall behavior. The DT classifier algorithm gave the best value of 0.98 for the recall performance metric in third feature set.

The DT classifier algorithm gave the best value of 0.991 for the precision performance metric. The C4.5 is the second algorithm with 0.985 precision value. SVM, K-NN, NB and ANN algorithms followed, respectively, the C4.5 algorithm according to precision (Table 3, Figure 5).

Additionally, the DT classifier algorithm gave the best value of 0.991 for the F-measure performance metric, which is the harmonic mean of the precision and recall metrics (Table 4). The C4.5 is the second algorithm with 0.985 F-measure value. SVM, K-NN, NB and ANN algorithms followed,

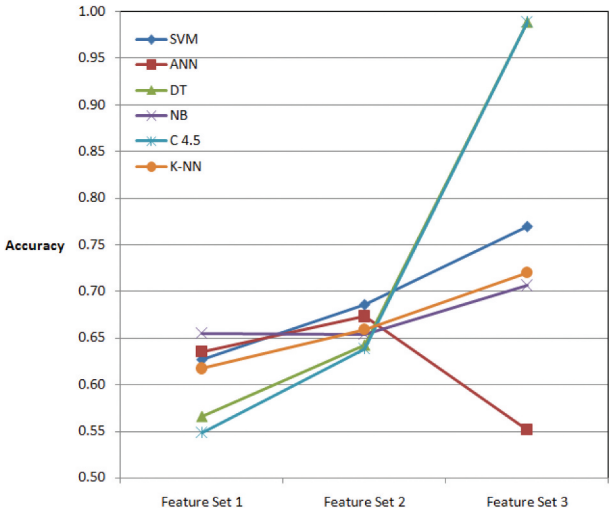

Figure 2. Diagrammatic presentation of accuracies in the experiments.

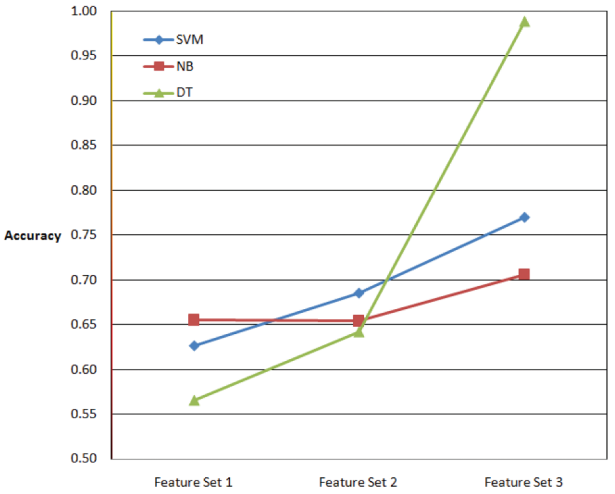

Figure 3. Accuracies behavior of SVM, NB, and DT.

Table 2. Average recall results of a 10-fold cross validation as a function of the number of selected features. Best results are in boldface.

Classifiers	Number of features		
	25	100	1800
Recall			
SVM	61.48%	**67.52%**	77.91%
ANN	**63.3%**	66.3%	50.1%
NB	62.1%	64.2%	70.3%
DT	54.8%	59.5%	**98.9%**
C 4.5	53.7%	62.7%	98.4%
K-NN	61.1%	65.6%	72.7%

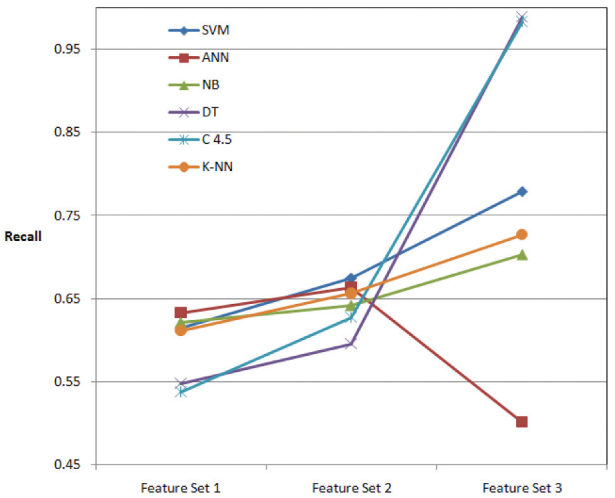

Figure 4. Diagrammatic presentation of recall in the experiments.

