UTM UNIVERSITI TEKNOLOGI MALAYSIA
RESEARCH UNIVERSITY

International Journal
of
Innovative Computing

Journal Homepage: http://se.cs.utm.my/ijic

A Novel Feature Reduction Method in Sentiment Analysis

Alireza Yousefpour, Roliana Ibrahim[1] and Haza Nuzly Abdull Hamed[2]
Faculty of Computing
Universiti Teknologi Malaysia
81310 UTM Skudai, Johor,
Malaysia
yalireza3@live.utm.my, {[1]roliana, [2]haza}@utm.my

***Abstract*— With the genesis of the Internet and the world wide web, we have seen an enormous growth of data and information on the web, as well as an increase in digital or textual opinions, sentiments and attitudes that have been remarked upon in reviews. More reviews in document-level have expressed a high-dimensional in feature space. The main task of feature selection and feature reduction is a reduction dimension in feature space while, at the same time, ensuring that is no loss in the minimum of accuracy. There are several factors to consider in reduction dimension of a term - document matrix of feature space. It can lead to removal of irrelevant and useless features; including as a result, more efficient categories, easier analysis more accurately of sentiment after reduction. For this aim, we have proposed a novel feature reduction method using standard deviation based on more variation or dispersion of features in feature space. We used three popular classifiers, namely: Naive Bayes, Maximum Entropy and Support Vector Machine for sentiment classification and ensemble of these classifiers. We then compared our proposed method with other feature reduction methods used on book and music reviews. Results show that classification by using the novel method improved the accuracy of sentiment classification.**

Keywords – sentiment analysis, feature reduction, standard deviation, sentiment classification.

I. INTRODUCTION

Nowadays, there is a need for customers and companies to make use of public opinions and sentiments in the decision-making process for their products. With the apparition of Web 2.0, as well as widespread internet and social media such as social networks, reviews, comments, twitter and forum discussions constantly occurring on the web, they can focus on collecting public opinions very efficiently due to the substantial amount of general information available. Thus, processing and extraction of information and opinions on the web and subsequently distilling them can be quite a formidable task. Sentiment analysis applications can be extracted from roughly every possible area such as services, financial services, political elections and customer products respectively. Special attention needs to be paid to the processes and understating of information by information retrieval methods (IR) and natural language processing methods (NLP). One of the main problems in this scope is that of sentiment analysis, whereby a review is classified into two classes, i.e. positive (thumbs up or favourable) or negative (thumbs down or unfavourable) opinions respectively [1].

Sentiment analysis (also known as opinion mining) is a field of computational study that processes opinions, attitudes, sentiments, emotions, and appraisals of people concerning products, movies, entities, events, issues, topics and their respective features.

The aim of sentiment analysis and opinion mining is to identify attitudes and emotions of a person concerning a certain subject, such as a movie or product etc. In fact, opinion mining denotes drawing or outing of subjective information from a text corpus or reviews; while sentiment analysis signifies the evaluation of this information when extracted. Various researches into sentiment analysis in more recent works have presented different techniques by which to extract and analyse sentiments. Since the last decade, there have been several works undertaken in this area [1-4].

In this paper, we assessed the relevance of features by looking only at the probability distribution of the features in feature space. The standard deviation of each feature as

feature weighting is calculated in feature space. The feature is reduced using filter method in feature space.

The remainder of the paper is organized as follows. Section 2 gives an overview and related works of sentiment analysis. Section 3 presents dimensionality reduction methods and proposed method. Section 4 shows the experimental results by comparing its classification performance with previous. Finally, the paper ends with discussion and conclusion in Section 4 and 5, respectively.

II. Related Works

The text feature extraction has three levels consisting of, namely: sentiment analysis, document level, sentence level, as well as entity and aspect level. A considerable amount of research work has been presented dealing with different methods, techniques and new ideas in relation to selection and extraction of sentiment words as features from corpus and reviews. They used techniques including: linguistic rules and patterns, part-of-speech (POS), fuzzy pattern matching stemming, document citation, stylistic measures, punctuation, percentage of appraisal groups as well as Ngram in order to extract features and sentiment lexicon form reviews as nouns, adjectives, adverbs and verbs respectively. The relationship among features can be helpful in understanding opinions while still using multi words as a feature by Ngram methods and dictionary [1-5].

