Chapter 12

The Today Tendency of Sentiment Classification

Vo Ngoc Phu and Vo Thi Ngoc Tran

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.74930

Abstract

Sentiment classification has already been studied for many years because it has had many crucial contributions to many different fields in everyday life, such as in political activities, commodity production, and commercial activities. There have been many kinds of the sentiment analysis such as machine learning approaches, lexicon-based approaches, etc., for many years. The today tendency of the sentiment classification is as follows: (1) Processing many big data sets with shortening execution times (2) Having a high accuracy (3) Integrating flexibly and easily into many small machines or many different approaches. We will present each category in more details.

Keywords: sentiment classification, machine learning approaches, lexicon-based approaches, today tendency of the sentiment classification, big data set

1. Introduction

Many different approaches have already been developed for sentiment analysis for many years because a lot of researchers have already desired to find many optimal algorithms and optimal approaches for many surveys and commercial applications.

The sentiment classification, called opinion mining, is the computational studies of opinions, sentiments, evaluations, attitudes, appraisal, affects, views, emotions, subjectivity, etc., expressed in texts (reviews, blogs, discussions, news, comments, feedbacks, etc.)

The different approaches have been used to cross-check with each other to reform their accuracies.

One document (one sentence or one phrase) is classified into the positive polarity, the negative polarity or the neutral polarity.

The positive polarity is a polarity of a word or a phrase (a sentence or a document) which performs aspects about good, nice, like, love, delicious, happiness, enthusiasm, kindness, etc. Examples of phrases: very good, very nice, etc. Examples of sentences: "He is very handsome"; "She is very beautiful." Examples of documents: "He is very handsome. He is also good at Sports."

The negative polarity is a polarity of a word or a phrase (a sentence or a document) which expresses aspects about bad, evil, poor, ugly, wrong, inclement, foul, shabby, sinister, rotten, ill, shoddy, etc. Examples of phrases: very bad, very evil, etc. Examples of sentences: "He is very bad"; "She is very wrong." Examples of documents: "She is very bad. She is very stupid."

The neutral polarity is a polarity of a word or a phrase (a sentence or a document) which is not both the positive polarity and the negative polarity. Examples of neutral words: eat, talk, drink, etc. Examples of phrases: a bucket of water, 1 kg, and etc. Examples of documents: "He eats a banana. He drinks a glass of water."

The polarity (positive, negative, or neutral) of a sentence or a document has been identified by using many machine learning algorithms in the surveys of the sentiment classification in [1–83].

The sentiment polarity of a word or phrase (a sentence or a document) is also expressed through a valence (sentiment score or sentiment value) of this word or this phrase (this sentence or this document).

The polarity and valence of a word or a phrase in English have been calculated by using many different approaches such as many sentiment dictionaries. Besides, the polarity and sentiment value of a word or a phrase have been identified by using many similarity measures in English and Vietnamese in [49, 50, 51, 52]. In addition, according to our opinion, the polarity and sentiment score of a word or phrase of all languages (Chinese, French, etc.) can be calculated easily by using the similarity coefficients.

If the valence of a word or phrase (a sentence or a document) is greater than 0, this word or phrase (this sentence or this document) is the positive polarity. A word or phrase (a sentence or document) is the neutral polarity if the sentiment score of this word or phrase (this sentence or this document) is as equal as 0. If the sentiment value of a word or phrase (a sentence or a document) is less than 0, this word or phrase (this sentence or this document) is the negative polarity.

Many machine learning algorithms have already had two kinds (supervised Learning and unsupervised learning) comprising a lot of algorithm groups such as: deep learning group, ensemble group, neural networks group, regularization group, rule system group, regression group, Bayesian group, decision tree group, dimensionality reduction group, instance based group, and clustering group.

The sentiment analysis has had many machine learning approaches and lexicon-based approaches.

The lexicon-based approaches comprise many dictionary-based approaches and corpus-based approaches. The corpus-based approaches include statistical and semantic.

In this chapter, we display the dictionary-based approaches and the corpus-based approaches of the sentiment classification basically; and we also present the today tendency of the sentiment analysis in more details as follows: (1) Processing many big data sets with shortening execution times (2) Having a high accuracy (3) Integrating flexibly and easily into many small machines or many different approaches, because there have been a lot of documents, reviews, discussions, blogs, news, comments, feedbacks, etc., on many websites, online news sites, and social networks.

There have also been many big corporations in the world. The corporations have had many branches in many different countries in the world. Each branch of a corporation has had thousands of employees. Therefore, the corporations have had a lot of big information and big data sets about their employees, their businesses, etc. Processing the big information and the big data sets is very difficult by using the old algorithms, the old surveys, and old applications; and sometimes the big information and the big data set cannot be processed successfully.

Thus, the researchers now find the approaches for the surveys and the commercial applications to process the big data set for shortening execution times, improve the accuracies of these approaches. In addition, they can flexibly be integrated, and easily into the small machines or the different approaches because these small machines can be used conveniently in anywhere, for any type of users, and for various purposes. In the near future, these small machines can be produced easily, and they can be very cheap and easy to carry in everywhere.

This chapter includes six sections: The first section is the Introduction section. The second section is the Approaches of the Sentiment Classification section. The third section is the Today Tendency of the Sentiment Analysis section. The fourth section is the Conclusion section. The fifth section is the Conflict of Interest section, and the sixth section is the References section.

2. The approaches of sentiment classification

This section comprises two sub-sections as follows: In the first Section 2.1, we present the lexicon-based approaches of the opinion analysis. The machine learning approaches are displayed in the second Section 2.2.

2.1. Lexicon-based approaches

The lexicon-based approaches are comprised of multiple approaches, both dictionary-based and corpus-based.

The dictionary-based approaches involve using a dictionary that contains synonyms and antonyms of a word: for example [1], this study used seed sentiment words from a dictionary.

