A SELF-ORGANIZING MAP ALGORITHM USING ONLY A TESTING DATA SET WITH THE ONE-DIMENSIONAL VECTORS AND AN ODDS RATIO COEFFICIENT FOR ENGLISH SENTIMENT CLASSIFICATION IN A PARALLEL SYSTEM

[1]DR.VO NGOC PHU, [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

E-mail: [1]vongocphu03hca@gmail.com, vongocphu@ntt.edu.vn, [2]vtntran@HCMUT.edu.vn

ABSTRACT

Many different approaches have already been studied for sentiment classification for many years because It has been significant in everyday life, such as in political activities, commodity production, and commercial activities. A new model using an unsupervised learning for big data sentiment classification has been proposed in this survey. We have used a Self-Organizing Map Algorithm (SOM) to cluster all sentences of one document of the testing data set comprising 8,500,000 documents, which are the 4,250,000 positive and the 4,250,000 negative in English, into either the positive polarity or the negative polarity certainly. In this survey, we do not use any data sets. We do not any one-dimensional vectors based on a vector space modeling (VSM). We also do not use any multi-dimensional vectors based on the VSM. We only use many one-dimensional vectors based on many sentiment lexicons of our basis English sentiment dictionary (bESD). The valences and the polarities of the sentiment lexicons of the bESD are calculated by using An Odds Ratio Coefficient (ORC) through a Google search engine with AND operator and OR operator. We also do not use many multi-dimensional vectors based on the sentiment lexicons of the bESD. With one document of the testing data set, the SOM is used to cluster all the sentences of this document into either the positive or the negative on a map. The sentiment classification of this document is identified based on this map completely. We have tested the proposed model in both a sequential environment and a distributed network system. We have achieved 88.14% accuracy of the testing data set. The execution of the proposed model in the sequential system is greater than that in the parallel network environment. Many applications and research of the sentiment classification can widely use the results of the proposed model.

Keywords: *English Sentiment Classification; Distributed System; Parallel System; Odds Ratio Similarity Coefficient; Cloudera; Hadoop Map And Hadoop Reduce; Clustering Technology; Self-Organizing Map*

1. INTRODUCTION

Many algorithms of the different fields such as a data mining, a computer science, a natural language processing and etc. have already been developed more and more. They are also used in applying to sentiment classification. The data mining and the natural language processing have had many significant relationships for many years. About clustering technologies of the data mining, a set of objects is processed into classes of similar objects, called clustering data. 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.

Many approaches based on many sentiment lexicons for the sentiment classification have been developed for many years. There are the reseaches related the sentiment lexicons in [4-14].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

According to the surveys related the Self-Organizing Map Algorithm (SOM) in [20-24], a self-organizing map (SOM) or self-organizing feature map (SOFM) is a type of artificial neural network (ANN) that is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a map, and is therefore a method to do dimensionality reduction. Self-organizing maps differ from other artificial neural networks as they apply competitive learning as opposed to error-correction learning (such as backpropagation with gradient descent), and in the sense that they use a neighborhood function to preserve the topological properties of the input space.

Also based on the SOM in [20-24], the advantages of the SOM are as follows: It is an unsupervised learning. We do not need any training data sets in English for the SOM. It shows many multi-dimensional data sets into either the one-dimensional data sets or the two-dimensional data sets, etc.

We build the basic principles for our new model as follows:

- We assume that each English sentence has m English words (or English phrases).
- We assume that the maximum number of one English sentence is m_max terms (words or phrases); it means that m is less than m_max or m is equal to m_max.
- We assume that each English document has n English sentences.
- We assume that the maximum number of one English document is n_max sentences; it means that n is less than n_max or n is equal to n_max.
- We transfer one sentence into one one-dimensional vector in English. Thus, the length of the vector is m. If m is less than m_max then each element of the vector from m to m_max-1 is 0 (zero).
- All the sentences of one document of the testing data set are transfer on many sentiment lexicons of our basis English sentiment dictionary (bESD).

Based on our opinion, the motivation of this new model is as follows: Many algorithms in the data mining field can be applied to natural language processing, specifically semantic classification for processing millions of English documents. An Odds Ratio Coefficient (ORC) and the SOM of the clustering technologies of the data mining filed can be applied to the sentiment classification in both a sequential environment and a parallel network system. This will result in many discoveries in scientific research, hence the motivation for this study.

The novelty of the proposed approach is as follows: the Odds Ratio Coefficient (ORC) and the SOM are applied to sentiment analysis. This can also be applied to identify the sentiments (positive, negative, or neutral) of millions of many documents. This survey can be applied to other parallel network systems. Hadoop Map (M) and Hadoop Reduce (R) are used in the proposed model. Therefore, we will study this model in more detail.

With the purpose of this survey, we always try to find a new approach to reform many accuracies of the results of the sentiment classification and to shorten many execution times of the proposed model with a low cost.

To get higher accuracy and shorten execution time of the sentiment classification, we dot not use any data sets. We do not any one-dimensional vectors based on a vector space modeling (VSM) in [1-3]. We also do not use any multi-dimensional vectors based on the VSM [1-3]. We only use many one-dimensional vectors based on many sentiment lexicons of our basis English sentiment dictionary (bESD). We also do not use many multi-dimensional vectors based on the sentiment lexicons of the bESD. We create the sentiment lexicons of the bESD by using An Odds Ratio Coefficient (ORC) through a Google search engine with AND operator and OR operator. All the n_max sentences of one document of the testing data set are transferred into the n_max one-dimensional vectors of the document. We use the SOM to cluster the n_max one-dimensional vectors of the document into either the positive polarity or the negative polarity with the input of the SOM is the n_max one-dimensional vectors of this document. The document is identified the sentiment classification based on the results of the n_max one-dimensional vectors clustered into either the positive or the negative.

Our proposed model is performed as follows: Firstly, we calculate the valences and the polarities of the sentiment lexicons of the bESD using the ORC through the Google search engine with AND operator and OR operator. We transfer all the n_max sentences of one document of the testing data set into the n_max one-dimensional vectors of this document. All the n_max one-dimensional vectors of this document are clustered into either the positive polarity or the negative polarity by using the SOM with the input is the n_max one-dimensional vectors. We set an initialization of the SOM with its map in Fig. 1 as follows:

Journal of Theoretical and Applied Information Technology
31st August 2018. Vol.96. No 16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

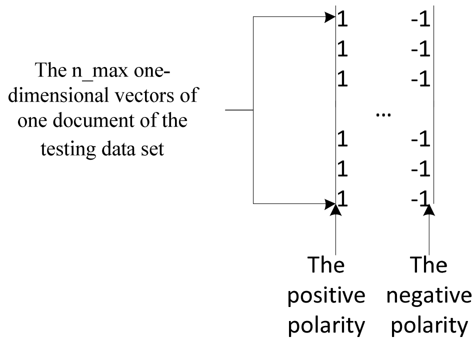

Fig. 1: An Initialization Of The SOM – The Map

Then, after the SOM is implemented completely, we have the Map in Fig. 2 as follows:

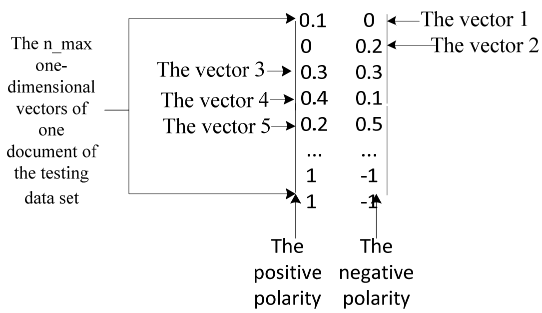

Fig. 2: The Final Map – The Result Of Clustering By Using The SOM

In Fig. 2, we have the vector 1 (0.1, 0), the vector 2 (0, 0.2), the vector 3 (0.3, 0.3), the vector 4 (0.4, 0.1) , and the vector 5 (0.2, 0.5). With the vector 1, the column of the positive polarity is 0.1 and the column of the negative polarity is 0, Thus, the value of the column of the positive polarity is greater than the value of the column of the negative polarity, therefore this vector is clustered into the positive. With the vector 2, the column of the positive polarity is 0 and the column of the negative polarity is 0.2. Therefore, the value of the column of the positive polarity is less than the value of the column of the negative polarity, thus this vector is clustered into the negative. With the vector 3, the column of the positive polarity is 0.3 and the column of the negative polarity is 0.3. So, the value of the column of the positive polarity is as equal as the value of the column of the negative polarity, therefore this vector is not clustered into both the positive and the negative. It meas that this vector is clustered into the neutral polarity. With the vector 4 (0.4, 0.1), the column of the positive polarity is 0.4 and the column of the negative polarity is 0.1. Thus, the value of the column of the positive polarity is greater than the value of the column of the negative polarity, so this vector is clustered into the positive. With the vector 5, the column of the positive polarity is 0.2 and the column of the negative polarity is 0.5. Therefore, the value of the column of the positive polarity is less than the column of the negative polarity, thus this vector is clustered into the negative. One document is clustered into the positive if the number of the one-dimensional vectors (corresponding to the sentences of this document) clustered into the positive is greater than the number of the one-dimensional vectors (corresponding to the sentences of this document) clustered into the negative in the document. The document is clustered into the negative if the number of the one-dimensional vectors clustered into the positive is less than the number of the one-dimensional vectors clustered into the negative in the document. The document is clustered into the neutral if the number of the one-dimensional vectors clustered into the positive is as equal as the number of the one-dimensional vectors clustered into the negative in the document. Finally, the sentiment classification of all the documents of the testing data set is identified completely.

Firstly, all the above things are performed in the sequential system to get an accuracy of the result of the sentiment classification and an execution time of the result of the sentiment classification of the proposed model. Secondly, we implement all the above things in the parallel network environment to shorten the execution times of the proposed model to get the accuracy of the results of the sentiment classification and the execution times of the results of the sentiment classification of our new model

The crucial contributions of our new model can be applied 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.

3)The algorithms are built in the proposed model.

4)This survey can certainly be applied to other languages easily.

5)The results of this study can significantly be applied to the types of other words in English.

6)Many crucial contributions are listed in the Future Work section.

7)The algorithm of data mining is applicable to semantic analysis of natural language processing.

8)This study also proves that different fields of scientific research can be related in many ways.

9)Millions of English documents are successfully processed for emotional analysis.

10)The semantic classification is implemented in the parallel network environment.

11)The principles are proposed in the research.
12)The Cloudera distributed environment is used in this study.
13)The proposed work can be applied to other distributed systems.
14)This survey uses Hadoop Map (M) and Hadoop Reduce (R).
15)Our proposed model can be applied to many different parallel network environments such as a Cloudera system
16)This study can be applied to many different distributed functions such as Hadoop Map (M) and Hadoop Reduce (R).
17)The SOM – related algorithms are proposed in this survey.
18)The ORC – related algorithms are built in this work.

This study contains 6 sections. Section 1 introduces the study; Section 2 discusses the related works about the vector space modeling (VSM), Self-Organizing Map Algorithm (SOM) and Odds Ratio Coefficient (ORC), 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.

2. RELATED WORK

We summarize many researches which are related to our research.

There are the works related to vector space modeling (VSM) in [1-3]. In this study [1], the authors examined the Vector Space Model, an Information Retrieval technique and its variation. In this survey [2], the authors considered multi-label text classification task and apply various feature sets. The authors considered a subset of multi-labeled files from the Reuters-21578 corpus. The authors used 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 [3] introduced a new weighting method based onstatistical estimation of the importance of a word for a specific categorization problem. This method also had the benefit to makefeature selectionimplicit, since uselessfeatures for the categorization problemconsidered getavery small weight.

The latest researches of the sentiment classification are [4-14]. In the research [4], the authors presented 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 [5] discussed an approach where an exposed stream of tweets from the Twitter micro blogging site were preprocessed and classified based on their sentiments. In sentiment classification system the concept of opinion subjectivity has been accounted. In the stedudy, the authors present opinion detection and organization subsystem, which have already been integrated into our larger question-answering system, etc.

