American Journal of Engineering and Applied Sciences

Original Research Paper

English Sentiment Classification using Only the Sentiment Lexicons with a JOHNSON Coefficient in a Parallel Network Environment

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

[1]Nguyen Tat Thanh University, 300A Nguyen Tat Thanh Street,
Ward 13, District 4, Ho Chi Minh City, 702000, Vietnam
[2]School of Industrial Management (SIM),
Ho Chi Minh City University of Technology - HCMUT, Vietnam National University, Ho Chi Minh City, Vietnam

Article history
Received: 03-11-2017
Revised: 02-12-2017
Accepted: 20-12-2017

Corresponding Author:
Vo Ngoc Phu
Nguyen Tat Thanh University,
300A Nguyen Tat Thanh
Street, Ward 13, District 4, Ho
Chi Minh City, 702000,
Vietnam
Email: vongocphu@ntt.edu.vn
vongocphu03hca@gmail.com

Abstract: Sentiment classification is significant in everyday life, such as in political activities, commodity production and commercial activities. In this survey, we have proposed a new model for Big Data sentiment classification. We use many sentiment lexicons of our basis English Sentiment Dictionary (bESD) to classify 5,000,000 documents including 2,500,000 positive and 2,500,000 negative of our testing data set in English. We do not use any training data set in English. We do not use any one-dimensional vector in both a sequential environment and a distributed network system. We also do not use any multi-dimensional vector in both a sequential system and a parallel network environment. We use a JOHNSON Coefficient (JC) through a Google search engine with AND operator and OR operator to identify many sentiment values of the sentiment lexicons of the bESD in English. One term (a word or a phrase in English) is clustered into either the positive polarity or the negative polarity if this term is very close to either the positive or the negative by using many similarity measures of the JC. It means that this term is very similar to either the positive or the negative. We tested the proposed model in both a sequential environment and a distributed network system. We achieved 87.56% accuracy of the testing data set. The execution time of the model in the parallel network environment is faster than the execution time of the model in the sequential system. Our new model can classify sentiment of millions of English documents based on the sentiment lexicons of the bESD in a parallel network environment. The proposed model is not depending on both any special domain and any training stage. This survey used many similarity coefficients of a data mining field. The results of this work can be widely used in applications and research of the English sentiment classification.

Keywords: English Sentiment Classification, Distributed System, Parallel System, JOHNSON Coefficient, Cloudera, Hadoop Map and Hadoop Reduce, Sentiment Lexicons

Introduction

Clustering data is to process a set of objects into classes of similar objects. One cluster is a set of data objects which 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 following experience or can be automatically identified as part of clustering method.

The aim of this survey is to find a new approach to improve the accuracy of the sentiment classification results and to shorten the execution time of the proposed model with a low cost.

The motivation of this new model is as follows: JOHNSON Coefficient (JC) can be applied to a distributed network environment. Cosine, Ochiai, Sorensen, Tanimoto, PMI and Jaccard measures are used popularly to calculate the emotional values of the words. Thus, other similar

measures can be used to identify the semantic scores of the words. Many algorithms in the data mining field can be applied to natural language processing, specifically semantic classification for processing millions of English documents. This will result in many discoveries in scientific research, hence the motivation for this study.

We show the novelty and originality of our novel approach as follows:

1. The JOHNSON Coefficient (JC) was applied to the sentiment classification
2. This can also be applied to identify the sentiments of millions of documents.
3. We did not use any training data set in English
4. We only used a testing data set in English
5. We did not use any one-dimensional vector in both a sequential environment and a distributed network system
6. We did use any multi-dimensional vector in both a sequential system and a parallel network environment
7. We used a JOHNSON Coefficient (JC) through a Google search engine with AND operator and OR operator to identify many sentiment values of the sentiment lexicons of the bESD in English
8. The input of this survey is the documents of the testing data set in English. We studied to transfer the documents into the formats for the novel model which can process them
9. This survey can be applied to other parallel network systems such as a Cloudera distributed environment
10. The Cloudera system, Hadoop Map (M) and Hadoop Reduce (R) were used in the proposed model
11. The novel model can be applied to other parallel functions such as Hadoop Map (M) and Hadoop Reduce (R)
12. We tested the proposed model in both a sequential environment and a distributed network system
13. The JC - related equations were built in this survey
14. We proposed the algorithms in both a sequential environment and distributed network system

Therefore, we have studied this model in more details.

To get higher accuracy of the results of the sentiment classification and shorten execution time of the sentiment classification, we did not transfer one sentence into one one-dimensional vector based on VSM (Singh and Singh, 2015; Carrera-Trejo *et al*., 2015; Soucy and Mineau, 2015) in both the sequential system and the distributed system. We also do not transfer one sentence into one one-dimensional vector based on many sentiment lexicons of our basis English Sentiment Dictionary (bESD). We did not transfer one document into one multi-dimensional vector based on VSM (Singh and Singh, 2015; Carrera-Trejo *et al*., 2015; Soucy and Mineau, 2015). We also did not transfer one document into one multi-dimensional vector based on the sentiment lexicons of our basis English Sentiment Dictionary (bESD). We create many sentiment lexicons of our basis English Sentiment Dictionary (bESD) and the valences and the sentiment polarity of the sentiment lexicons of the bESD are calculated by using the JC through a Google search engine with AND operator and OR operator. One term (one word or phrase in English) is the positive polarity if this term is very close to the positive (the term is very similar to the positive). One term is the negative polarity if this term is very close to the negative (the term is very similar to the negative). One term is the neutral polarity if this term is not very close to both the positive and the negative (the term is not very similar to both the positive and the negative).

The term is very close to the positive if a similarity measure (by using the JC) between this term and the positive polarity is greater than a similarity measure (by using the JC) between this term and the negative polarity. Thus, the term is clustered to the positive.

The term is very close to the negative if a similarity coefficient (by using the JC) between this term and the positive polarity is less than a similarity coefficient (by using the JC) between this term and the negative polarity. Therefore, the term is clustered into the negative.

The term is very close to the neutral if a similarity measure (by using the JC) between this term and the positive polarity is as equal as a similarity measure (by using the JC) between this term and the negative polarity. Thus, the term is not clustered to both the positive and the negative. The term is certainly the neutral polarity.

One sentence in English is the positive polarity if a total of terms clustered into the positive is greater than a total of terms clustered into the negative in this sentence.

One sentence in English is the negative polarity if a total of terms clustered into the positive is less than a total of terms clustered into the negative in this sentence.

One sentence in English is the neutral polarity if a total of terms clustered into the positive is as equal as a total of terms clustered into the negative in this sentence.

One document in English is the positive polarity if the number of sentences clustered into the positive is greater than the number of sentences clustered into the negative in this document.

One document in English is the negative polarity if the number of sentences clustered into the positive is less than the number of sentences clustered into the negative in this document.

One document in English is the neutral polarity if the number of sentences clustered into the positive is as equal as the number of sentences clustered into the negative in this document.

We perform the proposed model as follows: First of all, we calculate the valences of the sentiment lexicons of the bESD by using the JC through the Google search engine with AND operator and OR operator. With each sentence of one document of the testing data set, we split this sentence into the meaningful terms (meaningful words

or meaningful phrases) based on the bESD. Each term in the meaningful terms of one sentence of one document of the testing data set, we calculate the sentiment score of this term based on the bESD. This term belongs to the positive group if this valence is greater than 0. The term belongs to the negative group if the sentiment value is less than 0. The term belongs to the neutral group if the sentiment score is as equal as 0. One sentence is clustered into the positive group if a total of the valences of all the meaningful terms is greater than 0 in this sentence. One sentence is clustered into the negative group if a total of the sentiment scores of all the meaningful terms is less than 0 in this sentence. One sentence is clustered into the neutral group if a total of the sentiment values of all the meaningful terms is as equal as 0 in this sentence. One document of the testing data set is clustered into the positive group if the number of the sentences clustered into the positive is greater than the number of the sentences clustered into the negative in the document. One document of the testing data set is clustered into the negative group if the number of the sentences clustered into the positive is less than the number of the sentences clustered into the negative in the document. One document of the testing data set is clustered into the neutral group if the number of the sentences clustered into the positive is as equal as the number of the sentences clustered into the negative in the document.

We perform all the above things in the sequential system firstly. To shorten execution time of the proposed model, we implement all the above things in the distributed environment secondly.

Our model has many significant applications to many areas of research as well as commercial applications as follows:

1. Many surveys and commercial applications can use the results of this work in a significant way
2. JC is used in identifying opinion scores of the English verb phrases and words through the Google search on the internet
3. The formulas are proposed in the paper
4. The algorithms are built in the proposed model
5. This survey can certainly be applied to other languages easily
6. The results of this study can significantly be applied to the types of other words in English
7. Many crucial contributions are listed in the Future Work section
8. The algorithm of data mining is applicable to semantic analysis of natural language processing
9. This study also proves that different fields of scientific research can be related in many ways
10. Millions of English documents are successfully processed for emotional analysis
11. The semantic classification is implemented in the parallel network environment
12. The principles are proposed in the research
13. The Cloudera distributed environment is used in this study
14. The proposed work can be applied to other distributed systems
15. This survey uses Hadoop Map (M) and Hadoop Reduce (R)
16. Our proposed model can be applied to many different parallel network environments such as a Cloudera system
17. This study can be applied to many different distributed functions such as Hadoop Map (M) and Hadoop Reduce (R)

We also compare this novel model's results with the latest sentiment classification models in (Agarwal and Mittal, 2016a; 2016b; Canuto *et al*., 2016; Ahmed and Danti, 2016; Phu and Tuoi, 2014; Tran *et al*., 2014; Dat *et al*., 2017; Phu *et al*., 2017f; 2017g; 2017h)

This study contains 6 sections. Section 1introduces the study; section 2 discusses the related works about the JOHNSON Coefficient (JC), etc.; section 3 is about the English data set; section 4 represents the methodology of our proposed model; section 5 represents the experiment. Section 6 provides the conclusion. The References section comprises all the reference documents; all tables are shown in the Appendices section.

Related Work

We summarize many researches which are related to our research. By far, we know that Pointwise Mutual Information (PMI) equation and Sentiment Orientation (SO) equation are used for determining polarity of one word (or one phrase) and strength of sentiment orientation of this word (or this phrase). Jaccard Measure (JM) is also used for calculating polarity of one word and the equations from this Jaccard measure are also used for calculating strength of sentiment orientation this word in other research. PMI, Jaccard, Cosine, Ochiai, Tanimoto and Sorensen measure are the similarity measure between two words; from those, we prove that the JOHNSON Coefficient (JC) is also used for identifying valence and polarity of one English word (or one English phrase). Finally, we identify the sentimental values of English verb phrases based on the basis English semantic lexicons of the basis English Emotional Dictionary (bESD).

There are the works related to PMI measure in (Bai *et al*., 2014; Turney and Littman, 2002; Malouf and Mullen, 2017; Scheible, 2010; Jovanoski *et al*., 2015; Htait *et al*., 2016; Wan *et al*., 2009; Brooke *et al*., 2009; Jiang *et al*., 2015; Brooke *et al*., 2009; Hernández-Ugalde *et al*., 2011; Ponomarenko *et al*., 2002; Meyer *et al*., 2004; Mladenović Drinić *et al*., 2008; Tamás *et al*., 2001). In the research (Bai *et al*., 2014), the authors generate several Norwegian sentiment lexicons by

extracting sentiment information from two different types of Norwegian text corpus, namely, news corpus and discussion forums. The methodology is based on the Point wise Mutual Information (PMI). The authors introduce a modification of the PMI that considers small "blocks" of the text instead of the text as a whole. The study in (Turney and Littman, 2002) introduces a simple algorithm for unsupervised learning of semantic orientation from extremely large corpora, etc.

Two studies related to the PMI measure and Jaccard measure are in (Feng *et al*., 2013; An and Hagiwara, 2014). In the survey (Feng *et al*., 2013), the authors empirically evaluate the performance of different corpora in sentiment similarity measurement, which is the fundamental task for word polarity classification. The research in (An and Hagiwara, 2014) proposes a new method to estimate impression of short sentences considering adjectives. In the proposed system, first, an input sentence is analyzed and preprocessed to obtain keywords. Next, adjectives are taken out from the data which is queried from Google N-gram corpus using keywords-based templates.

The works related to the Jaccard measure are in (Shikalgar and Dixit, 2014; Ji *et al*., 2015; Omar *et al*., 2013; Mao *et al*., 2014; Ren *et al*., 2014; Netzer *et al*., 2012; Ren *et al*., 2011). The survey in (Shikalgar and Dixit, 2014) investigates the problem of sentiment analysis of the online review. In the study (Ji *et al*., 2015), the authors are addressing the issue of spreading public concern about epidemics. Public concern about a communicable disease can be seen as a problem of its own, etc.

The surveys related the similarity coefficients to calculate the valences of words are in (Phu *et al*., 2017a; 2017b; 2017c; 2017d; 2017e).

The English dictionaries are (EDL, 2017; OED, 2017; CED, 2017a; LED, 2017; CED, 2017b; MMED, 2017) and there are more than 55,000 English words (including English nouns, English adjectives, English verbs, etc.) from them.

There are the works related to the JOHNSON Coefficient (JC) in (Choi *et al*., 2010; Wilk *et al*., 2002; Tulloss, 1997; Dalirsefat *et al*., 2009; Wijaya *et al*., 2016; Duarte *et al*., 1999). The authors in (Choi *et al*., 2010) collected 76 binary similarity and distance measures used over the last century and reveal their correlations through the hierarchical clustering technique, etc.