Sentiment analysis was first introduced by [2]; however, there were some previous works that proposed a sentiment classification task by [1, 4]. [2] exploited the concept of a *supervised approach*, while another study [10] proposed the *unsupervised approach*. In the supervised learning approach, a classifier is first trained based on a large feature set, which consists of labelled data. This classifier is then used to identify and classify unlabelled test data to two classes (or more) of positive and negative sentiments. Some researchers used several feature sets to attempt to improve the classification accuracy [6-8].

Some researchers proposed a shallow parsing to select an appropriate feature set. Other feature selection techniques, such as IG, were used to achieve better performance. Further, some works tried to combine several approaches with hybrid classifiers [9]. Most of the existing studies define sentiment classification as a supervised classification problem and try to train a classifier from a large amount of labelled data for this task [6, 10]. Our research relies on the evaluation of supervised learning methods on reviews.

In the literature [11] proposed and compared four feature reduction methods, namely the Document Frequency (DF) method, the Principal Component Analysis (PCA) method, the category Frequency-Document Frequency(CF-DF) method, and TF-IDF method for the neural network classifier. Their opinion, aim of feature reduction is to minimize features set loss while maximizing reduction in dimensionality. The filter method was applied to reduce the feature space with high reduction rate of 98.9% with PCA method.

In some literatures [12, 13] used five feature selection methods, namely Information Gain (IG), Chi-square (CHI), Document Frequency (DF), Term Strength (TS), and Information Mutual (IM) methods in text classification. They obtained better result through DF, IG, and CHI methods.

High dimensional problem and high computational cost to select optimal feature subset and dependency of feature subset selection to classifier in wrapper method. In this paper, we propose a method for feature selection in sentiment analysis based on dispersion and distribution of features in feature space. In other words, the best of features are selected by its higher value of standard deviation for construction of the optimal feature subset in a filter method.

III. Dimensionality reduction methods

Further reviews in document-level have many more features. In fact, a considerable amount of features express a high dimensional space. The main task of feature selection and feature reduction/extraction is reduction dimension in feature space while ensuring the minimum of accuracy. It causes the removal of irrelevant features and results in the following outcomes: more efficient categories; easier analysis of sentiment after reduction; visualisation of results; and there may be a better perception of low dimension. Sentiment analysis includes the following steps, specifically: *pre-processing* to extract tokens from reviews and removal of stop words; feature selection that identifies features from tokens by filter, wrapper and embedded methods; feature reduction to reduce feature space as document frequency; and, finally, my proposed method, namely, standard deviation method. The last step a classifier is used to predict polarity sentiment as sentiment classification.

A. Preprocessing phase

We performed a simple preprocessing in two steps, as described below.

1) Tokenizing

In the pre-processing phase, reviews are scanned to extract tokens consisting of words and numbers. It is perhaps a result of errors in reviews that they are not always lexically and syntactically well-formed [14].

2) Removal Stop Words

Stop words in reviews can play a negative role. They can appear in training sets of both positive and negative factors; as a result, they increase more instances of ambiguity in sentiment classification since stop words do not have any sentiment information [15]. We have removed stop words in the pre-processing step; including words with a length less than 2 and words such as: she, he, at, about, at, the, etc.

B. Feature Selection

Using N-gram as opposed to a single word has an advantage in that there are some dependencies between certain words and importance is placed on individual phrases. There are three popular forms of N-gram, specifically: unigram, bigram and trigram. [15] uses a unigram model when performing comparisons with the feature-based model and the tree kernel-based model for Twitter data. [1] applies unigram and bigram, as well as a combination of these two for extracting features of movie reviews. In addition, some of these works have also used Ngram [12-15].