Table 3. Average precision results of a 10-fold cross validation as a function of the number of selected features. Best results are in boldface.

Classifiers	Number of features		
	25	100	1800
Precision			
SVM	67.54%	**72.21%**	78.63%
ANN	64.6%	69.6%	52.5%
NB	**71.2%**	68.9%	71.2%
DT	62.3%	68.7%	**99.1%**
C 4.5	62.7%	65.6%	98.5%
K-NN	61.7%	66.1%	75.1%

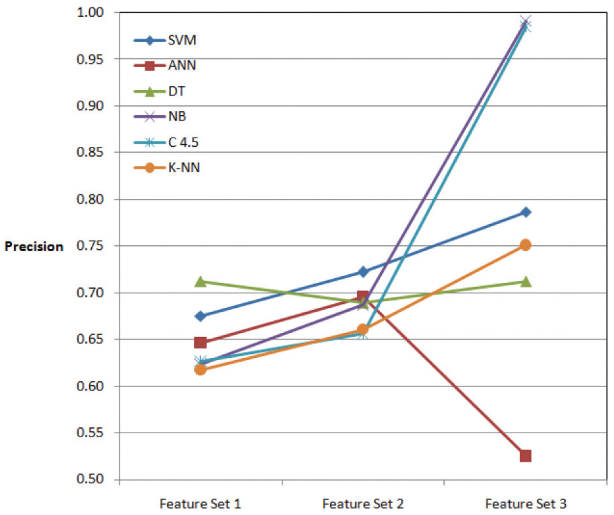

Figure 5. Diagrammatic presentation of precision in the experiments.

Table 4. F-measure results of a 10-fold cross validation as a function of the number of selected features. Best results are in boldface.

Classifiers	Number of features		
	25	100	1800
F-measure			
SVM	64.37%	**69.79%**	78.27%
ANN	63.94%	67.91%	51.27%
NB	**66.34%**	66.47%	70.75%
DT	58.31%	63.77%	**99.00%**
C 4.5	57.85%	64.12%	98.45%
K-NN	61.40%	65.85%	73.88%

respectively, the C4.5 algorithm according to F-measure values (Figure 6). Therefore, ANN is not suitable for large feature set. Figure 6 and Figure 7 summarized the performance of machine learning methods in terms of F-measure.

According to TF-IDF, main factors which generate customer dissatisfaction through DT are shown in Figure 8.

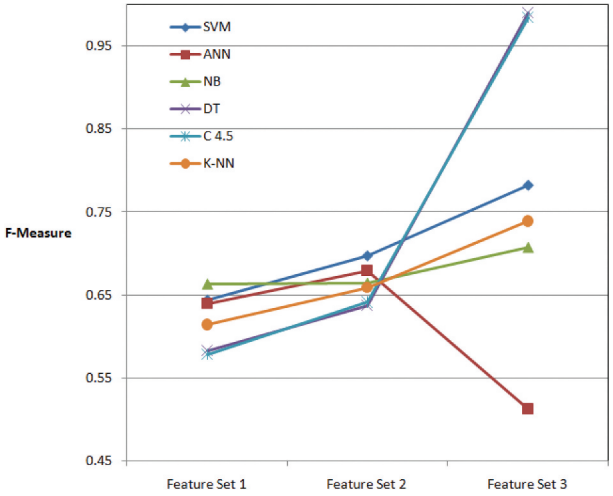

Figure 6. The F-measure of six different methods.

