Journal of Theoretical and Applied Information Technology
15th June 2018. Vol.96. No 11

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

ENGLISH SENTIMENT CLASSIFICATION USING AN YULEQ SIMILARITY MEASURE AND THE ONE-DIMENSIONAL VECTORS IN A PARALLEL NETWORK ENVIRONMENT

[1]DR.VO NGOC PHU, [2]DR.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

Sentiment classification is significant in everyday life, such as in political activities, commodity production, and commercial activities. In this study, we have proposed a new model for Big Data sentiment classification. We use a YULEQ coefficient (YC) of the clustering technologies of a data mining field to cluster one document of our English testing data set, which is 7,000,000 documents comprising the 3,500,000 positive and the 3,500,000 negative, into either the positive polarity or the negative polarity based on our English training data set which is 5,000,000 documents including the 2,500,000 positive and the 2,500,000 negative. We do not use any sentiment lexicons in English. We do not use any multi-dimensional vector based on both a vector space modeling (VSM) and the sentiment lexicons. We only use many one-dimensional vectors based on VSM. One one-dimensional vector is clustered into either the positive or the negative if this vector is very close to either the positive or the negative by using many similarity coefficients of the YC. It means that this vector is very similar to either the positive or the negative. One document of the testing data set is clustered into the sentiments (positive, negative, or neutral) based on many one-dimensional vectors. We tested the proposed model in both a sequential environment and a distributed network system. We achieved 87.85% accuracy of the testing data set. The execution time of the model in the parallel network environment is faster than the execution time of the model in the sequential system. This survey used many similarity coefficients of the data mining field. The results of this work can be widely used in applications and research of the English sentiment classification.

Keywords: *English Sentiment Classification; Distributed System; Parallel System; YULEQ Similarity Measure; Cloudera; Hadoop Map And Hadoop Reduce; Clustering Technology.*

1. INTRODUCTION

Opinion mining relates to emotional researches that are expressed in documents. Sentiment analysis has a wide range of applications in the fields of business, organizations, governments and individuals.

Clustering data is to process a set of objects into classes of similar objects. One cluster is a set of data objects which are similar to each other and are not similar to objects in other clusters. A number of data clusters can be clustered, which can be identified following experience or can be automatically identified as part of clustering method.

To implement our new model, we propose the following basic principles:

- Assuming that each English sentence has m English words (or English phrases).
- Assuming that the maximum number of one English sentence is m_max; it means that m is less than m_max or m is equal to m_max.
- Assuming that each English document has n English sentences.

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

- Assuming that the maximum number of one English document is n_max; it means that n is less than n_max or n is equal to n_max.

- Each English sentence is transferred into one vector (one-dimensional). 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).

- Each English document is transferred into one multi-dimensional vector. Therefore, the multi-dimensional vector has n rows and m columns. If n is less than n_max then each element of the multi-dimensional vector from n to n_max-1 is 0 (zero vector).

- All the sentences of one document of both the testing data set and the training data set are transferred into the one-dimensional vectors based on the vector space modeling (VSM).

- All the positive sentences of the documents of the training data set are transferred the positive one-dimensional vectors.

- All the negative sentences of the documents of the training data set are transferred the negative one-dimensional vectors

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

The motivation of this new model is as follows: Many algorithms in the data mining field can be applied to natural language processing, specifically semantic classification for processing millions of English documents. A YULEQ similarity measure (YC) 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 that the YULEQ similarity measure (YC) is applied to sentiment analysis. This algorithm can also be applied to identify the emotions of millions of 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.

To get higher accuracy of the results of the sentiment classification and shorten execution time of the sentiment classification, we do not use any sentiment lexicons in English. We do not use any multi-dimensional vector based on both a vector space modeling (VSM) in [1-3] and the sentiment lexicons. We only use many one-dimensional vectors based on VSM[1-3]. All the sentences of each document of the documents of both the testing data set and training data set are transferred into the one-dimensional vectors based on VSM[1-3]. All the sentences of the positive documents of the training data set are transferred into the positive one-dimensional vectors based on VSM[1-3], called the positive vector group. All the sentences of the negative documents of the training data set are transferred into the negative one-dimensional vectors based on VSM[1-3], called the negative vector group. We calculate total_positive_measure which is a total of many similarity measures of the YC between one one-dimensional vector A (corresponding to one sentence) of the one-dimensional vectors (corresponding to all the sentences) of one document B of the testing data set and all the one-dimenional vectors of the positive vector group. We calculate total_negative_measure which is a total of many similarity measures of the YC between one one-dimensional vector A (corresponding to one sentence) of the one-dimensional vectors (corresponding to all the sentences) of one document B of the testing data set and all the one-dimenional vectors of the negative vector group. This vector A is clustered into the positive if total_positive_measure is greater than total_negative_measure. This vector A is clustered into the negative if total_positive_measure is less than total_negative_measure. This vector A is clustered into the neutral if total_positive_measure is as equal as total_negative_measure. This document B is clustered into the positive if the number of the one-dimensional vectors clustered into the positive is greater than the number of the one-dimensional vectors clustered into the negative in the document B. This document B 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 B. This document B 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 B.

We perform the proposed model as follows: First of all, we transfer all the sentences of one document into the one-dimensional vectors based on VSM[1,2,3]. We transfer all the sentences of all the documents of bith the testing data set and the

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

training data set into the one-dimensional vectors based on VSM[1-3]. All the sentences of the positive documents of the training data set are transferred into the all the positive one-dimensional vectors based on VSM[1-3], called the positive vector group of the training data set. All the sentences of the negative documents of the training data set are transferred into the all the negative one-dimensional vectors based on VSM[1-3], called the negative vector group of the training data set. We calculate total_positive_measure which is a total of many similarity measures of the YC between one one-dimensional vector A (corresponding to one sentence) of the one-dimensional vectors (corresponding to all the sentences) of one document B of the testing data set and all the one-dimenional vectors of the positive vector group. We calculate total_negative_measure which is a total of many similarity measures of the YC between one one-dimensional vector A (corresponding to one sentence) of the one-dimensional vectors (corresponding to all the sentences) of one document B of the testing data set and all the one-dimenional vectors of the negative vector group. This vector A is clustered into the positive if total_positive_measure is greater than total_negative_measure. This vector A is clustered into the negative if total_positive_measure is less than total_negative_measure. This vector A is clustered into the neutral if total_positive_measure is as equal as total_negative_measure. This document B is clustered into the positive if the number of the one-dimensional vectors clustered into the positive is greater than the number of the one-dimensional vectors clustered into the negative in the document B. This document B 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 B. This document B 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 B. Finally, the sentiment classification of all the documents of the testing data set is identified certainly.

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

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

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 YC – related algorithms are proposed in this survey.

This study contains 6 sections. Section 1 introduces the study; Section 2 discusses the related works about the vector space modeling (VSM), YULEQ similarity measure (YC), 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.

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

2. RELATED WORK

We summarize many researches which are related to our research

There are the works related to vector space modeling 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 consider multi-label text classification task and apply various feature sets. The authors consider a subset of multi-labeled files from the Reuters-21578 corpus. The authors use traditional tf-IDF values of the features and tried both considering and ignoring stop words. The authors also tried several combinations of features, like bigrams and unigrams. The authors in [3] introduce a new weighting method based on statistical estimation of the importance of a word for a specific categorization problem. This method also has the benefit to make feature selection implicit, since useless features for the categorization problem considered to get a very small weight.

The latest researches of the sentiment classification are [4-14]. In the research [4], the authors present their machine learning experiments with regard to sentiment analysis in blog, review and forum texts found on the World Wide Web and written in English, Dutch and French. The survey in [5] discusses an approach where an exposed stream of tweets from the Twitter micro blogging site are preprocessed and classified based on their sentiments. In sentiment classification system the concept of opinion subjectivity has been accounted. In the study, the authors present opinion detection and organization subsystem, which have already been integrated into our larger question-answering system, etc.

The surveys related to the YULEQ 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, etc.

3. DATA SET

In Figure 1 below, the English testing data set includes 7,000,000 English documents in the movie field, which contains 3,500,000 positive English documents and 3,500,000 negative English documents. All English sentences 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.

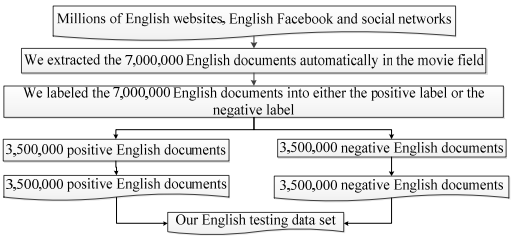

Figure 1: Our English testing data set.

In Figure 2 below, the English training data set includes 5,000,000 English documents in the movie field, which contains 2,500,000 positive English documents and 2,500,000 negative English documents. All English sentences 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.

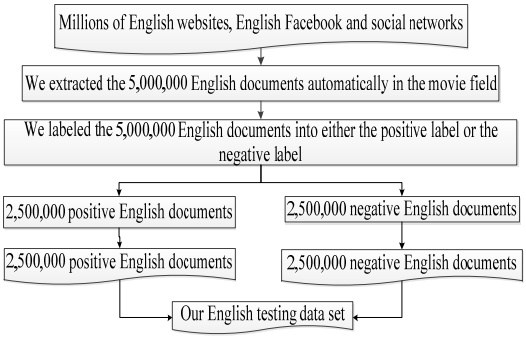

Figure 2: Our English training data set.