The approaches based on the corpus find opinion words with context-specific orientations according to a seed list of opinion words, to find other opinion words in a large corpus. There are two approaches within the category of corpus-based approaches:

a. Statistical Approach (example in [2]): If a word appears intermittently amid positive texts, then its polarity is positive. If it appears frequently among negative texts, then its polarity can be considered negative. If it has equal frequencies in positive and negative texts, then it can be considered a neutral word. Seed opinion words can be found using statistical techniques. Most state-of-the-art methods are based on the observation that similar opinion words often appear together in a corpus. Thus, if two words appear together frequently within the same context, then the probability is high that they have same polarity. Therefore, the polarity of an unknown word can be determined by calculating the relative frequency of co-occurrence with another word.

b. Semantic approach (example in [3]): This principle assigns similar sentiment values to semantically-close words. These semantically-close words can be obtained by getting a list of sentiment words, iteratively expanding the initial set with synonyms and antonyms, and then determining the sentiment polarity for an unknown word by the relative count of positive and negative synonyms of this word.

A lexicon-based method was used for the sentiment classification of Twitter data in [4]. The approaches were used to identify and extract sentiments from emotions and hashtags. Also used in [4] was the practice of converting non-grammatical words to grammatical words, and normalizing non-root to root words to extract sentiments.

The survey in [5] used lexicon–based classification and included two techniques: a method-of-moments estimator for word, and a Bayesian adjustment for repeated counts of the same word.

A structured approach was used in [6] for domain-dependent sentiment analysis, using lexicon expansion aided by emoticons.

The survey [7] introduced was a new approach to lexicon extraction, which can be successfully used for sentiment polarity assignment. It has been shown that the accuracy obtained from such lexicons outperforms other lexicon-based approaches.

The lexicon–based approach that [8] used was the Semantic Orientation CALculator (SO-CAL), which includes dictionaries of words annotated with their semantic orientation (polarity and strength), and incorporates intensification and negation.

The survey in [9] proposes a framework for sentiment analysis using dictionary-based approach. An approach to sentiment analysis is proposed that uses dictionary-based approach incorporating fuzzy logic.

In the research in [10], a lexicon-based approach was proposed to calculate reputation scores from Twitter. A Saudi-dialect lexicon was developed from Saudi tweets, to improve addressing the sentiment of the Arabic tweets.

The authors of [11] propose a lexicon-based approach to sentiment classification of Twitter posts. Their approach is based on the exploitation of widespread lexical resources such as SentiWordNet, WordNet-Affect, MPQA, and SenticNet.

The lexical or lexicon-based approach is a method for a teaching dictionary-based approach described by Mechael Lewis in the early 1990s in [12]. The basic concept and methods of this approach represent an idea that signifies how education involves the understanding and production of lexical phrases. This pattern of language has grammar as well as a meaningful collection of words.

Sentiment analysis performs a role in the lexicon-based approach in [13]. It plays a significant role in determining classes such as positive, negative, and neutral.

Lexicon based approach [14] is to extract and handle the sentiment as no-slang words.

The sentiments are as followed in many dictionaries which are named as lexicon based dictionaries which are: (1) Bing Liu's opinion lexicon. (2) MPQA subjectivity lexicon. (3) SentiWordNet lexicon. (4) Semantic Evaluation (SemEval).

The acronym dictionary included in [15, 16] is very helpful in expanding tweets and improve overall sentiments scores.

In [17, 18, 19], the emoticons have a different combination of symbols as different abbreviations.

The lexicon-based antonym dictionary in [20] contains set of well-lexicons, such as WordNet dictionary in English. WordNet dictionary maintains the set of lexical datasets for English words and also keeps record of semantic relationship between works.

The authors in [21–35] use the equations determining Pointwise Mutual Information (PMI) between two words wi and wj as follows:

$$PMI(wi, wj) = log_2\left(\frac{P(wi, wj)}{P(wi)xP(wj)}\right) \tag{1}$$

They use the equations determining SO (sentiment orientation) of word wi as follows:

$$SO\ (wi)\ =\ PMI(wi, positive) - PMI(wi, negative) \tag{2}$$

In [21–28], the positive and the negative of Eq. (2) in English are: positive = {good, nice, excellent, positive, fortunate, correct, superior} and negative = {bad, nasty, poor, negative, unfortunate, wrong, inferior}.

The AltaVista search engine (AVSE) is used in the PMI equations of [22, 23, 25], and the Google search engine (GSE) is used in the PMI equations of [24, 26, 28]. In addition, the authors of [24] also use German, the authors of [25] also use Macedonian, the authors of [26] also use Arabic, the authors of [27] also use Chinese, and the authors of [28] also use Spanish. In addition, the Bing search engine (BSE) is also used in [26].

With [29–32], the PMI equations are used in Chinese, not English, and Tibetan is also added in [29]. In terms of the search engine, AVSE is used in [31], and the authors of [32] use three search engines: GSE, the Yahoo search engine (YSE), and the Baidu search engine (BSE). The PMI equations are also used in Japanese with GSE in [33]. The authors in [34, 35] also use the PMI equations and Jaccard equations with GSE in English.

The Jaccard equations with GSE in English are used in [34, 35, 37]. The authors in [36, 41] use the Jaccard equations in English. The authors in [40, 42] use the Jaccard equations in Chinese. The authors in [38] use the Jaccard equations in Arabic. The Jaccard equations with the Chinese search engine (CSE) in Chinese are used in [39].

The authors in [48] use the Ochiai Measure through GSE with the AND and OR operators, to calculate the sentiment values of the words in Vietnamese. The authors in [49] use the Cosine Measure through GSE with the AND and OR operators, to identify the sentiment scores of the words in English. The authors in [50] use the Sorensen Coefficient through GSE with the AND and OR operators, to calculate the sentiment values of the words in English. The authors in [51] use the Jaccard Measure through GSE with the AND and OR operators, to calculate the sentiment values of the words in Vietnamese. The authors in [52] use the Tanimoto Coefficient through GSE with the AND and OR operators, to identify the sentiment scores of the words in English.

With the above proofs of the surveys in [21–52], according to our evaluation, all the similarity coefficients (or the similarity measures) can be applied with certainty to identify valences (or the sentiment scores) of all the words in many different languages.

2.2. Machine-learning approaches

The supervised learning algorithms and the unsupervised learning algorithms of the machine learning algorithms have been developed for the sentiment classification in **Figure 1**.

For the deep learning group of the sentiment analysis, deep learning (also known as deep structured learning or hierarchical learning) is based on learning data representations. Learning can be supervised, semi-supervised, or unsupervised. Examples of deep learning include deep neural networks, deep belief networks, and recurrent neural networks. They have been applied to many fields, including computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics, and drug design.