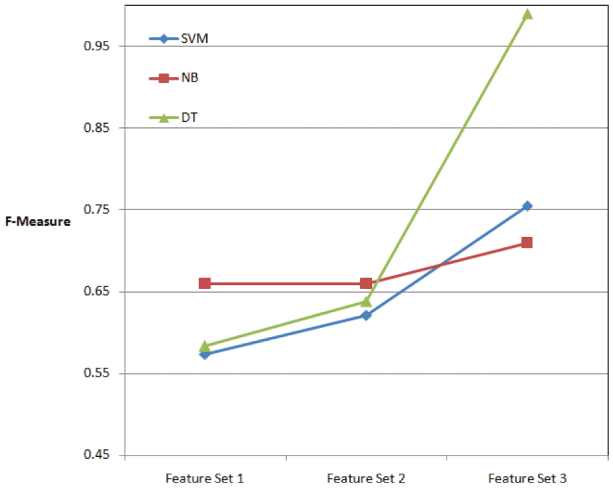

Figure 7. The F-measure of three methods.

Meanwhile in the third feature set, the C4.5 algorithm gave the best result with a value of 0.979 for the ROC area performance measure, which expresses the residual under the ROC curve (Table 5). DT, SVM, NB, K-NN, and ANN algorithms, respectively, followed the C4.5 algorithm.

Accordingly, it is seen that DT algorithm gives more favorable results than the other 5 algorithms when the accuracy, precision, recall, and F-measure metrics are taken into consideration. The C4.5 algorithm yielded more favorable results for the area under the ROC curve only. In summary, both DT and C4.5 algorithms can be used for classification purposes on the large feature set. The DT and C4.5 classifiers may be to maximize the accuracy, to maximize the

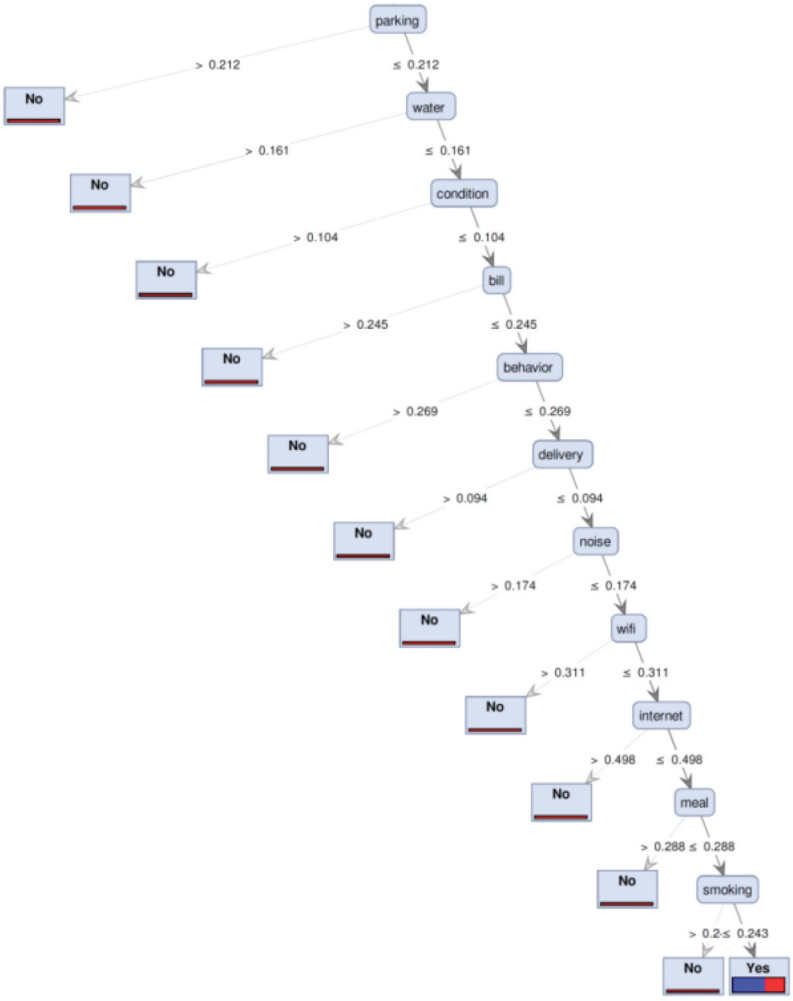

Figure 8. Voice of customers tree.