4. METHODOLOGY

This section comprises two parts: The first part is to use the YULEQ similarity coefficient of the clustering technologies of the data mining field in a sequential environment in the sub-section (4.1). The second part is to use the YULEQ similarity coefficient of the clustering technologies of the data mining field in a distributed network environment in the sub-section (4.2).

4.1 Using the YULEQ similarity coefficient of the clustering technologies of the data mining field in a sequential environment

In Figure 3, we use the YULEQ similarity coefficient of the clustering technologies of the data mining field in a sequential environment as follows:

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

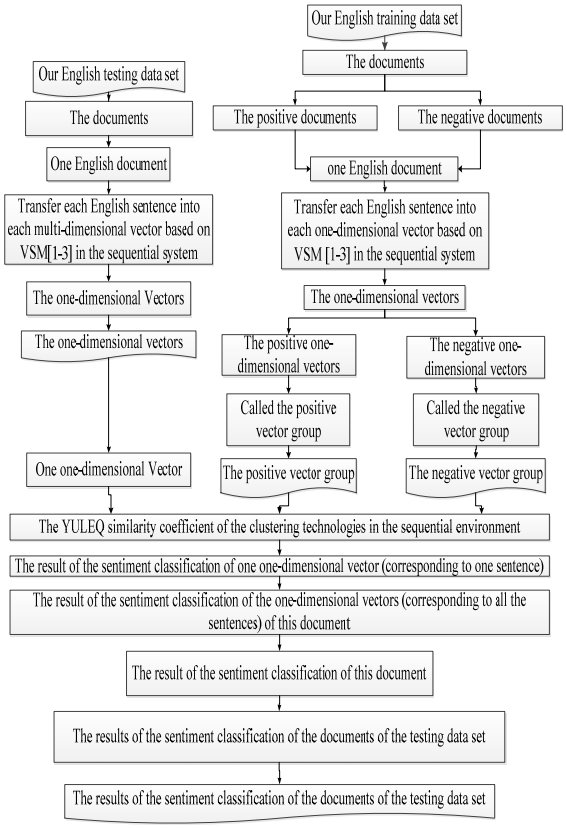

Figure 3: Overview of using the YULEQ similarity coefficient of the clustering technologies of the data mining field in a sequential environment

In Figure 3, this section is implemented in the sequential system as follows: we transfer all the sentences of one document into the one-dimensional vectors based on VSM[1,2,3]. Then, we transfer all the sentences of all the documents of both the testing data set and the training data set into the one-dimensional vectors based on VSM[1-3]. All the sentences of the positive documents of the training data set are transferred into the all the positive one-dimensional vectors based on VSM[1-3], called the positive vector group of the training data set. All the sentences of the negative documents of the training data set are transferred into the all the negative one-dimensional vectors based on VSM[1-3], called the negative vector group of the training data set. We calculate total_positive_measure which is a total of many similarity measures of the YC between one one-dimensional vector A (corresponding to one sentence) of the one-dimensional vectors (corresponding to all the sentences) of one document B of the testing data set and all the one-dimenional vectors of the positive vector group. We calculate total_negative_measure which is a total of many similarity measures of the YC between one one-dimensional vector A (corresponding to one sentence) of the one-dimensional vectors (corresponding to all the sentences) of one document B of the testing data set and all the one-dimenional vectors of the negative vector group. This vector A is clustered into the positive if total_positive_measure is greater than total_negative_measure. This vector A is clustered into the negative if total_positive_measure is less than total_negative_measure. This vector A is clustered into the neutral if total_positive_measure is as equal as total_negative_measure. This document B is clustered into the positive if the number of the one-dimensional vectors clustered into the positive is greater than the number of the one-dimensional vectors clustered into the negative in the document B. This document B 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 B. This document B 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 B. Finally, we cluster all the documents of the testing data set into either the positive or the negative certainly.

We build the algorithm 1 to transfer one English document into the one-dimensional vectors based on VSM[1-3] in the sequential environment. Each document is split into many sentences. Each sentence in each document is transferred to one one-dimensional vector based on VSM [1, 2, 3] in the sequential environment. We insert each one-dimensional vector into all the one-dimensional vectors of this document.

The main ideas of the algorithm 1 are as follows:

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: Each sentence in the n sentences of this document, do repeat:

Step 3: Transfer this sentence into one one-dimensional vector based on VSM [1, 2, 3];

Step 4: Add the transferred vector into the one-dimensional vectors;

Step 5: End Repeat – End Step 2

Step 6: Return the one-dimensional vectors;

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

We propose the algorithm 2 to create the positive vector group of the training data set in the sequential system. Each positive document of the positive documents of the English training data set is split into many sentences. Each sentence of the document is transferred to one one-dimensional vector based on VSM [1, 2, 3] in the sequential environment. We insert each one-dimensional vector into all the positive one-dimensional vectors, called the positive vector group of the training data set.
The main ideas of the algorithm 2 are as follows:
Input: the positive English documents of the English training data set.
Output: the positive vector group PositiveVectorGroup
Step 1:Each document in the positive documents of the training data set, do repeat:
Step 2: TheOne-dimensionalVectors := Call Algorithm 1 with the positive English document in the English training data set;
Step 3: Add TheOne-dimensionalVectors into PositiveVectorGroup;
Step 4: End Repeat – End Step 1
Step 5: Return PositiveVectorGroup;

We propose the algorithm 3 to create the negative vector group of the training data set in the sequential system. Each negative document of the negative documents of the English training data set is split into many sentences. Each sentence of the document is transferred to one one-dimensional vector based on VSM [1, 2, 3] in the sequential environment. We insert each one-dimensional vector into all the positive one-dimensional vectors, called the negative vector group of the training data set.
The main ideas of the algorithm 3 are as follows:
Input: the negative English documents of the English training data set.
Output: the negative vector group NegativeVectorGroup
Step 1:Each document in the negative documents of the training data set, do repeat:
Step 2: TheOne-dimensionalVectors := Call Algorithm 1 with the negative English document in the English training data set;
Step 3: Add TheOne-dimensionalVectors into NegativeVectorGroup;
Step 4: End Repeat – End Step 1
Step 5: Return NegativeVectorGroup;

According to the surveys related the YULEQ coefficient in [15- 19], we have an equation as follows:

$$\text{YULEQ Coefficient}(a,b) = \text{YULEQ Measure}(a,b) = \text{YC}(a,b) = \frac{(a \cap b) * (\neg a \cap \neg b) - (\neg a \cap b) * (a \cap \neg b)}{(a \cap b) * (\neg a \cap \neg b) + (\neg a \cap b) * (a \cap \neg b)} \quad (1)$$

with a and b is two vectors.
We build the algorithm 4 to cluster one one-dimensional vector (corresponding to one sentence of one document of the English testing data set) into either the positive vector group or the negative vector group by using the YULEQ similarity coefficient of the clustering technologies in the sequential system.
The main ideas of the algorithm 4 are as follows:
Input: one one-dimensional vector A (corresponding to one sentence of one document of the English testing data set), the positive vector group and the negative vector group;
Output: the sentiment polarity (positive, negative, neutral);
Step 1: Set total_positive := 0 and total_negative := 0;
Step 2: Each vector B in the positive vector group, do repeat:
Step 3: total_positive := total_positive + Similarity measure from Eq. (1) between the vector A and the vector B;
Step 4: End Repeat – End Step 2;
Step 5: Each vector C in the negative vector group, do repeat:
Step 6: total_negative := total_negative + Similarity measure from Eq. (1) between the vector A and the vector C;
Step 7: End Repeat – End Step 2;
Step 8: If total_positive is greater than total_negative The Return positive;
Step 9: Else If total_positive is less than total_negative The Return negative;
Step 10: Return neutral;

We propose the algorithm 5 to cluster one document of the testing data set into either positive or the negative by using the YULEQ similarity coefficient of the clustering technologies in the sequential system.
The main ideas of the algorithm 5 are as follows:
Input: one document of the testing data set, the positive vector group and the negative vector group.
Output: the result of the sentiment classification of this document
Step 1: Set count_positive := 0 and count_negative := 0;
Step 2: Split this document into the sentences;

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

Step 3: Each sentence in the sentences of the testing data set, do repeat:
Step 4: OneResult := the algorithm 4 to cluster one one-dimensional vector (corresponding to one sentence of one document of the English testing data set) into either the positive vector group or the negative vector group by using the YULEQ similarity coefficient of the clustering technologies in the sequential system with the input is this document;
Step 5: End Repeat – End Step 2;
Step 6: If OneResult is the positive Then count_positive := count_positive +1;
Step 7: Else If OneResult is the negative Then count_negative := count_negative +1;
Step 8: End Repeat – End Step 3;
Step 9: If count_positive is greater than count_negative Then Return positive;
Step 10: Else If count_positive is less than count_negative Then Return negative;
Step 11: Return the neutral;