In the survey in [54], the deep learning techniques showed promising accuracy in this domain on English tweet corpus. The authors conducted the first study that applies deep learning

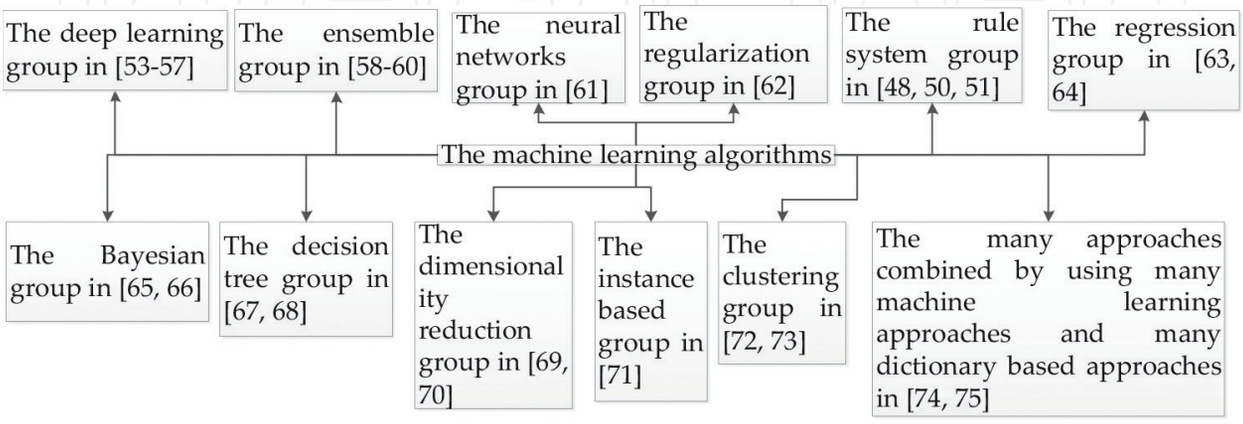

Figure 1. The machine learning algorithms.

techniques to classifying sentiment of Thai Twitter data. Two deep-learning techniques are included in the study: Long Short Term Memory (LSTM) and Dynamic Convolutional Neural Network (DCNN).

The authors of [55] used a new model to initialize the parameter weights of the convolutional neural network. They also used an unsupervised neural language model to train initial words.

Deep learning and micro-blog sentiment analysis were proposed in [56].

The authors in [57] fine-tuned a convolutional neural network (CNN) for image sentiment analysis and train a paragraph vector model for textual sentiment analysis. The authors conducted extensive experiments on both machine weakly-labeled and manually-labeled image tweets.

Ensemble approaches in statistics and machine learning use multiple learning algorithms to get better predictive performance than constituent learning algorithms. A machine learning ensemble, unlike a statistical ensemble in statistical mechanics, comprises only a concrete, finite set of alternative models, but typically allows for much more flexible structures to exist among those alternatives.

A comparative study of the effectiveness of ensemble technique for sentiment classification was proposed in [58]. This survey used the ensemble framework for sentiment classification, with the aim of efficiently integrating different feature sets and classification algorithms in order to synthesize a more accurate classification procedure. The research in [59] presents an ensemble learning method for sentiment classification of reviews. The ensemble learning framework, or stacking generalization, is introduced based on different algorithms with different settings, and compared with the majority voting. An ensemble sentiment classification strategy in [60] was applied based on Majority Vote principle of multiple classification methods, including Naive Bayes, SVM, Bayesian Network, C4.5 Decision Tree, and Random Forest algorithms.

The simplest definition of a neural network—more properly referred to as an "artificial" neural network (ANN)—is provided by the inventor of one of the first neurocomputers, Dr. Robert Hecht-Nielsen. The neural networks (NN)-based method in [61] combines the BPN and SO indexes to classify bloggers' sentiment. The NN-based method can reduce training time when classifying textual data. The NN-based method outperforms the traditional sentiment classification methods (BPN and SO index) in experimental results.

In mathematics, statistics, and computer science—particularly in the fields of machine learning and inverse problems—regularization is the process of introducing additional information in order to solve an ill-posed problem or to prevent over-fitting. The authors in [62] discussed a relation between Learning Theory and Regularization of linear ill-posed inverse problems. The authors showed that a notion of regularization (defined according to what is usually done for ill-posed inverse problems) allows derivation of learning algorithms that are consistent and that provide a fast convergence rate.

The authors in [48, 50, 51] used the rules of rule systems for the sentiment classification in Vietnamese and English.

Regression analysis in statistical modeling is a set of statistical processes for estimating the relationships among variables, and it comprises many techniques for modeling and analyzing

several variables. In regression analysis, we can see how the typical value of the dependent variable (or "criterion variable") changes when any one of the independent variables is varied while the other independent variables are held fixed. Regression analysis is a form of predictive modeling technique, which investigates the relationship between a dependent (target) and an independent variable (s) (predictor). The study in [63] analyzed the effect of using regression on sentiment classification of Twitter data.

Sentiment analysis was used in [64] to predict the Indonesian stock market. This study used the Naive Bayes and Random Forest algorithms to calculate sentiment regarding a company. The results of sentiment analysis were used to predict the company stock price. A linear regression method was used to build the prediction model.

Naïve Bayes classifiers in machine learning are a family of simple probabilistic classifiers according to Bayes' theorem, with strong (naive) independence assumptions between the features. Naïve Bayes was developed in 1950, and it was introduced under a different name to the text retrieval community in the early 1960s. It remains a popular (baseline) method for text categorization, considering the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.), with word frequencies as the features. It is competitive in this domain, with more advanced methods including support vector machines, and it also finds application in automatic medical diagnosis.

The authors in [65] explored different methods of improving the accuracy of a Naive Bayes classifier for sentiment analysis. The supervised learning algorithm was used to classify a review document as either positive or negative in [66]. The authors also improved the Naïve Bayes algorithm.

A decision tree is a tool supporting a decision, and it uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Operation research commonly uses decision trees, specifically in decision analysis, to help identify a strategy that is most likely to reach a goal; it is a popular tool in machine learning.