The surveys related the Odds Ratio coefficient are in [15-19]. The authors in [15] collected 76 binary similarity and distance measures used over the last century and reveal their correlations through the hierarchical clustering technique. In [16], because the odds ratio has many desirable properties, and some investigators may find the odds ratio is easier to interpret, the authors discussed modelling the association between binary responses at pairs of times with the odds ratio, etc.

There are the researches related the Self-Organizing Map Algorithm (SOM) in [20-24]. In [20], the self-organized map, an architecture suggested for artificial neural networks, was explained by presenting simulation experiments and practical applications. The self-organizing map had the property of effectively creating spatially organized internal representations of various features of input signals and their abstractions. In [21], the Kohonen Self-Organizing Map (SOM) was one of the most well-known neural network with unsupervised learning rules; it performed a topology-preserving projection of the data space onto a regular two-dimensional space. Its achievement has already been demonstrated in various areas, but this approach is not yet widely known and used by ecologists. The present work described how SOM can be used for the study of ecological communities, etc.

By far, we know that PMI (Pointwise Mutual Information) equation and SO (Sentiment Orientation) 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 ODDS RATIO coefficient (ORC) 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 [25-37]. In the research [25], the authors generated several Norwegian sentiment lexicons by extracting sentiment information from two different types of Norwegian text corpus, namely, news corpus and discussion forums. The methodology was based on the Point wise Mutual Information (PMI). The authors introduced a modification of the PMI that considers small "blocks" of the text instead of the text as a whole. The study in [26] introduced 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 [38-39]. In the survey [38], the authors empirically evaluated the performance of different corpora in sentiment similarity measurement, which is the fundamental task for word polarity classification. The research in [39] proposed a new method to estimate impression of short sentences considering adjectives. In the proposed system, first, an input sentence was 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 [40-46]. The survey in [40] investigated the problem of sentiment analysis of the online review. In the study [41], the authors were 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 to the similarity coefficients to calculate the valences of words are in [52-56].

The English dictionaries are [57-62] and there are more than 55,000 English words (including English nouns, English adjectives, English verbs, etc.) from them.

3. DATA SET

Based on Fig. 1 below, we built our the testing data set including the 8,500,000 documents in the movie field, which contains the 4,250,000 positive and 4,250,000 negative in English. All the documents in our English training data set are automatically extracted from English Facebook, English websites and social networks; then we labeled positive and negative for them.

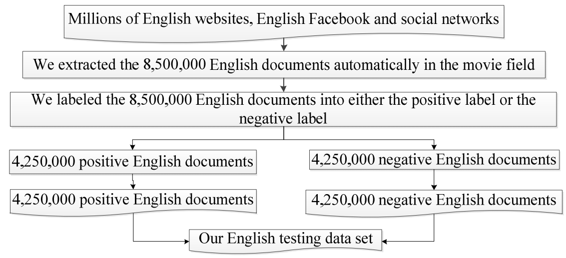

Fig. 3: Our Testing Data Set In English.

4. METHODOLOGY

An overview of the proposed model is shown in Fig. 4. This section comprises two parts. The first part is the sub-section (4.1) which we create the sentiment lexicons of the bESD in both a sequential environment and a parallel network system. The second part is the sub-section (4.2) which we use the SOM to cluster the documents of the testing data set into either the positive or the negative.

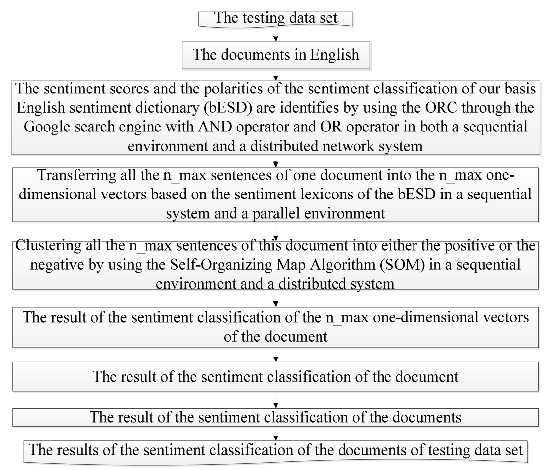

Fig. 4: Overview Of Our New Model.

The sub-section (4.1) has three parts. The first part is the sub-section (4.1.1) which we calculate the valence and the polarity of one term (word or phrase) in English by using the ORC through the Google search engine with AND operator and OR operator. The second part is the sub-section (4.1.2) which we identify the valences and the polarities of the sentiment lexicons of the bESD in a sequential system. The third part is the sub-section (4.1.3) which we calculate the valences and the polarities of the sentiment lexicons of the bESD in a parallel network environment.

The sub-section (4.2) comprise two parts. The first part is the sub-sectin (4.2.1) which we use the SOM to cluster the documents of the testing data set into either the positive or the negative in a sequential environment. The second part is the sub-section

(4.2.2) which we use the SOM to cluster the documents of the testing data set into either the positive or the negative in a parallel network system.

4.1 The sentiment lexicons in English

This section is used to create the sentiment lexicons in English in both a sequential environment and a distributed system.

The section comprises three parts. We identify a sentiment value of one word (or one phrase) in English in the first sub-section (4.1.1). We create a basis English sentiment dictionary (bESD) in a sequential system in the second sub-section (4.1.2). We also create a basis English sentiment dictionary (bESD) in a parallel environment in the third sub-section (4.1.3).

4.1.1 A valence of one word (or one phrase) in English

In this part, the valence and the polarity of one English word (or phrase) by using the ORC through a Google search engine with AND operator and OR operator are calculated, as the following diagram in Fig. 5 below shows.

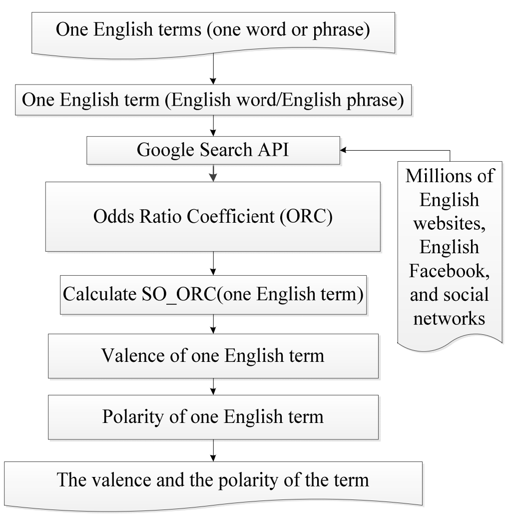

Fig. 5: Overview Of Identifying The Valence And The Polarity Of One Term In English Using An Odds Ratio Coefficient (ORC)

We have an equation about Pointwise Mutual Information (PMI) between two words wi and wj based on [24-39] as follows:

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

We also have an equation about SO (sentiment orientation) of word wi according to [1-15] as follows:

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

In eq. (2), according to [24-32], the positive is identified as follows: positive = {good, nice, excellent, positive, fortunate, correct, superior}

In eq. (2), based on [24-32], the negative is shown as follows: negative = {bad, nasty, poor, negative, unfortunate, wrong, inferior}.

The PMI equations in [26, 27, 29] use the AltaVista search engine and the PMI equations in [28, 30, 32] use the Google search engine.

In addition, German is used in [28]. Macedonian is used in [29]. Arabic is used in [30]. Chinese is used in [31] and Spanish is used in [32].

The Bing search engine is also used in [30]. Chinesese is used in the PMI equations of [33-36] and Tibetan is also added in [33].

About the search engine, the AltaVista search engine is used in [35, 36]. The survey [36] uses three search engines, such as the Google search engine, the Yahoo search engine and the Baidu search engine.

Japanese with the Google search engine is used in the PMI equations of [37].

English is used in PMI equations and Jaccard equations with the Google search engine of [38, 39]. We have an equation about Jaccard between two words wi and wj according to [38-46] as follows:

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

Based on [38-46], we have other type of the Jaccard equation between two words wi and wj as follows:

$$\mathrm{Jaccard(wi, wj) = J(wi, wj) = sim(wi, wj)} = \frac{\mathrm{F(wi, wj)}}{\mathrm{F(wi) + F(wj) - F(wi, wj)}} \qquad (4)$$

and we also have an equation about SO (sentiment orientation) of word wi as follows:

$$\mathrm{SO(wi)} = \sum \mathrm{Sim(wi, positive)} - \sum \mathrm{Sim(wi, positive)} \qquad (5)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

In eq. (5), according to [38-45], the positive and the negative in English are identified as follows: positive = {good, nice, excellent, positive, fortunate, correct, superior} and negative = {bad, nasty, poor, negative, unfortunate, wrong, inferior}. English is used in the Jaccard equations with the Google search engine in [38, 39, 41].

English is also used in the Jaccard equations in [40, 45].

Chinese is used in the Jaccard equations in [44, 46].

Arabic is used in the Jaccard equations in [42] and Chinese is used in the Jaccard equations with the Chinese search engine in [43].

Vietnamese is used in the Ochiai Measure through the Google search engine with AND operator and OR operator to calculate the sentiment values of the words in [52].

English is used in the Cosine Measure through the Google search engine with AND operator and OR operator to identify the sentiment scores of the words in [53].

The Sorensen Coefficient through the Google search engine with AND operator and OR operator is used to calculate the sentiment values of the words in English in [54].

The Jaccard Measure through the Google search engine with AND operator and OR operator is used to calculate the sentiment values of the words in Vietnamese in [55]

The Tanimoto Coefficient through the Google search engine with AND operator and OR operator is used to identify the sentiment scores of the words in English in [56]

With the above proofs, we have 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 Cosine and Sorensen are used with the Google in English.

Based on [24-56], PMI, Jaccard, Cosine, Ochiai, Sorensen, Tanimoto and ODDS RATIO coefficient (ORC) are the similarity measures between two words, and they can perform the same functions and with the same characteristics. Therefore, ORC is used in calculating the valence of the words.

In addition, we also prove that ORC can be used in identifying the valence of the English word through the Google search with the AND operator and OR operator.

We have an equation of the ORC based on the ODDS RATIO coefficient (ORC) in [15-19] as follows:

$$\begin{aligned}\text{ODDS RATIO Coefficient }(a,b) &= \text{ODDS RATIO Measure}(a,b)\\ &= \text{ORC}(a,b)\\ &= \frac{(a \cap b) * (\neg a \cap \neg b)}{(\neg a \cap b) * (a \cap \neg b)}\end{aligned} \quad (6)$$

with a and b are the vectors.

According to he eq. (1), (2), (3), (4), (5), (6), we build many new equations of the ORC 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 eq. (6), when a has only one element, a is a word. When b has only one element, b is a word. In eq. (6), a is replaced by w1 and b is replaced by w2.

$$\begin{aligned}&\text{ODDS RATIO Measure}(w1,w2)\\ &= \text{ODDS RATIO Coefficient}(w1,w2)\\ &= \text{ORC }(w1,w2)\\ &= \frac{P(w1,w2) * P(\neg w1, \neg w2)}{P(\neg w1, w2) * P(w1, \neg w2)}\end{aligned} \quad (7)$$

Eq. (7) is similar to eq. (1). In eq. (2), eq. (1) is replaced by eq. (7). We have eq. (8)

$$\begin{aligned}&\text{Valence}(w) = \text{SO\_ORC}(w)\\ &= \text{ORC}(w, \text{positive\_query})\\ &- \text{ORC}(w, \text{negative\_query})\end{aligned} \quad (8)$$

In eq. (7), w1 is replaced by w and w2 is replaced by position_query. We have eq. (9). Eq. (9) is as follows:

$$\text{ORC}(w, \text{positive\_query}) = \frac{A9}{B9} \quad (9)$$

with $A9 = P(w, \text{positive\_query}) * P(\neg w, \neg \text{positive\_query})$

$B9 = P(\neg w, \text{positive\_query}) * P(w, \neg \text{positive\_query})$

In eq. (7), w1 is replaced by w and w2 is replaced by negative_query. We have eq. (10). Eq. (10) is as follows:

$$\text{ORC}(w, \text{negative\_query}) = \frac{A10}{B10} \quad (10)$$

with $A10 = P(w, \text{negative\_query}) * P(\neg w, \neg \text{negative\_query})$

$B10 = P(\neg w, \text{negative\_query}) * P(w, \neg \text{negative\_query})$ We have the information about w, w1, w2, etc. as follows:

1)w, w1, w2 : are the English words (or the English phrases)

Journal of Theoretical and Applied Information Technology
31st August 2018. Vol.96. No 16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2)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).