There are the works related to vector space modeling in (Singh and Singh, 2015; Carrera-Trejo *et al*., 2015; Soucy and Mineau, 2015). In this study (Singh and Singh, 2015), the authors will be Examining the Vector Space Model, an Information Retrieval technique and its variation. In this survey (Carrera-Trejo *et al*., 2015), the authors consider multi-label text classification task and apply various feature sets. The authors consider a subset of multi-labeled files from the Reuters-21578 corpus. The authors use traditional tf-IDF values of the features and tried both considering and ignoring stop words. The authors also tried several combinations of features, like bigrams and unigrams. The authors in (Soucy and Mineau, 2015) introduce a new weighting method based on statistical estimation of the importance of a word for a specific categorization problem. This method also has the benefit to make feature selection implicit, since useless features for the categorization problem considered to get a very small weight.

The latest researches of the sentiment classification are (Agarwal and Mittal, 2016a; 2016b; Canuto *et al*., 2016; Ahmed and Danti, 2016; Phu and Tuoi, 2014; Tran *et al*., 2014; Dat *et al*., 2017; Phu *et al*., 2016). In the research (Agarwal and Mittal, 2016a), the authors present their machine learning experiments with regard to sentiment analysis in blog, review and forum texts found on the World Wide Web and written in English, Dutch and French. The survey in (Agarwal and Mittal, 2016b) discusses an approach where an exposed stream of tweets from the Twitter micro blogging site are preprocessed and classified based on their sentiments. In sentiment classification system the concept of opinion subjectivity has been accounted. In the study, the authors present opinion detection and organization subsystem, which have already been integrated into our larger question-answering system, etc.

Data Set

In Fig. 1, the testing data set includes 5,000,000 documents in the movie field, which contains 2,500,000 positive documents and 2,500,000 negative documents in English. All the documents in our English testing data set are automatically extracted from English Facebook, English websites and social networks; then we labeled positive and negative for them.

Methodology

This section comprises two parts: The first part is to create the sentiment lexicons in English in both a sequential environment and a distributed system in the sub-section (4.1). The second part is to use the sentiment-lexicons with the JC to classify the documents of the testing data set into either the positive vector group or the negative vector group in both a sequential environment and a distributed system in the sub-section (4.2).

In the sub-section (4.1), the section includes three parts. The first sub-section of this section is to identify a sentiment value of one word (or one phrase) in English in the sub-section (4.1.1). The second part of this section is to create a basis English Sentiment Dictionary (bESD) in a sequential system in the sub-section (4.1.2). The third sub-section of this section is to create a basis English Sentiment Dictionary (bESD) in a parallel environment in the sub-section (4.1.3).

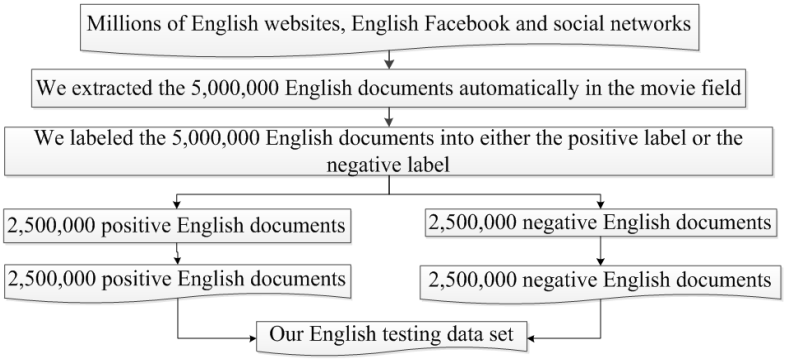

Fig. 1: Our English testing data set

In the sub-section (4.2), the section comprises two parts. The first part of this section is to use the sentiment-lexicons with the JC to classify the documents of the testing data set into either the positive vector group or the negative vector group in a sequential environment in the sub-section (4.2.1). The second part of this section is to use the sentiment-lexicons with the JC to classify the documents of the testing data set into either the positive vector group or the negative vector group in a distributed system in the sub-section (4.2.2).

Creating the Sentiment Lexicons in English

The section includes three parts: The first sub-section of this section is to identify a sentiment value of one word (or one phrase) in English in the sub-section (4.1.1). The second part of this section is to create a basis English Sentiment Dictionary (bESD) in a sequential system in the sub-section (4.1.2). The third sub-section of this section is to create a basis English sentiment dictionary (bESD) in a parallel environment in the sub-section (4.1.3).

Calculating a Valence of One Word (or One Phrase) in English

In this part, we calculate the valence and the polarity of one English word (or phrase) by using the JC through a Google search engine with AND operator and OR operator, as the following diagram in Fig. 2.

According to (Bai *et al*., 2014; Turney and Littman, 2002; Malouf and Mullen, 2017; Scheible, 2010; Jovanoski *et al*., 2015; Htait *et al*., 2016; Wan *et al*., 2009; Brooke *et al*., 2009; Jiang *et al*., 2015; Brooke *et al*., 2009; Hernández-Ugalde *et al*., 2011; Ponomarenko *et al*., 2002; Meyer *et al*., 2004; Mladenović Drinić *et al*., 2008; Tamás *et al*., 2001), Pointwise Mutual Information (PMI) between two words *wi* and *wj* has the equation:

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

and sentiment orientation (SO) of word *wi* has the equation:

$$SO(wi) = PMI(wi, positive) - PMI(wi, negative) \quad (2)$$

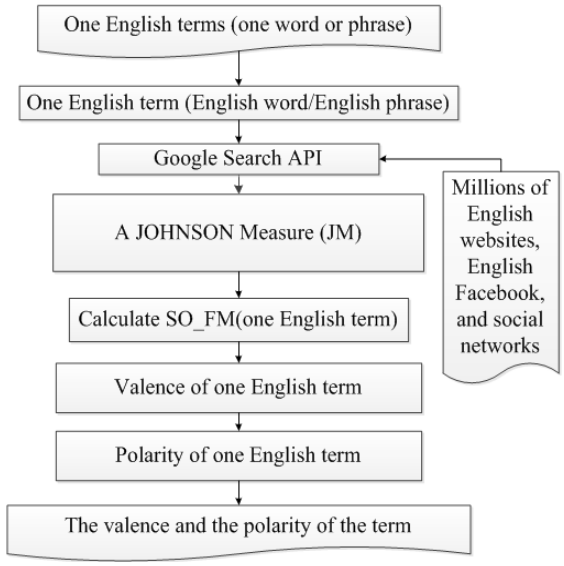

Fig. 2: Overview of identifying the valence and the polarity of one term in English using a JOHNSON Coefficient (JC)

Bai *et al*. (2014; Turney and Littman, 2002; Malouf and Mullen, 2017; Scheible, 2010; Jovanoski *et al*., 2015; Htait *et al*., 2016; Wan *et al*., 2009; Brooke *et al*., 2009) the positive and the negative of Equation 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 is used in the PMI equations of (Turney and Littman, 2002; Malouf and Mullen, 2017; Jovanoski *et al*., 2015) and the Google search engine is used in the PMI equations of (Scheible, 2010; Htait *et al*., 2016; Brooke *et al*., 2009). Besides, (Scheible, 2010) also uses German, (Jovanoski *et al*., 2015) also uses Macedonian, (Htait *et al*., 2016) also uses Arabic, (Wan *et al*., 2009) also uses Chinese and (Brooke *et al*., 2009) also uses Spanish. In addition, the Bing search engine is also used in (Htait *et al*., 2016).

With (Jiang *et al*., 2015; Tan and Zhang, 2007; Du *et al*., 2010; Zhang *et al*., 2010), the PMI equations

are used in Chinese, not English and Tibetan is also added in (Jiang *et al*., 2015). About the search engine, the AltaVista search engine is used in (Du *et al*., 2010) and (Zhang *et al*., 2010) and uses three search engines, such as the Google search engine, the Yahoo search engine and the Baidu search engine. The PMI equations are also used in Japanese with the Google search engine in (Wang and Araki, 2007). Feng *et al*. (2013) and (An and Hagiwara, 2014) also use the PMI equations and Jaccard equations with the Google search engine in English.

According to (Feng *et al*., 2013; An and Hagiwara, 2014; Shikalgar and Dixit, 2014; Ji *et al*., 2015; Omar *et al*., 2013; Mao *et al*., 2014; Ren *et al*., 2014; Netzer *et al*., 2012; Ren *et al*., 2011), Jaccard between two words *wi* and *wj* has the equations:

$$Jaccard(wi,wj) = J(wi,wj) = \frac{|wi \cap wj|}{|wi \cup wj|} \quad (3)$$

and other type of the Jaccard equation between two words *wi* and *wj* has the equation:

$$\begin{aligned} &Jaccard(wi,wj) = J(wi,wj) \\ &= sim(wi,wj) = \frac{F(wi,wj)}{F(wi)+F(wj)-F(wi,wj)} \end{aligned} \quad (4)$$

and Sentiment Orientation (SO) of word *wi* has the equation:

$$SO(wi) = \sum Sim(wi, positive) - \sum Sim(wi, positive) \quad (5)$$

In (Feng *et al*., 2013; An and Hagiwara, 2014; Shikalgar and Dixit, 2014; Ji *et al*., 2015; Omar *et al*., 2013; Mao *et al*., 2014; Ren *et al*., 2014; Netzer *et al*., 2012) the positive and the negative of Equation 5 in English are: Positive = {good, nice, excellent, positive, fortunate, correct, superior} and negative = {bad, nasty, poor, negative, unfortunate, wrong, inferior}.

The Jaccard equations with the Google search engine in English are used in (Feng *et al*., 2013; An and Hagiwara, 2014; Ji *et al*., 2015). (Shikalgar and Dixit, 2014) and (Netzer *et al*., 2012) use the Jaccard equations in English. (Ren *et al*., 2014) and (Ren *et al*., 2011) use the Jaccard equations in Chinese. (Omar *et al*., 2013) uses the Jaccard equations in Arabic. The Jaccard equations with the Chinese search engine in Chinese are used in (Mao *et al*., 2014).

The authors in (Phu *et al*., 2017a) used the Ochiai Measure through the Google search engine with AND operator and OR operator to calculate the sentiment values of the words in Vietnamese. The authors in (Phu *et al*., 2017b) used the Cosin Measure through the Google search engine with AND operator and OR operator to identify the sentiment scores of the words in English. The authors in (Phu *et al*., 2017c) used the Sorensen Coefficient through the Google search engine with AND operator and OR operator to calculate the sentiment values of the words in English. The authors in (Phu *et al*., 2017d) used the Jaccard Measure through the Google search engine with AND operator and OR operator to calculate the sentiment values of the words in Vietnamese. The authors in (Phu *et al*., 2017e) used the Tanimoto Coefficient through the Google search engine with AND operator and OR operator to identify the sentiment scores of the words in English

With the above proofs, we have the information as follows: PMI is used with AltaVista in English, Chinese and Japanese with the Google in English; Jaccard is used with the Google in English, Chinese and Vietnamse. The Ochiai is used with the Google in Vietnamese. The Cosin and Sorensen are used with the Google in English.

According to (Bai *et al*., 2014; Turney and Littman, 2002; Malouf and Mullen, 2017; Scheible, 2010; Jovanoski *et al*., 2015; Htait *et al*., 2016; Wan *et al*., 2009; Brooke *et al*., 2009; Jiang *et al*., 2015; Hernández-Ugalde *et al*., 2011; Ponomarenko *et al*., 2002; Meyer *et al*., 2004; Mladenović Drinić *et al*., 2008; Tamás *et al*., 2001), PMI, Jaccard, Cosine, Ochiai, Sorensen, Tanimoto and JOHNSON Coefficient (JC) are the similarity measures between two words and they can perform the same functions and with the same characteristics; so JC is used in calculating the valence of the words. In addition, we prove that JC can be used in identifying the valence of the English word through the Google search with the AND operator and OR operator.

With the JOHNSON Coefficient (JC) in (Choi *et al*., 2010; Wilk *et al*., 2002; Tulloss, 1997; Dalirsefat *et al*., 2009; Wijaya *et al*., 2016; Duarte *et al*., 1999), we have the equation of the JC:

$$\begin{aligned} &JOHNSON\ Coefficient(a,b) \\ &= JOHNSON\ Measure(a,b) = JC(a,b) \\ &= \frac{(a \cap b)}{(a \cap b)+(\neg a \cap b)} + \frac{(a \cap b)}{(a \cap b)+(a \cap \neg b)} \end{aligned} \quad (6)$$

with *a* and *b* are the vectors.

From the Equation 1 to 6, we propose many new equations of the JC to calculate the valence and the polarity of the English words (or the English phrases) through the Google search engine as the following equations below.