The authors in [67] proposed a new model using C4.5 Algorithm of a decision tree to classify semantics (positive, negative, neutral) for the English documents. A novel model using an ID3 algorithm of a decision tree was used to classify sentiments for the documents in English in [68]. This survey was based on many rules which are generated by applying the ID3 algorithm to 115,000 English sentences of our English training data set.

Dimensionality reduction, or dimension reduction in machine learning and statistics, is the process of reducing the number of random variables under consideration by obtaining a set of principal variables. Dimensionality reduction comprises feature selection and feature extraction.

Naive Bayes and Support Vector Machine were used in [69] to analyze the sentiments of huge amount of tweets generated from Twitter users (they are stored in Twitter database). Unigram and bigram as feature extractors along with Chi2 and Singular Value Decomposition were also used for dimensionality reduction.

A novel, semi-supervised Laplacian eigenmap (SS-LE) was proposed in [70]. Redundant features were removed by decreasing its detection errors of sentiments. It enabled visualization of documents in perceptible, low-dimensional embedded space, to provide a useful tool for

text analytics. The authors evaluated the novel approach by comparing it to other dimensionality reduction methods.

Instance-based learning (memory-based learning) in machine learning is a family of learning algorithms that, instead of performing explicit generalization, compare new problem instances with instances seen in training, which have been stored in memory.

Naive Bayes, Instance Based Learning, Decision Tree, SVM, and IB1 (Instance Based Learning 1) were implemented for sentiment classification of the class of reviews from Rotten Tomatoes in [71].

Clustering data concerns a set of objects processed into classes of similar objects. One cluster is a set of data objects that are similar to each other and are not similar to objects in other clusters. A number of data clusters can be clustered, which can be identified by following experience or can be automatically identified as part of the clustering method. The authors of [72] proposed a new model for big-data sentiment classification in the parallel network environment. The authors' proposed model used the Fuzzy C-Means (FCM) method for English sentiment classification, with Hadoop MAP (M) /REDUCE (R) in Cloudera. The authors [73] proposed a new model for Big Data sentiment classification in the parallel network environment. Our new model uses the STING Algorithm (SA) (in the data mining field) for English document-level sentiment classification with Hadoop Map (M)/Reduce (R), based on the 90,000 English sentences of the training data set in a Cloudera parallel network environment—a distributed system.

Furthermore, many approaches have combined several machine-learning and dictionary-based approaches. The authors in [74] proposed a system for sentiment analysis and classification using NLP, machine-learning technique, and dictionary-based approach; our proposed methodology classifies peoples' sentiments into different polarity classes (positive, negative, and neutral).The main objective of the proposed system is to address and solve the polarity shift problem and to provide feasible solutions to the BOW model in sentiment classification; we achieved that objective by Detecting, Eliminating, and Modifying negation polarity shifter from a given text.

Two main approaches (lexical approach and machine learning) were applied to sentiment analysis in [75]. The lexicon-based method was used to create emotional dictionaries for each domain, as well as the algorithm that calculates the weight of texts. The Maximum Entropy method and the Support Vectors Machines were used in the machine learning approach to create a dictionary and an algorithm for the construction of the feature vector for the Maximum Entropy method.

3. The today tendency of the sentiment analysis

According to a testing data set and a training data set, the opinion classification has been classified into different categories in **Figure 2**.

With the category (1), the authors [49] used two testing data sets in English and they did not use any training data set. Each testing data set has the 25,000 English documents. The authors

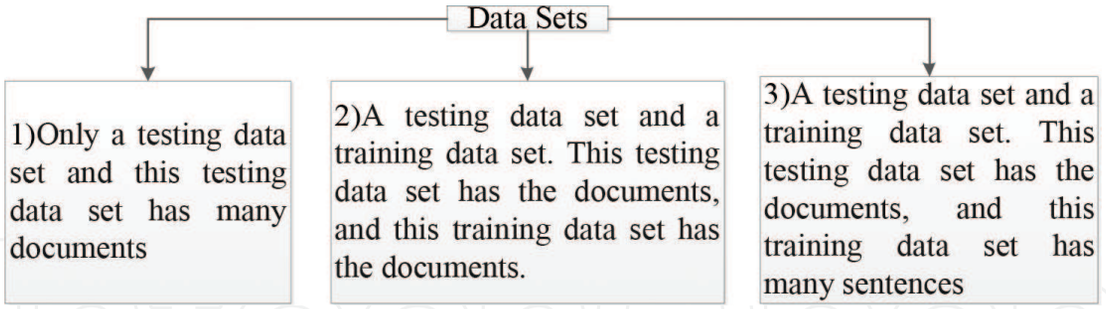

Figure 2. Categories of the sentiment classification based on the data sets.

[51] used one testing data set in Vietnamese and they did not use any training data set. The testing data set has the 30,000 Vietnamese documents. The survey [83] used one testing data set in English and it did not use any training data set. The testing data set has the 5,000,000 English documents.

The category (1) uses the lexicon-based approaches in [1–52, 77]. In addition, category (1) uses a Self-Organizing Map Algorithm—The Self-Organizing Map is based on unsupervised learning.

a. With one document of the testing data set, the SOM is used to cluster all the sentences of this document into either the positive or the negative sections on a map. The sentiment classification of this document is identified completely based on this map. There is no training data set in this category.

b. With many documents of the testing data set, the SOM is used to cluster all the documents into either the positive or the negative sections on a map. The sentiment classification of all the documents is identified completely based on this map. There is no training data set in this category.

Category (1) uses many similarity coefficients (or similarity measures) to classify one document of the testing data set into either the positive polarity or the negative polarity. According to our opinion, all the similarity measures can be used for the sentiment analysis of category (1).

In addition, category (1) also uses many rules for the sentiment classification in [48- 52], in many different languages.