3)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.

4)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.

5)Valence(W) = SO_ORC(w): valence of English word (or English phrase) w; is SO of word (or phrase) by using the ODDS RATIO coefficient (ORC)

6)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.

7)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.

8)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)

9)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)

10)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

11)P(¬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).

12)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)]).

13)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 ((not w) and [not (positive_query)]).

14)P(¬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).

15)P(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 (w and (not (negative_query))).

16)P(¬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 ORC (w, positive_query) is as equal as 1.

The English word w is the farthest of positive_query if ORC(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 ORC(w, positive_query) > 0 and ORC(w, positive_query) ≤ 1.

The English word w is the nearest of negative_query if ORC(w, negative_query) is as equal as 1.

The English word w is the farthest of negative_query if ORC(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 ORC(w, negative_query) > 0 and ORC(w, negative_query) ≤ 1. So, the valence of the English word w is the value of ORC(w, positive_query) substracting the value of ORC(w, negative_query) and the eq. (8) is the equation of identifying the valence of the English word w.

We have the information about ORC, SO_ORC, etc. as follows:

1)ORC(w, positive_query) ≥ 0 and ORC(w,

positive_query) ≤ 1.
2)ORC(w, negative_query) ≥ 0 and ORC (w, negative_query) ≤ 1
3)If ORC (w, positive_query) = 0 and ORC (w, negative_query) = 0 then SO_ORC (w) = 0.
4)If ORC (w, positive_query) = 1 and ORC (w, negative_query) = 0 then SO_ORC (w) = 0.
5)If ORC (w, positive_query) = 0 and ORC (w, negative_query) = 1 then SO_ORC (w) = -1.
6)If ORC (w, positive_query) = 1 and ORC (w, negative_query) = 1 then SO_ORC(w) = 0.
So, SO_ORC (w) ≥ -1 and SO_ORC (w) ≤ 1.

The polarity of the English word w is positive polarity If SO_ORC (w) > 0. The polarity of the English word w is negative polarity if SO_ORC (w) < 0. The polarity of the English word w is neutral polarity if SO_ORC (w) = 0. In addition, the semantic value of the English word w is SO_ORC (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 eq. (8), eq. (9), and eq. (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.

We compare this result of the proposed model with the surveys in the tables as follows: Table 8, Table 9, Table 12, and Table 13.

In Table 8, we show the comparisons of our model with the researches related to the Odds Ratio Coefficient (ORC) in [15- 19]

The comparisons of our model’s positives and negatives the surveys related to the Odds Ratio Coefficient (ORC) in [15-19] are displayed in Table 9.

In Table 12, we present the comparisons of our model’s results with the works related to [1-32].

The comparisons of our model’s advantages and disadvantages with the works related to [1-32] are displayed in Table 13.

4.1.2 A basis English sentiment dictionary (bESD) in a sequential environment

At least 55,000 terms, including nouns , verbs, adjectives, etc. in English are based on [57-62]. The valence and the polarity of the English words or phrases for our basis English sentiment dictionary (bESD) by using the ORC are identified in a sequential system, as the following diagram in Fig. 6 below shows.

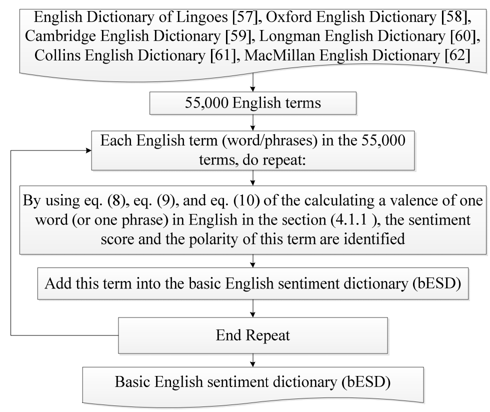

Fig. 6: Overview Of Creating A Basis English Sentiment Dictionary (Besd) In A Sequential Environment

The algorithm 1 is proposed to perform this section

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 eq. (8), eq. (9), and eq. (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 ORC 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;

More 55,000 English words (or English phrases) of our basis English sentiment dictionary (bESD) are stored in Microsoft SQL Server 2008 R2.

4.1.3 A basis English sentiment dictionary (bESD) in a distributed system

In this part, the valence and the polarity of the English words or phrases for our basis English sentiment dictionary (bESD) by using the ORC are calculated in a parallel network environment from at least 55,000 English terms, including nouns, verbs, adjectives, etc. based on [57-62], as the following diagram in Fig. 7 below shows.

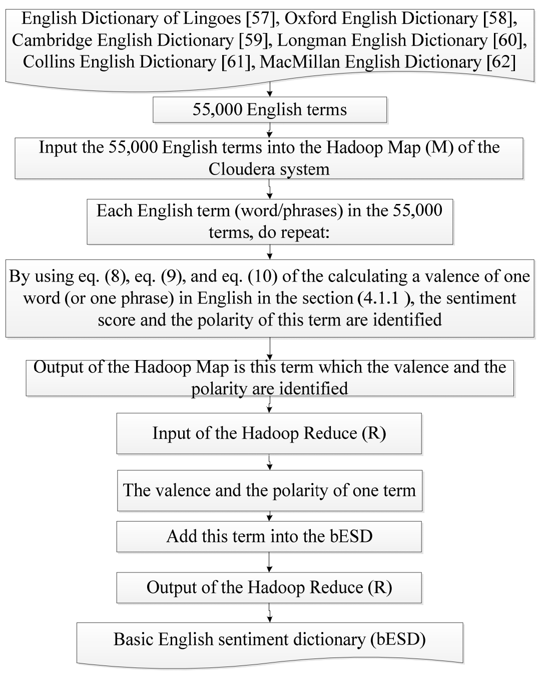

Fig. 7: Overview Of Creating A Basis English Sentiment Dictionary (Besd) In A Distributed Environment

In Fig. 7, there are two phases in this section as follows: 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 [57-62]. 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).

The algorithm 2 is built to implement the Hadoop Map phase.

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 eq. (8), eq. (9), and eq. (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 ORC through the Google search engine with AND operator and OR operator.

Step 3: Return this term;

The algorithm 3 is proposed to perform thie Hadoop Reduce phase. The algorithm 3 has the main ideas 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;

At least 55,000 English words (or English phrases) of our basis English sentiment dictionary (bESD) are stored in Microsoft SQL Server 2008 R2.

4.2 Using Self-Organizing Map Algorithm to cluster the documents of the testing data set into either the positive or the negative in both a sequential environment and a distributed system

The section comprise two parts. The first part is the sub-sectin (4.2.1) which we use the SOM to cluster the documents of the testing data set into either the positive or the negative in a sequential environment. The second part is the sub-section (4.2.2) which we use the SOM to cluster the documents of the testing data set into either the positive or the negative in a parallel network system.

4.2.1 Using Self-Organizing Map Algorithm to cluster the documents of the testing data set into either the positive or the negative in a sequential environment

In Fig. 8, we use Self-Organizing Map Algorithm to cluster the documents of the testing data set into either the positive or the negative in a sequential environment

In Fig. 8, this section is implemented in the sequential system as follows: we calculate the valences and the polarities of the sentiment lexicons of the bESD according to a basis English sentiment dictionary (bESD) in a sequential environment (4.1.2). We transfer all the n_max sentences of one document of the testing data set into the n_max one-dimensional vectors of this document. All the n_max one-dimensional vectors of this document are clustered into either the positive polarity or the negative polarity by using the SOM with the input is the n_max one-dimensional vectors.

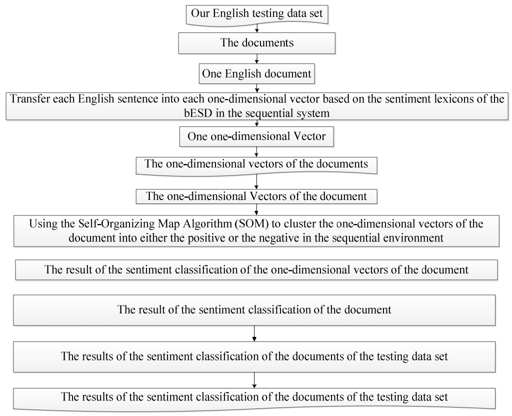

Fig. 8: Overview Of Using Self-Organizing Map Algorithm To Cluster The Documents Of The Testing Data Set Into Either The Positive Or The Negative In A Sequential Environment

We set an initialization of the SOM with its map in Fig. 9 as follows:

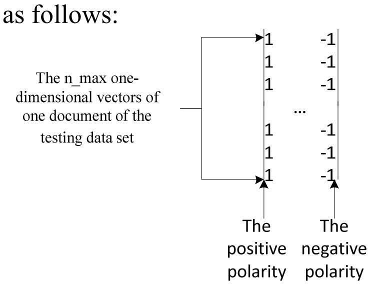

Fig. 9: An Initialization Of The SOM – The Map

Then, after the SOM is implemented completely, we have the Map in Fig. 10 as follows:

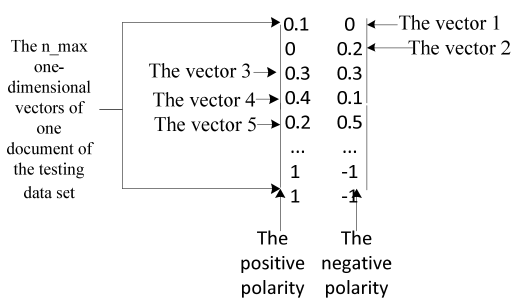

Fig. 10: The Final Map – The Result Of Clustering By Using The SOM

In Figue 10, we have the vector 1 (0.1, 0), the vector 2 (0, 0.2), the vector 3 (0.3, 0.3), the vector 4 (0.4, 0.1) , and the vector 5 (0.2, 0.5). With the vector 1, the column of the positive polarity is 0.1 and the column of the negative polarity is 0, Thus, the value of the column of the positive polarity is greater than the value of the column of the negative polarity, therefore this vector is clustered into the positive. With the vector 2, the column of the positive polarity is 0 and the column of the negative polarity is 0.2. Therefore, the value of the column of the positive polarity is less than the value of the column of the negative polarity, thus this vector is clustered into the negative. With the vector 3, the column of the positive polarity is 0.3 and the column of the negative polarity is 0.3. So, the value of the column of the positive polarity is as equal as the value of the column of the negative polarity, therefore this vector is not clustered into both the positive and the negative. It meas that this vector is clustered into the neutral polarity. With the vector 4 (0.4, 0.1), the column of the positive polarity is 0.4 and the column of the negative polarity is 0.1. Thus, the value of the column of the positive polarity is greater than the value of the column of the negative polarity, so this vector is clustered into the positive. With the vector 5, the column of the positive polarity is 0.2 and the column of the negative polarity is 0.5. Therefore, the value of the column of the positive polarity is less than the column of the negative polarity, thus this vector is clustered into the negative. One document is clustered into the positive if the number of the one-dimensional vectors (corresponding to the sentences of this document) clustered into the positive is greater than the number of the one-dimensional vectors (corresponding to the sentences of this document) clustered into the negative in the document. The document is clustered into the negative if the number of the one-dimensional vectors clustered into the positive is less than the number of the one-dimensional vectors clustered into the negative in the document. The document is clustered into the neutral if the number of the one-dimensional vectors clustered into the positive is as equal as the number of the one-dimensional vectors clustered into the negative in the document. Finally, the sentiment classification of all the documents of the testing data set is identified completely.