We build the algorithm 6 to cluster the documents of the testing data set into either positive or the negative by using the YULEQ similarity coefficient of the clustering technologies in the sequential system.
The main ideas of the algorithm 6 are as follows:
Input: the testing data set and the training data set
Output: the results of the sentiment classification of the documents of the testing data set;
Step 1: the positive vector group := The algorithm 2 to create the positive vector group of the training data set in the sequential system with the input is the positive document of the training data set;
Step 2: the negative vector group := The algorithm 3 to create the negative vector group of the training data set in the sequential system with the input is the negative document of the training data set;
Step 3: Each document in the documents of the testing data set, do repeat:
Step 4: OneResult := The algorithm 5 to cluster one document of the testing data set into either positive or the negative by using the YULEQ similarity coefficient of the clustering technologies in the sequential system with the input is this document, the positive vector group and the negative vector group;
Step 5: Add OneResult into the results of the sentiment classification of the documents of the testing data set;
Step 6: End Repeat – End Step 3;
Step 7: Return the results of the sentiment classification of the documents of the testing data set;

4.2 Using the YULEQ similarity coefficient of the clustering technologies of the data mining field in a distributed network environment

In Figure 4, we use the YULEQ similarity coefficient of the clustering technologies of the data mining field in a sequential environment as follows:

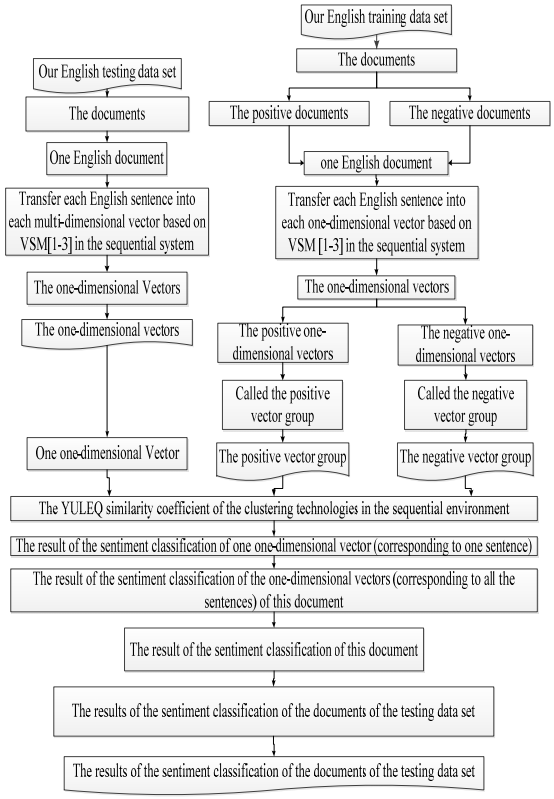

Figure 4: Overview of using the YULEQ similarity coefficient of the clustering technologies of the data mining field in a parallel network environment

In Figure 4, this section is implemented in the distributed network system as follows: we transfer all the sentences of one document into the one-dimensional vectors based on VSM[1,2,3]. Then, we transfer all the sentences of all the documents of both the testing data set and the training data set into the one-dimensional vectors based on VSM[1-3]. All the sentences of the positive documents of the training data set are transferred into the all the positive one-dimensional vectors based on VSM[1-3], called the positive vector group of the training data set. All the sentences of the negative documents of the training data set are transferred into the all the negative one-dimensional vectors based on VSM[1-3], called the negative vector group of the training data set. We calculate

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

total_positive_measure which is a total of many similarity measures of the YC between one one-dimensional vector A (corresponding to one sentence) of the one-dimensional vectors (corresponding to all the sentences) of one document B of the testing data set and all the one-dimenional vectors of the positive vector group. We calculate total_negative_measure which is a total of many similarity measures of the YC between one one-dimensional vector A (corresponding to one sentence) of the one-dimensional vectors (corresponding to all the sentences) of one document B of the testing data set and all the one-dimenional vectors of the negative vector group. This vector A is clustered into the positive if total_positive_measure is greater than total_negative_measure. This vector A is clustered into the negative if total_positive_measure is less than total_negative_measure. This vector A is clustered into the neutral if total_positive_measure is as equal as total_negative_measure. This document B is clustered into the positive if the number of the one-dimensional vectors clustered into the positive is greater than the number of the one-dimensional vectors clustered into the negative in the document B. This document B 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 B. This document B 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 B. Finally, we cluster all the documents of the testing data set into either the positive or the negative certainly.

In Figure 5, we transfer all the sentences of one document into the one-dimensional vectors of the document based on VSM[1-3] in the parallel network environment. 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 se. The output of the Hadoop Reduce is one one-dimensional vector (corresponding to one sentence) of this document. The input of the Hadoop Reduce is the output of the Hadoop Map, thus, the input of the Hadoop Reduce is one one-dimensional vector (corresponding to one sentence) of this document. The output of the Hadoop Reduce is the one-dimensional vectors (corresponding to all the sentences) of this document.

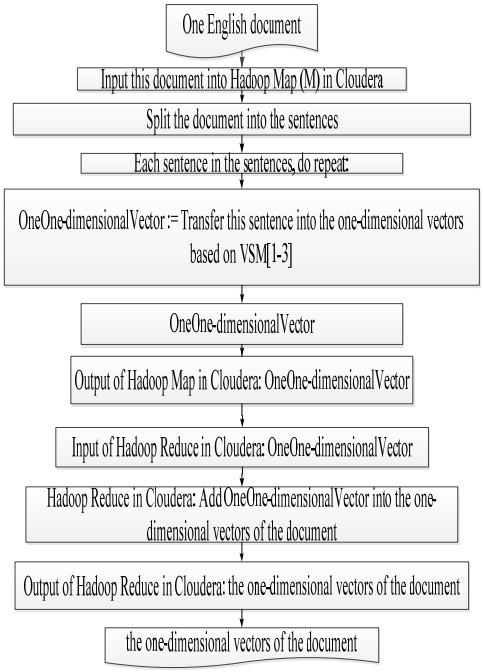

Figure 5: Overview of transferring all the sentences of one document into the one-dimensional vectors of the document based on VSM[1-3] in the parallel network environment

We build the algorithm 7 to perform the Hadoop Map phase of transferring all the sentences of one document into the one-dimensional vectors of the document based on VSM[1-3] in the Cloudera parallel network environment . The main ideas of the algorithm 7 are as follows:

Input: one document of the testing data set;

Output: one one-dimensional vector of this document

Step 1: Input this document into the Hadoop Map in the Cloudera system;

Step 2: Split this document into the sentences;

Step 3: Each sentence in the sentences, do repeat:

Step 4: one one-dimensional vector := The transforming one English sentence into one one-dimensional vector based on VSM [1-3] with the input is this sentence

Step 5: Return one one-dimensional vector; //the output of the Hadoop Map phase.

We propose the algorithm 8 to perform the Hadoop Reduce phase of transferring all the sentences of one document into the one-dimensional vectors of the document based on VSM [1-3] in the Cloudera parallel network environment. The main ideas of the algorithm 8 are as follows:

Input: one one-dimensional vector of this document

Output: the one-dimensional vectors of this document

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

Step 1: Receive one one-dimensional vector;
Step 2: Add this one-dimensional vector into the one-dimensional vectors of this document;
Step 3: Return the one-dimensional vectors of this document;

In Figure 6, we transfer the positive documents the training data set into the positive one-dimensional vectors (called the positive vector group of the training data set) in the distributed system.

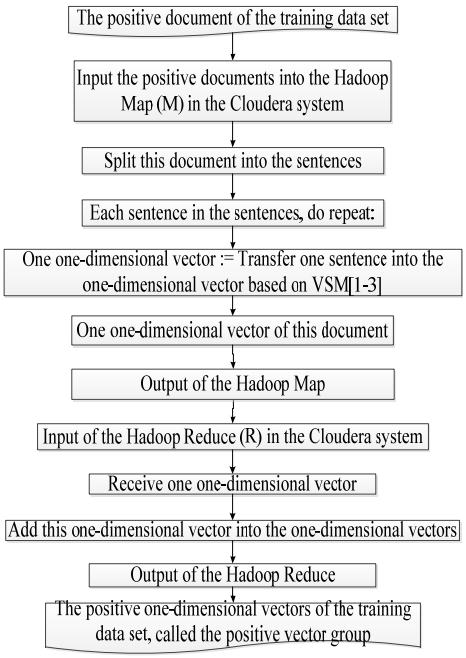

Figure 6: Overview of transferring the positive documents of the training data set into the positive one-dimensional vectors (called the positive vector group of the training data set) in the distributed system.