There are several factors in selection and feature reduction of dimension in a term-document matrix of feature space. It causes the removal of irrelevant features and can result in, specifically: more efficient categories, easier analysis of sentiment after reduction, visualisation of results, and better perception of low dimension. Two main approaches to achieve the appropriate size of dimension for classifier are as follows:

a) *Feature selection:* This refers to identification and extraction of features that lead to class separability. Using the univariate approach for rating and the multivariate approach for optimization represents a criterion function that may enable a better chip in classification.
b) *Feature extraction:* This refers to the reduction of high-dimensional space in order to lower feature space via a linear or nonlinear transformation. This transformation can take three forms, namely: supervised learning, semi-supervised learning and unsupervised learning methods.

[16] stated that there are four basic steps in any feature extraction method. These are, namely: the *generation* that results in generating the candidate feature subset; while the next step is *evaluation.* In this step, produced subsets having relevancy value are evaluated by an evaluation function or classifier. The s*topping criteria* occur in two states. In one of these states, if the subset is optimal, it signifies the end of the process. However, if it is in use, the generation process calls again for creation of the next subset of features. The last stage to emerge from the feature selection process is *validation.* The validation step survey can verify a selected feature subset in order to confirm that it is a valid subset as required.

Generation

In the generation stage, a subset of features is created from a feature set. In each of the iterations, a subset is created and evaluated in the next stage until an optimal subset is achieved. There are different techniques to create subsets, i.e. creating feature subsets completely, randomly and heuristically; whereas producing subsets differs from one method to another [19].

- *Complete generation:* complete or exhaustive generation, or a combination all the feature subsets is tested. If 'n' is the number of features, the subsets are defined as O (2^n). The order of search space is large. Even if we were able to find the best or optimal feature subset, it was either too expensive or often not very practical for commercial use.
- *Heuristic generation:* the other method to produce a subset is by a heuristic method, i.e. searching for algorithms in a forward and backward selection that attempts to find the best of subsets. The algorithm adds features one by one to the candidate feature subset until the evaluation function returns the target relevancy value. There is one problem facing such a heuristic approach; a high order combination of relevant feature subsets may exist because some relevant feature subset may have been omitted {f1, f2}.
- *Random generation:* In a random state, feature subsets are created randomly without having any criterion or algorithm. As a result, the number of optimal feature subsets can be identified by the number of users trying to create subsets.

Evaluation of Subsets

In the evaluation stage, produced subsets are evaluated by appraising evaluation function within filter and classifier within wrapper method, where some relevancy value will be computed. Such value is then compared with the previously known best value.

Similarly with the generation step, we are able to categorise a different feature selection method according to the manner in which evaluation is carried out. There currently exist 5 different evaluation methods as shown below [16].

TABLE I. A COOPERATION OF EVALUATION FUNCTIONS

Approach	Examples	Generality	Time	Accuracy	Type
Distance	Euclidean Distance Measure	yes	low	-	Filter
Information	Entropy, Information Gain, etc.	yes	low	-	Filter
Dependency	Correlation Coefficient	yes	low	-	Filter
Consistency	Min-Features Bias	yes	moderate	-	Filter
Classifier error-rate	The Classifiers Themselves	no	high	very high	Wrapper

Table I shows a comparison of different evaluation functions regardless of the kind of procedure used. The '-' in the accuracy column means that information about the accuracy of the corresponding evaluator cannot be concluded. Following is a brief discussion of each of these types of evaluation functions:

- In *distance measure*, we compute the physical distance. Features that can support instances/records of the class in order to stay together are selected. The key concept is the assumption that instances of the same class must be closer than those in a different class.
- *Information measure* refers to the selection of a feature subset that can yield maximal information gain.
- *Dependency measure*, this assesses the correlation between a feature and a class label. If feature A is more highly related to the class than feature B, then we select feature A. It measures how closely a feature can be related to the outcome of the class label. A slight variation of the definition can be used to measure the degree of redundancy between features. For example, if feature A is heavily dependent upon feature B, then feature A is redundant. Since the correlation is merely a measure of relationship, we need some kind of physical measure in order to define such relationship.
- *Consistency measure:* Two instances are considered in-consistence if such a situation occurs; having all matching feature values, except its class. It is to be selected only if there is no such case. It makes use of the Min-feature bias, where the Find minimally-sized subset

satisfies the acceptable inconsistency rate (i.e. defined by the user). This bias may lead to problems when one feature alone guarantees no inconsistency. The IC value is unique for all instances. It is impossible to find two people with the same IC number (i.e. two feature values that are the same).