In Equation 6, when a has only one element, *a* is a word. When *b* has only one element, *b* is a word. In Equation 6, *a* is replaced by *w*1 and *b* is replaced by *w*2:

$$\begin{aligned} &JOHNSON\ Measure(w1,w2) \\ &= JOHNSON\ Coefficient(w1,w2) = JC(w1,w2) \\ &= \frac{P(w1,w2)}{P(w1,w2)+P(\neg w1,w2)} + \frac{P(w1,w2)}{P(w1,w2)+P(w1,\neg w2)} \end{aligned} \quad (7)$$

DOI: 10.3844/ajeassp.2018.38.65

Equation 7 is similar to Equation 1. In Equation 2, Equation 1 is replaced by Equation 7. We have Equation 8 as follows:

$$\begin{aligned} &Valence(w) = SO\_JC(w) \\ &= JC(w, positive\_query) - JC(w, negative\_query) \end{aligned} \tag{8}$$

In Equation 7, $w1$ is replaced by w and $w2$ is replaced by position_query. We have Equation 9 as follows:

$$\begin{aligned} &JC(w, positive\_query) \\ &= \frac{P(w, positive\_query)}{P(w, positive\_query) + P(\neg w, positive\_query)} \\ &+ \frac{P(w, positive\_query)}{P(w, positive\_query) + P(w, \neg positive\_query)} \end{aligned} \tag{9}$$

In Equation 7, $w1$ is replaced by w and $w2$ is replaced by negative_query. We have Equation 10) as follows:

$$\begin{aligned} &JC(w, negative\_query) \\ &= \frac{P(w, positive\_query)}{P(w, positive\_query) + P(\neg w, positive\_query)} \\ &+ \frac{P(w, positive\_query)}{P(w, positive\_query) + P(w, \neg positive\_query)} \end{aligned} \tag{10}$$

with:

- w, $w1$, $w2$: Are the English words (or the English phrases)
- $P(w1, w2)$: Number of returned results in Google search by keyword ($w1$ and $w2$). We use the Google Search API to get the number of returned results in search online Google by keyword ($w1$ and $w2$)
- $P(w1)$: number of returned results in Google search by keyword $w1$. We use the Google Search API to get the number of returned results in search online Google by keyword $w1$
- $P(w2)$: Number of returned results in Google search by keyword $w2$. We use the Google Search API to get the number of returned results in search online Google by keyword $w2$
- Valence(W) = SO_JC(w): Valence of English word (or English phrase) w; is SO of word (or phrase) by using the JOHNSON Coefficient (JC)
- Positive_query: {active or good or positive or beautiful or strong or nice or excellent or fortunate or correct or superior}
- With the positive query is the a group of the positive English words
- Negative_query: {passive or bad or negative or ugly or week or nasty or poor or unfortunate or wrong or inferior}
- With the negative_query is the a group of the negative English words
- $P(w$, positive_query): Number of returned results in Google search by keyword (positive_query and w). We use the Google Search API to get the number of returned results in search online Google by keyword (positive_query and w)
- $P(w$, negative_query): Number of returned results in Google search by keyword (negative_query and w). We use the Google Search API to get the number of returned results in search online Google by keyword (negative_query and w)
- $P(w)$: number of returned results in Google search by keyword w. We use the Google Search API to get the number of returned results in search online Google by keyword w
- $P(\neg$w,positive_query): Number of returned results in Google search by keyword ((not w) and positive_query). We use the Google Search API to get the number of returned results in search online Google by keyword ((not w) and positive_query)
- $P(w$, ¬positive_query): Number of returned results in the Google search by keyword (w and (not (positive_query))). We use the Google Search API to get the number of returned results in search online Google by keyword (w and [not (positive_query)])
- $P(\neg w$, ¬positive_query): Number of returned results in the Google search by keyword (w and (not (positive_query))). We use the Google Search API to get the number of returned results in search online Google by keyword ((not w) and [not (positive_query)])
- $P(\neg$w,negative_query): Number of returned results in Google search by keyword ((not w) and negative_query). We use the Google Search API to get the number of returned results in search online Google by keyword ((not w) and negative_query)
- $P(w,\neg$negative_query): Number of returned results in the Google search by keyword (w and (not (negative_query))). We use the Google Search API to get the number of returned results in search online Google by keyword (w and (not (negative_query)))
- $P(\neg$w,¬negative_query): Number of returned results in the Google search by keyword (w and (not (negative_query))). We use the Google Search API to get the number of returned results in search online Google by keyword ((not w) and (not (negative_query)))

As like Cosine, Ochiai, Sorensen, Tanimoto, PMI and Jaccard about calculating the valence (score) of the word, we identify the valence (score) of the English word w based on both the proximity of positive_query with w and the remote of positive_query with w; and the proximity of negative_query with w and the remote of

negative_query with *w*. The English word w is the nearest of positive_query if JC(w, positive_query) is as equal as 1. The English word w is the farthest of positive_query if JC(w, positive_query) is as equal as 0. The English word w belongs to positive_query being the positive group of the English words if JC(w, positive_query) > 0 and JC(w, positive_query) ≤ 1.The English word w is the nearest of negative_query if JC(w, negative_query) is as equal as 1. The English word w is the farthest of negative_query if JC(w, negative_query) is as equal as 0. The English word w belongs to negative_query being the negative group of the English words if JC(w, negative_query) > 0 and JC(w, negative_query) ≤ 1. So, the valence of the English word w is the value of JC(w, positive_query) substracting the value of JC(w, negative_query) and the Equation 8 is the equation of identifying the valence of the English word *w*.

We have the information about JC as follows:

- JC(w, positive_query) ≥ 0 and JC(w, positive_query) ≤ 1
- JC(w, negative_query) ≥ 0 and JC (w, negative_query) ≤ 1
- If JC(w, positive_query) = 0 and JC (w, negative_query) = 0 then SO_JC (w) = 0
- If JC(w, positive_query) = 1 and JC (w, negative_query) = 0 then SO_JC (w) = 0
- If JC(w, positive_query) = 0 and JC (w, negative_query) = 1 then SO_JC (w) = -1
- If JC(w, positive_query) = 1 and JC (w, negative_query) = 1 then SO_JC(w) = 0

So, SO_JC (w) ≥ -1 and SO_JC (w) ≤ 1.

The polarity of the English word w is positive polarity If SO_JC (w) > 0. The polarity of the English word w is negative polarity if SO_JC (w) < 0. The polarity of the English word w is neutral polarity if SO_JC (w) = 0. In addition, the semantic value of the English word w is SO_JC (w).

We calculate the valence and the polarity of the English word or phrase w using a training corpus of approximately one hundred billion English words - the subset of the English Web that is indexed by the Google search engine on the internet. AltaVista was chosen because it has a NEAR operator. The AltaVista NEAR operator limits the search to documents that contain the words within ten words of one another, in either order. We use the Google search engine which does not have a NEAR operator; but the Google search engine can use the AND operator and the OR operator. The result of calculating the valence w (English word) is similar to the result of calculating valence *w* by using AltaVista. However, AltaVista is no longer.

In summary, by using Equation 8 to 10, we identify the valence and the polarity of one word (or one phrase) in English by using the SC through the Google search engine with AND operator and OR operator.

In Table 1, we present the comparisons of our model's advantages and disadvantages with the works related to (Bai *et al*., 2014; Turney and Littman, 2002; Malouf and Mullen, 2017; Scheible, 2010; Jovanoski *et al*., 2015; Htait *et al*., 2016; Wan *et al*., 2009; Brooke *et al*., 2009; Jiang *et al*., 2015).

The comparisons of our model's benefits and drawbacks with the studies related to the JOHNSON coefficient (JC) in (Choi *et al*., 2010; Wilk *et al*., 2002; Tulloss, 1997; Dalirsefat *et al*., 2009; Wijaya *et al*., 2016; Duarte *et al*., 1999) are displayed in Table 2.

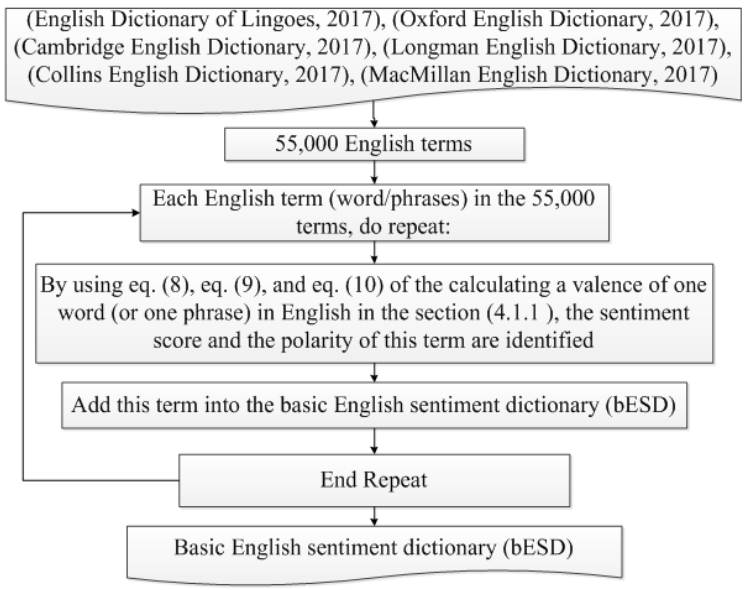

Fig. 3: Overview of creating a basis English Sentiment Dictionary (bESD) in a sequential environment

Table 1: Comparisons of our model's advantages and disadvantages with the works related to (Bai *et al*., 2014; Turney and Littman, 2002; Malouf and Mullen, 2017; Scheible, 2010; Jovanoski *et al*., 2015; Htait *et al*., 2016; Wan *et al*., 2009; Brooke *et al*., 2009; Jiang *et al*., 2015; Brooke *et al*., 2009; Hernández-Ugalde *et al*., 2011; Ponomarenko *et al*., 2002; Meyer *et al*., 2004; Mladenović Drinić *et al*., 2008; Tamás *et al*., 2001)

Surveys	Approach	Advantages	Disadvantages
Bai *et al*. (2014)	Constructing sentiment lexicons in Norwegian from a large text corpus	Through the authors'PMI computations in this survey they used a distance of 100 words from the seed word, but it might be that other lengths that generate better sentiment lexicons. Some of the authors' preliminary research showed that 100 gave a better result.	The authors need to investigate this more closely to find the optimal distance. Another factor that has not been investigated much in the literature is the selection of seed words. Since they are the basisfor PMI calculation, it might be a lot to gain by finding better seed words. The authors would like to explore the impact that different approaches to seed word selection have on the performance of the developed sentiment lexicons.
Turney and Littman (2002)	Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus.	This survey has presented a general strategy for learning semantic orientation from semantic association, SO-A. Two instances of this strategy have been empirically evaluated, SO-PMI-IR andSO-LSA. SO-PMI-IR requires a large corpus, but it is simple,easy to implement, unsupervised and it is not restricted to adjectives.	No Mention
Malouf and Mullen (2017)	Graph-based user classification for informal online political discourse	The authors describe several experiments in identifying the political orientation of posters in an informal environment. The authors' results indicate that the most promising approach is to augment text classification methods by exploiting information about how posters interact with each other	There is still much left to investigate in terms of optimizing the linguistic analysis, beginning with spelling correction and working up to shallow parsing and co-reference identification. Likewise, it will also be worthwhile to further investigate exploiting sentiment values of phrases and clauses, taking cues from methods
Scheible (2010)	Anovel, graph-based approach using SimRank.	The authors presented a novel approach to the translation of sentiment information that outperforms SOPMI, an established method. In particular, the authors could show that SimRank outperforms SO-PMI for values of the threshold x in an interval that most likely leads to the correct separation of positive, neutral and negative adjectives.	The authors' future work will include a further examination of the merits of its application for knowledge-sparse languages.
Jovanoski *et al*. (2015)	Analysis in Twitter for Macedonian	The authors' experimental results show an F1-score of 92.16, which is very strong and is on par with the best results for English, which were achieved in recent SemEval competitions.	In future work, the authors are interested in studying the impact of the raw corpus size, e.g., the authors could only collect half a million tweets for creating lexicon sand analyzing/ evaluating the system. Moreover, the authors are interested not only in quantity but also in quality, i.e., in studying the quality of the individual words and phrases used as seeds.
Htait *et al*. (2016)	Using Web Search Engines for English and Arabic Unsupervised Sentiment Intensity Prediction	- For the General English sub-task, the authors' system has modest but interesting results. - For the Mixed Polarity English sub-task, the authors' system results achieve the second place. - For the Arabic phrases sub-task, the authors' system has very interesting results since they applied the unsupervised method only	Although the results are encouraging, further investigation is required, in both languages, concerning the choice of positive and negative words which once associated to a phrase, they make it more negative or more positive.

Table 1: Continue

Wan *et al*. (2009)	Co-Training for Cross-Lingual Sentiment Classification	The authors propose a co-training approach to making use of unlabeled Chinese data. Experimental results show the effectiveness of the proposed approach, which can outperform the standard inductive classifiers and the transductive classifiers.	In future work, the authors will improve the sentiment classification accuracy in the following two ways: (1) The smoothed co-training approach will be adopted for sentiment classification. (2) The authors will employ the Structural Correspondence Learning (SCL) domain adaption algorithm for linking the translated text and the natural text.
Brooke *et al*. (2009)	Cross-Linguistic Sentiment Analysis: From English to Spanish	Our Spanish SO Calculator (SOCAL) is clearly inferior to the authors' English SO-CAL, probably the result of a number of factors, including a small, preliminary dictionary and a need for additional adaptation to a new language. Translating our English dictionary also seems to result in significant semantic loss, at least for original Spanish texts.	No Mention
Jiang *et al*. (2015)	Micro-blog Emotion Orientation Analysis Algorithm Based on Tibetan and Chinese Mixed Text	By emotion orientation analyzing and studying of Tibetan microblog which is concerned in Sina, making Tibetan Chinese emotion dictionary, Chinese sentences, Tibetan part of speech sequence and emotion symbol as emotion factors and using expected cross entropy combined fuzzy set to do feature selection to realize a kind of microblog emotion orientation analyzing algorithm based on Tibetan and Chinese mixed text. The experimental results showed that the method can obtain better performance in Tibetan and Chinese mixed Microblog orientation analysis.	No Mention
Tan and Zhang (2007)	An empirical study of sentiment analysis for Chinese documents	Four feature selection methods (MI, IG, CHI and DF) and five learning methods (centroid classifier, K-nearest neighbor, winnow classifier, Naïve Bayes and SVM) are investigated on a Chinese sentiment corpus with a size of 1021 documents. The experimental results indicate that IG performs the best for sentimental terms selection and SVM exhibits the best performance for sentiment classification. Furthermore, the authors found that sentiment classifiers are severely dependent on domains or topics.	No Mention
Du *et al*. (2010)	Adapting Information Bottleneck Method for Automatic Construction of Domain-oriented Sentiment Lexicon	The authors' theory verifies the convergence property of the proposed method. The empirical results also support the authors' theoretical analysis. In their experiment, it is shown that proposed method greatly outperforms the baseline methods in the task of building out-of-domain sentiment lexicon.	In this study, only the mutual information measure is employed to measure the three kinds of relationship. In order to show the robustness of the framework, the authors' future effort is to investigate how to integrate more measures into this framework.
Zhang *et al*. (2010)	Sentiment Classification for Consumer Word-of-Mouth in Chinese: Comparison between Supervised and Unsupervised Approaches	This study adopts three supervised learning approaches and a web-based semantic orientation approach, PMI-IR, to Chinese reviews. The results show that SVM outperforms naive bayes and N-gram model on various sizes of training examples, but does not obviously exceeds the semantic orientation approach when the number of training examples is smaller than 300.	No Mention