In Figure 6, the stage includes two phases: the Hadoop Map (M) phase and the Hadoop Reduce (R) phase. The input of the Hadoop Map phase is the positive documents of the training data set. The output of the Hadoop Map phase is one one-dimensional vector (corresponding to one sentence) of one positive document of the training data set. The input of the Hadoop Reduce phase is the output of the Hadoop Map phase, thus, the input of the Hadoop Reduce phase is one one-dimensional vector (corresponding to one sentence) of the positive document of the training data set. The output of the Hadoop Reduce phase is the positive one-dimensional vectors, called the positive vector group of the training data set

We propose the algorithm 9 to transfer the positive documents of the training data set into the positive one-dimensional vectors (called the positive vector group of the training data set) in the distributed system in Figure 6. The main ideas of the algorithm 9 are as follows:

Input: the positive documents of the training data set
Output: one one-dimensional vector of one positive document of the training data set
Step 1: Input the positive documents into the Hadoop Map in the Cloudera system.
Step 2: Each document in the positive document, do repeat:
Step 3: TheOne-dimentionalVectorsOfOneDocument := The transferring all the sentences of one document of the testing data set into the one-dimensional vectors of the document of testing data set based on VSM[1-3] in the parallel network environment in Figure 5 with the input is this document;
Step 4: Return TheOne-dimentionalVectorsOfOneDocument ;

We propose the algorithm 10 to implement the Hadoop Reduce phase of transferring the positive sentences of the training data set into the positive multi-dimensional vectors in the distributed system in Figure 6. The main ideas of the algorithm 10 are as follows:

Input:TheOne-dimentionalVectorsOfOneDocument - the one-dimensional vectors of the positive document of the training data set
Output: the positive one-dimensional vectors, called the positive vector group of the training data set
Step 1: Receive TheOne-dimentionalVectorsOfOneDocument ;
Step 2: Add TheOne-dimentionalVectorsOfOneDocument into PositiveVectorGroup;
Step 3: Return PositiveVectorGroup - the positive one-dimensional vectors, called the positive vector group of the training data set;

In Figure 7, we transfer the negative documents the training data set into the negative one-dimensional vectors (called the negative vector group of the training data set) in the distributed system.

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

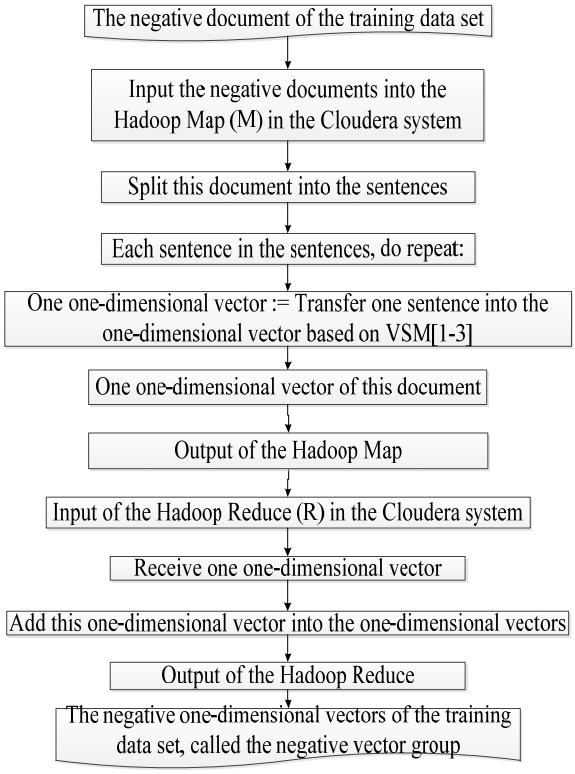

Figure 7: Overview of transferring the negative documents of the training data set into the negative one-dimensional vectors (called the negative vector group of the training data set) in the distributed system.

In Figure 7, the stage includes two phases: the Hadoop Map (M) phase and the Hadoop Reduce (R) phase. The input of the Hadoop Map phase is the negative documents of the training data set. The output of the Hadoop Map phase is one one-dimensional vector (corresponding to one sentence) of one negative document of the training data set. The input of the Hadoop Reduce phase is the output of the Hadoop Map phase, thus, the input of the Hadoop Reduce phase is one one-dimensional vector (corresponding to one sentence) of the negative document of the training data set. The output of the Hadoop Reduce phase is the negative one-dimensional vectors, called the negative vector group of the training data set

We propose the algorithm 11 to transferring the negative documents of the training data set into the negative one-dimensional vectors (called the negative vector group of the training data set) in the distributed system in Figure 6. The main ideas of the algorithm 11 are as follows:

Input: the negative documents of the training data set

Output: one one-dimensional vector of one negative document of the training data set

Step 1: Input the negative documents into the Hadoop Map in the Cloudera system.

Step 2: Each document in the negative document, do repeat:

Step 3: TheOne-dimentionalVectorsOfOneDocument := The transferring all the sentences of one document of the testing data set into the one-dimensional vectors of the document of testing data set based on VSM[1-3] in the parallel network environment in Figure 5 with the input is this document;

Step 4: Return TheOne-dimentionalVectorsOfOneDocument ;

We propose the algorithm 12 to implement the Hadoop Reduce phase of transferring the negative sentences of the training data set into the negative multi-dimensional vectors in the distributed system in Figure 6. The main ideas of the algorithm 12 are as follows:

Input: TheOne-dimentionalVectorsOfOneDocument - the one-dimensional vectors of the negative document of the training data set

Output: the negative one-dimensional vectors, called the negative vector group of the training data set

Step 1: Receive TheOne-dimentionalVectorsOfOneDocument ;

Step 2: Add TheOne-dimentionalVectorsOfOneDocument into NegativeVectorGroup;

Step 3: Return NegativeVectorGroup - the negative one-dimensional vectors, called the negative vector group of the training data set;

According to the surveys related the YULEQ coefficient in [15- 19], we have an equation as follows:

$$\text{YULEQ Coefficient}(a,b) = \text{YULEQ Measure}(a,b) = \text{YC}(a,b) = \frac{(a \cap b) * (\neg a \cap \neg b) - (\neg a \cap b) * (a \cap \neg b)}{(a \cap b) * (\neg a \cap \neg b) + (\neg a \cap b) * (a \cap \neg b)} \quad (1)$$

with a and b is two vectors.

In Figure 8, we cluster one one-dimensional vector (corresponding to one sentence of one document of the English testing data set) into either the positive vector group or the negative vector group by using the YULEQ similarity coefficient of the clustering technologies in the distributed network system.

Journal of Theoretical and Applied Information Technology
15th June 2018. Vol.96. No 11

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

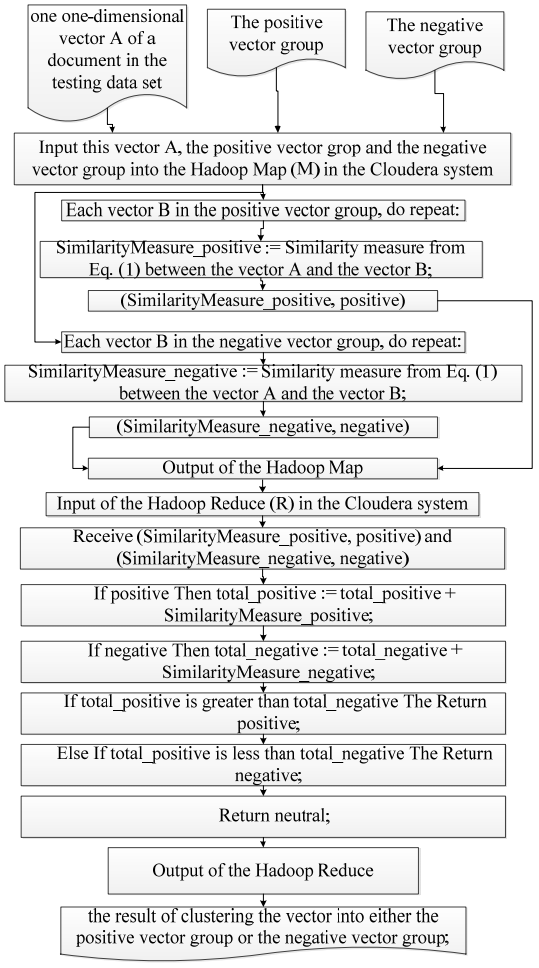

Figure 8: Overview of clustering one one-dimensional vector (corresponding to one sentence of one document of the English testing data set) into either the positive vector group or the negative vector group by using the YULEQ similarity coefficient of the clustering technologies in the parallel network system.

In Figure 8, this stage has two phases: the Hadoop Map phase and the Hadoop Reduce phase. The input of the Hadoop Map is one one-dimensional vector (corresponding one sentence of one document of the testing data set), the positive vector grop and the negative vector group the training data set. The output of the Hadoop Map is (SimilarityMeasure_positive, positive) and (SimilarityMeasure_negative, negative). The input of the Hadoop Reduce is the output of the Hadoop Map, thus the input of the Hadoop Reduce is (SimilarityMeasure_positive, positive) and (SimilarityMeasure_negative, negative). The output of the Hadoop Reduce is the the result of the sentiment classification of this vector.

We propose the algorithm 13 to perform the Hadoop Map phase of clustering one one-dimensional vector (corresponding to one sentence of one document of the English testing data set) into either the positive vector group or the negative vector group by using the YULEQ similarity coefficient of the clustering technologies in the parallel network system. The main ideas of the algorithm 13 are as follows:

Input: one one-dimensional vector A of a document in the testing data set; the positive vector group and the negative vector group of the training data set.