- *Classifier error rate:* In this approach, feature selection loss is a generality, but gains accuracy towards the classification task. From a computational aspect, it is very costly.

Table II shows that some literature studies have categorised the first four as a *filter approach* and the final one as a *wrapper approach*. In [17], feature extraction methods are classified into three classes as follows:

- Filter techniques
- Wrapper techniques
- Embedded techniques

a) Filter techniques

Filter methods are independent of an inductive algorithm. Filter methods select the best of features based on an intrinsic properties criterion, e.g. using their Euclidean Distance measure that is, choosing features to stay with the same proximity by the instance of the same class. It is important to assume that samples of the same class should be nearest to those in other classes. In fact, filter techniques can select related features that have high-scoring attributes and remove features with low-scoring attributes. The best of the features subsets are then sent to a classifier [17].

b) Wrapper techniques

Wrapper methods describe inductive algorithms such as an evaluator. These methods select the best of feature subsets by generation and evaluation of different subsets in space of states. The selection and evaluation of a specific features subset is achieved from a classifier by means of training and testing algorithms. Hence, wrapped algorithms search the space of all subsets of features as classification methods. Heuristic techniques can help to search for optimal subsets [17].

Fig. 1 shows two different types of feature extraction. Filter methods have several benefits including: being fast, simple and independent of classifiers, as well as being easily scaled to very high-dimensional. However, they have a common problem in that classifiers are relinquished and the majority of suggested methods are univariate, which may reduce the accuracy of the classifier. Multivariate filter methods have been introduced to attempt to reduce the above problem, i.e. ignore feature affiliations. In fact they connect affiliations of features to the same degree.

The benefit of wrapper methods is a semantic relationship among model selections and subsets of feature of search. Also, there is an interrelation with the classification algorithms. Nevertheless, they are a higher over-fitting risk and, as a result, have complexity of computation and cost [20].

a) Embedded techniques

Embedded methods look for an optimal subset of features via a search in hypotheses and space of feature subset. In fact, this method creates the classifier construction. Embedded methods are a special technique for utilising existing classifiers with learning algorithms. The benefit of this method is far less complex than wrapper methods. At the same time, there is interplay with classifier and dependence upon features. For example, [18] and [19] used the weight vector of each feature in SVM as a linear classifier. The weights express a relationship of multivariate features and, as a result, cause the cancellation of features with lightweight.

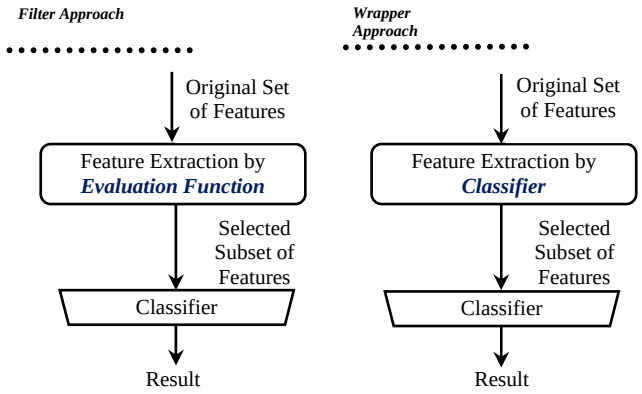

Fig.1 Comparing two approaches based on type of feature extraction functions

Table II shows a taxonomy and categorisation of feature extraction methods. According to [17], for each feature extraction, have expressed a set of characters, which can help to select a suitable and better method to the goals via listing of the advantages and disadvantages of each method.