Table 1: Continue

Wang and Araki (2007)	Modifying SO-PMI for Japanese Weblog Opinion Mining by Using a Balancing Factor and Detecting Neutral Expressions	After these modifications, the authors achieved a well-balanced result: Both positive and negative accuracy exceeded 70%. This shows that the authors' proposed approach not only adapted the SO-PMI for Japanese, but also modified it to analyze Japanese opinions more effectively.	In the future, the authors will evaluate different choices of words for the sets of positive and negative reference words. The authors also plan to appraise their proposal on other languages.
Feng *et al*. (2013)	In this survey, the authors empirically evaluate the performance of different corpora in sentiment similarity measurement, which is the fundamental task for word polarity classification.	Experiment results show that the Twitter data can achieve a much better performance than the Google, Web1T and Wikipedia based methods.	No Mention
An and Hagiwara (2014)	Adjective-Based Estimation of Short Sentence's Impression	The adjectives are ranked and top an adjectives are considered as an output of system. For example, the experiments were carried out and got fairly good results. With the input "it is snowy", the results are white (0.70), light (0.49), cold (0.43), solid (0.38) and scenic (0.37)	In the authors' future work, they will improve more in the tasks of keyword extraction and semantic similarity methods to make the proposed system working well with complex inputs.
Shikalgar and Dixit (2014)	Jaccard Index based Clustering Algorithm for Mining Online Review	In this study, the problem of predicting sales performance using sentiment information mined from reviews is studied and a novel JIBCA Algorithm is proposed and mathematically modeled. The outcome of this generates knowledge from mined data that can be useful for forecasting sales.	For future work, by using this framework, it can extend it to predicting sales performance in the other domains like customer electronics, mobile phones, computers based on the user reviews posted on the websites, etc.
Ji *et al*. (2015)	Twitter sentiment classification for measuring public health concerns	Based on the number of tweets classified as Personal Negative, the authors compute a Measure of Concern (MOC) and a timeline of the MOC. We attempt to correlate peaks of the MOC timeline to the peaks of the News (Non-Personal) timeline. The authors' best accuracy results are achieved using the two-step method with a Naïve Bayes classifier for the Epidemic domain (six datasets) and the Mental Health domain (three datasets).	No Mention
Omar *et al*. (2013)	Ensemble of Classification algorithms for Subjectivity and Sentiment Analysis of Arabic Customers' Reviews	The experimental results show that the ensemble of the classifiers improves the classification effectiveness in terms of macro-F1 for both levels. The best results obtained from the subjectivity analysis and the sentiment classification in terms of macro-F1 are 97.13% and 90.95% respectively.	No Mention
Mao *et al*. (2014)	Automatic Construction of Financial Semantic Orientation Lexicon from Large-Scale Chinese News Corpus	Semantic orientation lexicon of positive and negative words is indispensable for sentiment analysis. However, many lexicons are manually created by a small number of human subjects, which are susceptible to high cost and bias. In this survey, the authors propose a novel idea to construct a financial semantic orientation lexicon from large-scale Chinese news corpus automatically...	No Mention
Ren *et al*. (2014)	Sentiment Classification in Under-Resourced Languages UsingGraph-based Semi-supervised Learning Methods	In particular, the authors found that choosing initially labeled vertices in a JC or dance with their degree and PageRank score can improve the performance. However, pruning unreliable edges will make things more difficult to predict. The authors believe that other people who are interested in this field can benefit from their empirical findings.	As future work, first, the authors will attempt to use a sophisticated approach to induce better sentiment features. The authors consider such elaborated features improve the classification performance, especially in the book domain. The authors also plan to exploit a much larger amount of unlabeled data to fully take advantage of SSL algorithms

Table 1: Continue

Netzer *et al*. (2012)	A text-mining approach and combine it with semantic network analysis tools	In summary, the authors hope the text-mining and derived market-structure analysis presented in this study provides a first step in exploring the extremely large, rich and useful body of consumer data readily available on Web 2.0.	No Mention
Ren *et al*. (2011)	Sentiment Classification in Resource-Scarce Languages by using Label Propagation	The authors compared our method with supervised learning and semi-supervised learning methods on real Chinese reviews classification in three domains. Experimental results demonstrated that label propagation showed a competitive performance against SVM or Transductive SVM with best hyper-parameter settings. Considering the difficulty of tuning hyper-parameters in a resource scarce setting, the stable performance of parameter-free label propagation is promising.	The authors plan to further improve the performance of LP in sentiment classification, especially when the authors only have a small number of labeled seeds. The authors will exploit the idea of restricting the label propagating steps when the available labeled data is quite small.
Phu *et al*. (2017a)	A Vietnamese adjective emotion dictionary based on exploitation of Vietnamese language characteristics	The Vietnamese adjectives often bear emotion which values (or semantic scores) are not fixed and are changed when they appear indifferent contexts of these phrases. Therefore, if the Vietnamese adjectives bring sentiment and their semantic values (or their sentiment scores) are not changed in any context, then the results of the emotion classification are not high accuracy. The authors propose many rules based on Vietnamese language characteristics to determine the emotional values of the Vietnamese adjective phrases bearing sentiment in specific contexts. The authors' Vietnamese sentiment adjective dictionary is widely used in applications and researches of the Vietnamese semantic classification.	not calculating all Vietnamese words completely; not identifying all Vietnamese adjective phrases fully, etc.
Phu *et al*. (2017b)	A Valences-Totaling Model for English Sentiment Classification	The authors present a full range of English sentences; thus, the emotion expressed in the English text is classified with more precision. The authors new model is not dependent on a special domain and training data set-it is a domain-independent classifier. The authors test our new model on the Internet data in English. The calculated valence (and polarity) of English semantic words in this model is based on many documents on millions of English Web sites and English social networks.	It has low accuracy; it misses many sentiment-bearing English words; it misses many sentiment-bearing English phrases because sometimes the valence of a English phrase is not the total of the valences of the English words in this phrase; it misses many English sentences which are not processed fully; and it misses many English documents which are not processed fully.
Phu *et al*. (2017c)	Shifting Semantic Values of English Phrases for Classification	The results of the sentiment classification are not high accuracy if the English phrases bring the emotions and their semantic values (or their sentiment scores) are not changed in any context. For those reasons, the authors propose many rules based on English language grammars to calculate the sentimental values of the English phrases bearing emotion in their specific contexts. The results of this work are widely used in applications and researches of the English semantic classification.	This survey is only applied to the English adverb phrases. The proposed model is needed to research more and more for the different types of the English words such as English noun, English adverbs, etc.

Table 1: Continue

Phu *et al*. (2017d)	A Valence-Totaling Model for Vietnamese Sentiment Classification	The authors have used the VTMfV to classify 30,000 Vietnamese documents which include the 15,000 positive Vietnamese documents and the 15,000 negative Vietnamese documents. The authors have achieved accuracy in 63.9% of the authors' Vietnamese testing data set. VTMfV is not dependent on the special domain. VTMfV is also not dependent on the training data set and there is no training stage in this VTMfV. From the authors' results in this study, our VTMfV can be applied in the different fields of the Vietnamese natural language processing. In addition, the authors' TCMfV can be applied to many other languages such as Spanish, Korean, etc. It can also be applied to the big data set sentiment classification in Vietnamese and can classify millions of the Vietnamese documents	it has a low accuracy.
Phu *et al*. (2017e)	Semantic Lexicons of English Nouns for Classification	The proposed rules based on English language grammars to calculate the sentimental values of the English phrases bearing emotion in their specific contexts. The results of the sentiment classification are not high accuracy if the English phrases bring the emotions and their semantic values (or their sentiment scores) are not changed in any context. The valences of the English words (or the English phrases) are identified by using Tanimoto Coefficient (TC) through the Google search engine with AND operator and OR operator. The emotional values of the English noun phrases are based on the English grammars (English language characteristics)	This survey is only applied in the English noun phrases. The proposed model is needed to research more and more about the different types of the English words such as English adverbs, etc.
Our work	-We use the sentiment-lexicons with the JC to classify one document of the testing data set into either the positive polarity or the negative polarity in both the sequential environment and the distributed system. -The advantages and disadvantages of this survey are shown in the Conclusion section.		

Table 2: Comparisons of our model's benefits and drawbacks with the studies related to the JOHNSON Coefficient (JC) in (Choi *et al*., 2010; Wilk *et al*., 2002; Tulloss, 1997; Dalirsefat *et al*., 2009; Wijaya *et al*., 2016; Duarte *et al*., 1999).

Surveys	Approach	Benefits	Drawbacks
Choi *et al*. (2010)	A Survey of Binary Similarity and Distance Measures	Applying appropriate measures results in more accurate data analysis. Not with standing, few comprehensive surveys on binary measures have been conducted. Hence the authors collected 76 binary similarity and distance measures used over the last century and reveal their correlations through the hierarchical clustering technique	No mention
Wilk *et al*. (2002)	Test-Retest Stability of the Repeatable Battery for the Assessment of Neuropsychological Status in Schizophrenia	The RBANS demonstrated reasonable intraclass correlation coefficient test- retest reliability for both schizophrenia patients and healthy comparison subjects. Confidence intervals are comparable to those previously published for the WAIS-R and Wechsler Memory Scale-Revised, suggesting that retest measurement error is not dramatically increased in the RBANS, despite the brevity of the test. These data may serve as an informative guide for using the RBANS to evaluate neuropsychological change on the level of the individual subject.	No mention
Tulloss (1997)	Assessment of Similarity Indices for Undesirable Properties and a new Tripartite Similarity Index Based on Cost Functions	The purpose of this study is to motivate, describe and offer an implementation for, a working similarity index that avoids the difficulties noted for the others.	No mention

Table 2: Continue

Dalirsefat *et al*. (2009)	Comparison of Similarity Coefficients used for Cluster Analysis with Amplified Fragment Length Polymorphism Markers in the Silkworm, Bombyx mori	The results demonstrated that for almost all methodologies, the Jaccard and Sorensen-Dice coefficients revealed extremely close results, because both of them exclude negative co-occurrences. Due to the fact that there is no guarantee that the DNA regions with negative co occurrences between two strains are indeed identical, the use of coefficients such as Jaccard and Sorensen-Dice that do not include negative co-occurrences was imperative for closely related organisms.	No mention
Wijaya *et al*. (2016)	Finding an appropriate equation to measure similarity between binary vectors: case studies on Indonesian and Japanese herbal medicines	The selection of binary similarity and dissimilarity measures for multivariate analysis is data dependent. The proposed method can be used to find the most suitable binary similarity and dissimilarity equation wisely for a particular data. Our finding suggests that all four types of matching quantities in the Operational Taxonomic Unit (OTU) table are important to calculate the similarity and dissimilarity coefficients between herbal medicine formulas. Also, the binary similarity and dissimilarity measures that include the negative match quantity *d* achieve better capability to separate herbal medicine pairs compared to equations that exclude *d*.	No mention
Duarte *et al*. (1999)	Comparison of similarity coefficients based on RAPD markers in the common been	The employment of different similarity coefficients caused few alterations in cultivar classification, since correlations among genetic distances were larger than0.86. Nevertheless, the different similarity coefficients altered the projection efficiency in a two-dimensional space and formed different numbers of groups by Tocher's optimization procedure. Among these coefficients, Russel and Rao's was the most discordant and the Sorensen- Dice was considered the most adequate due to a higher projection efficiency in a two-dimensional space. Even though few structural changes were suggested in the most different groups, these coefficients altered some relationships between cultivars with high genetic similarity.	No mention
Our work	-We use the sentiment-lexicons with the JC to classify one document of the testing data set into either the positive polarity or the negative polarity in both the sequential environment and the distributed system. -The advantages and disadvantages of this survey are shown in the Conclusion section.		

Creating a basis English Sentiment Dictionary (bESD) in a Sequential Environment

According to (EDL, 2017; OED, 2017; CED, 2017a; LED, 2017; CED, 2017b; MMED, 2017), we have at least 55,000 English terms, including nouns, verbs, adjectives, etc. In this part, we calculate the valence and the polarity of the English words or phrases for our basis English Sentiment Dictionary (bESD) by using the JC in a sequential system, as the following diagram in Fig. 3.

We propose the algorithm 1 to perform this section. The main ideas of the algorithm 1 are as follows:

Input: the 55,000 English terms; the Google search engine
Output: a basis English sentiment dictionary (bESD)
Step 1: Each term in the 55,000 terms, do repeat:
Step 2: By using Equation 8 to 10 of the calculating a valence of one word (or one phrase) in English in the section (4.1.1), the sentiment score and the polarity of this term are identified. The valence and the polarity are calculated by using the JC through the Google search engine with AND operator and OR operator.
Step 3: Add this term into the basis English Sentiment Dictionary (bESD);
Step 4: End Repeat - End Step 1;
Step 5: Return bESD;

Our basis English Sentiment Dictionary (bEED) has more 55,000 English words (or English phrases) and bESD is stored in Microsoft SQL Server 2008 R2.

Creating a basis English Sentiment Dictionary (bESD) in a Distributed System

According to (EDL, 2017; OED, 2017; CED, 2017a; LED, 2017; CED, 2017b; MMED, 2017), we have at least 55,000 English terms, including nouns, verbs, adjectives, etc. In this part, we calculate the valence and the polarity of the English words or phrases for our basis English Sentiment Dictionary (bESD) by using the JC in a parallel network environment, as the following diagram in Fig. 4.