We proposed the algorithm 4 to transfer one English sentence into the one-dimensional vector based on the sentiment lexicons of the bESD the sequential environment.
Input: one English sentence
Output: one one-dimensional vector
Step 1: Split this sentence into the meaningful terms (meaningful words or meaningful phrases);
Step 2: One-dimensionalVector := null;
Step 3: Each term in the terms, do repeat:

Step 4: Get the valence of this term based on the sentiment lexicons of the bESD;
Step 5: Add this term into One-dimensionalVector;
Step 6: End Repeat – End Step 2
Step 7: Return One-dimensionalVector;

We built the algorithm 5 to transfer one English document into the one-dimensional vectors of the document in the sequential environment.
Input: one English document
Output: the one-dimensional vectors of this document
Step 1: Split the English document into many separate sentences based on "." Or "!" or "?";
Step 2: Ste TheOne-dimensionalVectors := null;
Step 3: Each sentence in the sentences of this document, do repeat:
Step 4: One-dimensionalVector := the algorithm 1 to transfer one English sentence into the one-dimensional vector based on the sentiment lexicons of the bESD the sequential environment with the input is this sentence;
Step 5: Add One-dimensionalVector into TheOne-dimensionalVectors;
Step 6: End Repeat – End Step 2
Step 7: Return TheOne-dimensionalVectors;

We proposed the algorithm 6 to cluster one document of the testing data set into either the positive or the negative by using the SOM in the sequential environment.
Input: one document
Output: positive, negative, neutral;
Step 1: Set Matrix := {}{} with the n_max rows, the 2 columns
Step 2: Set i:= 0;
Step 3: Each i in the 2 columns -1, do repeat:
Step 4: Set j := 0;
Step 5: Ech j in the n_max rows -1, do repeat:
Step 6: If i is as equal as 0 Then Matrix[j][i] :=1;
Step 7: If i is as equal as 1 Then Matrix[j][i] :=-1;
Step 8: End Repeat – End Step 5
Step 9: End Repeat – End Step 3
Step 10: Set Learning rate := 0.9;
Step 11: Set R := 0;
Step 12: While stopping condition false do step 13 to 19
Step 13: For each input vector x do step 14 to 16
Step 14: For ech j neuron, compute the Euclidean distance D(j)
Step 15: Find the index J such D(j) is a minimum
Step 16: For all neurons j within a specified neighbourhood of J and for all i: wji (new)= wij(old)+ learning rate * (xi - wij (old))
Step 17: Update learning rate. It is a decreasing function of the number of epochs: learning rate (t+1) = [learning rate(t)]/2;
Step 18: Reduce radius of topolofical neighbourhood at specified times
Step 19: Test stop condition. Typically this is a small value of the learning rate with which the weight updates are insignificant.
Step 20: Set count_positive := 0 and count_negative := 0;
Step 20: Ech j in the n_max rows -1, do repeat:
Step 21: If Matrix[j][0] is greater than Matrix[j][1] Then count_positive := count_positive +1;
Step 22: If Matrix[j][0] is less than Matrix[j][1] Then count_negative := count_negative +1;
Step 23: End Repeat – End Step 20;
Step 24: If count_positive is greater than count_negative Then Return positive;
Step 25: Else If count_positive is less than count_negative Then Return negative;
Step 26: Return neutral;

We proposeD the algorithm 7 to cluster the documents of the testing data set into either positive or the negative by using the SOM in the sequential system.
Input: the testing data set
Output: the results of the sentiment classification of the testing data set
Step 1: Set TheResults := null;
Step 2: Each document in the documents of the testing data set, do repeat:
Step 3: OneResult := the algorithm 3 to cluster one document of the testing data set into either the positive or the negative by using the SOM in the sequential environment with the input is this document;
Step 4: Add OneResult into TheResults;
Step 5: End Repeat – End Step 2;
Step 6: Return TheResults;

4.2.2 Using Self-Organizing Map Algorithm to cluster the documents of the testing data set into either the positive or the negative in a distributed system

In Fig. 11, we use Self-Organizing Map Algorithm to cluster the documents of the testing data set into either the positive or the negative in a distributed environment.

In Fig. 11, this section is implemented in the distributed system as follows: we calculate the valences and the polarities of the sentiment lexicons of the bESD based on a basis English sentiment dictionary (bESD) in a distributed system (4.1.3). We transfer all the n_max sentences of one

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

document of the testing data set into the n_max one-dimensional vectors of this document. All the n_max one-dimensional vectors of this document are clustered into either the positive polarity or the negative polarity by using the SOM with the input is the n_max one-dimensional vectors.

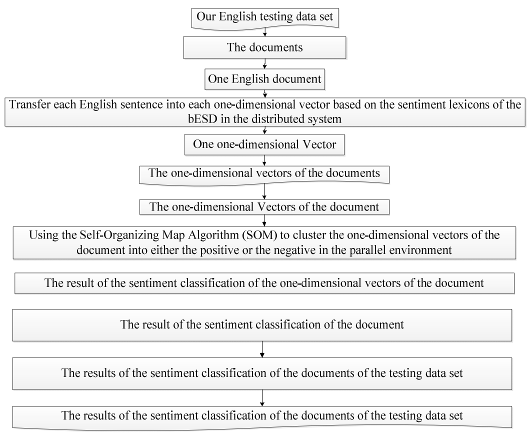

Fig. 11. Overview Of Using Self-Organizing Map Algorithm To Cluster The Documents Of The Testing Data Set Into Either The Positive Or The Negative In A Parallel Environment

We set an initialization of the SOM with its map in Fig. 12 as follows:

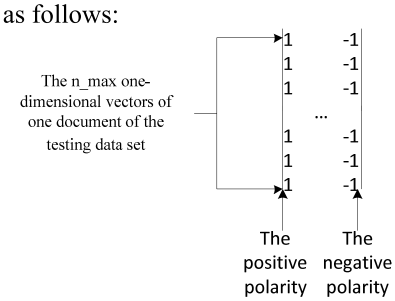

Fig. 12: An Initialization Of The SOM – **the Map**

Then, after the SOM is implemented completely, we have the Map in Fig. 13 as follows:

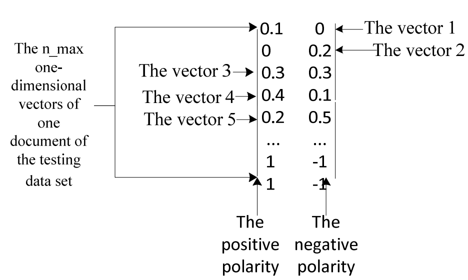

Fig. 13: The Final Map – The Result Of Clustering By Using The SOM

In Figue 13, we have the vector 1 (0.1, 0), the vector 2 (0, 0.2), the vector 3 (0.3, 0.3), the vector 4 (0.4, 0.1) , and the vector 5 (0.2, 0.5). With the vector 1, the column of the positive polarity is 0.1 and the column of the negative polarity is 0, Thus, the value of the column of the positive polarity is greater than the value of the column of the negative polarity, therefore this vector is clustered into the positive. With the vector 2, the column of the positive polarity is 0 and the column of the negative polarity is 0.2. Therefore, the value of the column of the positive polarity is less than the value of the column of the negative polarity, thus this vector is clustered into the negative. With the vector 3, the column of the positive polarity is 0.3 and the column of the negative polarity is 0.3. So, the value of the column of the positive polarity is as equal as the value of the column of the negative polarity, therefore this vector is not clustered into both the positive and the negative. It meas that this vector is clustered into the neutral polarity. With the vector 4 (0.4, 0.1), the column of the positive polarity is 0.4 and the column of the negative polarity is 0.1. Thus, the value of the column of the positive polarity is greater than the value of the column of the negative polarity, so this vector is clustered into the positive. With the vector 5, the column of the positive polarity is 0.2 and the column of the negative polarity is 0.5. Therefore, the value of the column of the positive polarity is less than the column of the negative polarity, thus this vector is clustered into the negative. One document is clustered into the positive if the number of the one-dimensional vectors (corresponding to the sentences of this document) clustered into the positive is greater than the number of the one-dimensional vectors (corresponding to the sentences of this document) clustered into the negative in the document. The document is clustered into the negative if the number of the one-dimensional vectors clustered into the positive is less than the number of the one-dimensional vectors clustered into the negative in the document. The document is clustered into the neutral if the number of the one-dimensional vectors clustered into the positive is as equal as the number of the one-dimensional vectors clustered into the negative in the document. Finally, the sentiment classification of all the documents of the testing data set is identified completely.

In Fig. 14, we transfer one sentence into one one-dimensional vector based on the sentiment lexicons of the bESD in the parallel system as follows:

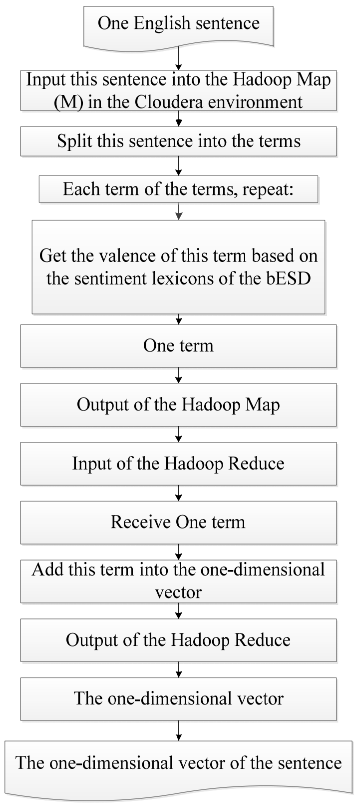

Fig. 14: Overview Of Transferring One Sentence Into One One-Dimensional Vector Based On The Sentiment Lexicons Of The Besd In The Parallel System

In Fig. 14, this stage includes two phases: the Hadoop Map phase and the Hadoop Reduce phase. The input of the Hadoop Map phase is one sentence. The output of the Hadoop Mp is one term. The input of the Hadoop Reduce is the Hadoop Map, thus, the input of the Hadoop Reduce is one term. The output of the Hadoop Reduce is the one-dimensional vector of this sentence.