C. Feature Reduction

There are several popular methods of reduction, including Document Frequency (DF) and Term Frequency-Inverse Document Frequency (TF-IDF). In addition, there is my proposed method of Standard Deviation (SD). All of these Methods use score in terms of extraction and selection of the size of the predefined set of characters.

1) Document Frequency

In the Document Frequency (DF) method, features are ordered by document frequency for each feature in a whole document. This method is the simplest measure for feature reduction and has a linear time complexity capable of scaling a large dataset.

2) Standard Deviation

Standard deviation is a statistical and probability method that calculates the amount of variation or diversity of data from the average (or mean) existing. It is normally used to measure assurance in a statistical conclusion to express the dispersion of a population. Standard deviation is defined as follows:

,

σ c (2)(1)

Where $\{x_1, x_2, ..., x_N\}$are features, N is the number of features, $\bar{x}$ is average or mean value, and σ is standard deviation. In fact, a high standard deviation shows that features extend out over a large range of values. Conversely, a low standard deviation explained that feature points lead very close to the main. As a result, using features with higher

standard deviation can better predict better sentiment classification.

TABLE II. A CATEGORISATION OF PROS AND CONS OF FEATURE EXTRACTION METHODS

Type Search		Positive	Negative	Example
Filter methods	**Univariate**	. Quick . Gradable . No dependence upon classifier	. Relinquish dependence upon feature . Relinquish interplay with classifier	. Information Gain (IG) . x^2 – CHI . *t* - test
	Multivariate	. Dependence upon feature . No dependence upon classifier . Better time complexity than wrapper	. Slower than univariate methods . Less gradable than univariate methods . Relinquish interplay with classifier	. Correlation-based feature selection (CFS) . Markov blanket filter (MBF) . Fast correlation-based feature selection (FCBF)
Wrapper methods	**Deterministic**	. Simple . Dependence upon feature . Interplay with classifier . Slower than Randomize	. High risk of overfitting . More chance of entrapment with local optimum than Randomize . Classifier dependent selection	. Sequential forward selection (SFS) . Sequential backward elimination (SBE) . Beam search
	Randomize	. Dependence upon feature . Less entrapment with local optimum . Interplay with classifier	. Classifier dependent selection . More risk of being over fitting than deterministic	. Simulated Annealing . Randomized hill climbing . Genetic algorithms . Estimation of distribution algorithms
Embedded methods		. Dependence to feature . Interplay with classifier . Better time complexity than wrapper	. Classifier dependent selection	. Decision trees . Weighted Naive Bayes . Feature selection using the weight vector of SVM

D. Classification Techniques

Sentiment classification can be divided into two types of classification forms. The first is a binary sentiment classification (also known as a polarity sentiment classification) that includes both positive and negative classes. Lastly, there is the multi-class sentiment classification which uses rating marks; for example, five-class, i.e. 1 to 5 stars or classes of {strong positive, positive, neutral, negative, strong negative} respectively instead of a two-class of negative and positive classes.

a) Naive Bayes

The Bayesian classification is a statistical method underlying a probabilistic model and supervised learning algorithms. Naive Bayes (NB) uses a features vector matrix to determine a document depending upon polarity classes (i.e. positive and negative classes) by probability. It attaches a document to the relevant class with the highest probability [5]. The probability is calculated as follows:

(3)

Where *P(c)* is the prior probability of category *c; P(d)* is the prior probability of training data *d*, *P(c|d)* is probability *c* given *d*, and *P(d|c)* is the probability of *d* given *c*.

b) Support Vector Machine

Support Vector machine (SVM) is a most popular algorithm that can classify data as either linear or nonlinear. It can also map input data to high-dimensional feature spaces, in addition to classifiers' SVM support regression, binary and multiclass classification respectively. For example, let us suppose the SVM classifier on binary classification is trying to find a decision surface that can separate data into two classes and determine a result to make a decision based on this support vector (Yang and Pedersen, 1997, Tan and Zhang, 2007). Following is an equation that should succeed in minimizing optimization for SVM:

$$\vec{\alpha}^* = \arg\min\{-\sum_{i=1}^{n}\alpha_i + \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j (\vec{x}_i$$