The category (2) has used a testing data set and a training data set. This testing data set has the documents, and this training data set has the documents. The authors [82] used one testing data set including 1,000,000 documents and one training data set comprising 2,000,000 documents in English. This category has used many machine learning algorithms (supervised learning, unsupervised learning, semi-supervised learning, etc.). The authors in [78] use a Machine Learning algorithm, Support Vector Machines, for their sentiment classification. Latent semantic analysis (LSA) has proven to be extremely useful in information retrieval in [79]. A novel approach based on LSA and support vector machine (SVM) aims to improve the sentiment classification performance. Three machine learning approaches (Naive Bayes, maximum entropy classification, and support vector machines) were used for sentiment

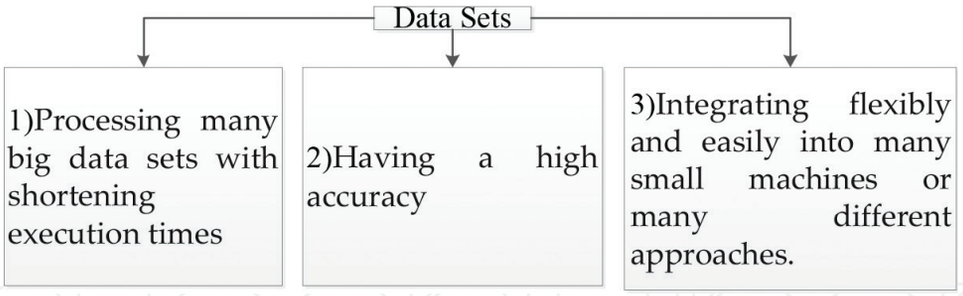

Figure 3. The today tendency of the sentiment analysis.

classification with movie reviews in [80]. The vote algorithm in [81] was used in conjunction with three classifiers, namely Naive Bayes, Support Vector Machine (SVM), and Bagging.

The category (3) uses a testing data set and a training data set. This testing data set has the documents, and this training data set has many sentences. The authors in [67] used one training data set that included 140,000 sentences and two testing data sets in English. Each testing data set has 25,000 documents. The research in [68] used one training data set that included 115,000 sentences and two testing data sets in English. Each testing data set has 25,000 documents. The authors in [72] used one training data set that included 60,000 sentences and two testing data sets in English. Each testing data set had 25,000 documents. The survey in [73] used one training data set that included 90,000 sentences and two testing data sets in English. Each testing data set had 25,000 documents.

This category also uses many machine-learning algorithms (supervised learning, unsupervised learning, semi-supervised learning, etc.). The authors in [67] used a decision tree—a C4.5 algorithm to generate many association rules for English sentiment classification. The authors in [68] also used a decision tree—an ID3 algorithm to generate many association rules for English sentiment classification. The authors in [72, 73] used the clustering algorithms of machine learning to cluster the documents of the testing data set into either the positive polarity or the negative polarity, based on the training data set. The authors in [76] used a SVM algorithm of machine learning to classify the documents of the testing data set into either the positive polarity or the negative polarity, according to the sentences of the training data set.

Paying attention to the current statuses of the economies of the world (we have presented information about big corporations, many documents, etc., in the Introduction section), we show the today tendency of the opinion analysis in **Figure 3**.

1. Processing many big data sets with shortened execution times: As we have presented the information about big corporations, many documents, etc., in the Introduction section, many old approaches (methods or models) cannot process the big data sets with certainty, or they can process the big data sets but only with long times and high costs. The processing of big data sets can be implemented in many parallel network systems. The authors' proposed model in [72] used the Fuzzy C-Means (FCM) method for English sentiment classification, with Hadoop MAP (M) /REDUCE (R) in Cloudera, a parallel network environment. The authors in [73] used a STING Algorithm for English Sentiment Classification

in A Parallel Environment. The authors of [76] used a SVM algorithm for English Semantic Classification in Parallel Environment. Furthermore, lexicon-based approaches can be performed in the distributed network systems with certainty. In the near future, there will be many small machines that can implement the parallel systems. The execution time of the proposed model is dependent on many factors: (1) the parallel network environment, such as the Cloudera system; (2) the distributed functions, such as Hadoop Map (M) and Hadoop Reduce (R); (3) the algorithms in the approach; (4) the performance of the distributed network system; (5) the number of nodes of the parallel network environment; (6) the performance of each node (each server) of the distributed environment; and (7) the sizes of the training data set and the testing data set.

2. Having high accuracy: A high accuracy is crucial for surveys and commercial applications. We can use the works of sentiment classification to cross-check in order to improve their accuracies. The accuracy of the proposed model is dependent on several factors: (1) the algorithms in the approach; (2) the testing data set and the training data set; (3) whether the documents of the testing data set are standardized carefully; and (4) whether the documents (or the sentences) of the training data set are standardized carefully.

3. Integrating flexibly and easily into many small machines or many different approaches: This category is very important for surveys, researchers, and commercial applications. The small machines used in many different fields can be conveniently used anywhere, for any type of users, and for various purposes. These small machines can be produced easily, and can be very cheap and easy to carry. The easy and flexible integration of sentiment classification into the small machines helps save a lot of time and cost. The lexicon-based approaches and the rules-based approaches can be integrated into the small machines, because the small machines have the space to store their data. In addition, the lexicons and the rules can be implemented easily in the small machines. We will not spend much time studying and implementing the surveys that currently exist.

4. Conclusion

In summary, we have presented the dictionary-based approaches and the corpus-based approaches of the sentiment classification basically; and we have also shown the today tendency of the sentiment analysis in more details.

We have displayed the information about the surveys in each section of this chapter. We have also displayed the advantages of the studies in more details.

According to the above proofs and our opinion, three tendencies of the sentiment classification will strongly have developed more and more in the near future because they have the advantages in the different fields and commercial applications.

There will be the surveys developed for the sentiment analysis.

Conflict of interest

We declare that we have no conflict of interest in this chapter.

Notes/Thanks/Other declarations

Thank Dr. Marco Antonio Aceves-Fernandez so much for inviting us to contribute this chapter to the book "Artificial Intelligence."