We proposed the algorithm 8 to implement the Hadoop Map phase
Input: one sentence;
Output: one term;
Step 1: Input this document into the Hadoop Map in the Cloudera system.
Step 2: Split this sentence into the meaningful terms;
Step 3: Each term in the terms, do repeat:
Step 4:Get the valence of this term based on the sentiment lexicons of the bESD;
Step 5: Return this term;
Step 6: The output of the Hadoop Map is this term;

We built the algorithm 9 to implement the Hadoop Reduce phase
Input: one term of the Hadoop Map (the input of the Hadoop Reduce is the output of the Hadoop Map)
Output: the one-dimensional vector of the English sentence – One-dimensionalVector;
Step 1: Receive one termp;
Step 2: Add this term into One-dimensionalVector;
Step 3: Return One-dimensionalVector;

In Fig. 15, we transfer one document into the one-dimensional vectors of the document based on the sentiment lexicons of the bESD in the parallel system as follows:

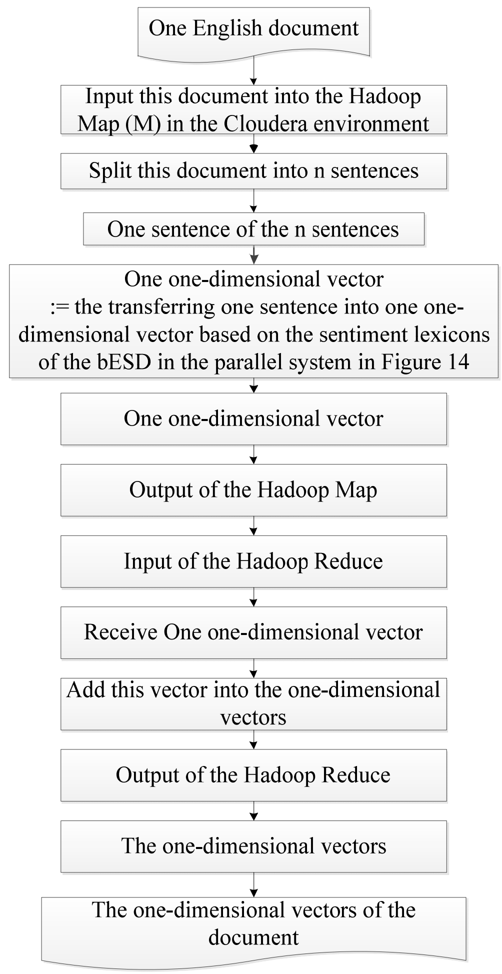

Fig. 15: Overview Of Transferring One Document Into The One-Dimensional Vectors Of The Document Based On The Sentiment Lexicons Of The Besd In The Parallel System

In Fig. 15, this stage includes two phases: the Hadoop Map phase and the Hadoop Reduce phase. The input of the Hadoop Map phase is one document. The output of the Hadoop Mp is one one-dimensional vector. The input of the Hadoop Reduce is the Hadoop Map, thus, the input of the

Hadoop Reduce is one one-dimensional vector. The output of the Hadoop Reduce is the one-dimensional vectors of this dcoument.

We proposed the algorithm 10 to implement the Hadoop Map phase
Input: one document;
Output: one one-dimensional vector (corresponding to one sentence)
Step 1: Input this document into the Hadoop Map in the Cloudera system.
Step 2: Split this document into the n sentences;
Step 3: Each sentence in the n sentences, do repeat:
Step 4: the one-dimensional vector := the transferring one sentence into one one-dimensional vector based on the sentiment lexicons of the bESD in the parallel system in Fig. 14 with the input is this sentence;
Step 5: Return this one-dimensional vector;
Step 6: The output of the Hadoop Map is this one-dimensional vector;

We proposed the algorithm 11 to implement the Hadoop Reduce phase
Input: one one-dimensional vector of the Hadoop Map (the input of the Hadoop Reduce is the output of the Hadoop Map)
Output: the one-dimensional vectors of the English document
Step 1: Receive one one-dimensional vector of the Hadoop Map
Step 2: Add this one-dimensional vector into the one-dimensional vectors of the English document
Step 3: Return the one-dimensional vectors of the English document;

In Fig. 16, we use the Self-Organizing Map Algorithm (SOM) to cluster one document into either the positive or the negative in the distributed system. In Fig. 16, this stage comprises two phases: the Hadoop Map phase and the Hadoop Reduce phase. The input of the Hadoop Map is one document of the testing data set. The output of the Hadoop Map is the result of the sentiment classification of this document. The input of the Hadoop Reduce is the output of the Hadoop Map, thus, the input of the Hadoop Reduce is the result of the sentiment classification of this document. The output of the Hadoop Reduce is the result of the sentiment classification of this document

We proposed the algorithm 12 to perform the Hadoop Map phase
Input: one document
Output: the result of the sentiment classification of this document
Step 1: Set Matrix := {}{} with the n_max rows, the 2 columns
Step 2: Set i:= 0;
Step 3: Each i in the 2 columns -1, do repeat:
Step 4: Set j := 0;
Step 5: Ech j in the n_max rows -1, do repeat:
Step 6: If i is as equal as 0 Then Matrix[j][i] :=1;
Step 7: If i is as equal as 1 Then Matrix[j][i] :=-1;
Step 8: End Repeat – End Step 5
Step 9: End Repeat – End Step 3
Step 10: Set Learning rate := 0.9;
Step 11: Set R := 0;
Step 12: While stopping condition false do step 13 to 19
Step 13: For each input vector x do step 14 to 16
Step 14: For ech j neuron, compute the Euclidean distance D(j)
Step 15: Find the index J such D(j) is a minimum
Step 16: For all neurons j within a specified neighbourhood of J and for all i: wji (new)= wij(old)+ learning rate * (xi - wij (old))
Step 17: Update learning rate. It is a decreasing function of the number of epochs: learning rate (t+1) = [learning rate(t)]/2;
Step 18: Reduce radius of topolofical neighbourhood at specified times
Step 19: Test stop condition. Typically this is a small value of the learning rate with which the weight updates are insignificant.
Step 20: Set count_positive := 0 and count_negative := 0;
Step 20: Ech j in the n_max rows -1, do repeat:
Step 21: If Matrix[j][0] is greater than Matrix[j][1] Then count_positive := count_positive +1;
Step 22: If Matrix[j][0] is less than Matrix[j][1] Then count_negative := count_negative +1;
Step 23: End Repeat – End Step 20;
Step 24: If count_positive is greater than count_negative Then Return positive;
Step 25: Else If count_positive is less than count_negative Then Return negative;
Step 26: Return neutral;

We built the algorithm 13 to implement the Hadoop Reduce phase
Input the result of the sentiment classification of this document (the input of the Hadoop Reduce is the output of the Hadoop Map)
Output: the result of the sentiment classification of this document
Step 1: Receive the result of the sentiment classification of this document
Step 2: Return the result of the sentiment classification of this document;

Journal of Theoretical and Applied Information Technology
31st August 2018. Vol.96. No 16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

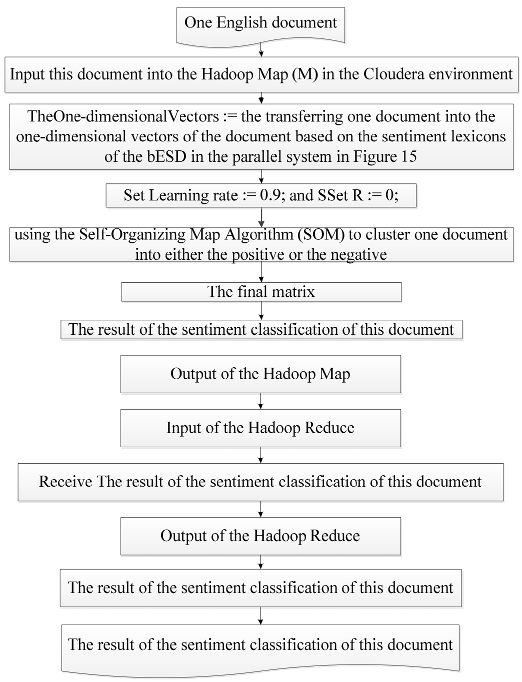

Fig. 16: Overview Of Using The Self-Organizing Map Algorithm (SOM) To Cluster One Document Into Either The Positive Or The Negative In The Distributed System.

In Fig. 17, we cluster the documents of the testing data set into either positive or the negative by using the Self-Organizing Map Algorithm (SOM) in the parallel system. In Fig. 17, this stage includes two phases: the Hadoop Map phase and the Hadoop Reduce phase. The input of the Hadoop Map phase is the documents of the testing data set. The output of the Hadoop Mp is one result of the sentiment classification of one document of the testing data set. The input of the Hadoop Reduce is the Hadoop Map, thus, the input of the Hadoop Reduce is one result of the sentiment classification of one document of the testing data set. The output of the Hadoop Reduce is the results of the sentiment classification of the testing data set

We proposed the algorithm 14 to implement the Hadoop Map phase
Input: the documents of the testing data set and the training data set;
Output: one result of the sentiment classification of one document of the testing data set;
Step 1: The sentiment lexicons of the bESD are created based on a basis English sentiment dictionary (bESD) in a distributed system (4.1.3);
Step 2: Each document of the documents of the testing data set, do repeat:
Step 3: OneResult := the using the Self-Organizing Map Algorithm (SOM) to cluster one document into either the positive or the negative in the distributed system in Fig. 16 with the input is this document;
Step 4: Return this OneResult;
Step 5: The output of the Hadoop Map is this OneResult;

We built the algorithm 15 to implement the Hadoop Reduce phase
Input: one result of the sentiment classification of one document of the testing data set;
Output: the results of the sentiment classification of the testing data set;
Step 1: Receive OneResult of the Hadoop Map
Step 2: Add this OnResult 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;

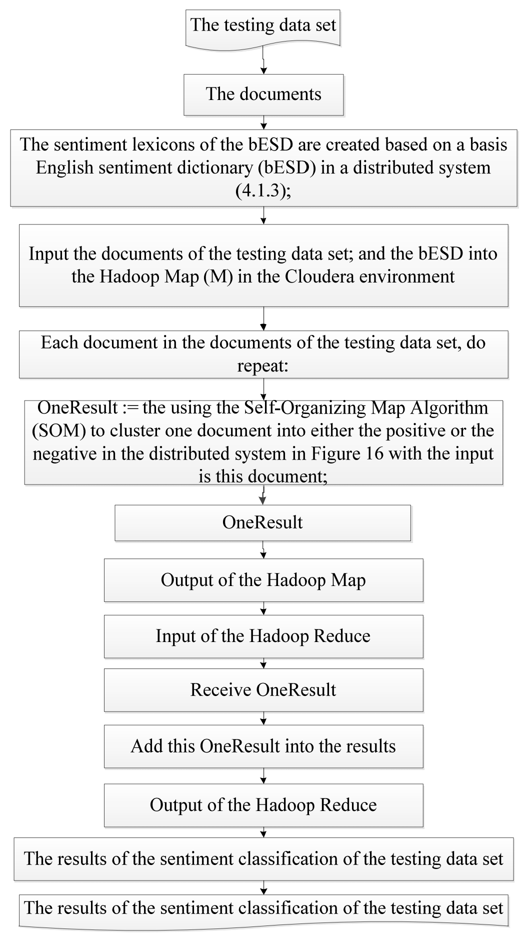

Fig. 17: Overview Of Clustering The Documents Of The Testing Data Set Into Either Positive Or The Negative By Using The Self-Organizing Map Algorithm (SOM)

5. EXPERIMENT

We have measured Accuracy (A) to calculate the accuracy of the results of emotion classification.

We used a Java programming language for programming to save data sets, implementing our proposed model to classify the 8,500,000 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 proposed model was implemented in both the sequential system and the distributed network environment.

Our model related to the Self-Organizing Map algorithm, a testing data set with the one-dimensional vectors and An Odds Ratio coefficient is implemented in the sequential environment with the configuration as follows: The sequential environment in this research includes 1 node (1 server). The configuration of the server in the sequential environment is: Intel® Server Board S1200V3RPS, Intel® Pentium® Processor G3220 (3M Cache, 3.00 GHz), 2GB CC3-10600 ECC 1333 MHz LP Unbuffered DIMMs. The operating system of the server is: Cloudera. The Java language is used in programming our model related to the Self-Organizing Map algorithm, a testing data set with the one-dimensional vectors and An Odds Ratio coefficient.

The proposed model related to the Self-Organizing Map algorithm, a testing data set with the one-dimensional vectors and An Odds Ratio coefficient is performed in the Cloudera parallel network environment with the configuration as follows: This Cloudera system includes 9 nodes (9 servers). The configuration of each server in the Cloudera system is: Intel® Server Board S1200V3RPS, Intel® Pentium® Processor G3220 (3M Cache, 3.00 GHz), 2GB CC3-10600 ECC 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. The Java language is used in programming the application of the proposed model related to the Odds Ratio similarity coefficient of the clustering technologies in the Cloudera

In Table 1, the results of the documents of the English testing data set to test are presented.

The accuracy of the sentiment classification of the documents in the English testing data set is shown in Table 2 below.

The average time of the classification of our new model for the English documents in testing data set are displayed in Table 3.