In Fig. 4, this section includes two phases: The Hadoop Map (M) phase and the Hadoop Reduce (R) phase. The input of the Hadoop Map phase is the 55,000 terms in English in (EDL, 2017; OED, 2017; CED, 2017a; LED, 2017; CED, 2017b; MMED, 2017). The output of the Hadoop Map phase is one term which the sentiment score and the polarity are identified. The output of the Hadoop Map phase is the input of the Hadoop Reduce phase. Thus, the input of the Hadoop Reduce phase is one term which the sentiment score and the polarity are identified. The output of the Hadoop Reduce phase is the basis English Sentiment Dictionary (bESD).

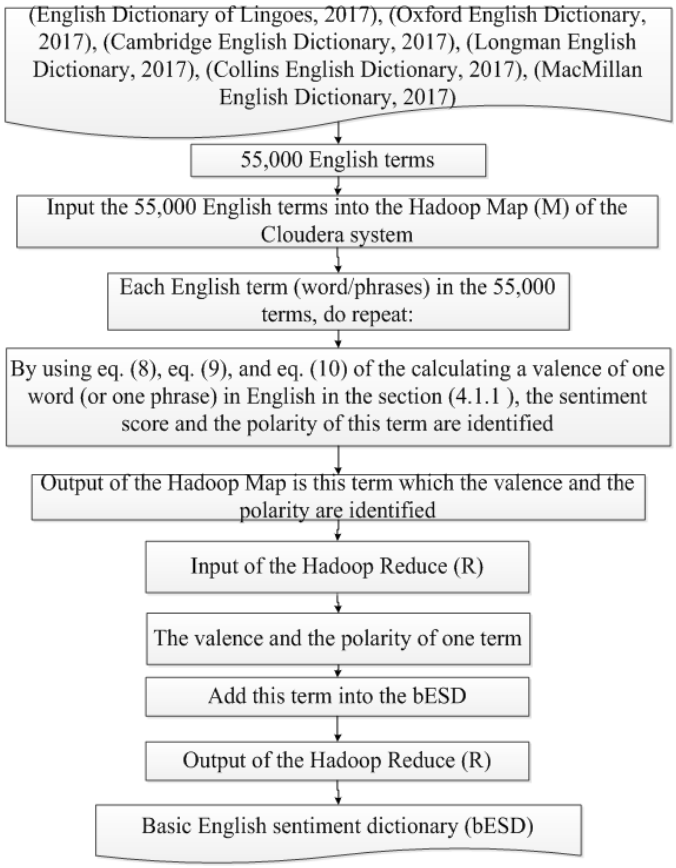

Fig. 4: Overview of creating a basis English Sentiment Dictionary (bESD) in a distributed environment

We build the algorithm 2 to implement the Hadoop Map phase. The main ideas of the algorithm 2 are as follows:

Input: the 55,000 English terms; the Google search engine
Output: one term which the sentiment score and the polarity are identified.
Step 1: Each term in the 55,000 terms, do repeat:
Step 2: By using Equation 8 to 10 of the calculating a valence of one word (or one phrase) in English in the section (4.1.1), the sentiment score and the polarity of this term are identified. The valence and the polarity are calculated by using the JC through the Google search engine with AND operator and OR operator.
Step 3: Return this term;

We proposed the algorithm 3 to perform the Hadoop Reduce phase. The main ideas of the algorithm 3 are as follows:

Input: one term which the sentiment score and the polarity are identified – The output of the Hadoop Map phase.
Output: a basis English sentiment dictionary (bESD)
Step 1: Add this term into the basis English sentiment dictionary (bESD);
Step 2: Return bESD;

Our basis English sentiment dictionary (bEED) has more 55,000 English words (or English phrases) and bESD is stored in Microsoft SQL Server 2008 R2.

Using the Sentiment-Lexicons with the JC to Classify the Documents of the Testing Data Set into Either the Positive Polarity or the Negative Polarity

This section comprises two parts. The first part of this section is to use the sentiment-lexicons with the JC to classify the documents of the testing data set into either the positive polarity or the negative polarity in a sequential environment in the sub-section (4.2.1). The second part of this section is to use the sentiment-lexicons with the JC to classify the documents of the testing data set into either the positive polarity or the negative polarity in a distributed system in the sub-section (4.2.2).

Using the Sentiment-Lexicons with the JC to Classify the Documents of the Testing Data Set into Either the Positive Polarity or the Negative Polarity in a Sequential Environment

In Fig. 5, we use the sentiment-lexicons with the JC to classify the documents of the testing data set into either the positive polarity or the negative polarity in the sequential environment as follows.

DOI: 10.3844/ajeassp.2018.38.65

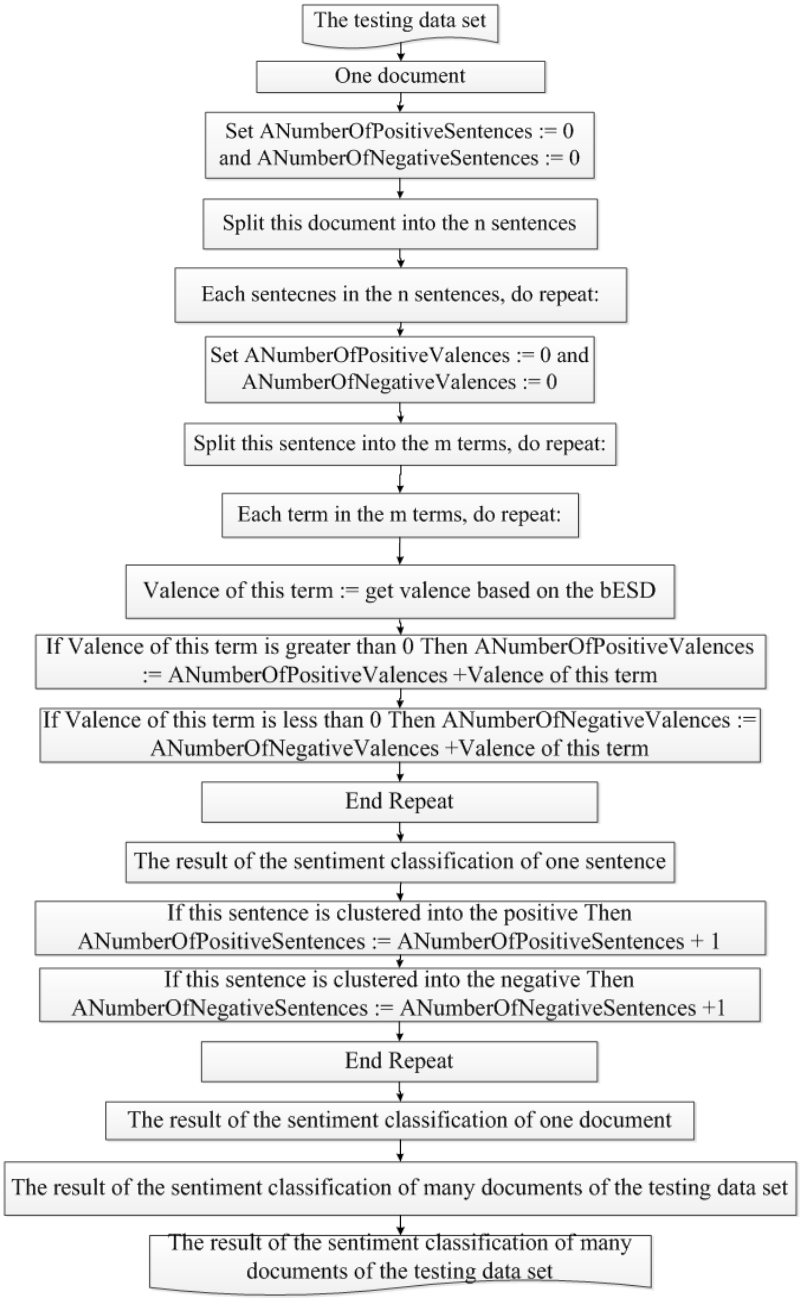

Fig. 5: Overview of using the sentiment-lexicons with the JC to classify the documents of the testing data set into either the positive polarity or the negative polarity in the sequential environment

This section is performed in the sequential system as follows: Firstly, we create the sentiment lexicons of the basis English Sentiment Dictionary (bESD) based on the creating a basis English Sentiment Dictionary (bESD) in a sequential environment in (4.1.2). Each document in the documents of the testing data set, we split this document into the n sentences. Each sentence in the n sentences, we split this sentence into the m meaningful terms based on the bESD. Each term in the m terms, we identify the sentiment score of this term based on the bESD. The sentiment polarity of this sentence is based on a total of all the valences of all the terms in the sentence. This sentence is clustered into the positive if a total of all the sentiment values of all the terms clustered into the positive is greater than a total of all the sentiment scores of all the terms clustered into the negative in the sentence. This sentence is clustered into the negative if a total of all the sentiment values of all

the terms clustered into the positive are less than a total of all the sentiment scores of all the terms clustered into the negative in the sentence. This sentence is clustered into the neutral if a total of all the sentiment values of all the terms clustered into the positive are as equal as a total of all the sentiment scores of all the terms clustered into the negative in the sentence. The sentiment polarity of this document is based on the number of all the sentences clustered into either the positive or the negative. The document is clustered into the positive if the number of the sentences clustered into the positive is greater than the number of the sentences clustered into the negative in the document. The document is clustered into the negative if the number of the sentences clustered into the positive is less than the number of the sentences clustered into the negative in the document. The document is clustered into the neutral if the number of the sentences clustered into the positive is as equal as the number of the sentences clustered into the negative in the document.

We propose the algorithm 4 to cluster one sentence into either the positive or the negative in the sequential system. The main ideas of the algorithm 4 are as follows:

Input: one sentence
Output: the sentiment polarity (positive, negative, neutral)
Step 1: Split this sentence into m meaningful terms (meaningful words or meaningful phrases) based on the bESD;
Step 2: Set ANumberOfPositiveValences := 0 and ANumberOfNegativeValences := 0
Step 3: Each term in the m terms, do repeat:
Step 4: Valence := get valence of this term based on the bESD;
Step 5: If Valence is greater than 0 Then ANumberOfPositiveValences := ANumberOfPositiveValences + Valence;
Step 6: Else If Valence is less than 0 Then ANumberOfNegativeValences := ANumberOfNegativeValences + Valence;
Step 7: End Repeat – End Step 3;
Step 8: If ANumberOfPositiveValences is greater than ANumberOfPositiveValences Then Return positive;
Step 9: Else If ANumberOfPositiveValences is less than ANumberOfPositiveValences Then Return negative;
Step 10: Return neutral;

We propose the algorithm 5 to cluster on document into either the positive or the negative in the sequential environment. The main ideas of the algorithm 5 are as follows:

Input: one document
Output: the sentiment polarity (positive, negative, neutral)
Step 1: Split this document into n sentences;
Step 2: Set ANumberOfPositiveSentences := 0 and ANumberOfNegativeSentences := 0
Step 3: Each sentence in the n sentences terms, do repeat:
Step 4: Polarity := Algorithm 4 with the input is this sentence;
Step 5: If Polarity is positive Then ANumberOfPositiveSentences := ANumberOfPositiveSentences + 1;
Step 6: Else If Polartiy is negative Then ANumberOfNegativeSentences := ANumberOfNegativeSentences + 1;
Step 7: End Repeat – End Step 3;
Step 8: If ANumberOfPositiveSentences is greater than ANumberOfNegativeSentences Then Return positive;
Step 9: Else If ANumberOfPositiveSentences is less than ANumberOfNegativeSentences Then Return negative;
Step 10: Return neutral;

We propose the algorithm 6 to cluster all the documents of the testing data set into either the positive or the negative in the sequential system. The main ideas of the algorithm 6 are as follows:

Input: the testing data set
Output: the sentiment polarity (positive, negative, neutral)
Step 1: Each document in the documents of the testing data set, do repeat:
Step 2: Polarity := Algorithm 5 with the input is this document;
Step 3: End Repeat – End Step 1;

Using the Sentiment-Lexicons with the JC to Classify the Documents of the Testing Data Set into Either the Positive Polarity or the Negative Polarity in a Parallel System

In Fig. 6, we use the sentiment-lexicons with the JC to classify the documents of the testing data set into either the positive polarity or the negative polarity in the distributed environment as follows.

This section is performed in the parallel system as follows: Firstly, we create the sentiment lexicons of the basis English sentiment dictionary (bESD) based the creating a basis English sentiment dictionary (bESD) in a distributed system in (4.1.3). Each document in the documents of the testing data set, we split this document into the n sentences. Each sentence in the n sentences, we split this sentence into the m meaningful terms based on the bESD. Each term in the m terms, we identify the sentiment score of this term based on the bESD. The sentiment polarity of this sentence is based on a total of all the valences of all the terms in the sentence. This sentence is clustered into the positive if a total of all the sentiment values of all the terms clustered into the positive is greater than a total of all the sentiment scores

of all the terms clustered into the negative in the sentence. This sentence is clustered into the negative if a total of all the sentiment values of all the terms clustered into the positive is less than a total of all the sentiment scores of all the terms clustered into the negative in the sentence. This sentence is clustered into the neutral if a total of all the sentiment values of all the terms clustered into the positive is as equal as a total of all the sentiment scores of all the terms clustered into the negative in the sentence. The sentiment polarity of this document is based on the number of all the sentences clustered into either the positive or the negative. The document is clustered into the positive if the number of the sentences clustered into the positive is greater than the number of the sentences clustered into the negative in the document. The document is clustered into the negative if the number of the sentences clustered into the positive is less than the number of the sentences clustered into the negative in the document. The document is clustered into the neutral if the number of the sentences clustered into the positive is as equal as the number of the sentences clustered into the negative in the document.

In Fig 7, we propose the algorithm 7 and the algorithm 8 to cluster one sentence into either the positive or the negative in the parallel system. This stage includes two phases: the Hadoop Map (M) phase and the Hadoop Reduce (R). The input of the Hadoop Map is one sentence and the bESD. The output of the Hadoop Map phase is one term which the valence is identified based on the bESD. The input of the Hadoop Reduce (R) is the output of the Hadoop Map phase, thus, the input of the Hadoop Reduce is one term. The output of the Hadoop Reduce phase is the sentiment polarity (positive, negative, neutral) of this sentence.