Output: (SimilarityMeasure_positive, positive) and (SimilarityMeasure_negative, negative)

Step 1: Input this vector A, the positive vector group and the negative vector group into the Hadoop Map in the Cloudera system;

Step 2: Each vector B in the positive vector group, do repeat:

Step 3: SimilarityMeasure_positive := Similarity measure from Eq. (1) between the vector A and the vector B;

Step 4: Return (SimilarityMeasure_positive, positive); //the output of the Hadoop Map

Step 5: Each vector C in the negative vector group, do repeat:

Step 6: SimilarityMeasure_negative := Similarity measure from Eq. (1) between the vector A and the vector C;

Step 7: Return (SimilarityMeasure_negative, negative);//the output of the Hadoop Map

We build the algorithm 14 to implement the Hadoop Reduce phase of clustering one one-dimensional vector (corresponding to one sentence of one document of the English testing data set) into either the positive vector group or the negative vector group by using the YULEQ similarity coefficient of the clustering technologies in the parallel network system. The main ideas of the algorithm 14 are as follows:

Input: (SimilarityMeasure_positive, positive) and (SimilarityMeasure_negative, negative) – the output of the Hadoop Map

Output: the result of clustering the vector into either the positive vector group or the negative vector group.

Step 1: Receive (SimilarityMeasure_positive, positive) and (SimilarityMeasure_negative, negative);

Step 2: If positive Then total_positive := total_positive + SimilarityMeasure_positive;

Step 6: If negative Then total_negative := total_negative + SimilarityMeasure_negative;

Step 7: If total_positive is greater than total_negative The Return positive;

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

Step 8: Else If total_positive is less than total_negative The Return negative;
Step 9: Return neutral;
Step 10: Return the result of clustering the vector into either the positive vector group or the negative vector group;

In Figure 9, we use the YULEQ coefficient (YC) of the clustering technologies of the data mining field to cluster one document of the testing data set into either the positive or the negative in the distributed environment. The input of the Hadoop Map is one document of the testing data set, the positive vector group and the negative vector group of the training data set. The output of the Hadoop Map is the result of the sentiment classification of one one-dimensional vector (corresponding to one sentence of this document) into either the positive vector group or the negative vector group. 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 one one-dimensional vector (corresponding to one sentence of this document) into either the positive vector group or the negative vector group. The output of the Hadoop Reduce is the result of the sentiment classification of this document.

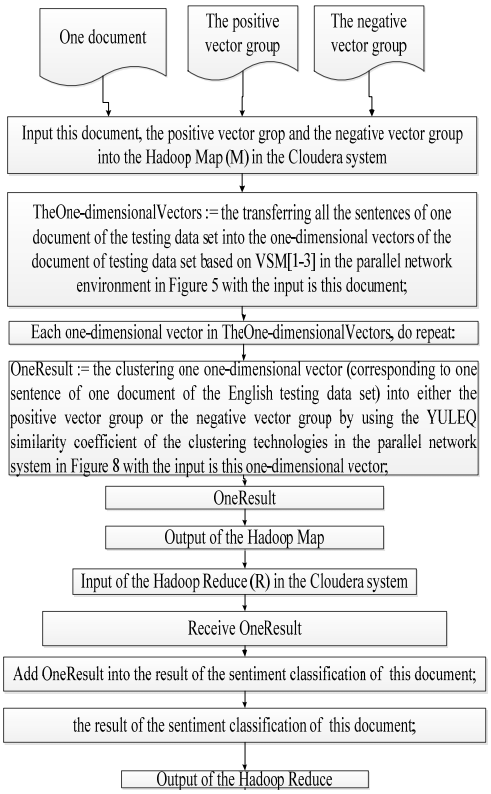

Figure 9: Overview of using the YULEQ coefficient (YC) of the clustering technologies of the data mining field to cluster one document of the testing data set into either the positive or the negative in the distributed environment

We propose the algorithm 15 to perform the Hadoop Map phase of using the YULEQ coefficient (YC) of the clustering technologies of the data mining field to cluster one document of the testing data set into either the positive or the negative in the distributed environment. The main ideas of the algorithm 15 are as follows:
Input: one document of the testing data set; the positive vector group and the negative vector group of the training data set.
Output: the result of the sentiment classification of one one-dimensional vector (corresponding to one sentence of this document) into either the positive vector group or the negative vector group
Step 1: Input this document, the positive vector group and the negative vector group into the Hadoop Map in the Cloudera system.
Step 2: TheOne-dimensionalVectors := transferring all the sentences of one document of the testing data set into the one-dimensional vectors of the document of testing data set based on VSM[1-3] in the parallel network environment in Figure 5 with the input is this document;
Step 3: Each one-dimensional vector in TheOne-dimensionalVectors, do repeat:
Step 4: OneResult := the clustering one one-dimensional vector (corresponding to one sentence of one document of the English testing data set) into either the positive vector group or the negative vector group by using the YULEQ similarity coefficient of the clustering technologies in the parallel network system in Figure 8 with the input is this one-dimensional vector;
Step 5: Return OneResult; // the output of the Hadoop Map
We propose the algorithm 16 to perform the Hadoop Reduce phase of using the YULEQ coefficient (YC) of the clustering technologies of the data mining field to cluster one document of the testing data set into either the positive or the negative in the distributed environment. The main ideas of the algorithm 16 are as follows:
Input: OneResult - the result of the sentiment classification of one one-dimensional vector (corresponding to one sentence of this document) into either the positive vector group or the negative vector group
Output: the result of the sentiment classification of this document.
Step 1: Receive OneResult - the result of the sentiment classification of one one-dimensional vector (corresponding to one sentence of this

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

document) into either the positive vector group or the negative vector group;
Step 2: Add OneResult into the result of the sentiment classification of this document;
Step 3: Return the result of the sentiment classification of this document;

In Figure 10, we cluster the documents of the testing data set into either positive or the negative by using the YULEQ similarity coefficient of the clustering technologies in the parallel network system. This stage comprises two phases: the Hadoop Map phase and the Hadoop Reduce phase. The input of the Hadoop Map is the documents of the testing data set and the training data set. The output of the Hadoop Map is the result of the sentiment classification of one document of the testing data set. 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 one document of the testing data set. The output of the Hadoop Reduce is the results of the sentiment classification of the documents of the testing data set.

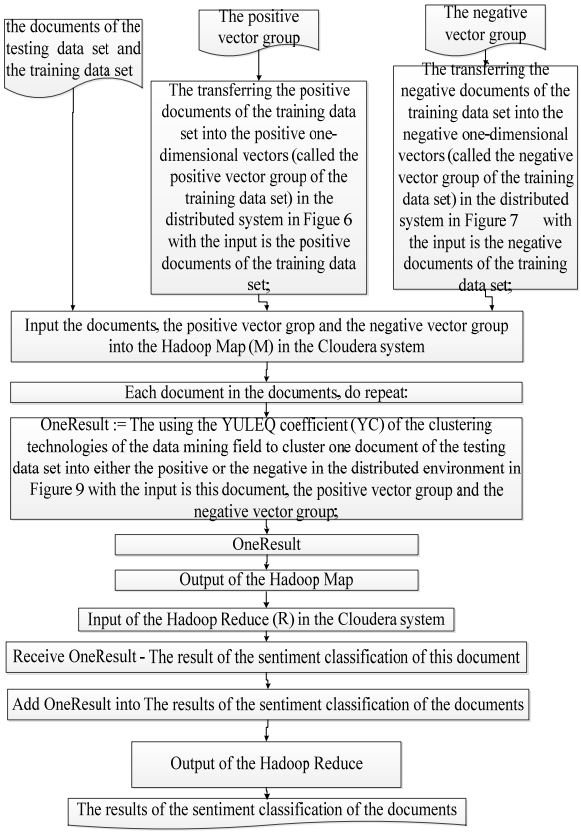

Figure 10: Overview of clustering the documents of the testing data set into either positive or the negative by using the YULEQ similarity coefficient of the clustering technologies in the parallel network system.

We build the algorithm 17 to implement the Hadoop Map phase of clustering the documents of the testing data set into either the positive or the negative in the sequential environment. The main ideas of the algorithm 17 are as follows:
Input: the documents of the testing data set and the training data set
Output: the result of the sentiment classification of one document of the testing data set;
Step 1: The transferring the positive documents of the training data set into the positive one-dimensional vectors (called the positive vector group of the training data set) in the distributed system in Figure 6 with the input is the positive documents of the training data set;
Step 2: The transferring the negative documents of the training data set into the negative one-dimensional vectors (called the negative vector group of the training data set) in the distributed system in Figure 7 with the input is the negative documents of the training data set;
Step 3: Input the documents of the testing data set, the positive vector group and the negative vector group into the Hadoop Map in the Cloudera system
Step 4: Each document in the documents of the testing data set, do repeat:
Step 5: OneResult := The using the YULEQ coefficient (YC) of the clustering technologies of the data mining field to cluster one document of the testing data set into either the positive or the negative in the distributed environment in Figure 9 with the input is this document, the positive vector group and the negative vector group;
Step 6: Return OneResult - the result of the sentiment classification of one document of the testing data set;//the output of the Hadoop Map
We build the algorithm 18 to perform the Hadoop Reduce phase of clustering the documents of the testing data set into either the positive or the negative in the parallel environment. The main ideas of the algorithm 18 are as follows:
Input: OneResult - the result of the sentiment classification of one document of the testing data set;//the output of the Hadoop Map
Output: the results of the sentiment classification of the documents of the testing data set;
Step 1: Receive OneResult ;
Step 2: Add OneResult into the results of the sentiment classification of the documents of the testing data set;
Step 3: Return the results of the sentiment classification of the documents of the testing data set;

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

5. EXPERIMENT

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

Java programming language is used for programming to save data sets, implementing our proposed model to classify the 7,000,000 documents of the testing data set. To implement the proposed model, we have already used Java programming language to save the English testing data set and to save the results of emotion classification.