6. CONCLUSION

In this survey, a new model has been proposed to classify sentiment of many documents in English using the Self-Organizing Map algorithm, a testing data set with the one-dimensional vectors and An Odds Ratio coefficient with Hadoop Map (M) /Reduce (R) in the Cloudera parallel network environment. Based on our proposed new model, we have achieved 88.14% accuracy of the testing data set in Table 2. 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 the sentiments (positive, negative, or neutral) in text.

The proposed model can be applied to other languages although our new model has been tested on our English data set. Our model can be applied to larger data sets with millions of English documents in the shortest time although our model has been tested on the documents of the testing data set in which the data sets are small in this survey.

According to Table 3, the average time of the sentiment classification of using the Self-Organizing Map algorithm, a testing data set with the one-dimensional vectors and An Odds Ratio coefficient in the sequential environment is 34,734,059 seconds / 8,500,000 English documents and it is greater than the average time of the sentiment classification of using the Self-Organizing Map algorithm, a testing data set with the one-dimensional vectors and An Odds Ratio coefficient in the Cloudera parallel network environment with 3 nodes which is 10,244,686 seconds / 8,500,000 English documents. The average time of the sentiment classification of using the Self-Organizing Map algorithm, a testing data set with the one-dimensional vectors and An Odds Ratio coefficient in the Cloudera parallel network environment with 9 nodes is 3,881,562 seconds / 8,500,000 English documents, and It is the shortest time in the table. Besides, The average time of the sentiment classification of using the Self-Organizing Map algorithm, a testing data set with the one-dimensional vectors and An Odds Ratio coefficient in the Cloudera parallel network environment with 6 nodes is 5,922,343 seconds / 8,500,000 English documents

The execution time of using the Self-Organizing Map algorithm, a testing data set with the one-dimensional vectors and An Odds Ratio coefficient 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 accuracy of the proposed model is depending on many factors as follows:

1)The Odds Ratio Coefficient (ORC)

2)The SOM – related algorithms

3)The testing data set

4)The documents of the testing data set must be standardized carefully.

5)Transferring one document into one one-dimensional vector

The execution time of the proposed model is depending on many factors as follows:

1)The parallel network environment such as the Cloudera system.

2)The distributed functions such as Hadoop Map (M) and Hadoop Reduce (R).

3)The ORC – related algorithms

4)The performance of the distributed network system.

5)The number of nodes of the parallel network environment.

6)The performance of each node (each server) of the distributed environment.

7)The sizes of the training data set and the testing data set.

8)Transferring one document into one one-dimensional vector.

The proposed model has many advantages and disadvantages. Its positives are as follows: It uses the Odds Ratio similarity coefficient of the clustering technologies 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.

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

In Table 4, the comparisons of our model's results with the works in [1-3] are shown.

The comparisons of our model's advantages and disadvantages with the works in [1-3] are presented in Table 5.

In Table 6, the comparisons of our model with the latest sentiment classification models (or the latest sentiment classification methods) in [4-14] are displayed.

The comparisons of our model's positives and negatives with the latest sentiment classification models (or the latest sentiment classification methods) in [4-14] are shown in Table 7.

In Table 8, the comparisons of our model with the rsearches related to the Odds Ratio Coefficient (ORC) in [15- 19] are presented.

The comparisons of our model's positives and negatives the surveys related to the Odds Ratio Coefficient (ORC) in [15-19] are displayed in Table 9.

In Table 10, we show the comparisons of our model with the rsearches related to Self-Organizing Map Algorithm (SOM) in [20- 24].

The comparisons of our model's positives and negatives the surveys related to the Self-Organizing Map Algorithm (SOM) in [20-24] are displayed in Table 11.

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.

REFRENCES:

[1] Vaibhav Kant Singh, Vinay Kumar Singh, "Vector Space Model: An Information Retrieval System", Int. J. Adv. Engg. Res. Studies/IV/II/Jan.-March,2015/141-143, 2015

[2] Víctor Carrera-Trejo, Grigori Sidorov, Sabino Miranda-Jiménez, Marco Moreno Ibarra and Rodrigo Cadena Martínez, "*Latent Dirichlet Allocation complement in the vector space model for Multi-Label Text Classification*", International Journal of Combinatorial Optimization Problems and Informatics, Vol. 6, No. 1, pp. 7-19, 2015

[3] Pascal Soucy, Guy W. Mineau, "*Beyond TFIDF Weighting for Text Categorization in the Vector Space Model*", Proceedings of the 19th International Joint Conference on Artificial Intelligence, pp. 1130-1135, USA, 2015

[4] Basant Agarwal, Namita Mittal, "*Machine Learning Approach for Sentiment Analysis*", Prominent Feature Extraction for Sentiment Analysis, Print ISBN 978-3-319-25341-1, 10.1007/978-3-319-25343-5_3, 21-45, 2016

[5] Basant Agarwal, Namita Mittal, "*Semantic Orientation-Based Approach for Sentiment*

Analysis", Prominent Feature Extraction for Sentiment Analysis, Print ISBN 978-3-319-25341-1, 10.1007/978-3-319-25343-5_6, 77-88, 2016

[6] Sérgio Canuto, Marcos André, Gonçalves, Fabrício Benevenuto, "*Exploiting New Sentiment-Based Meta-level Features for Effective Sentiment Analysis*", Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM '16), 53-62, New York USA, 2016

[7] Shoiab Ahmed, Ajit Danti, "*Effective Sentimental Analysis and Opinion Mining of Web Reviews Using Rule Based Classifiers*", Computational Intelligence in Data Mining, Volume 1, Print ISBN 978-81-322-2732-8, DOI 10.1007/978-81-322-2734-2_18, 171-179, India, 2016

[8] Vo Ngoc Phu, Phan Thi Tuoi, "*Sentiment classification using Enhanced Contextual Valence Shifters*", International Conference on Asian Language Processing (IALP), 224-229, 2014

[9] Vo Thi Ngoc Tran, Vo Ngoc Phu and Phan Thi Tuoi, "*Learning More Chi Square Feature Selection to Improve the Fastest and Most Accurate Sentiment Classification*", The Third Asian Conference on Information Systems (ACIS 2014), 2014

[10] Nguyen Duy Dat, Vo Ngoc Phu, Vo Thi Ngoc Tran, Vo Thi Ngoc Chau, "*STING Algorithm used English Sentiment Classification in A Parallel Environment*", International Journal of Pattern Recognition and Artificial Intelligence, January 2017.

[11] Vo Ngoc Phu, Nguyen Duy Dat, Vo Thi Ngoc Tran, Vo Thi Ngoc Tran, "*Fuzzy C-Means for English Sentiment Classification in a Distributed System*", International Journal of Applied Intelligence (APIN), DOI: 10.1007/s10489-016-0858-z, 1-22, November 2016.

[12] Vo Ngoc Phu, Chau Vo Thi Ngoc, Tran Vo THi Ngoc, Dat Nguyen Duy, "*A C4.5 algorithm for english emotional classification*", Evolving Systems, pp 1-27, doi:10.1007/s12530-017-9180-1, April 2017.

[13] Vo Ngoc Phu, Vo Thi Ngoc Chau, Vo Thi Ngoc Tran, "*SVM for English Semantic Classification in Parallel Environment*", International Journal of Speech Technology (IJST), 10.1007/s10772-017-9421-5, 31 pages, May 2017.

[14] Vo Ngoc Phu, Vo Thi Ngoc Tran, Vo Thi Ngoc Chau, Nguyen Duy Dat, Khanh Ly Doan Duy, "*A Decision Tree using ID3 Algorithm for English Semantic Analysis*", International Journal of Speech Technology (IJST), DOI: 10.1007/s10772-017-9429-x, 23 pages, 2017

[15] Seung-Seok Choi, Sung-Hyuk Cha, Charles C. Tappert, "*A Survey Of Binary Similarity And Distance Measures*", Systemics, Cybernetics And Informatics, Issn: 1690-4524, Volume 8 - Number 1, 2010

[16] Stuart R. Lipsitz, Nan M. Laird, David P. Harrington, "*Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association*", Biometrika, Volume 78, Issue 1, 1 March 1991, Pages 153–160, https://doi.org/10.1093/biomet/78.1.153, 1991

[17] Afina S. Glas, Jeroen G. Lijmer, Martin H. Prins, Gouke J. Bonsel, Patrick M.M. Bossuyt, "*The diagnostic odds ratio: a single indicator of test performance*", Journal of Clinical Epidemiology, Volume 56, Issue 11, November 2003, Pages 1129-1135, https://doi.org/10.1016/S0895-4356(03)00177-X, 2003

[18] Szilard Nemes, Junmei Miao Jonasson, Anna Genell, Gunnar Steineck, "*Bias in odds ratios by logistic regression modelling and sample size*", BMC Medical Research Methodology 2009 9:56, https://doi.org/10.1186/1471-2288-9-56, 2009

[19] Margaret Sullivan Pepe, Holly Janes, Gary Longton, Wendy Leisenring, Polly Newcomb, "*Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic, Prognostic, or Screening Marker*", American Journal of Epidemiology, Volume 159, Issue 9, 1 May 2004, Pages 882–890, https://doi.org/10.1093/aje/kwh101, 2004

[20] T. Kohonen, "*The self-organizing map*", Proceedings of the IEEE, Volume: 78, Issue: 9,DOI: 10.1109/5.58325, 1990

[21] J.L. Giraudel, S. Lek, "*A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination*", Ecological Modelling, Volume 146, Issues 1–3, Pages 329-339, https://doi.org/10.1016/S0304-3800(01)00324-6, 2001

[22] J. Vesanto; E. Alhoniemi, "*Clustering of the self-organizing map*", IEEE Transactions on Neural Networks, Volume: 11, Issue: 3,DOI: 10.1109/72.846731, 2000

[23] Roussinov, Dmitri G.; Chen, Hsinchun, "*A Scalable Self-organizing Map Algorithm for*

Textual Classification: A Neural Network Approach to Thesaurus Generation", Communication and Cognition in Artificial Intelligence Journal, 15(1-2):81-111, 1998

[24] E. Berglund; J. Sitte, "*The parameterless self-organizing map algorithm*", IEEE Transactions on Neural Networks, Volume: 17, Issue: 2, DOI: 10.1109/TNN.2006.871720, 2006

[25] Aleksander Bai, Hugo Hammer, "*Constructing sentiment lexicons in Norwegian from a large text corpus*", 2014 IEEE 17th International Conference on Computational Science and Engineering, 2014

[26] P.D.Turney, M.L.Littman, "*Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus*", arXiv:cs/0212012, Learning (cs.LG); Information Retrieval (cs.IR), 2002

[27] Robert Malouf, Tony Mullen, "*Graph-based user classification for informal online political discourse*", In proceedings of the 1st Workshop on Information Credibility on the Web, 2017

[28] Christian Scheible, "*Sentiment Translation through Lexicon Induction*", Proceedings of the ACL 2010 Student Research Workshop, Sweden, pp 25–30, 2010

[29] Dame Jovanoski, Veno Pachovski, Preslav Nakov, "*Sentiment Analysis in Twitter for Macedonian*", Proceedings of Recent Advances in Natural Language Processing, Bulgaria, pp 249–257, 2015

[30] Amal Htait, Sebastien Fournier, Patrice Bellot, "*LSIS at SemEval-2016 Task 7: Using Web Search Engines for English and Arabic Unsupervised Sentiment Intensity Prediction*", Proceedings of SemEval-2016, California, pp 481–485, 2016

[31] Xiaojun Wan, "*Co-Training for Cross-Lingual Sentiment Classification*", Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, Singapore, pp 235–243, 2009

[32] Julian Brooke, Milan Tofiloski, Maite Taboada, "*Cross-Linguistic Sentiment Analysis: From English to Spanish*", International Conference RANLP 2009 - Borovets, Bulgaria, pp 50–54., 2009