Where, $\sum_{i=1}^{n}\alpha_i y_i = 0$, (4)

The support vector can be either linear or nonlinear. In the nonlinear classification of SVM, results can be perfect if the kernel trick is used, and if the kernel used is Gaussian Radial Basis Function (RBF). The feature space will then be a Hilbert space of infinite dimension, since, classifiers of maximum border are well regulated. Hence, infinite dimension does not destroy the results. Some of the kernels are, specifically: *a Gaussian Radial Basis Function (RBF), polynomial (homogeneous), polynomial (inhomogeneous),* and *hyperbolic tangent.* RBF kernel is a popular function used as SVM classification. SVM with RBF kernel is closely related to RBF neural networks, with the centres of the radial basis functions automatically chosen for SVM.

c) Maximum Entropy

Maximum Entropy (ME) classifier is one of the machine learning methods used for natural language processing applications, as it is implemented using a multinomial logit model as the classifier rule. ME is a kind of statistical inference that can be used to estimate any probability distributions on the partial knowledge.

$$P(c|d) = \frac{1}{F(d)}\exp\left(\sum_{i=1}^{N}\right.$$ (5)

And $$X_{i,c}(d,c') = \begin{cases} 1 & if\ n_i(d) > \\ 0 & else \end{cases}$$

Where $X_{i,c}$ is a function of feature/class for *i* feature and *c* class, $W_{i,c}$ is weight of features, and F(d) is a normalization function respectively.

d) Classifier Ensemble

Network ensembles are trained to find solutions for the same problems in a parallel independent. There are several classifiers for sentiment classification that have both advantages and disadvantages. The aim of a classifier ensemble is to combine these classifiers while at the same time gathering their benefits, thus improving performance.

We used a classifier ensemble including three classifiers, namely: Naive Bayes (NB), Support Vector Machine (SVM), and Maximum Entropy model (ME) as base-level algorithms, in addition to using majority voting high-level.

IV. Experiment Results

To evaluate the real performance of the proposed method for feature reduction in sentiment analysis, we have exploited the method by using a real dataset.

A. Datasets

In this work, two different data sets are used to investigate the performance of the proposed models. These datasets consist of several review documents in different domains and in different languages. This collection consists of two review datasets, i.e. Book and Music in English. Each review document is considered as a bag of words and is represented by a feature vector. All unigrams and bigrams are used as features and feature weight is set to term presence.

B. The performance measure

Generally, the performance of sentiment classification is evaluated by using index accuracy in order that this index can be computed through the following equation:

$$Accuracy \tag{6}$$

The common way for computing this index is based on the confusion matrix shown in Table III.

TABLE III. THE CONFUSION MATRIX

		Predicted	
		positives	***negatives***
Actual	***positives***	# of True Positive instances (TP)	# of False Negative instances (FN)
	negatives	# of False Positive instances (FP)	# of True Negative instances (TN)

C. Results

Review datasets are used in this experiment. This dataset consists of review documents from two different domains of book reviews and music reviews. To enable the feature reduction, we applied the following steps:

- Extraction of tokens from reviews
- Removal of stop words
- Using Uni-gram and Bi-gram to determine features and create the feature vector
- Weighting by using document frequency (and subsequently standard deviation so as to compare)
- Sorting feature vector in descending order
- Pruning part of the vector from the first vector
- Sending vector to classifiers

The dataset was divided into two sets, i.e. a training data set and testing dataset with 10-fold cross validation. Table IV below shows results of the classification on reviews.

Results show that standard deviation improved accuracy in more time; in particular, when we used the classifier ensemble in sentiment classification.