Author details

Vo Ngoc Phu[1]* and Vo Thi Ngoc Tran[2]

*Address all correspondence to: vongocphu03hca@gmail.com

1 Institute of Research and Development, Duy Tan University – DTU, Da Nang, Vietnam

2 School of Industrial Management (SIM), Ho Chi Minh City University of Technology - HCMUT, Vietnam National University, Ho Chi Minh City, Vietnam

References

[1] Goyal A, Daume III H. Generating semantic orientation lexicon using large data and thesaurus. WASSA '11 Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, Portland, Oregon; 2011. pp. 37-43

[2] Turney P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia; July 2002. p. 417-424

[3] Alena N, Helmut P, Mitsuru I. Recognition of affect, judgment, and appreciation in Text. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Beijing; 2010. pp. 806-14

[4] Palanisamy P, Yadav V, Serendio HE. Simple and practical lexicon based approach to sentiment analysis. Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013), Atlanta, Georgia, June 14-15; 2013. pp. 543-548

[5] Eisenstein J. Unsupervised learning for lexicon-based classification. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), the Hilton San Francisco, San Francisco, California, USA; 2017. pp. 3188-3194

[6] Zhou Z, Zhang X, Sanderson M. (2014) sentiment analysis on twitter through topic-based lexicon expansion. In: Wang H, Sharaf MA, editors. Databases Theory and Applications. ADC 2014. Lecture Notes in Computer Science. Vol. 8506. Cham: Springer; 2014

[7] Augustyniak L, Kajdanowicz T, Szymanski P, Tuligłowicz W, Kazienko P, Alhajj R, Szymanski B. Simpler is better? lexicon-based ensemble sentiment classification beats supervised methods. International Workshop on Curbing Collusive Cyber-gossips in Social Networks (C3-2014). Proc. IEEE/ACM Int. Conf. Advances in Social Network Analysis and Mining, ASONAM, Beijing, China; August 17, 2014

[8] Taboada M, Brooke J, Tofiloski M, Voll K, Stede M. Lexicon-based methods for sentiment analysis. Computational Linguistics. 2010;**37**(2):267-307

[9] Hardeniya T, Borikar DA. An approach to sentiment analysis using lexicons with comparative analysis of different techniques. IOSR Journal of Computer Engineering (IOSR-JCE). 2016;**18**(3):53-57. Ver. I; e-ISSN: 2278-0661,p-ISSN: 2278-8727

[10] Al-Hussaini H, Al-Dossari H. A lexicon-based approach to build service provider reputation from Arabic tweets in twitter. (IJACSA) International Journal of Advanced Computer Science and Applications. 2017;**8**(4)

[11] Musto C, Semeraro G, Polignano M. A comparison of Lexicon-based approaches for Sentiment Analysis of microblog posts. Proceedings of the 8th International Workshop on Information Filtering and Retrieval, Pisa, Italy; December 10th 2014

[12] Hamdan H, Bellot P, Bechet F. lsislif: Feature extraction and label weighting for sentiment in twitter. Proceedings of the 9th International Workshop on Semantic Evaluation, At Denver, Colorado, USA; 2015. p. 568-573

[13] Pan Y, Li X, Shi H, Liu H. Research of methods in sentiment orientation analysis of text based on domain sentiment lexicon. Information Technology Journal. 2014;**13**(9): 1612-1621. DOI: 10.3923/itj.2014.1612.1621

[14] Park S, Kim Y. Building thesaurus lexicon using dictionary-based approach for sentiment classification. 14th IEEE International Conference on Software Engineering Research, Management and Applications (SERA), Towson, MD, USA; 2016. pp. 39-44

[15] Ren F, Matsumoto K. Semi-automatic creation of youth slang corpus and its application to affective computing. IEEE Transactions on Affective Computing. 2016;**7**(2):176-189

[16] Xing L, Yuan L, Qinglin W, Yu L. An approach to sentiment analysis of short Chinese text based on SVMs. 34th IEEE Chinese Control Conference (CCC), China; 2015. pp. 9115-9120

[17] Kundi FM, Ahmed S, Khan A, Asghar MZ. Detection and scoring of internet slangs for sentiment analysis using SentiWordNet. Life Science Journal. 2014;**11**:66-72

[18] Huang S, Han W, Que X, Wang W. Polarity identification of sentiment words based on emoticons. 9th Conference on Computational Intelligence and Security, Emei Mountain, Sichuan Province, China; 2013. pp. 134-138

[19] Dayalani GG. Emoticon based unsupervised sentiment classifier for polarity analysis in tweets. International Journal of Engineering Research and General Science. 2014;**2**:438-445

[20] Xia R, Xu F, Zong C, Li Q, Qi Y, Li T. Dual sentiment analysis: Considering two sides of one review. IEEE Transactions on Knowledge and Data Engineering. 2015;**27**(8):2120-2133

[21] Bai A, Hammer H. Constructing sentiment lexicons in Norwegian from a large text corpus. 2014 IEEE 17th International Conference on Computational Science and Engineering, Chengdu, China; 2014

[22] Turney PD, Littman ML. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. arXiv:cs/0212012, Learning (cs.LG); Information Retrieval (cs.IR); 2002

[23] Malouf R, Mullen T. Graph-based user classification for informal online political discourse. In: Proceedings of the 1st Workshop on Information Credibility on the Web; 2017

[24] Scheible C. Sentiment translation through lexicon induction. Proceedings of the ACL 2010 Student Research Workshop, Sweden; 2010. pp. 25-30

[25] Jovanoski D, Pachovski V, Nakov P. Sentiment analysis in twitter for Macedonian. Proceedings of Recent Advances in Natural Language Processing, Bulgaria; 2015. pp. 249-257

[26] Htait A, Fournier S, Bellot P. LSIS at SemEval-2016 Task 7: Using Web Search Engines for English and Arabic Unsupervised Sentiment Intensity Prediction. Proceedings of SemEval-2016, 2016, California, p. 481-485

[27] Wan X. Co-training for cross-lingual sentiment classification. Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, Singapore; 2009. p. 235-243

[28] Brooke J, Tofiloski M, Taboada M. Cross-linguistic sentiment analysis: From English to Spanish. International Conference RANLP 2009, Borovets, Bulgaria; 2009. pp. 50-54

[29] Jiang T, Jiang J, Dai Y, Li A. Micro–blog emotion orientation analysis algorithm based on Tibetan and Chinese mixed text. International Symposium on Social Science (ISSS 2015); 2015

[30] Tan S, Zhang J. An empirical study of sentiment analysis for Chinese documents. Expert Systems with Applications. 2007;**34**(4):2622-2629. DOI: 10.1016/j.eswa.2007.05.028

[31] Du W, Tan S, Cheng X, Yun X. Adapting Information Bottleneck Method for Automatic Construction of Domain-oriented Sentiment Lexicon. WSDM'10, New York, USA; 2010