[33] Tao Jiang, Jing Jiang, Yugang Dai, Ailing Li, "*Micro–blog Emotion Orientation Analysis Algorithm Based on Tibetan and Chinese Mixed Text*", International Symposium on Social Science (ISSS 2015), 2015

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

[35] Weifu Du, Songbo Tan, Xueqi Cheng, Xiaochun Yun, "*Adapting Information Bottleneck Method for Automatic Construction of Domain-oriented Sentiment Lexicon*", WSDM'10, New York, USA, 2010

[36] Ziqing Zhang, Qiang Ye, Wenying Zheng, Yijun Li, "*Sentiment Classification for Consumer Word-of-Mouth in Chinese: Comparison between Supervised and Unsupervised Approaches*", The 2010 International Conference on E-Business Intelligence, 2010

[37] Guangwei Wang, Kenji Araki, "*Modifying SO-PMI for Japanese Weblog Opinion Mining by Using a Balancing Factor and Detecting Neutral Expressions*", Proceedings of NAACL HLT 2007, Companion Volume, NY, pp 189–192, 2007

[38] Shi Feng, Le Zhang, Binyang Li Daling Wang, Ge Yu, Kam-Fai Wong, "*Is Twitter A Better Corpus for Measuring Sentiment Similarity?* ", Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, USA, pp 897–902, 2013

[39] Nguyen Thi Thu An, Masafumi Hagiwara, "*Adjective-Based Estimation of Short Sentence's Impression*", (KEER2014) Proceedings of the 5th Kanesi Engineering and Emotion Research; International Conference; Sweden, 2014

[40] Nihalahmad R. Shikalgar, Arati M. Dixit, "*JIBCA: Jaccard Index based Clustering Algorithm for Mining Online Review",* International Journal of Computer Applications (0975 – 8887), Volume 105 – No. 15, 2014

[41] Xiang Ji, Soon Ae Chun, Zhi Wei, James Geller, "*Twitter sentiment classification for measuring public health concerns*", Soc. Netw. Anal. Min. (2015) 5:13, DOI 10.1007/s13278-015-0253-5, 2015

[42] Nazlia Omar, Mohammed Albared, Adel Qasem Al-Shabi, Tareg Al-Moslmi, "*Ensemble of Classification algorithms for Subjectivity and Sentiment Analysis of Arabic Customers' Reviews*", International Journal of Advancements in Computing Technology(IJACT), Volume5, 2013

[43] Huina Mao, Pengjie Gao, Yongxiang Wang, Johan Bollen, "*Automatic Construction of Financial Semantic Orientation Lexicon from*

Large-Scale Chinese News Corpus", 7th Financial Risks International Forum, Institut Louis Bachelier, 2014

[44] Yong REN, Nobuhiro KAJI, Naoki YOSHINAGA, Masaru KITSUREGAW, "*Sentiment Classification In Under-Resourced Languages Using Graph-Based Semi-Supervised Learning Methods*", IEICE TRANS. INF. & SYST., VOL.E97–D, NO.4, DOI: 10.1587/Transinf.E97.D.1, 2014

[45] Oded Netzer, Ronen Feldman, Jacob Goldenberg, Moshe Fresko, "*Mine Your Own Business: Market-Structure Surveillance Through Text Mining*", Marketing Science, Vol. 31, No. 3, pp 521-543, 2012

[46] Yong Ren, Nobuhiro Kaji, Naoki Yoshinaga, Masashi Toyoda, Masaru Kitsuregawa, "*Sentiment Classification in Resource-Scarce Languages by using Label Propagation*", Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation, Institute of Digital Enhancement of Cognitive Processing, Waseda University, pp 420 – 429, 2011

[47] José Alfredo Hernández-Ugalde, Jorge Mora-Urpí, Oscar J. Rocha, "*Genetic relationships among wild and cultivated populations of peach palm (Bactris gasipaes Kunth, Palmae): evidence for multiple independent domestication events*", Genetic Resources and Crop Evolution, Volume 58, Issue 4, pp 571-583, 2011

[48] Julia V. Ponomarenko, Philip E. Bourne, Ilya N. Shindyalov, "*Building an automated classification of DNA-binding protein domains*", BIOINFORMATICS, Vol. 18, pp S192-S201, 2002

[49] Andréia da Silva Meyer, Antonio Augusto Franco Garcia, Anete Pereira de Souza, Cláudio Lopes de Souza Jr, "*Comparison of similarity coefficients used for cluster analysis with dominant markers in maize (Zea maysL)* ", Genetics and Molecular Biology, 27, 1, 83-91, 2004

[50] Snežana Mladenović Drinić, Ana Nikolić, Vesna Perić, "*Cluster Analysis of Soybean Genotypes Based on RAPD Markers*", Proceedings. 43rd Croatian and 3rd International Symposium on Agriculture. Opatija. Croatia, 367- 370, 2008

[51] Tamás, Júlia; Podani, János; Csontos, Péter, "*An extension of presence/absence coefficients to abundance data:a new look at absence*", Journal of Vegetation Science 12: 401-410, 2001

[52] Vo Ngoc Phu, Vo Thi Ngoc Chau, Vo Thi Ngoc Tran, Nguyen Duy Dat, "*A Vietnamese adjective emotion dictionary based on exploitation of Vietnamese language characteristics*", International Journal of Artificial Intelligence Review (AIR), doi:10.1007/s10462-017-9538-6, 67 pages, 2017

[53] Vo Ngoc Phu, Vo Thi Ngoc Chau, Nguyen Duy Dat, Vo Thi Ngoc Tran, Tuan A. Nguyen, "*A Valences-Totaling Model for English Sentiment Classification",* International Journal of Knowledge and Information Systems, DOI: 10.1007/s10115-017-1054-0, 30 pages, 2017

[54] Vo Ngoc Phu, Vo Thi Ngoc Chau, Vo Thi Ngoc Tran, "*Shifting Semantic Values of English Phrases for Classification*", International Journal of Speech Technology (IJST), 10.1007/s10772-017-9420-6, 28 pages, 2017

[55] Vo Ngoc Phu, Vo Thi Ngoc Chau, Vo Thi Ngoc Tran, Nguy Duy Dat, Khanh Ly Doan Duy, "*A Valence-Totaling Model for Vietnamese Sentiment Classification*", International Journal of Evolving Systems (EVOS), DOI: 10.1007/s12530-017-9187-7, 47 pages, 2017

[56] Vo Ngoc Phu, Vo Thi Ngoc Chau, Vo Thi Ngoc Tran, Nguyen Duy Dat, Khanh Ly Doan Duy, "Semantic Lexicons of English Nouns for Classification", International Journal of Evolving Systems, DOI: 10.1007/s12530-017-9188-6, 69 pages, 2017

[57] English Dictionary of Lingoez, *http://www.lingoes.net/*, 2017

[58] Oxford English Dictionary, *http://www.oxforddictionaries.com/,* 2017

[59] Cambridge English Dictionary, *http://dictionary.cambridge.org/*, 2017

[60] Longman English Dictionary, *http://www.ldoceonline.com/*, 2017

[61] Collins English Dictionary, *http://www.collinsdictionary.com/dictionary/english*, 2017

[62] MacMillan English Dictionary, *http://www.macmillandictionary.com/*, 2017

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

APPENDICES:

Table 1: The results of the English documents in the testing data set.

	Testing Dataset	Correct Classification	Incorrect Classification
Negative	4,250,000	3,737,947	512,053
Positive	4,250,000	3,753,953	496,047
Summary	8,500,000	7,491,900	1,008,100

Table 2: The accuracy of our new model for the English documents in the testing data set.

Proposed Model	Class	Accuracy
Our new model	Negative	88.14%
	Positive	

Table 3: Average time of the classification of our new model for the English documents in testing data set.

	Average time of the classification / 8,500,000 English documents.
The proposed approach in the sequential environment	34,734,059 seconds
The proposed approach in the Cloudera distributed system with 3 nodes	10,244,686 seconds
The proposed approach in the Cloudera distributed system with 6 nodes	5,922,343 seconds
The proposed approach in the Cloudera distributed system with 9 nodes	3,881,562 seconds

Table 4: Comparisons of our model's results with the works in [1-3]

Clustering technique: CT.
Parallel network system: PNS (distributed system).
Special Domain: SD.
Depending on the training data set: DT.
Vector Space Model: VSM
No Mention: NM
English Language: EL.

Studies	ORC	CT	Sentiment Classification	PNS	SD	DT	Language	VSM
[1]	No	No	No	No	Yes	No	EL	Yes
[2]	No	No	Yes	No	Yes	No	EL	Yes
[3]	No	No	Yes	No	Yes	Yes	EL	Yes
Our work	Yes	Yes	Yes	Yes	No	No	EL	Yes

Table 5: Comparisons of our model's advantages and disadvantages with the works in [1-3]

Researches	Approach	Advantages	Disadvantages
[1]	Examining the vector space model, an information retrieval technique and its variation	In this work, 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	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.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

		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.	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.
[2]	+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 work, 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
[3]	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 Self-Organizing Map Algorithm using Only A Testing Data Set with The One-Dimensional Vectors and An Odds Ratio Coefficient 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 6: Comparisons of our model with the latest sentiment classification models (or the latest sentiment classification methods) in [4-14]

Studies	ORC	CT	Sentiment Classification	PNS	SD	DT	Language	VSM
[4]	No	No	Yes	NM	Yes	Yes	Yes	vector
[5]	No	No	Yes	NM	Yes	Yes	NM	NM
[6]	No	No	Yes	NM	Yes	Yes	EL	NM
[7]	No	No	Yes	NM	Yes	Yes	NM	NM
[8]	No	No	Yes	No	No	No	EL	No
[9]	No	No	Yes	No	No	No	EL	No
Our work	Yes	Yes	Yes	Yes	No	No	Yes	Yes

Table 7: Comparisons of our model's positives and negatives the latest sentiment classification models (or the latest sentiment classification methods) in [4-14]

Studies	Approach	Positives	Negatives
[4]	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 work 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
[5]	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
[6]	Exploiting New Sentiment-Based Meta-Level	Experiments performed with a substantial number of datasets (nineteen) demonstrate that	A line of future research would

	Features for Effective Sentiment Analysis	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.	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 according to their related reviews.
[7]	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
[8]	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	No mention

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

		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.	
[9]	Naive Bayes Model with N-GRAM Method, Negation Handling Method, Chi-Square Method and Good-Turing Discounting, etc.	The authors have explored the Naive 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 Self-Organizing Map Algorithm using Only A Testing Data Set with The One-Dimensional Vectors and An Odds Ratio Coefficient 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.		

Table 8: Comparisons of our model with the rsearches related to the Odds Ratio Coefficient (ORC) in [15- 19]

Studies	ORC	CT	Sentiment Classification	PNS	SD	DT	Language	VSM
[15]	Yes	Yes	Yes	NM	Yes	Yes	Yes	vector
[16]	Yes	No	Yes	NM	Yes	Yes	NM	NM
[17]	Yes	No	Yes	NM	Yes	Yes	EL	NM
[18]	Yes	No	Yes	NM	Yes	Yes	NM	NM
[19]	Yes	No	Yes	No	No	No	EL	No
Our work	Yes	Yes	Yes	Yes	No	No	Yes	Yes

Table 9: Comparisons of our model's positives and negatives the surveys related to the Odds Ratio Coefficient (ORC) in [15-19]

Studies	Approach	Positives	Negatives
[15]	A Survey of Binary Similarity and Distance Measures	Applying appropriate measures results in more accurate data analysis. Notwithstanding, 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
[16]	Generalized estimating equations for correlated binary data: Using the odds ratio as a meas	The authors discuss modelling the association between binary responses at pairs of times with the odds ratio. The authors then modify the estimating equations of Prentice to estimate the odds ratios. In simulations, the parameter estimates for the logistic regression model for the marginal probabilities appear slightly more efficient when using	No mention

	ure of association	the odds ratio parameterization.	
[17]	The diagnostic odds ratio: a single indicator of test performance	The authors propose the use of the odds ratio as a single indicator of diagnostic performance. The diagnostic odds ratio is closely linked to existing indicators, it facilitates formal meta-analysis of studies on diagnostic test performance, and it is derived from logistic models, which allow for the inclusion of additional variables to correct for heterogeneity. A disadvantage is the impossibility of weighing the true positive and false positive rate separately. In this study the application of the diagnostic odds ratio in test evaluation is illustrated.	No mention
[18]	Bias in odds ratios by logistic regression modelling and sample size	If several small studies are pooled without consideration of the bias introduced by the inherent mathematical properties of the logistic regression model, researchers may be mislead to erroneous interpretation of the results.	No mention
[19]	Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic, Prognostic, or Screening Marker	The authors illustrate that a single measure of association such as an odds ratio does not meaningfully describe a marker's ability to classify subjects. Appropriate statistical methods for assessing and reporting the classification power of a marker are described. In addition, the serious pitfalls of using more traditional methods based on parameters in logistic regression models are illustrated.	No mention
Our work	-We use Self-Organizing Map Algorithm using Only A Testing Data Set with The One-Dimensional Vectors and An Odds Ratio Coefficient 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.		