The sequential environment in this research includes 1 node (1 server). The Java language is used in programming our model related to the YULEQ similarity coefficient of the clustering Technologies. The configuration of the server in the sequential environment is: Intel® Server Board S1200V3RPS, Intel® Pentium® Processor G3220 (3M Cache, 3.00 GHz), 2GB YC3-10600 EYC 1333 MHz LP Unbuffered DIMMs. The operating system of the server is: Cloudera. We perform the proposed model related to the YULEQ similarity coefficient of the clustering technologies in the Cloudera parallel network environment; this Cloudera system includes 9 nodes (9 servers). The Java language is used in programming the application of the proposed model related to the YULEQ similarity coefficient of the clustering technologies in the Cloudera. The configuration of each server in the Cloudera system is: Intel® Server Board S1200V3RPS, Intel® Pentium® Processor G3220 (3M Cache, 3.00 GHz), 2GB YC3-10600 EYC 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 results of the documents of the English testing data set to test are presented in Table 1 below.

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

In Table 3 below, the average time of the classification of our new model for the English documents in testing data set are displayed

6. CONCLUSION

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

In this work, we have proposed a new model to classify sentiment of English documents using the Mahalanobis distance and the sentiment-lexicons-based multi-dimensional vectors with Hadoop Map (M) /Reduce (R) in the Cloudera parallel network environment. With our proposed new model, we have achieved 87.85% accuracy of the testing data set in Table 6. Until now, not many studies have shown that the clustering methods can be used to classify data. Our research shows that clustering methods are used to classify data and, in particular, can be used to classify emotion in text.

In Table 3, the average time of the semantic classification of using the YULEQ similarity coefficient of the clustering technologies in the sequential environment is 36,197,523 seconds / 7,000,000 English documents and it is greater than the average time of the emotion classification of using the YULEQ similarity coefficient of the clustering technologies in the Cloudera parallel network environment – 3 nodes which is 10,732,507 seconds /7,000,000 English documents. The average time of the emotion classification of using the YULEQ similarity coefficient of the clustering technologies in the Cloudera parallel network environment – 9 nodes, which is 4,013,269 seconds / 7,000,000 English documents, is the shortest time. Besides, the average time of the emotion classification of using the sentiment-lexicons with the YC in the Cloudera parallel network environment – 6 nodes is 6,146,253 seconds / 7,000,000 English documents

The execution time of using the YULEQ similarity coefficient of the clustering technologies in the Cloudera is dependent on the performance of the Cloudera parallel system and also dependent on the performance of each server on the Cloudera system.

The proposed model has many advantages and disadvantages. Its positives are as follows: It uses the YULEQ 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.

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

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 researches related to the YULEQ similarity coefficient in [15- 19] are presented.

The comparisons of our model's positives and negatives with the surveys related to the YULEQ similarity coefficient in [15-19] are displayed in Table 9.

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, 2015, pp. 7-19.

[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*, USA, 2015, pp. 1130-1135

[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, 2016, 21-45.

[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, 2016, 77-88

[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)* , New York USA, 2016, 53-62

[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, India, 2016, 171-179

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

[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:

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

10.1007/s10489-016-0858-z, November 2016, 1-2212.

[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*, doi:10.1007/s12530-017-9180-1, April 2017, pp 1-27

[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, May 2017, 31 pages

[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, 2017, 23 pages

[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] Hidenao Abe, Shusaku Tsumoto, "Analyzing Behavior of Objective Rule Evaluation Indices Based on a Correlation Coefficient", *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems KES 2008: Knowledge-Based Intelligent Information and Engineering Systems*, 2008, pp 758-765

[17] Hidenao Abe, Shusaku Tsumoto, "Analyzing Behavior of Objective Rule Evaluation Indices Based on Pearson Product-Moment Correlation Coefficient", *International Symposium on Methodologies for Intelligent Systems, ISMIS 2008: Foundations of Intelligent Systems*, 2008, pp 84-89

[18] Hidenao Abe, ShusakuTsumoto, "A Method to Characterize Dataset Based on Objective Rule Evaluation Indices", *Proc. of SPIE* Vol. 7344 73440D-1, doi: 10.1117/12.820319, 2009

[19] Sony Hartono Wijaya, Farit Mochamad Afendi, Irmanida Batubara, Latifah K. Darusman, Md Altaf-Ul-Amin, Shigehiko Kanaya, "Finding an appropriate equation to measure similarity between binary vectors: case studies on Indonesian and Japanese herbal medicines", *BMC BioinformaticsBMC series – open, inclusive and trusted* 2016 17:520, https://doi.org/10.1186/s12859-016-1392-z, 2016.

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

APPENDICES:

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

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

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

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

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

Table 6: Comparisons of our model with the latest sentiment classification models (or the latest sentiment classification methods) in [4-14]

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

Table 8: Comparisons of our model with the researches related to the YULEQ similarity coefficient in [15-19]

Table 9: Comparisons of our model's positives and negatives the surveys related to the YULEQ similarity coefficient in [15-19]

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

	Testing Dataset	**Correct Classification**	**Incorrect Classification**
Negative	3,00,000	3,084,762	415,238
Positive	3,500,000	3,064,738	435,262
Summary	7,000,000	6,149,500	850,500

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	87.85 %
	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 /7,000,000 English documents.
The YULEQ similarity coefficient of the clustering technologies in the sequential environment	36,197,523 seconds
The YULEQ similarity coefficient of the clustering technologies in the Cloudera distributed system – 3 nodes	10,732,507 seconds
The YULEQ similarity coefficient of the clustering technologies in the Cloudera distributed system – 6 nodes	6,146,253 seconds
The YULEQ similarity coefficient of the clustering technologies in the Cloudera distributed system – 9 nodes	4,013,269 seconds

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

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

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	YC	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	Yes	Yes	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 should be cost effective (i.e., should follow the space/time constraint. It is also very popular. Although the system has all these properties, it is facing some major drawbacks.	The drawbacks are that the system yields no theoretical findings. Weights associated with the vectors are very arbitrary, and this system is an independent system, thus requiring separate attention. Though it is a promising technique, the current level of success of the vector space model techniques used for information retrieval are not able to satisfy user needs and need extensive attention.
[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	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	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

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

	Cloudera distributed system.	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.	advocate its use in further experiments.
Our work	-We use the YULEQ similarity coefficient of the clustering technologies 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	YC	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	Yes	Yes	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 Features for Effective Sentiment	Experiments performed with a substantial number of datasets (nineteen) demonstrate that the effectiveness of the proposed sentiment-based meta-level features is not only superior to the traditional bag-of-words	A line of future research would be to explore the authors' meta features with other classification algorithms and feature

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

	Analysis	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 aYCount 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.	selection techniques in different sentiment analysis tasks such as scoring movies or products aYCording 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 69.224% on the training dataset. All of these methods are implemented to classify the reviews based on our new dictionary and the Internet Movie Database data set.	No mention
[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 the YULEQ similarity coefficient of the clustering technologies 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.		

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

Table 8: Comparisons of our model with the researches related to the YULEQ similarity coefficient in [15-19]

Studies	YC	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	Yes	Yes	Yes	Yes

Table 9: Comparisons of our model's positives and negatives the surveys related to the YULEQ similarity coefficient 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]	Analyzing Behavior of Objective Rule Evaluation Indices Based on a Correlation Coefficient	In this analysis, the authors calculated average values of each index using bootstrap method on 32 classification rule sets learned with information gain ratio. Then, the authors found the following relationships based on the correlation coefficient values: similar pairs, discrepant pairs, and independent indices. With regarding to this result, the authors discuss about relative functional relationships between each group of objective indices.	No mention
[17]	Analyzing Behavior of Objective Rule Evaluation Indices Based on Pearson Product-Moment Correlation Coefficient	In this analysis, the authors calculated average values of each index using bootstrap method on 32 classification rule sets learned with information gain ratio. Then, the authors found the following relationships based on the correlation coefficient values: similar pairs, discrepant pairs, and independent indices. With regarding to this result, we discuss about relative functional relationships between each group of objective indices.	No mention
[18]	A Method to Characterize Dataset Based on Objective Rule Evaluation Indices	The authors consider about a method to reuse objective rule evaluation indices of classification rules. Objective rule evaluation indices such as support, precision and recall are calculated by using a rule set and a validation dataset. This data-driven approach is often used to filter out not useful rules from obtained rule set by a rule learning algorithm. At thesame time, these indices can detect differences between two validation datasets by using the rule set and the indices, because the definitions of indices independent on both of a rule and a dataset. In this survey, the authors present a method to characterize given datasets based on objective rule evaluation indices by using differences of correlation coefficients between each index. By comparing the differences, the authors describe the results of similar/dissimilar groups of the datasets.	No mention
[19]	Finding an appropriate equation to measure similarity between binary vectors: case studies on	The selection of binary similarity and dissimilarity measures for multivariate analysis is data dependent. The proposed method can be used to find the most suitable binary similarity and dissimilarity equation	No mention

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

	Indonesian and Japanese herbal medicines	wisely for a particular data. The authors' finding suggests that all four types of matching quantities in the Operational Taxonomic Unit (OTU) table are important to calculate the similarity and dissimilarity coefficients between herbal medicine formulas. Also, the binary similarity and dissimilarity measures that include the negative match quantity d achieve better capability to separate herbal medicine pairs compared to equations that exclude d.	
Our work	-We use the YULEQ similarity coefficient of the clustering technologies 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.		