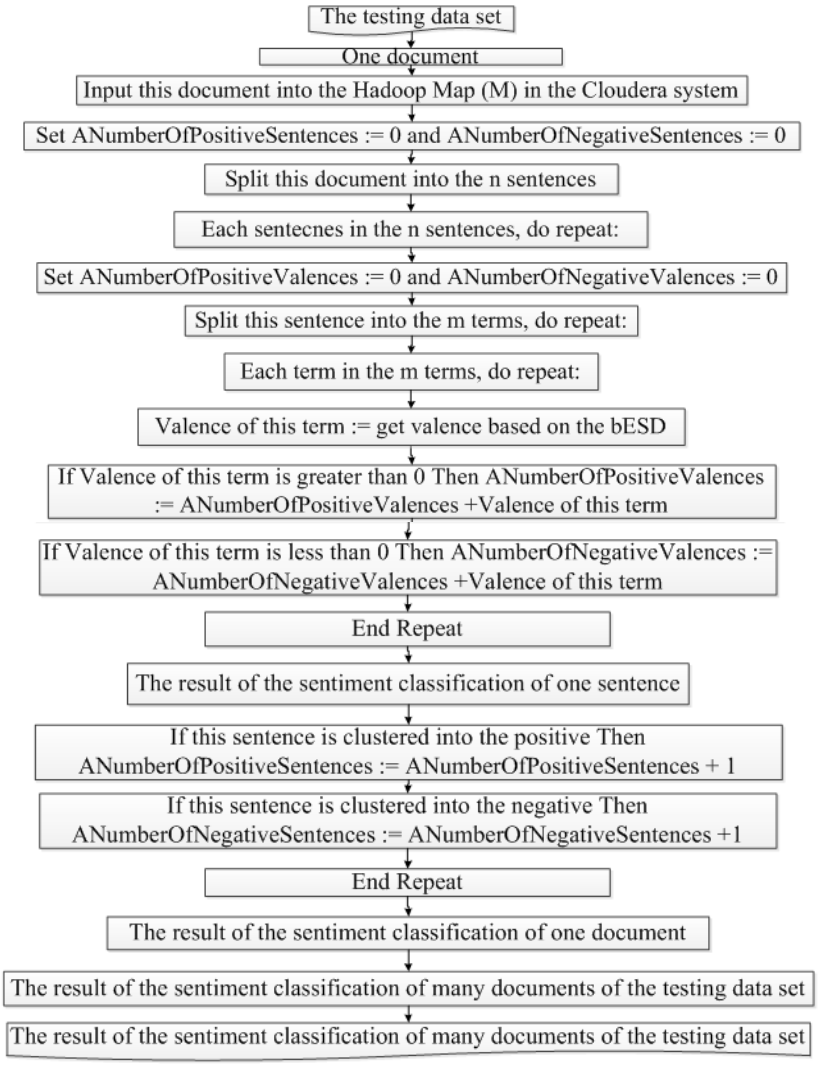

Fig. 6: Overview of using the sentiment-lexicons with the JC to classify the documents of the testing data set into either the positive polarity or the negative polarity in the distributed environment

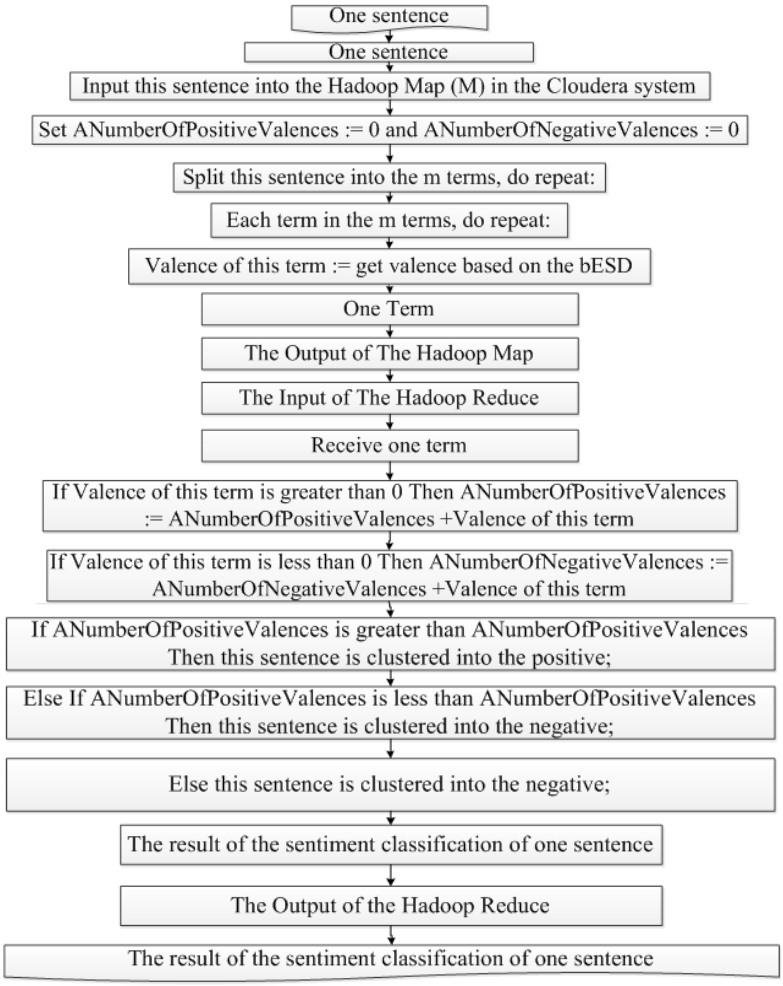

Fig. 7: Overview of clustering one sentence into either the positive or the negative in the parallel system

We use the algorithm 7 to perform the Hadoop Map phase of clustering one sentence into either the positive or the negative in the parallel system. The main ideas of the algorithm 7 are as follows:

Input: one sentence
Output: one term which the valence is identified based on the bESD.
Step 1: Input this sentence and the bESD into the Hadoop Map in the Cloudera system.
Step 2: Split this sentence into m meaningful terms (meaningful words or meaningful phrases) based on the bESD;
Step 3: Each term in the m terms, do repeat:
Step 4: Valence := get valence of this term based on the bESD;
Step 5: Return this term; //the output of the Hadoop Map

We use the algorithm 8 to perform the Hadoop Reduce phase of clustering one sentence into either the positive or the negative in the parallel system. The main ideas of the algorithm 8 are as follows:

Input: one term which the valence is identified based on the bESD – the output of the Hadoop Map
Output: the sentiment polarity (positive, negative, neutral)
Step 1: Receive one term;
Step 2: If Valence is greater than 0 Then ANumberOfPositiveValences := ANumberOfPositiveValences + Valence;
Step 3: Else If Valence is less than 0 Then ANumberOfNegativeValences := ANumberOfNegativeValences + Valence;
Step 4: End Repeat – End Step 3;
Step 5: If ANumberOfPositiveValences is greater than ANumberOfPositiveValences Then Return positive;
Step 6: Else If ANumberOfPositiveValences is less than ANumberOfPositiveValences Then Return negative;
Step 7: Return neutral;

DOI: 10.3844/ajeassp.2018.38.65

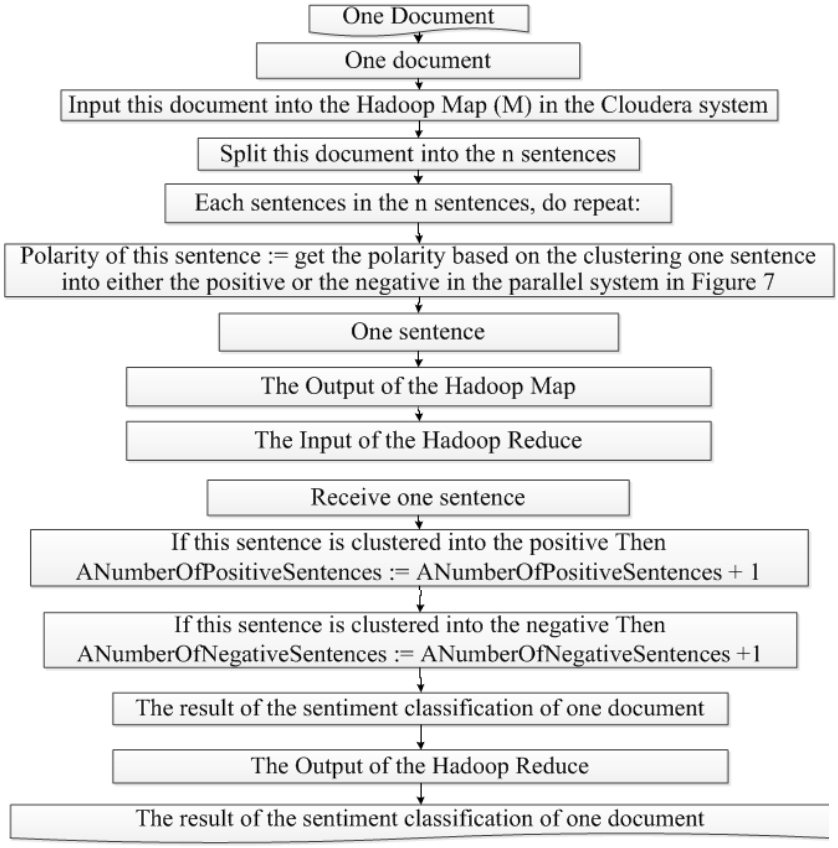

Fig. 8: Overview of clustering one document into either the positive or the negative in the distributed environment

In Fig. 8, we propose the algorithm 9 and the algorithm 10 to cluster one document into either the positive or the negative in the distributed environment. This stage includes two phases: The Hadoop Map (M) phase and the Hadoop Reduce (R). The input of the Hadoop Map is one document and the bESD. The output of the Hadoop Map phase is one sentence which the polarity is identified. The input of the Hadoop Reduce (R) is the output of the Hadoop Map phase, thus, the input of the Hadoop Reduce is one sentence. The output of the Hadoop Reduce phase is the sentiment polarity (positive, negative, neutral) of this document.

We propose the algorithm 9 to perform the Hadoop Map phase of cluster one document into either the positive or the negative in the distributed environment. The main ideas of the algorithm 9 are as follows:

Input: one document
Output: One sentence which the polarity is identified
Step 1: Input this document into the Hadoop Map in the Cloudera system
Step 2: Each sentence in the n sentences terms, do repeat:
Step 4: Polarity := get the polarity of this sentence based on the clustering one sentence into either the positive or the negative in the parallel system in Fig. 7;
Step 5: Return this sentence;//the output of the Hadoop Map

We propose the algorithm 10 to perform the Hadoop Reduce phase of cluster one document into either the positive or the negative in the distributed environment. The main ideas of the algorithm 10 are as follows:

Input: one sentence which the polarity is identified
Output: the sentiment polarity (positive, negative, neutral)
Step 1:Receive one sentence;
Step 2: If Polarity is positive Then ANumberOfPositiveSentences := ANumberOfPositiveSentences + 1;
Step 3: Else If Polartiy is negative Then ANumberOfNegativeSentences := ANumberOfNegativeSentences + 1;
Step 4: If ANumberOfPositiveSentences is greater than ANumberOfNegativeSentences Then Return positive;
Step 5: Else If ANumberOfPositiveSentences is less than ANumberOfNegativeSentences Then Return negative;
Step 6: Return neutral;

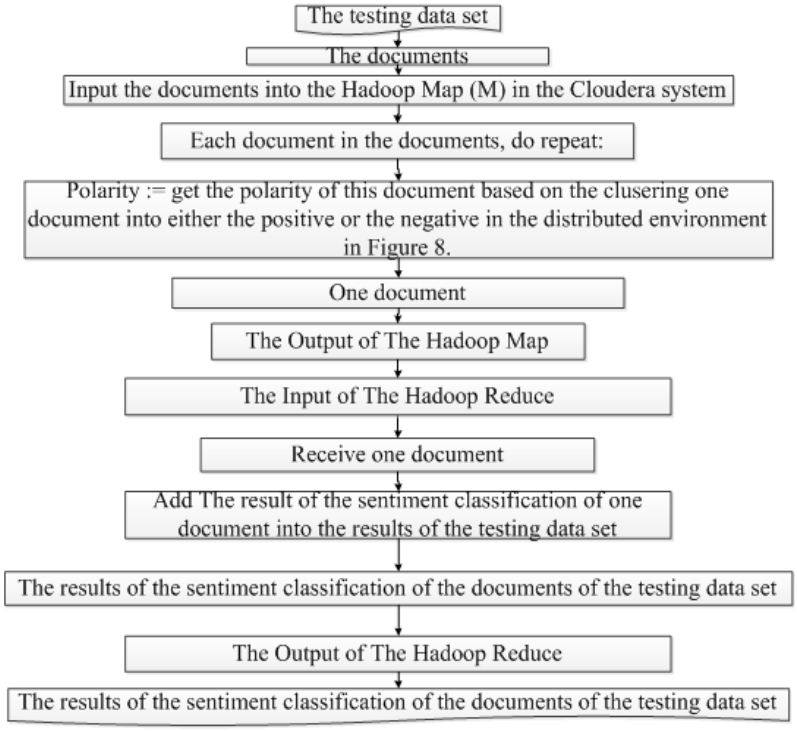

Fig. 9: Overview of clustering all the documents of the testing data set into either the positive or the negative in the parallel system

In Fig. 9, we propose the algorithm 11 and the algorithm 12 to cluster all the documents of the testing data set into either the positive or the negative in the parallel system. This stage includes two phases: The Hadoop Map (M) phase and the Hadoop Reduce (R). The input of the Hadoop Map is the testing data set and the bESD. The output of the Hadoop Map phase is one document which the polarity is identified. The input of the Hadoop Reduce (R) is the output of the Hadoop Map phase, thus, the input of the Hadoop Reduce is one document. The output of the Hadoop Reduce phase is the sentiment polarity (positive, negative, neutral) of the testing data set.