Table 10: Comparisons of our model with the rsearches related to Self-Organizing Map Algorithm (SOM) in [20-24]

Studies	ORC	CT	Sentiment Classification	PNS	SD	DT	Language	VSM
[20]	Yes	Yes	No	NM	No	No	Yes	vector
[21]	Yes	Yes	No	NM	No	No	NM	NM
[22]	Yes	Yes	No	NM	No	No	EL	NM
[23]	Yes	Yes	No	NM	No	No	NM	NM
[24]	Yes	Yes	No	No	No	No	EL	No
Our work	Yes	Yes	Yes	Yes	No	No	Yes	Yes

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 11: Comparisons of our model's positives and negatives the surveys related to the Self-Organizing Map Algorithm (SOM) in [20-24]

Studies	Approach	Positives	Negatives
[20]	The self-organizing map	One result of this is that the self-organization process can discover semantic relationships in sentences. Brain maps, semantic maps, and early work on competitive learning are reviewed. The self-organizing map algorithm (an algorithm which order responses spatially) is reviewed, focusing on best matching cell selection and adaptation of the weight vectors. Suggestions for applying the self-organizing map algorithm, demonstrations of the ordering process, and an example of hierarchical clustering of data are presented. Fine tuning the map by learning vector quantization is addressed. The use of self-organized maps in practical speech recognition and a simulation experiment on semantic mapping are discussed.	No mention
[21]	A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination	After the presentation of SOM adapted to ecological data, SOM was trained on popular example data; upland forest in Wisconsin (USA). The SOM results were compared with classical statistical techniques. Similarity between the results may be observed and constitutes a validation of the SOM method. SOM algorithm seems fully usable in ecology, it can perfectly complete classical techniques for exploring data and for achieving community ordination.	No mention
[22]	Clustering of the self-organizing map	In this study, different approaches to clustering of the SOM are considered. In particular, the use of hierarchical agglomerative clustering and partitive clustering using K-means are investigated. The two-stage procedure-first using SOM to produce the prototypes that are then clustered in the second stage-is found to perform well when compared with direct clustering of the data and to reduce the computation time.	No mention
[23]	A Scalable Self-organizing Map Algorithm for Textual Classification: A Neural Network Approach to	The authors' proposed data structure and algorithm took advantage of the sparsity of coordinates in the document input vectors and reduced the SOM computational complexity by several order of magnitude. The proposed Scaleable SOM (SSOM)	No mention

Journal of Theoretical and Applied Information Technology
31st August 2018. Vol.96. No 16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

	Thesaurus Generation	algorithm makes large-scale textual categorization tasks a possibility. Algorithmic intuition and the mathematical foundation of the authors' research are presented in detail. The authors also describe three benchmarking experiments to examine the algorithm's performance at various scales: classification of electronic meeting comments, Internet homepages, and the Compendex collection.	
[24]	The parameterless self-organizing map algorithm	The authors discuss the relative performance of the PLSOM and the SOM and demonstrate some tasks in which the SOM fails but the PLSOM performs satisfactory. Finally the authors discuss some example applications of the PLSOM and present a proof of ordering under certain limited conditions.	No mention
Our work	-We use Self-Organizing Map Algorithm using Only A Testing Data Set with The One-Dimensional Vectors and An Odds Ratio Coefficient 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.		

Table 12: Comparisons of our model's results with the works related to [24-56].

Odds Ratio Coefficient (ORC)

Semantic classification, sentiment classification: SC

Studies	PMI	JM	Language	SD	DT	ORC	SC	Other measures	Search engines
[24]	Yes	No	English	Yes	Yes	No	Yes	No	No Mention
[25]	Yes	No	English	Yes	No	No	Yes	Latent Semantic Analysis (LSA)	Alta Vista
[26]	Yes	No	English	Yes	Yes	No	Yes	Baseline; Turney-inspired; NB; Cluster + NB; Human	Alta Vista
[27	Y	N	Engli	Y	Y	N	Y	Si	G

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

]	es	o	sh German	es	es	o	es	mRank	oogle search engine
[28]	Yes	No	English Macedonian	Yes	Yes	No	Yes	No Mention	AltaVista search engine
[29]	Yes	No	English Arabic	Yes	No	No	Yes	No Mention	Google search engine Bing search engine
[30]	Yes	No	English Chinese	Yes	Yes	No	Yes	SVM(CN); SVM(EN); SVM(ENCN1); SVM(ENCN2); TSVM(CN); TSVM(EN); TSVM(ENCN1); TSVM(ENCN2); CoTrain	No Mention
[31]	Yes	No	English Spanish	Yes	Yes	No	Yes	SOCalculation SVM	Google
[32]	Yes	No	Chinese Tibet	Yes	Yes	No	Yes	-Feat	No M

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

			an					ure selection - Expectation Cross Entropy - Information Gain	ention
[33]	Yes	No	Chinese	Yes	Yes	No	Yes	DF, CHI, MI andIG	No Mention
[34]	Yes	No	Chinese	Yes	No	No	Yes	Information Bottleneck Method (IB); LE	AltaVista
[35]	Yes	No	Chinese	Yes	Yes	No	Yes	SVM	Google Yahoo Baidu
[36]	Yes	No	Japanese	No	No	No	Yes	Harmonic – Mean	Google and replaced the NEAR operator with the AND operator inthe SO formula.
[36]	Yes	Yes	English	Yes	Yes	No	Yes	Dice; NGD	Google sear

									chengine
[37]	Yes	Yes	English	Yes	No	No	Yes	Dice; Overlap	Google
[38]	No	Yes	English	Yes	Yes	No	Yes	A Jaccard index based clustering algorithm (JIBCA)	No Mention
[39]	No	Yes	English	Yes	Yes	No	Yes	Naive Bayes, Two-Step Multinomial Naive Bayes, and Two-Step Polynomial-Kernel Support Vector Machine	Google
[40]	No	Yes	Arabic	No	No	No	Yes	Naive Bayes (NB); Support Vector Machines (SVM); RoORCh	No Mention

Journal of Theoretical and Applied Information Technology
31st August 2018. Vol.96. No 16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

								io; Cosine	
[41]	No	Yes	Chinese	Yes	Yes	No	Yes	A new score–Economic Value (EV), etc.	Chinese search
[42]	No	Yes	Chinese	Yes	Yes	No	Yes	Cosine	No Mention
[43]	No	Yes	English	No	Yes	No	Yes	Cosine	No Mention
[44]	No	Yes	Chinese	No	Yes	No	Yes	Dice; overlap; Cosine	No Mention
[45]	No	No	Vietnamese	No	No	No	Yes	Ochiai Measure	Google
[46]	No	No	English	No	No	No	Yes	Cosine coefficient	Google
[47]	No	No	English	No	No	No	Yes	Sorensen measure	Google
[48]	No	Yes	Vietnamese	No	No	No	Yes	Jaccard	Google
[49]	No	No	English	No	No	No	Yes	Tanimoto coefficient	Google
Our work	No	No	English Language	No	No	Yes	Yes	No	Google search engine

Table 13: Comparisons of our model's advantages and disadvantages with the works related to [24-56].

Surveys	**Approach**	**Advantages**	**Disadvantages**
[24]	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 need to investigate this more closely to find

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

		the authors' preliminary research showed that 100 gave a better result.	the optimal distance. Another factor that has not been investigated much in the literature is the selection of seed words. Since they are the basis for PMI calculation, it might be a lot to gain by finding better seed words. The authors would like to explore the impact that differ

			ent approaches to seed word selection have on the performance of the developed sentiment lexicons.
[25]	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 and SO-LSA. The accuracy of SO-PMI-IR is comparable to the accuracy of HM, the algorithm of Hatzivassiloglou and McKeown (1997). SO-PMI-IR requires a large corpus, but it is simple, easy to implement, unsupervised, and it is not restricted to adjectives.	No Mention
[26]	Graph-based user classificati	The authors describe several experiments in	There is still much

	on for informal online political discourse	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	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
[27]	A novel, 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.
[28]	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

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

			collect half a million tweets for creating lexicons and analyzing/evaluating the system, while Kiritchenko et al. (2014) built their lexicon on million tweets and evaluated their system on 135 million English tweets. Moreover, the authors are interested not only in quanti

			ty but also in quality, i.e., in studying the quality of the individual words and phrases used as seeds.
[29]	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 phras

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

			e, they make it more negative or more positive.
[30]	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 used in (Mihalcea, 2004) will be adopted for sentiment classification. 2) The
			authors will employ the structural correspondence learning (SCL) domain adaptation algorithm used in (Blitzer et al., 2007) for linking the translated text and the natural text.
[31]	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	No Mention

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

		significant semantic loss, at least for original Spanish texts.	
[32]	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
[33]	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	No Mention

		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.	
[34]	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.
[35]	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
[36]	Modifying SO-PMI for Japanese Weblog Opinion Mining by Using a Balancing Factor and	After these modifications, the authors achieved a well-balanced result: both positive and negative accuracy exceeded 70%. This shows that the authors'	In the future, the authors will evaluate different

	Detecting Neutral Expressions	proposed approach not only adapted the SO-PMI for Japanese, but also modified it to analyze Japanese opinions more effectively.	choices of words for the sets of positive and negative reference words. The authors also plan to appraise their proposal on other languages.
[36]	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
[37]	Adjective-Based Estimation of Short	The adjectives are ranked and top na adjectives are considered as	In the authors'

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

	Sentence's Impression	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)	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.
[38]	Jaccard Index based Clustering Algorithm for Mining Online Review	In this work, 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 frame work, 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.
[39]	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
[40	Ensemble	The experimental	No

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

]	of Classification algorithms for Subjectivity and Sentiment Analysis of Arabic Customers' Reviews	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.	Mention
[41]	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
[42]	Sentiment Classification in Under-Resourced Languages Using Graph-based	In particular, the authors found that choosing initially labeled vertices in aORCordance with their degree and PageRank score can improve the performance.	As future work, first, the authors will attem

	Semi-supervised Learning Methods	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.	pt 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
[43]	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 paper provides a first step in exploring the extremely large, rich, and useful body of consumer data readily available on Web 2.0.	No Mention
[44]	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 resourcescarce 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.
[45]	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 in different 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	not calculating all Vietnamese words completely; not identifying all Vietnamese adjective phrases fully, etc.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

		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.	
[46]	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.
[47]	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	This survey is only applied to the English adverb phrases.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

		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.	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
[48]	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 work, 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.
[49]	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	This survey is only applied in the English noun phrases. The proposed model is needed to research more and more about the different

		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)	types of the English words such as English English adverbs, etc.
Our work	-We use Self-Organizing Map Algorithm using Only A Testing Data Set with The One-Dimensional Vectors and An Odds Ratio Coefficient 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.		