TABLE IV. COMPARISON OF RESULTS OF CLASSIFIER METHODS ON MUSIC AND BOOK REVIEW DATASET with 10-FOLD CROSS VALIDATION IN 10 ITRATIONS

Classifier	**Feature Reduction Method**	**Number of Feature**	**Accuracy (%)**	
			Music Review	**Book Review**
SVM	DF	1000	79.15	79.75
	SD	1000	79.30	80.65
ME	DF	1000	79.45	80.25
	SD	1000	78.50	80.55
NB	DF	1000	81.75	81.20
	SD	1000	81.20	80.65
Classifier Ensemble	DF	1000	87.40	87.60
	SD	1000	87.90	87.75
Ensemble of All classifier and All Feature Set		**1000**	**90.22**	**90.91**

V. Discussion

We used filter method for feature reduction because 1) they are more scalable to very high dimensional datasets, 2) their bias is different from the classifiers, and 3) computational is very simple and fast.

We obtained good results using standard deviation method in comparison with the DF method. But, there is a challenge in using from standard deviation method for feature weighting in sentiment analysis. Distribution of features in feature space can be two form of concave or convex, as calculating the mean or average in standard deviation is much related to features space form and dispersion of them. In this paper, feature space is assumed as a convex form.

VI. Conclusion and Future Work

We proposed a statistical and probabilistic method using standard deviation of features for high-dimensional problem on feature space in a sentiment analysis, while at the same time improving the accuracy of sentiment polarity forecasting. We used three popular classifiers, specifically: Naive Bayes, Maximum Entropy and Support Vector Machine for sentiment classification and compared our proposed method with other feature reduction methods on book and music reviews. Results have shown that classification by using this novel method has improved the accuracy of sentiment classification. In future work, we will try to find the optimal feature subset using clustering on feature space based on diversity of features in an embedded approach.

References

[1] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, 2002, pp. 79-86.

[2] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in *Proceedings of the 2nd international conference on Knowledge capture*, 2003, pp. 70-77.

[3] H. Tang, S. Tan, and X. Cheng, "A survey on sentiment detection of reviews," *Expert Systems with Applications*, vol. 36, pp. 10760-10773, 2009.

[4] S. Padmaja and S. S. Fatima, "Opinion Mining and Sentiment Analysis–An Assessment of Peoples' Belief: A Survey," *International Journal*, 2013.

[5] A. Abbasi, H. Chen, and A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums," *ACM Transactions on Information Systems (TOIS)*, vol. 26, p. 12, 2008.

[6] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th annual meeting on association for computational linguistics*, 2002, pp. 417-424.

[7] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, pp. 1-167, 2012.

[8] S. Zhou, Q. Chen, and X. Wang, "Active deep learning method for semi-supervised sentiment classification," *Neurocomputing*, 2013.

[9] R. Prabowo and M. Thelwall, "Sentiment analysis: A combined approach," *Journal of Informetrics*, vol. 3, pp. 143-157, 2009.

[10] P. P. Xu, H. L. Jin, H. X. Shi, and W. Chen, "An Unsupervised Sentiment Information Identification Approach," *Applied Mechanics and Materials*, vol. 263, pp. 3330-3334, 2013.

[11] S. L. Lam and D. L. Lee, "Feature reduction for neural network based text categorization," in *Database Systems for Advanced Applications, 1999. Proceedings., 6th International Conference on*, 1999, pp. 195-202.

[12] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *ICML*, 1997, pp. 412-420.

[13] M. Rogati and Y. Yang, "High-performing feature selection for text classification," in *Proceedings of the eleventh international conference on Information and knowledge management*, 2002, pp. 659-661.

[14] E. Boiy and M.-F. Moens, "A machine learning approach to sentiment analysis in multilingual Web texts," *Information retrieval*, vol. 12, pp. 526-558, 2009.

[15] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in *Proceedings of the Workshop on Languages in Social Media*, 2011, pp. 30-38.

[16] M. Dash and H. Liu, "Feature selection for classification," *Intelligent data analysis*, vol. 1, pp. 131-156, 1997.

[17] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *bioinformatics*, vol. 23, pp. 2507-2517, 2007.

[18] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, pp. 389-422, 2002.

[19] X. Zhang, X. Lu, Q. Shi, X.-q. Xu, H.-c. Leung, L. Harris, *et al.*, "Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data," *BMC bioinformatics*, vol. 7, p. 197, 2006.