Table 5. Average results of a 10-fold cross validation as a function of the number of selected features. Best results are in boldface.

Classifiers	Number of features		
	25	100	1800
ROC area			
SVM	0.708	0.733	0.863
ANN	0.683	**0.745**	0.585
NB	**0.703**	0.714	0.675
DT	0.581	0.513	0.942
C 4.5	0.544	0.656	**0.979**
K-NN	0.500	0.500	0.500

area under the ROC-curve, or to minimize the square root of the mean square error. As a result, with the DT, 98.9% accuracy value of the process of

sentiment classification in the hotel sector was obtained. This value is quite high for real-life data.

Yet, the SVM and NB algorithms are preferred in small and medium sized feature sets in this study because it has high accuracy and F-measure as well as being interpretable by the user. The results show that data mining can be used to generate an objective measurement system that can be used for the sentiment classification process of customers in the hotel industry.

Figure 9 shows results as a function of the number of selected terms that resulted in the best accuracy for DT and C4.5, respectively. Considering our results, we observed the following:

(1) Although DT classifier produced the best absolute values of accuracy, recall and precision, results indicated that such behavior results from a good performance on a large feature set. In this context, C4.5 produced better results in ROC area.
(2) Figure 2 shows that all machine learning methods have an increasing trend, except for the ANN that has a decreasing trend (Figures 2 and 6).
(3) The DT and C4.5 methods started from the lowest value of accuracy and reached the highest level of accuracy. In fact, they were most susceptible to the number of features (Figure 2).
(4) The NB had the best performance with small feature set (Figure 2).
(5) NB showed the least sensitivity on feature size variation. In the F-measure criterion, the same behavior was repeated (Figures 2 and 6).
(6) In Figure 9, it should be noted that second feature slightly outperforms first feature set. Moreover, it should be noted that results vary from third feature context in which DT and C4.5 produce quite similar values.

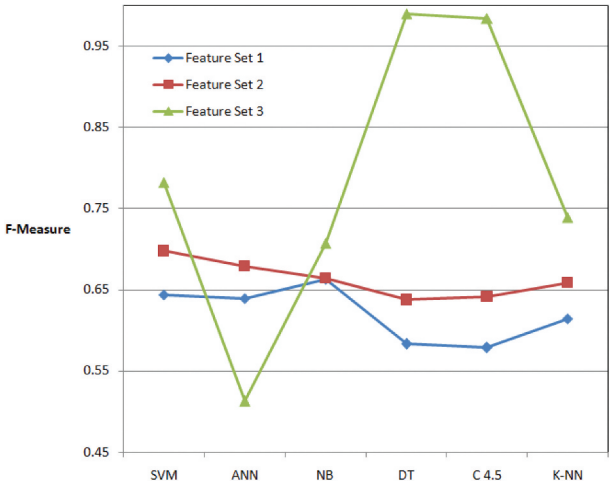

Figure 9. Classification results between different methods and feature sets.

(7) All machine learning methods increase their accuracy as the number of features increases, but ANN behaved differently, and its accuracy decrease as the number of features increases.
(8) DT has a better performance of sentiment classification compared to the other popular classifiers in large feature set and NB and SVM have better performance of sentiment classification compared to the other popular classifiers in smaller feature set.

Conclusion

Exploring and analyzing the invaluable hidden data of customer reviews has become a major prerequisite for effective and successful marketing analytics. The large volume of data and Big Data has become a fundamental feature of today's society, yet its ability to analyze, solidify and learn from them, has become a useful element for competition as well as supporting the growth of productivity and innovation. Although big data can certainly be considered as good for decision, large data does not lead to better marketing because they are related to some of the key challenges and issues (i.e. the lack of feature engineering, need for useful analysis, and so on). Machine learning techniques play an important role in the field of social media analysis (Ducange, Pecori, and Mezzina 2018).