APPENDIX OF CODES:

Algorithm 1: transferring one english document into one multi-dimensional vector.
Algorithm 2: creating the positive vector group of the training data set in the sequential system
Algorithm 3: creating the negative vector group of the training data set in the sequential system.
Algorithm 4: clustering one one-dimensional vector (corresponding to one sentence of one document of the english testing data set) into either the positive vector group or the negative vector group by using the YULEQ similarity coefficient of the clustering technologies in the sequential system.
Algorithm 5: clustering one document of the testing data set into either positive or the negative by using the YULEQ similarity coefficient of the clustering technologies in the sequential system.
Algorithm 6: clustering the documents of the testing data set into either positive or the negative by using the YULEQ similarity coefficient of the clustering technologies in the sequential system.
Algorithm 7: performing the hadoop map phase of transferring all the sentences of one document into the one-dimensional vectors of the document based on VSM[1-3] in the cloudera parallel network environment
Algorithm 8: performing the hadoop reduce phase of transferring all the sentences of one document into the one-dimensional vectors of the document based on VSM [1-3] in the cloudera parallel network environment
Algorithm 9: transferring the positive documents of the training data set into the positive one-dimensional vectors (called the positive vector group of the training data set) in the distributed system
Algorithm 10: implementing the hadoop reduce phase of transferring the positive sentences of the training data set into the positive multi-dimensional vectors in the distributed system
algorithm 11: transferring the negative documents of the training data set into the negative one-dimensional vectors (called the negative vector group of the training data set) in the distributed system
algorithm 12: implementing the hadoop reduce phase of transferring the negative sentences of the training data set into the negative multi-dimensional vectors in the distributed system
Algorithm 13: performing the hadoop map phase of clustering one one-dimensional vector (corresponding to one sentence of one document of the english testing data set) into either the positive vector group or the negative vector group by using the YULEQ similarity coefficient of the clustering technologies in the parallel network system
Algorithm 14: implementing the hadoop reduce phase of clustering one one-dimensional vector (corresponding to one sentence of one document of the english testing data set) into either the positive vector group or the negative vector group by using the YULEQ similarity coefficient of the clustering technologies in the parallel network system
Algorithm 15: performing the hadoop map phase of using the YULEQ coefficient (YC) of the clustering technologies of the data mining field to cluster one document of the testing data set into either the positive or the negative in the distributed environment
Algorithm 16: performing the hadoop reduce phase of using the YULEQ coefficient (YC) of the clustering technologies of the data mining field to cluster one document of the testing data set into either the positive or the negative in the distributed environment.
Algorithm 17: implementing the hadoop map phase of clustering the documents of the testing data set into either the positive or the negative in the sequential environment
Algorithm 18: performing the hadoop reduce phase of clustering the documents of the testing data set into either the positive or the negative in the parallel environment

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

ALGORITHM 1: Transferring one English document into one multi-dimensional vector.

Input: one English document
Output: the one-dimensional vectors
Begin
Step 1: Set TheOne-dimensionalVectors := {};
Step 2: Set arraySentences := Split the English document into many separate sentences based on "." Or "!" or "?";
Step 3: For i = 0; i < arraySentences.length; i++, do:
Step 4: Set OneDimensionalVector := Transfer arraySentences[i] into one one-dimensional vector based on VSM [1, 2, 3];
Step 5: Add OneDimensionalVector into the one-dimensional vectors;
Step 6: End For – End Step 3;
Step 7: Return the one-dimensional vectors;
End;

ALGORITHM 2: creating the positive vector group of the training data set in the sequential system

Input: the positive English documents of the English training data set.
Output: the positive vector group PositiveVectorGroup
Begin
Step 1:Each document in the positive documents of the training data set, do repeat:
Step 2: TheOne-dimensionalVectors := Call Algorithm 1 with the positive English document in the English training data set;
Step 3: Add TheOne-dimensionalVectors into PositiveVectorGroup;
Step 4: End Repeat – End Step 1
Step 5: Return PositiveVectorGroup;
End;

ALGORITHM 3: creating the negative vector group of the training data set in the sequential system.

Input: the negative English documents of the English training data set.
Output: the negative vector group NegativeVectorGroup
Begin
Step 1:Each document in the negative documents of the training data set, do repeat:
Step 2: TheOne-dimensionalVectors := Call Algorithm 1 with the negative English document in the English training data set;
Step 3: Add TheOne-dimensionalVectors into NegativeVectorGroup;
Step 4: End Repeat – End Step 1
Step 5: Return NegativeVectorGroup;
End;

ALGORITHM 4: Clustering One One-Dimensional Vector (Corresponding To One Sentence Of One Document Of The English Testing Data Set) Into Either The Positive Vector Group Or The Negative Vector Group By Using The YULEQ Similarity Coefficient Of The Clustering Technologies In The Sequential System.

Input: one one-dimensional vector A (corresponding to one sentence of one document of the English testing data set), the positive vector group and the negative vector group;
Output: the sentiment polarity (positive, negative, neutral);
Begin
Step 1: Set total_positive := 0 and total_negative := 0;
Step 2: Each vector B in the positive vector group, do repeat:
Step 3: total_positive := total_positive + Similarity measure from Eq. (1) between the vector A and the vector B;
Step 4: End Repeat – End Step 2;
Step 5: Each vector C in the negative vector group, do repeat:
Step 6: total_negative := total_negative + Similarity measure from Eq. (1) between the vector A and the vector C;
Step 7: End Repeat – End Step 2;
Step 8: If total_positive is greater than total_negative The Return positive;
Step 9: Else If total_positive is less than total_negative The Return negative;
Step 10: Return neutral;
End;

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

ALGORITHM 5: Clustering One Document Of The Testing Data Set Into Either Positive Or The Negative By Using The YULEQ Similarity Coefficient Of The Clustering Technologies In The Sequential System.

Input: one document of the testing data set, the positive vector group and the negative vector group.
Output: the result of the sentiment classification of this document
Begin
Step 1: Set count_positive := 0 and count_negative := 0;
Step 2: Split this document into the sentences;
Step 3: Each sentence in the sentences of the testing data set, do repeat:
Step 4: OneResult := the algorithm 4 to cluster one one-dimensional vector (corresponding to one sentence of one document of the English testing data set) into either the positive vector group or the negative vector group by using the YULEQ similarity coefficient of the clustering technologies in the sequential system with the input is this document;
Step 5: End Repeat – End Step 2;
Step 6: If OneResult is the positive Then count_positive := count_positive +1;
Step 7: Else If OneResult is the negative Then count_negative := count_negative +1;
Step 8: End Repeat – End Step 3;
Step 9: If count_positive is greater than count_negative Then Return positive;
Step 10: Else If count_positive is less than count_negative Then Return negative;
Step 11: Return the neutral;
End;

ALGORITHM 6: clustering the documents of the testing data set into either positive or the negative by using the YULEQ similarity coefficient of the clustering technologies in the sequential system.