[32] Zhang Z, Ye Q, Zheng W, Li Y. Sentiment classification for consumer word-of-mouth in Chinese: Comparison between supervised and unsupervised approaches. The 2010 International Conference on E-Business Intelligence; 2010

[33] Wang G, Araki K. Modifying SO-PMI for Japanese weblog opinion mining by using a balancing factor and detecting neutral expressions. Proceedings of NAACL HLT 2007, Companion Volume, NY; 2007. pp. 189-192

[34] Feng S, Zhang L, Li B, Wang D, Yu v, Wong K-F. Is twitter a better corpus for measuring sentiment similarity? Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, USA; 2013. pp. 897-902

[35] An NTT, Hagiwara M. Adjective-based estimation of short sentence's impression. (KEER2014) Proceedings of the 5th Kanesi Engineering and Emotion Research; International Conference, Sweden; 2014

[36] Shikalgar NR, Dixit AM. JIBCA: Jaccard index based clustering algorithm for mining online review. International Journal of Computer Applications (0975-8887). 2014;**105**(15):1-6

[37] Ji X, Chun SA, Wei Z, Geller J. Twitter sentiment classification for measuring public health concerns. Social Network Analysis and Mining. 2015;**5**:13. DOI: 10.1007/s13278-015-0253-5

[38] Omar N, Albared M, Al-Shabi AQ, Al-Moslmi T. Ensemble of Classification algorithms for subjectivity and sentiment analysis of Arabic Customers' reviews. International Journal of Advancements in Computing Technology (IJACT). 2013;**5**

[39] Mao H, Gao P, Wang Y, Bollen J. Automatic construction of financial semantic orientation lexicon from large-scale Chinese news corpus. 7th financial risks international forum, Institut Louis Bachelier; 2014

[40] Ren Y, Kaji N, Yoshinaga N, Kitsuregaw M. Sentiment classification in under-resourced languages using graph-based semi-supervised learning methods. IEICE Transactions on Information and Systems. 2014;**E97-D**(4):790-797. DOI: 10.1587/Transinf.E97.D.1

[41] Netzer O, Feldman R, Goldenberg J, Fresko M. Mine your own business: Market-structure surveillance through text mining. Marketing Science. 2012;**31**(3):521-543

[42] Ren Y, Kaji N, Yoshinaga N, Toyoda M, Kitsuregawa M. Sentiment classification in resource-scarce languages by using label propagation. Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation, Institute of Digital Enhancement of Cognitive Processing, Waseda University; 2011. pp. 420-429

[43] Alfredo Hernández-Ugalde J, Mora-Urpí J, Rocha OJ. Genetic relationships among wild and cultivated populations of peach palm (Bactris gasipaes Kunth, Palmae): Evidence for multiple independent domestication events. Genetic Resources and Crop Evolution. 2011;**58**(4):571-583

[44] Ponomarenko JV, Bourne PE, Shindyalov IN. Building an automated classification of DNA-binding protein domains. Bioinformatics. 2002;**18**:S192-S201

[45] da Silva Meyer A, Garcia AAF, de Souza AP, de Souza CL Jr. Comparison of similarity coefficients used for cluster analysis with dominant markers in maize (*Zea mays* L). Genetics and Molecular Biology. 2004;**27**(1):83-91

[46] Drinić SM, Nikolić A, Perić V. Cluster analysis of soybean genotypes based on RAPD markers. Proceedings. 43rd Croatian And 3rd International Symposium On Agriculture, Opatija, Croatia; 2008. pp. 367-370

[47] Tamás J, Podani J, Csontos P. An extension of presence/absence coefficients to abundance data: A new look at absence. Journal of Vegetation Science. 2001;**12**:401-410

[48] Phu VN, Chau VTN, Tran VTN, Dat ND. A Vietnamese adjective emotion dictionary based on exploitation of Vietnamese language characteristics. International Journal of Artificial Intelligence Review (AIR). 2017;**47**:67. DOI: 10.1007/CZEKANOWSKI462-017-9538-6

[49] Phu VN, Chau VTN, Dat ND, Tran VTN, Nguyen TA. A valences-totaling model for English sentiment classification. International Journal of Knowledge and Information Systems. 2017;**53**(3):579-636. DOI: 10.1007/CZEKANOWSKI115-017-1054-0

[50] Phu VN, Chau VTN, Tran VTN. Shifting semantic values of English phrases for classification. International Journal of Speech Technology (IJST). 2017;**20**(3):509-533. DOI: 10.1007/CZEKANOWSKI772-017-9420-6

[51] Phu VN, Chau VTN, Tran VTN, Dat ND, Duy KLD. A valence-totaling model for Vietnamese sentiment classification. International Journal of Evolving Systems (EVOS). 2017;**8**:47. https://doi.org/10.1007/s12530-017-9187-7

[52] Vo NP, Vo TNC, Tran VTN, Dat ND, Duy KLD. Semantic lexicons of English nouns for classification. International Journal of Evolving Systems. 2017;**8**:69. DOI: 10.1007/s12530-017-9188-6

[53] Shirani-Mehr H. Applications of Deep Learning to Sentiment Analysis of Movie Reviews. Technical Report. Stanford University; 2014

[54] Vateekul P, Koomsubha T. A study of sentiment analysis using deep learning techniques on Thai twitter data. 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), Khon Kaen, Thailand; 2016. DOI: 10.1109/JCSSE.2016.7748849

[55] Severyn A, Moschitti A. Twitter sentiment analysis with deep convolutional neural networks. SIGIR '15 Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile; 2015. pp. 959-962

[56] Yanmei L, Yuda C. Research on Chinese micro-blog sentiment analysis based on deep learning. 8th Int. Symp. Comput. Intell. Des., Hangzhou, China; 2015. pp. 358-361

[57] You Q, Luo J, Jin H, Yang J. Joint visual-textual sentiment analysis with deep neural networks. MM '15 Proceedings of the 23rd ACM international conference on Multimedia, Brisbane, Australia; 2015. pp. 1071-1074

[58] Xia R, Zong C, Li S. Ensemble of feature sets and classification algorithms for sentiment classification. Information Sciences. 2015;**181**(6):1138-1152. DOI: 10.1016/j.ins.2010.11.023