We propose the algorithm 11 to implement the Hadoop Map phase of clustering all the documents of the testing data set into either the positive or the negative in the parallel system. The main ideas of the algorithm 11 are as follows:

Input: The testing data set
Output: One document which the polarity is identified
Step 1: Input the documents of the testing data set into the Hadoop Map in the Cloudera system;
Step 2: Each document into the documents, do repeat:
Step 3: Polarity := get the polarity of this document based on the clustering one document into either the positive or the negative in the distributed environment in Fig. 8;
Step 4: Return this document;//the output of the Hadoop Map

We propose the algorithm 12 to implement the Hadoop Reduce phase of clustering all the documents of the testing data set into either the positive or the negative in the parallel system. The main ideas of the algorithm 12 are as follows:

Input: One document which the polarity is identified
Output: The results of the sentiment classification of the testing data set
Step 1: Receive one document;
Step 2: Add the result of the sentiment classification of this document into the results of the sentiment classification of the testing data set;
Step 3: Return the results of the sentiment classification of the testing data set;

Experiment

We have measured an Accuracy (A) to calculate the accuracy of the results of emotion classification. A Java programming language is used for programming to save data sets, implementing our proposed model to classify the 5,000,000 documents of the testing data set. To implement the proposed model, we have already used Java programming language to save the English testing data set and to save the results of emotion classification.

The sequential environment in this research includes 1 node (1 server). The Java language is used

in programming our model related to the sentiment-lexicons with the JC. The configuration of the server in the sequential environment is: Intel® Server Board S1200V3RPS, Intel® Pentium® Processor G3220 (3M Cache, 3.00 GHz), 2 GB JC3-10600 EJC 1333 MHz LP Unbuffered DIMMs. The operating system of the server is: Cloudera. We perform the proposed model related to the sentiment-lexicons with the JC in the Cloudera parallel network environment; this Cloudera system includes 9 nodes (9 servers). The Java language is used in programming the application of the proposed model related to the sentiment-lexicons with the JC in the Cloudera. The configuration of each server in the Cloudera system is: Intel® Server Board S1200V3RPS, Intel® Pentium® Processor G3220 (3M Cache, 3.00 GHz), 2GB JC3-10600 EJC 1333 MHz LP Unbuffered DIMMs. The operating system of each server in the 9 servers is: Cloudera. All 9 nodes have the same configuration information.

In Table 3, we display the results of the documents in the testing data set.

The accuracy of our new model for the documents in the testing data set is shown in Table 4.

In Table 5, we present the average execution times of the classification of our new model for the documents in testing data set.

Results and Discussion

In this section, we show the results of this survey in the tables as follows: Table 3 to 5.

We show the results of the documents in the testing data set in Table 3.

The accuracy of the sentiment classification of the documents of the testing data set is presented in Table 4.

We display the average execution times of the classification of our novel model for the documents of the testing data set in Table 5.

In Table 3, we have had the 4,378,000documents of the correct classification of the testing data set comprising the 2,500,000 negative documents and the 2,500,000 positive documents. We have also had the 622,000 documents of the incorrect classification of the testing data set. The documents of the correct classification of the testing data set have comprises the 2,188,746 negative documents and the 2,189,254 positive documents. The documents of the incorrect classification of the testing data set have includes the 311,254 negative documents and the 310,746 positive documents.

In Table 4, we had achieved 87.56% accuracy of the testing data set.

In Table 5, the average time of the semantic classification of using the sentiment-lexicons with the JC in the sequential environment is 21,035,241 sec/5,000,000 English documents and it is greater than the average time of the emotion classification of using the sentiment-lexicons with the JC in the Cloudera parallel network environment with 3 nodes which is 7,485,069 sec/5,000,000 English documents. The average time of the emotion classification of using the sentiment-lexicons with the JC in the Cloudera parallel network environment with 9 nodes, which is 2,359,471 sec/5,000,000 English documents, is the shortest time. Besides, the average time of the emotion classification of using the sentiment-lexicons with the JC in the Cloudera parallel network environment with 6 nodes is 3,627,584 sec/5,000,000 English documents

Table 3: The results of the documents in the testing data set

	Testing dataset	Correct classification	Incorrect classification
Negative	2,500,000	2,188,746	311,254
Positive	2,500,000	2,189,254	310,746
Summary	5,000,000	4,378,000	622,000

Table 4: The accuracy of our new model for the documents in the testing data set

Proposed model	Class	Accuracy
Our novel model	Negative Positive	87.56%

Table 5: The average execution times of the classification of our new model for the documents in testing data set

	Average time of the classification /5,000,000 documents.
The JOHNSON Coefficient (JC) in the sequential environment	21,035,241 sec
The JOHNSON Coefficient (JC) in the Cloudera distributed system with 3 nodes	7,485,069 sec
The JOHNSON Coefficient (JC) in the Cloudera distributed system with 6 nodes	3,627,584 sec
The JOHNSON Coefficient (JC) in the Cloudera distributed system with 9 nodes	2,359,471 sec

Conclusion

Although our new model has been tested on our English data set, it can be applied to many other languages. In this study, our model has been tested on the 5,000,000 English documents of the testing data set in which the data sets are small. However, our model can be applied to larger data sets with millions of English documents in the shortest time.

In this study, we have proposed a new model to classify sentiment of English documents using the sentiment-lexicons with the JC with Hadoop Map (M) /Reduce (R) in the Cloudera parallel network environment. With our proposed new model, we have achieved 87.56% accuracy of the testing data set in Table 6. Until now, not many studies have shown that the clustering methods can be used to classify data. Our research shows that clustering methods are used to classify data and, in particular, can be used to classify emotion in text.

The execution time of using the sentiment-lexicons with the JC in the Cloudera is dependent on the performance of the Cloudera parallel system and also dependent on the performance of each server on the Cloudera system.

The proposed model has many advantages and disadvantages. Its positives are as follows: It uses using the sentiment-lexicons with the JC to classify semantics of English documents based on sentences. The proposed model can process millions of documents in the shortest time. This study can be performed in distributed systems to shorten the execution time of the proposed model. It can be applied to other languages. Its negatives are as follows: It has a low rate of accuracy. It costs too much and takes too much time to implement this proposed model.

Table 6: Comparisons of our model's advantages and disadvantages with the works in (Singh and Singh, 2015; Carrera-Trejo *et al*., 2015; Soucy and Mineau, 2015)

Researches	Approach	Advantages	Disadvantages
Singh and Singh (2015)	Examining the vector space model, an information retrieval technique and its variation	In this study, the authors have given an insider to the working of vector space model techniques used for efficient retrieval techniques. It is the bare fact that each system has its own strengths and weaknesses. What we have sorted out in the authors' work for vector space modeling is that the model is easy to understand and cheaper to implement, considering the fact that the system should be cost effective (i.e., should follow the space/time constraint. It is also very popular. Although the system has all these properties, it is facing some major drawbacks.	The drawbacks are that the system yields no theoretical findings. Weights associated with the vectors are very arbitrary and this system is an independent system, thus requiring separate attention. Though it is a promising technique, the current level of success of the vector space model techniques used for information retrieval are not able to satisfy user needs and need extensive attention.
Carrera-Trejo *et al*. (2015)	+Latent Dirichlet allocation (LDA). +Multi-label text classification tasks and apply various feature sets. +Several combinations of features, like bi-grams and uni-grams.	In this study, the authors consider multi-label text classification tasks and apply various feature sets. The authors consider a subset of multi-labeled files of the Reuters-21578 corpus. The authors use traditional TF-IDF values of the features and tried both considering and ignoring stop words. The authors also tried several combinations of features, like bi-grams and uni-grams. The authors also experimented with adding LDA results into vector space models as new features. These last experiments obtained the best results.	No mention
Soucy and Mineau (2015)	The K-Nearest Neighbors algorithm for English sentiment classification in the Cloudera distributed system.	In this study, the authors introduce a new weighting method based on statistical estimation of the importance of a word for a specific categorization problem. One benefit of this method is that it can make feature selection implicit, since useless features of the categorization problem considered get a very small weight. Extensive experiments reported in the work show that this new weighting method improves significantly the classification accuracy as measured on many categorization tasks.	Despite positive results in some settings, GainRatio failed to show that supervised weighting methods are generally higher than unsupervised ones. The authors believe that ConfWeight is a promising supervised weighting technique that behaves gracefully both with and without feature selection. Therefore, the authors advocate its use in further experiments.
Our work	-We use the sentiment-lexicons with the JC to classify one document of the testing data set into either the positive polarity or the negative polarity in both the sequential environment and the distributed system. -The advantages and disadvantages of the proposed model are shown in the Conclusion section.		

Table 7: Comparisons of our model's positives and negatives the latest sentiment classification models in (Agarwal and Mittal, 2016a; 2016b; Canuto *et al*., 2016; Ahmed and Danti, 2016; Phu and Tuoi, 2014; Tran *et al*., 2014; Dat *et al*., 2017; Phu *et al*., 2017f; 2017g; 2017h)

Studies	Approach	Positives	Negatives
Agarwal and Mittal (2016a)	The Machine Learning Approaches Applied to Sentiment Analysis-Based Applications	The main emphasis of this survey is to discuss the research involved in applying machine learning methods, mostly for sentiment classification at document level. Machine learning-based approaches work in the following phases, which are discussed in detail in this study for sentiment classification: (1) feature extraction, (2) feature weighting schemes, (3) feature selection and (4) machine-learning methods. This study also discusses the standard free benchmark datasets and evaluation methods for sentiment analysis. The authors conclude the research with a comparative study of some state-of-the -art methods for sentiment analysis and some possible future research directions in opinion mining and sentiment analysis.	No mention
Agarwal and Mittal (2016b)	Semantic Orientation-Based Approach for Sentiment Analysis	This approach initially mines sentiment-bearing terms from the unstructured text and further computes the polarity of the terms. Most of the sentiment-bearing terms are multi-word features unlike bag-of-words, e.g., "good movie," "nice cinematography," "nice actors," etc. Performance of semantic orientation-based approach has been limited in the literature due to inadequate coverage of multi-word features.	No mention
Canuto *et al*. (2016)	Exploiting New Sentiment-Based Meta-Level Features for Effective Sentiment Analysis	Experiments performed with a substantial number of datasets (nineteen) demonstrate that the effectiveness of the proposed sentiment-based meta-level features is not only superior to the traditional bag-of-words representation (by up to 16%) but also is also superior in most cases to state-of-art meta-level features previously proposed in the literature for text classification tasks that do not take into account any idiosyncrasies of sentiment analysis. The authors' proposal is also largely superior to the best lexicon-based methods as well as to supervised combinations of them. In fact, the proposed approach is the only one to produce the best results in all tested datasets in all scenarios.	A line of future research would be to explore the authors' meta features with other classification algorithms and feature selection techniques in different sentiment analysis tasks such as scoring movies or products a JCording to their related reviews.
Ahmed and Danti (2016)	Rule-Based Machine Learning Algorithms	The proposed approach is tested by experimenting with online books and political reviews and demonstrates the efficacy through Kappa measures, which have a higher accuracy of 97.4% and a lower error rate. The weighted average of different accuracy measures like Precision,Recall and TP-Rate depicts higher efficiency rate and lower FP-Rate. Comparative experiments on various rule-based machine learning algorithms have been performed through a ten-fold cross validation training model for sentiment classification.	No mention

Table 7: Continue

Phu and Tuoi (2014)	The Combination of Term-Counting Method and Enhanced Contextual Valence Shifters Method	The authors have explored different methods of improving the accuracy of sentiment classification. The sentiment orientation of a document can be positive (+), negative (-), or neutral (0). The authors combine five dictionaries into a new one with 21,137 entries. The new dictionary has many verbs, adverbs, phrases and idioms that were not in five dictionaries before. The study shows that the authors' proposed method based on the combination of Term-Counting method and Enhanced Contextual Valence Shifters method has improved the accuracy of sentiment classification. The combined method has accuracy 68.984% on the testing dataset and 69.224% on the training dataset. All of these methods are implemented to classify the reviews based on our new dictionary and the Internet Movie Database data set.	No mention
Tran *et al*. (2014)	Naive Bayes Model with N-GRAM Method, Negation Handling Method, Chi-Square Method and Good-Turing Discounting, etc.	The authors have explored the Naïve Bayes model with N-GRAM method, Negation Handling method, Chi-Square method and Good-Turing Discounting by selecting different thresholds of Good-Turing Discounting method and different minimum frequencies of Chi-Square method to improve the accuracy of sentiment classification.	No Mention
Our work	-We use the sentiment-lexicons with the JC to classify one document of the testing data set into either the positive polarity or the negative polarity in both the sequential environment and the distributed system. -The positives and negatives of the proposed model are given in the Conclusion section.		

To understand the scientific values of this research, we have compared our model's results with many studies in the tables below.

In Table 6, we show the comparisons of our model's advantages and disadvantages with the works in (Singh and Singh, 2015; Carrera-Trejo *et al*., 2015; Soucy and Mineau, 2015)

The comparisons of our model's positives and negatives the latest sentiment classification models in (Agarwal and Mittal, 2016a; 2016b; Canuto *et al*., 2016; Ahmed and Danti, 2016; Phu and Tuoi, 2014; Tran *et al*., 2014; Dat *et al*., 2017; Phu *et al*., 2016) are presented in Table 7.

Author's Contributions

Vo Ngoc Phu: He conceived the original research idea. He implemented surveys. He checked, fixed and wrote the draft documents finally.

Vo Thi Ngoc Tran: He built data sets. He wrote the draft documents of our manuscripts.

Conflict of Interest

The authors declare that they have no conflict of interest.