The sentiment classification, which make up a large part of the voice of customer in the hotel sector, has become one of the most important problems of the hospitality industry. This is because the estimation of whether the customers will be none satisfied is a rather complex problem for the hotels. Studies in the literature are predominantly related to the process for sentiment classification. However, there are very few studies on the process of feature identification, feature engineering and performance comparison of machine learning algorithm. For this reason, in order to fill the gap in the literature, feature identification and engineering for this sector has been studied. Then real-life data was collected from an international five star Iranian hotel chain. Moreover, the data were classified by using six different machine learning algorithms. According to the results, sentiment predictions can be done with 98.9% accuracy.

The experimental results indicated the following main conclusions:

(1) The highest accuracy of review classification was achieved by DT and C4.5.
(2) Comparing the six machine learning methods, DT and C4.5 performed the best with large feature set, SVM follows, and ANN performed the worst.

(3) Comparing the six machine learning methods, NB performed the best with fewer features.
(4) The performance of almost all methods increased with increasing features except ANN. In other words, increasing the features will certainly increase the accuracy of the model, except in the case of ANN.

Further research will be conducted in the following aspects: (1) reviews with a neutral sentiment (neither positive nor negative), clustered SVM, multiclass SVM and its contribution to sentiment classification will be discussed; (2) The effect that punctuation (e.g. '!', '?', etc.) has on sentiment classification will be analyzed; and (3) The ensemble learning (e.g., bagging and voting) in sentiment classification.

References

Ali, F., K. Kwak, and Y. Kim. 2016. Opinion mining based on fuzzy domain ontology and support vector machine: A proposal to automate online review classification. *Applied Soft Computing* 47:235–50. doi:10.1016/j.asoc.2016.06.003.

Al-Smadi, M., O. Qawasmeh, M. Al-Ayyoub, Y. Jararweh, and B. Gupta. 2018. Deep Recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews. *Journal of Computational Science* 27:386–93. doi:10.1016/j.jocs.2017.11.006.

Bahassine, S., A. Madani, M. Al-Sarem, and M. Kissi. 2018. Feature selection using an improved Chi-square for Arabic text classification. *Journal of King Saud University-Computer and Information Sciences 32(2):225-231*. https://doi.org/10.1016/j.jksuci.2018.05.010

Bi, J., Y. Liu, Z. Fan, and E. Cambria. 2019. Modelling customer satisfaction from online reviews using ensemble neural network and effect-based Kano model. *International Journal of Production Research* 57(22):1–21. https://doi.org/10.1080/00207543.2019.1574989

Cambria E., D. Das, S. Bandyopadhyay, and A. Feraco. 2017. Affective computing and sentiment analysis. In *A Practical Guide to Sentiment Analysis. Socio-Affective Computing vol 5.*, ed. Cambria E., Das D., Bandyopadhyay S., Feraco A., vol 5. Cham: Springer. https://doi.org/10.1007/978-3-319-55394-8_1

Cao, Q., W. Duan, and Q. Gan. 2011. Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. *Decision Support Systems* 50:511–21. doi:10.1016/j.dss.2010.11.009.

Cardot, H., and D. Degras. 2018. Online principal component analysis in high dimension: Which algorithm to choose? *International Statistical Review* 86:29–50. doi:10.1111/insr.12220.

Dadgar, S., M. Araghi, and M. Farahani 2016. A novel text mining approach based on TF-IDF and support vector machine for news classification. In A novel text mining approach based on TF-IDF and support vector machine for news classification. 2016 IEEE International Conference on Engineering and Technology (ICETECH), IEEE, Coimbatore, India. 112–16.

Dickinger, A., and J. Mazanec. 2015. Significant word items in hotel guest reviews: A feature extraction approach. *Tourism Recreation Research* 40:353–63. doi:10.1080/02508281.2015.1079964.

Duan, W., Y. Yu, Q. Cao, and S. Levy. 2016. Exploring the impact of social media on hotel service performance: A sentimental analysis approach. *Cornell Hospitality Quarterly* 57:282–96. doi:10.1177/1938965515620483.