[59] Su Y, Zhang Y, Ji D, Wang Y, Wu H. Ensemble learning for sentiment classification. In: Ji D, Xiao G, editors. Chinese Lexical Semantics. CLSW 2012. Lecture Notes in Computer Science. Vol. 7717. Berlin, Heidelberg: Springer; 2013

[60] Wan Y, Gao Q. An ensemble sentiment classification system of twitter data for airline services analysis. IEEE International Conference on Data Mining Workshop (ICDMW). Atlantic City, NJ, USA; 2015. DOI: 10.1109/ICDMW.2015.7

[61] Chen L-S, Liu C-H, Chiu H-J. A neural network based approach for sentiment classification in the blogosphere. Journal of Informetrics. 2011;**5**(2):313-322. DOI: 10.1016/j.joi.2011.01.003

[62] Bauer F, Pereverzev S, Rosasco L. On regularization algorithms in learning theory. Journal of Complexity. 2007;**23**(1):52-57. DOI: 10.1016/j.jco.2006.07.001

[63] Onal I, Ertugrul AM. Effect of using regression in sentiment analysis. Signal Processing and Communications Applications Conference (SIU), 2014 22nd, Trabzon, Turkey; 2014. DOI: 10.1109/SIU.2014.6830606

[64] Cakra YE, Trisedya BD. Stock price prediction using linear regression based on sentiment analysis. International Conference on Advanced Computer Science and Information Systems (ICACSIS), Depok, Indonesia; 2015. DOI: 10.1109/ICACSIS.2015.7415179

[65] Narayanan V, Arora I, Bhatia A. Fast and accurate sentiment classification using an enhanced naive bayes model. In: Yin H et al., editors. Intelligent Data Engineering and Automated Learning – IDEAL 2013. IDEAL 2013. Lecture Notes in Computer Science. Vol. 8206. Berlin, Heidelberg: Springer

[66] Kang H, Yoo SJ, Han D. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. Expert Systems with Applications. 2012;**39**(5):6000-6010. DOI: 10.1016/j.eswa.2011.11.107

[67] Phu VN, Ngoc CVT, Ngoc TVT, Duy DN. A C4.5 algorithm for english emotional classification. International Journal of Evolving Systems. 2017;**8**:1-27. DOI: 10.1007/s12530-017-9180-1

[68] Vo NP, Vo TNT, Vo TNC, Dat ND, Duy KLD. A decision tree using ID3 algorithm for English semantic analysis. International Journal of Speech Technology (IJST). 2017;**20**(3):593-613. DOI: 10.1007/s10772-017-9429-x

[69] Shyamasundar LB, Jhansi Rani P. Twitter sentiment analysis with different feature extractors and dimensionality reduction using supervised learning algorithms. India Conference (INDICON), 2016 IEEE Annual, Bangalore, India; 2016. DOI: 10.1109/INDICON.2016.7839075

[70] Kim K, Lee J. Sentiment visualization and classification via semi-supervised nonlinear dimensionality reduction. Pattern Recognition. 2014;**47**(2):758-768. DOI: 10.1016/j.patcog.2013.07.022

[71] Oswin Rahadiyan H, Virginia G, Antonius Rachmat C. Sentiment Classification of Film Reviews Using IB1. 7th International Conference on Intelligent Systems, Modeling and Simulation (ISMS). Bangkok, Thailand; 2016. DOI: 10.1109/ISMS.2016.38

[72] Phu VN, Dat ND, Tran VTN, Chau VTN, Nguyen TA. Fuzzy C-means for English sentiment classification in a distributed system. International Journal of Applied Intelligence (APIN). 2017;**46**(3):717-738. DOI: 10.1007/s10489-016-0858-z

[73] Dat ND, Phu VN, Chau VTN, Tran VTN, Nguyen TA. STING algorithm used English sentiment classification in a parallel environment. International Journal of Pattern Recognition and Artificial Intelligence. 2017;**31**(7):30. DOI: 10.1142/S0218001417500215

[74] Kolekar NV, Rao G, Dey S, Mane M, Jadhav V, Patil S. Sentiment analysis and classification using lexicon-based approach and addressing polarity shift problem. Journal of Theoretical and Applied Information Technology. 2016;**90**(1):1-8

[75] Blinov PD, Klekovkina MV, Kotelnikov EV, Pestov OA. Research of lexical approach and machine learning methods for sentiment analysis. Proceedings of. Dialogos. 2013;**2**:51-61

[76] Vo NP, Vo TNC, Vo TNT. SVM for English semantic classification in parallel environment. International Journal of Speech Technology (IJST). 2017;**20**(3):487-508. DOI: 10.1007/s10772-017-9421-5

[77] Vo NP, Phan TT. Sentiment classification using Enhanced Contextual Valence Shifters. International Conference on Asian Language Processing (IALP), Kuching, Malaysia; 2014. DOI: 10.1109/IALP.2014.6973485

[78] Kennedy A, Inkpen D. Sentiment classification of movie reviews using contextual valence shifters. Computational Intelligence. 2006;**22**(2):110-125. DOI: 10.1111/J.1467-8640.2006.00277.X

[79] Wang L, Wan Y. Sentiment classification of documents based on latent semantic analysis. In: Lin S, Huang X, editors. Advanced Research on Computer Education, Simulation and Modeling. Communications in Computer and Information Science. Vol. 176. Berlin, Heidelberg: Springer; 2011

[80] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. Proceedings of EMNLP. 2002:79-86

[81] Catal C, Nangir M. A sentiment classification model based on multiple classifiers. Applied Soft Computing. 2017;**50**:135-141. DOI: 10.1016/j.asoc.2016.11.022

[82] Vo NP, Vo TNT. A STING algorithm and multi-dimensional vectors used for English sentiment classification in a distributed system. American Journal of Engineering and Applied Sciences. 2017;**12**:1-19. DOI: 10.3844/ajeassp.2017

[83] Vo NP, Vo TNT. English sentiment classification using only the sentiment lexicons with a JOHNSON coefficient in a parallel network environment. American Journal of Engineering and Applied Sciences. 2017;**12**:1-28. DOI: 10.3844/ajeassp.201

IntechOpen