Future Work

Based on the results of this proposed model, many future projects can be proposed, such as creating full emotional lexicons in a parallel network environment to shorten execution times, creating many search engines, creating many translation engines, creating many applications that can check grammar correctly. This model can be applied to many different languages, creating applications that can analyze the emotions of texts and speeches and machines that can analyze sentiments.

References

Agarwal, B. and N. Mittal, 2016a. Machine Learning Approach for Sentiment Analysis. In: Prominent Feature Extraction for Sentiment Analysis, Agarwal, B. and N. Mittal (Eds.), Springer, ISBN-10: 3319253417, pp: 21-45.

Agarwal, B. and N. Mittal, 2016b. Semantic Orientation-Based Approach for Sentiment Analysis. In: Prominent Feature Extraction for Sentiment Analysis, Agarwal, B. and N. Mittal (Eds.), Springer, ISBN-10: 3319253417, pp: 77-88.

Ahmed, S. and A. Danti, 2016. Effective sentimental analysis and opinion mining of web reviews using rule based classifiers. Proceedings of the International Conference on Computational Intelligence in Data Mining, (IDM' 16), Springer, New Delhi, pp: 171-179. DOI: 10.1007/978-81-322-2734-2_18

An, N.T.T. and M. Hagiwara, 2014. Adjective-based estimation of short sentence's impression. Proceedings of the 5th Kanesi Engineering and Emotion Research, (EER' 14), Sweden.

Bai, A., H. Hamme, A. Yazidi and P. Engelstad, 2014. Constructing sentiment lexicons in Norwegian from a large text corpus. Proceedings of the IEEE 17th International Conference on Computational Science and Engineering, Dec. 19-21, IEEE Xplore Press, Chengdu, China, pp: 231-237. DOI: 10.1109/CSE.2014.73

Brooke, J., M. Tofiloski and M. Taboada, 2009. Cross-linguistic sentiment analysis: From English to Spanish. Proceedings of the International Conference RANLP, (RANLP' 09), Borovets, Bulgaria, pp: 50-54.

Canuto, S., M.A. Gonçalves and F. Benevenuto, 2016. Exploiting new sentiment-based meta-level features for effective sentiment analysis. Proceedings of the 9th ACM International Conference on Web Search and Data Mining, Feb. 22-25, ACM, USA, pp: 53-62. DOI: 10.1145/2835776.2835821

Carrera-Trejo, V., G. Sidorov, S. Miranda-Jiménez, M.M. Ibarra and R.C. Martínez, 2015. Latent dirichl*et al*location complement in the vector space model for multi-label text classification. Int. J. Combinat. Optimiz. Prob. Inform., 6: 7-19.

CED, 2017a. Cambridge English Dictionary.

CED, 2017b. Collins English Dictionary.

Choi, S.S., S.H. Cha and C.C. Tappert, 2010. A survey of binary similarity and distance measures. Syst. Cybernet. Inform.

Dalirsefat, S.B., A.D.S. Meyer and S.Z. Mirhoseini, 2009. Comparison of similarity coefficients used for cluster analysis with amplified fragment length polymorphism markers in the Silkworm, *Bombyx mori*. J. Insect. Sci., 9: 71-71. DOI: 10.1673/031.009.7101

Dat, N.D., V.N. Phu, V.T.N. Tran and V.T.N. Chau, 2017. STING algorithm used English sentiment classification in a parallel environment. Int. J. Patt. Recognit. Artif. Intell.

Du, W., S. Tan, X. Cheng and X. Yun, 2010. Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon. Proceedings of the 3rd ACM International Conference on Web Search and Data Mining, Feb. 04-06, ACM, New York, pp: 111-120. DOI: 10.1145/1718487.1718502

Duarte, J.M., J.B. dos Santos and L.C. Melo, 1999. Comparison of similarity coefficients based on RAPD markers in the common bean. Genet. Mol. Biol. DOI: 10.1590/S1415-47571999000300024

EDL, 2017. English Dictionary of Lingoes.

Feng, S., L. Zhang, B.L.D. Wang, G. Yu and K.F. Wong, 2013. Is Twitter a better corpus for measuring sentiment similarity? Proceedings of the Conference on Empirical Methods in Natural Language Processing, Oct. 18-21, Association for Computational Linguistics, USA, pp: 897-902.

Hernández-Ugalde, J.A., J. Mora-Urpí and O.J. Rocha, 2011. Genetic relationships among wild and cultivated populations of peach palm (*Bactris gasipaes* Kunth, Palmae): Evidence for multiple independent domestication events. Genetic Resources Crop Evolut., 58: 571-583. DOI: 10.1007/s10722-010-9600-6

Htait, A., S. Fournier and P. Bellot, 2016. LSIS at SemEval-2016 Task 7: Using web search engines for english and Arabic unsupervised sentiment intensity prediction. Proceedings of SemEval, Jun. 16-17, Association for Computational Linguistic, California, pp: 481-485.

Ji, X., S.A. Chun, Z. Wei and J. Geller, 2015. Twitter sentiment classification for measuring public health concerns. Soc. Netw. Anal. Min., 5: 13-13. DOI: 10.1007/s13278-015-0253-5

Jiang, T., J. Jiang, Y. Dai and A. Li, 2015. Micro-blog emotion orientation analysis algorithm based on Tibetan and Chinese mixed text. Proceedings of the International Symposium on Social Science, (SSS' 15), Atlantis Press.

Jovanoski, D., V. Pachovski and P. Nakov, 2015. Sentiment analysis in twitter for Macedonian. Proceedings of Recent Advances in Natural Language Processing, Sep. 7-9, Bulgaria, pp: 249-257.

LED, 2017. Longman English Dictionary.

Malouf, R. and T. Mullen. 2017. Graph-based user classification for informal online political discourse. Proceedings of the 1st Workshop on Information Credibility on the Web, (ICW' 17), At Miyazaki, Japan.

Mao, H., P. Gao, Y. Wang and J. Bollen, 2014. Automatic construction of financial semantic orientation lexicon from large-scale Chinese news corpus. Proceedings of the 7th Financial Risks International Forum, (RIF' 14), Institut Louis Bachelier.

Meyer, A.D.S., A.A.F. Garcia, A.P. de Souza and C.L. de Souza Jr., 2004. Comparison of similarity coefficients used for cluster analysis with dominant markers in maize (*Zea mays* L). Genet. Molecular Biol., 27: 83-91. DOI: 10.1590/S1415-47572004000100014

Mladenović Drinić, S., A. Nikolić and V. Perić, 2008. Cluster analysis of soybean genotypes based on RAPD markers. Proceedings of the 43rd Croatian and 3rd International Symposium on Agriculture, (ISA' 08), Opatija, Croatia, pp: 367-370.

MMED, 2017. MacMillan English Dictionary.

Netzer, O., R. Feldman, J. Goldenberg and M. Fresko, 2012. Mine your own business: Market-structure surveillance through text mining. Market. Sci., 31: 521-543. DOI: 10.1287/mksc.1120.0713

OED, 2017. Oxford English Dictionary.

Omar, N., M. Albared, A.Q. Al-Shabi and T. Al-Moslmi. 2013. Ensemble of classification algorithms for subjectivity and sentiment analysis of Arabic customers' reviews. Int. J. Advancements Comput. Technol., 5: 77-85.

Phu, V.N. and P.T. Tuoi, 2014. Sentiment classification using enhanced contextual valence shifters. Proceedings of the International Conference on Asian Language Processing, Oct. 20-22, IEEE Xplore Press, Kuching, Malaysia, pp: 224-229. DOI: 10.1109/IALP.2014.6973485

Phu, V.N., N.D. Dat, V.T.N. Tran and V.T.N. Tran, 2016. Fuzzy c-means for English sentiment classification in a distributed system. Int. J. Applied Intell., 46: 717-738. DOI: 10.1007/s10489-016-0858-z

Phu, V.N., V.T.N. Chau, V.T.N. Tran and N.D. Dat, 2017a. A Vietnamese adjective emotion dictionary based on exploitation of Vietnamese language characteristics. Artificial Intell. Rev. DOI: 10.1007/s10462-017-9538-6

Phu, V.N., V.T.N. Chau, N.D. Dat, V.T.N. Tran and T.A. Nguyen, 2017b. A valences-totaling model for English sentiment classification. Know. Inform. Syst., 53: 579-636. DOI: 10.1007/s10115-017-1054-0

Phu, V.N., V.T.N. Chau and V.T.N. Tran, 2017c. Shifting semantic values of English phrases for classification. Int. J. Speech Technol., 20: 509-533. DOI: 10.1007/s10772-017-9420-6

Phu, V.N., V.T.N. Chau, V.T.N. Tran, N.D. Dat and K.L.D. Duy, 2017d. A valence-totaling model for vietnamese sentiment classification. Int. J. Evolv. Syst. DOI: 10.1007/s12530-017-9187-7

Phu, V.N., V.T.N. Chau, V.T.N. Tran, N.D. Dat and K.L.D. Duy, 2017e. Semantic lexicons of english nouns for classification. Int. J. Evolv. Syst. DOI: 10.1007/s12530-017-9188-6

Phu, V.N., C.V.T. Ngoc, T.V.T. Ngoc and D.N. Duy, 2017f. A C4.5 algorithm for English emotional classification. Evol. Syst. DOI: 10.1007/s12530-017-9180-1

Phu, V.N., V.T.N. Chau and V.T.N. Tran, 2017g. SVM for English semantic classification in parallel environment. Int. J. Speech Technol., 20: 487-508. DOI: 10.1007/s10772-017-9421-5

Phu, V.N., V.T.N. Tran, V.T.N. Chau, N.D. Dat and K.L.D. Duy, 2017h. A decision tree using ID3 algorithm for English semantic analysis. Int. J. Speech Technol., 20: 593-613. DOI: 10.1007/s10772-017-9429-x

Ponomarenko, J.V., P.E. Bourne and I.N. Shindyalov, 2002. Building an automated classification of DNA-binding protein domains. Bioinformatics, 18: S192-S201. PMID: 12386003

Ren, Y., N. Kaji, N. Yoshinaga and M. Kitsuregaw, 2014. Sentiment classification in under-resourced languages using graph-based semi-supervised learning methods. IEICE Trans. Inf. Syst., E97: 1-10. DOI: 10.1587/transinf.E97.D.1

Ren, Y., N. Kaji, N. Yoshinaga, M. Toyoda and M. Kitsuregawa, 2011. Sentiment classification in resource-scarce languages by using label propagation. Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation, (LIC' 11), Institute of Digital Enhancement of Cognitive Processing, Waseda University, pp: 420-429.

Scheible, C., 2010. Sentiment translation through lexicon induction. Proceedings of the ACL Student Research Workshop, Jul. 13-13, Association for Computational Linguistics, Sweden, pp: 25-30.

Shikalgar, N.R. and A.M. Dixit, 2014. JIBCA: Jaccard Index Based Clustering Algorithm for mining online review. Int. J. Comput. Applic., 105: 23-28.

Singh, V.K. and V.K. Singh, 2015. Vector space model: An information retrieval system. Int. J. Adv. Eng. Res. Stud.

Soucy, P. and G.W. Mineau, 2015. Beyond TFIDF weighting for text categorization in the vector space model. Proceedings of the 19th International Joint Conference on Artificial Intelligence, Jul. 30-Aug. 05, Morgan Kaufmann Publishers Inc., Edinburgh, Scotland, pp: 1130-1135.

Tamás, J., J. Podani and P. Csontos, 2001. An extension of presence/absence coefficients to abundance data: A new look at absence. J. Vegetat. Sci., 12: 401-410. DOI: 10.2307/3236854

Tan, S. and J. Zhang, 2007. An empirical study of sentiment analysis for Chinese documents. Expert Syst. Applic., 34: 2622-2629. DOI: 10.1016/j.eswa.2007.05.028

Tran, V.T.N., V.N. Phu and P.T. Tuoi, 2014. Learning more chi square feature selection to improve the fastest and most AJCurate sentiment classification. Proceedings of the 3rd Asian Conference on Information Systems, (CIS' 14).

Tulloss, R.E., 1997. Assessment of Similarity Indices for Undesirable Properties and a new Tripartite Similarity Index Based on Cost Functions. In: MJCology in Sustainable Development: Expanding Concepts, Vanishing Borders, Palm, M.E. and I.H. Chapela (Eds.), Parkway Publishers, Boone, North Carolina, pp: 122-143.

Turney, P.D. and M.L. Littman, 2002. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. arXiv:cs/0212012, Learning (cs.LG); Information Retrieval (cs.IR).

Wan, X., 2009. Co-training for cross-lingual sentiment classification. Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, Aug. 02-07, Association for Computational Linguistics, Singapore, pp: 235-243. DOI: 10.3115/1687878.1687913

Wang, G. and K. Araki, 2007. Modifying SO-PMI for Japanese weblog opinion mining by using a balancing factor and detecting neutral expressions. Proceedings of the North American Chapter of the Association for Computational Linguistics, Apr. 22-27, Association for Computational Linguistics, Rochester, New York, pp: 189-192. DOI: 10.3115/1614108.1614156

Wijaya, S.H., F.M. Afendi, I. Batubara, L.K. Darusman and M. Altaf-Ul-Amin *et al*., 2016. Finding an appropriate equation to measure similarity between binary vectors: Case studies on Indonesian and Japanese herbal medicines. BMC Bioinformat., 17: 520-520. DOI: 10.1186/s12859-016-1392-z

Wilk, C.M., J.M. Gold, J.J. Bartko, F. Dickerson and W.S. Fenton *et al*., 2002. Test-retest stability of the repeatable battery for the assessment of neuropsychological status in schizophrenia. Am. J. Psychiatry, 159: 838-844. DOI: 10.1176/appi.ajp.159.5.838

Zhang, Z., Q. Ye, W. Zheng and Y. Li, 2010. Sentiment classification for consumer word-of-mouth in Chinese: Comparison between supervised and unsupervised approaches. Proceedings of the International Conference on E-Business Intelligence, (EBI' 10).