Input: the testing data set and the training data set
Output: the results of the sentiment classification of the documents of the testing data set;
Begin
Step 1: the positive vector group := The algorithm 2 to create the positive vector group of the training data set in the sequential system with the input is the positive document of the training data set;
Step 2: the negative vector group := The algorithm 3 to create the negative vector group of the training data set in the sequential system with the input is the negative document of the training data set;
Step 3: Each document in the documents of the testing data set , do repeat:
Step 4: OneResult := The algorithm 5 to cluster one document of the testing data set into either positive or the negative by using the YULEQ similarity coefficient of the clustering technologies in the sequential system with the input is this document, the positive vector group and the negative vector group;
Step 5: Add OneResult into the results of the sentiment classification of the documents of the testing data set;
Step 6: End Repeat – End Step 3;
Step 7: Return the results of the sentiment classification of the documents of the testing data set;
End;

ALGORITHM 7: performing the hadoop map phase of transferring all the sentences of one document into the one-dimensional vectors of the document based on VSM[1-3] in the cloudera parallel network environment

Input: one document of the testing data set;
Output: one one-dimensional vector of this document
Begin
Step 1: Input this document into the Hadoop Map in the Cloudera system;
Step 2: Split this document into the sentences;
Step 3: Each sentence in the sentences, do repeat:
Step 4: one one-dimensional vector := The transforming one English sentence into one one-dimensional vector based on VSM [1-3] with the input is this sentence
Step 5: Return one one-dimensional vector; //the output of the Hadoop Map phase.
End;

ALGORITHM 8: performing the hadoop reduce phase of transferring all the sentences of one document into the one-dimensional vectors of the document based on VSM [1-3] in the cloudera parallel network environment

Input: one one-dimensional vector of this document
Output: the one-dimensional vectors of this document
Begin
Step 1: Receive one one-dimensional vector;
Step 2: Add this one-dimensional vector into the one-dimensional vectors of this document;
Step 3: Return the one-dimensional vectors of this document;
End;

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

ALGORITHM 9: transferring the positive documents of the training data set into the positive one-dimensional vectors (called the positive vector group of the training data set) in the distributed system

Input: the positive documents of the training data set
Output: one one-dimensional vector of one positive document of the training data set
Begin
Step 1: Input the positive documents into the Hadoop Map in the Cloudera system.
Step 2: Each document in the positive document, do repeat:
Step 3: TheOne-dimentionalVectorsOfOneDocument := The transferring all the sentences of one document of the testing data set into the one-dimensional vectors of the document of testing data set based on VSM[1-3] in the parallel network environment in Figure 5 with the input is this document;
Step 4: Return TheOne-dimentionalVectorsOfOneDocument ;
End;

ALGORITHM 10: implementing the hadoop reduce phase of transferring the positive sentences of the training data set into the positive multi-dimensional vectors in the distributed system

Input: TheOne-dimentionalVectorsOfOneDocument - the one-dimensional vectors of the positive document of the training data set
Output: the positive one-dimensional vectors, called the positive vector group of the training data set
Begin
Step 1: Receive TheOne-dimentionalVectorsOfOneDocument ;
Step 2: Add TheOne-dimentionalVectorsOfOneDocument into PositiveVectorGroup;
Step 3: Return PositiveVectorGroup - the positive one-dimensional vectors, called the positive vector group of the training data set;
End;

ALGORITHM 11: transferring the negative documents of the training data set into the negative one-dimensional vectors (called the negative vector group of the training data set) in the distributed system

Input: the negative documents of the training data set
Output: one one-dimensional vector of one negative document of the training data set
Begin
Step 1: Input the negative documents into the Hadoop Map in the Cloudera system.
Step 2: Each document in the negative document, do repeat:
Step 3: TheOne-dimentionalVectorsOfOneDocument := The transferring all the sentences of one document of the testing data set into the one-dimensional vectors of the document of testing data set based on VSM[1-3] in the parallel network environment in Figure 5 with the input is this document;
Step 4: Return TheOne-dimentionalVectorsOfOneDocument ;
End;

ALGORITHM 12: implementing the hadoop reduce phase of transferring the negative sentences of the training data set into the negative multi-dimensional vectors in the distributed system

Input: TheOne-dimentionalVectorsOfOneDocument - the one-dimensional vectors of the negative document of the training data set
Output: the negative one-dimensional vectors, called the negative vector group of the training data set
Begin
Step 1: Receive TheOne-dimentionalVectorsOfOneDocument ;
Step 2: Add TheOne-dimentionalVectorsOfOneDocument into NegativeVectorGroup;
Step 3: Return NegativeVectorGroup - the negative one-dimensional vectors, called the negative vector group of the training data set;
End;

ALGORITHM 13: performing the hadoop map phase of clustering one one-dimensional vector (corresponding to one sentence of one document of the english testing data set) into either the positive vector group or the negative vector group by using the YULEQ similarity coefficient of the clustering technologies in the parallel network system

Input: one one-dimensional vector A of a document in the testing data set; the positive vector group and the negative vector group of the training data set.
Output: (SimilarityMeasure_positive, positive) and (SimilarityMeasure_negative, negative)
Begin

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

Step 1: Input this vector A, the positive vector group and the negative vector group into the Hadoop Map in the Cloudera system;
Step 2: Each vector B in the positive vector group, do repeat:
Step 3: SimilarityMeasure_positive := Similarity measure from Eq. (1) between the vector A and the vector B;
Step 4: Return (SimilarityMeasure_positive, positive); //the output of the Hadoop Map
Step 5: Each vector C in the negative vector group, do repeat:
Step 6: SimilarityMeasure_negative := Similarity measure from Eq. (1) between the vector A and the vector C;
Step 7: Return (SimilarityMeasure_negative, negative);//the output of the Hadoop Map
End;

ALGORITHM 14: implementing the hadoop reduce phase of clustering one one-dimensional vector (corresponding to one sentence of one document of the english testing data set) into either the positive vector group or the negative vector group by using the YULEQ similarity coefficient of the clustering technologies in the parallel network system

Input: (SimilarityMeasure_positive, positive) and (SimilarityMeasure_negative, negative) – the output of the Hadoop Map
Output: the result of clustering the vector into either the positive vector group or the negative vector group.
Begin
Step 1: Receive (SimilarityMeasure_positive, positive) and (SimilarityMeasure_negative, negative);
Step 2: If positive Then total_positive := total_positive + SimilarityMeasure_positive;
Step 6: If negative Then total_negative := total_negative + SimilarityMeasure_negative;
Step 7: If total_positive is greater than total_negative The Return positive;
Step 8: Else If total_positive is less than total_negative The Return negative;
Step 9: Return neutral;
Step 10: Return the result of clustering the vector into either the positive vector group or the negative vector group;
End;

ALGORITHM 15: performing the hadoop map phase of using the YULEQ coefficient (YC) of the clustering technologies of the data mining field to cluster one document of the testing data set into either the positive or the negative in the distributed environment

Input: one document of the testing data set; the positive vector group and the negative vector group of the training data set.
Output: the result of the sentiment classification of one one-dimensional vector (corresponding to one sentence of this document) into either the positive vector group or the negative vector group
Begin
Step 1: Input this document, the positive vector group and the negative vector group into the Hadoop Map in the Cloudera system.
Step 2: TheOne-dimensionalVectors := the transferring all the sentences of one document of the testing data set into the one-dimensional vectors of the document of testing data set based on VSM[1-3] in the parallel network environment in Figure 5 with the input is this document;
Step 3: Each one-dimensional vector in TheOne-dimensionalVectors, do repeat:
Step 4: OneResult := the clustering one one-dimensional vector (corresponding to one sentence of one document of the English testing data set) into either the positive vector group or the negative vector group by using the YULEQ similarity coefficient of the clustering technologies in the parallel network system in Figure 8 with the input is this one-dimensional vector;
Step 5: Return OneResult; // the output of the Hadoop Map
End;

ALGORITHM 16: performing the hadoop reduce phase of using the YULEQ coefficient (YC) of the clustering technologies of the data mining field to cluster one document of the testing data set into either the positive or the negative in the distributed environment.

Input: OneResult - the result of the sentiment classification of one one-dimensional vector (corresponding to one sentence of this document) into either the positive vector group or the negative vector group
Output: the result of the sentiment classification of this document.
Begin
Step 1: Receive OneResult - the result of the sentiment classification of one one-dimensional vector (corresponding to one sentence of this document) into either the positive vector group or the negative vector group;
Step 2: Add OneResult into the result of the sentiment classification of this document;
Step 3: Return the result of the sentiment classification of this document;
End;

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

ALGORITHM 17: implementing the hadoop map phase of clustering the documents of the testing data set into either the positive or the negative in the sequential environment

Input: the documents of the testing data set and the training data set
Output: the result of the sentiment classification of one document of the testing data set;
Begin
Step 1: The transferring the positive documents of the training data set into the positive one-dimensional vectors (called the positive vector group of the training data set) in the distributed system in Figue 6 with the input is the positive documents of the training data set;
Step 2: The transferring the negative documents of the training data set into the negative one-dimensional vectors (called the negative vector group of the training data set) in the distributed system in Figure 7 with the input is the negative documents of the training data set;
Step 3: Input the documents of the testing data set, the positive vector group and the negative vector group into the Hadoop Map in the Cloudera system
Step 4: Each document in the documents of the testing data set, do repeat:
Step 5: OneResult := The using the YULEQ coefficient (YC) of the clustering technologies of the data mining field to cluster one document of the testing data set into either the positive or the negative in the distributed environment in Figure 9 with the input is this document, the positive vector group and the negative vector group;
Step 6: Return OneResult - the result of the sentiment classification of one document of the testing data set;//the output of the Hadoop Map
End;

ALGORITHM 18: performing the hadoop reduce phase of clustering the documents of the testing data set into either the positive or the negative in the parallel environment

Input: OneResult - the result of the sentiment classification of one document of the testing data set;//the output of the Hadoop Map
Output: the results of the sentiment classification of the documents of the testing data set;
Begin
Step 1: Receive OneResult ;
Step 2: Add OneResult into the results of the sentiment classification of the documents of the testing data set;
Step 3: Return the results of the sentiment classification of the documents of the testing data set;
End;