Ducange, P., R. Pecori, and P. Mezzina. 2018. A glimpse on big data analytics in the framework of marketing strategies. *Soft Computing* 22:325–42. doi:10.1007/s00500-017-2536-4.

Fu, Y., J. Hao, X. Li, C. Hsu, X. Li, C. Hsu, and C. Hsu. 2018. Predictive accuracy of sentiment analytics for tourism: A metalearning perspective on Chinese travel news. *Journal of Travel Research* 57:0047287518772361. doi:10.1177/0047287517700317.

Gulsoy, N., and S. Kulluk. 2019. A data mining application in credit scoring processes of small and medium enterprises commercial corporate customers. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9:e1299.

Hu, Y., and K. Chen. 2016. Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings. *International Journal of Information Management* 36:929–44. doi:10.1016/j.ijinfomgt.2016.06.003.

Ma, E., M. Cheng, and A. Hsiao. 2018. Sentiment analysis–a review and agenda for future research in hospitality contexts. *International Journal of Contemporary Hospitality Management* 30:3287–308. doi:10.1108/IJCHM-10-2017-0704.

Manek, A., P. Shenoy, M. Mohan, and K. Venugopal. 2017. Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. *World Wide Web* 20:135–54. doi:10.1007/s11280-015-0381-x.

Moro, S., P. Rita, and J. Coelho. 2017. Stripping customers' feedback on hotels through data mining: The case of Las Vegas Strip. *Tourism Management Perspectives* 23:41–52. doi:10.1016/j.tmp.2017.04.003.

Sánchez-Franco, M., A. Navarro-García, and F. Rondán-Cataluña. 2019. A naive Bayes strategy for classifying customer satisfaction: A study based on online reviews of hospitality services. *Journal of Business Research 101:499-506.*

Schuckert, M., X. Liu, and R. Law. 2015. Hospitality and tourism online reviews: Recent trends and future directions. *Journal of Travel & Tourism Marketing* 32(5): 608–621. https://doi.org/10.1080/10548408.2014

Talón-Ballestero, P., L. González-Serrano, C. Soguero-Ruiz, S. Muñoz-Romero, and J. Rojo-Álvarez. 2018. Using big data from customer relationship management information systems to determine the client profile in the hotel sector. *Tourism Management* 68:187–97. doi:10.1016/j.tourman.2018.03.017.

Tran, T., H. Ba, and V. Huynh 2019. Measuring hotel review sentiment: An aspect-based sentiment analysis approach. In: Seki H., Nguyen C., Huynh V. N., Inuiguchi M. (eds) *Integrated uncertainty in knowledge modelling and decision making*. IUKM 2019. Lecture notes in computer science, vol 11471. Springer, Cham. https://doi.org/10.1007/978-3-030-14815-7_33

Tripathy, A., A. Agrawal, and S. Rath. 2016. Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications* 57:117–26. doi:10.1016/j.eswa.2016.03.028.

Uğuz, H. 2011. A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowledge-Based Systems* 24:1024–32. doi:10.1016/j.knosys.2011.04.014.

Uysal, A., and S. Gunal. 2014. The impact of preprocessing on text classification. *Information Processing & Management* 50:104–12. doi:10.1016/j.ipm.2013.08.006.

Wang, F., C. Li, J. Wang, J. Xu, and L. Li. 2015. A two-stage feature selection method for text categorization by using category correlation degree and latent semantic indexing. *Journal of Shanghai Jiaotong University (Science)* 20:44–50. doi:10.1007/s12204-015-1586-y.

Wang, H., P. Yin, J. Yao, and J. Liu. 2013. Text feature selection for sentiment classification of Chinese online reviews. *Journal of Experimental & Theoretical Artificial Intelligence* 25:425–39. doi:10.1080/0952813X.2012.721139.

Ye, Q., Z. Zhang, and R. Law. 2009. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications* 36:6527–35. doi:10.1016/j.eswa.2008.07.035.

Zainuddin, N., A. Selamat, and R. Ibrahim. 2018. Hybrid sentiment classification on twitter aspect-based sentiment analysis. *Applied Intelligence* 48:1218–1232.