American Journal of Engineering and Applied Sciences

Original Research Paper

A STING Algorithm and Multi-dimensional Vectors Used for English Sentiment Classification in a Distributed System

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

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

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

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

Abstract: Sentiment classification is significant in everyday life, such as in political activities, commodity production and commercial activities. Finding a fast, highly accurate solution to classify emotion has been a challenge for scientists. In this research, we have proposed a new model for Big Data sentiment classification in the parallel network environment - a Cloudera system with Hadoop Map (M) and Hadoop Reduce (R). Our new model has used a Statistical Information Grid Algorithm (STING) with multi-dimensional vector and 2,000,000 English documents of our English training data set for English document-level sentiment classification. Our new model can classify sentiment of millions of English documents based on many English documents in the parallel network environment. However, we tested our new model on our testing data set (including 1,000,000 English reviews, 500,000 positive and 500,000 negative) and achieved 83.92% accuracy.

Keywords: Sentiment Classification, English Sentiment Classification, Opinion Mining, English Document Opinion Mining, Statistical Information Grid, STING, Distributed System, Parallel System

Introduction

Sentiment classification is significant in everyday life, such as in political activities, commodity production and commercial activities. Finding a fast, highly accurate solution to classify emotion has been a challenge for scientists.

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:

- It is assumed that each English sentence has m English words (or English phrases)
- It is assumed 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
- It is assumed that each English document has n English sentences
- It is assumed 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 documents of the training data set are transferred into the multi-dimensional vectors. The positive documents of the training data set are transferred into the positive multi-dimensional vectors, the positive vector group. The negative documents of the training data set are transferred into the negative multi-dimensional vectors, the negative vector group

- All the documents of the testing data set are transferred into the multi-dimensional vectors
- One multi-dimensional vector (corresponding to one English document in the English testing data set) is the positive polarity if the vector is clustered into the positive vector group. One multi-dimensional vector (corresponding to one English document in the English testing data set) is the negative polarity if the vector is clustered into the negative vector group. One multi-dimensional vector (corresponding to one English document in the English testing data set) is the neutral polarity if the vector is not clustered into either the positive vector group or the negative vector group

In this study, we propose a new model by using the Statistical Information Grid Algorithm (STING) to classify emotions (positive, negative, neutral) of English documents in the parallel system. A study of semantic classification (emotional analysis) using the STING does not currently exist in the world.

According to the STING works in the world and in (Wang *et al*., 1997; Pilevar and Sukumar, 2005; Lin *et al*., 2008; Halkidi *et al*., 2001a; Joshi and Kaur, 2013; Madhulatha, 2012; Halkidi *et al*., 2001b; Murtagh and Contreras, 2011; Qian and Zhou, 2002; Amini *et al*., 2011; Wang *et al*., 1999; Yue, 2017; Lin *et al*., 2007a; Zhang *et al*., 2010), there is not any research related to the STING that is similar to our work. With the research related to the STING in the parallel system (or the STING in the distributed system) in the world, there are not any STING-related studies in the parallel system that is similar to our work.

According to the latest research on sentiment classification in the world and in (Agarwal and Mittal, 2016a; 2016b; Canuto *et al*., 2016; Ahmed and Danti, 2016; Phu and Tuoi, 2014; Tran *et al*., 2014; Phu *et al*., 2017a; Dat *et al*., 2017; Phu *et al*., 2016; 2017b; 2017c; 2017d; 2017e; 2017f; 2017g; 2017h; 2017i), there is no semantic classification work similar to our model.

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. The Statistical Information Grid Algorithm (STING) can be applied to the sentiment classification. The STING can be applied to the sentiment classification of millions of the documents in English in a parallel network environment. This will result in many discoveries in scientific research, hence the motivation for this study.

The novelty and originality of the proposed approach are as follows:

- The Statistical Information Grid Algorithm (STING) was applied to the sentiment classification of the survey
- The Vector Space Model (VSM) was used in the sentiment classification of the study
- We used the multi-dimensional vectors based on the VSM
- We used many sentiment lexicons
- The Cloudera system, Hadoop Map (M) and Hadoop Reduce (R) were used in the proposed model
- The input of this survey is the documents of the testing data set and the documents of the training data set in English. We studied to transfer the documents into the formats for the novel model which can process them
- We tested the proposed model in both a sequential environment and a distributed network system
- We proposed the STING - related algorithms in both a sequential system and a parallel network environment
- Therefore, we have studied this model in more details

Our model has many significant contributions to the different fields and commercial applications as follows:

- The algorithm of data mining is applicable to semantic analysis of natural language processing
- This study also proves that different fields of scientific research can be related in many ways
- Millions of the documents are successfully processed for emotional analysis
- Many studies and commercial applications can use the results of this survey
- The semantic classification is implemented in the parallel network environment
- The principles are proposed in the research
- The opinion classification of English documents is performed on English sentences
- The proposed model can be applied to other languages easily
- The Cloudera distributed environment is used in this study
- The proposed work can be applied to other distributed systems
- This survey uses Hadoop Map (M) and Hadoop Reduce (R)
- Our proposed model can be applied to many different parallel network environments such as a Cloudera system
- This study can be applied to many different distributed functions such as Hadoop Map (M) and Hadoop Reduce (R)
- The STING-related algorithms are proposed in this research

This study contains 6 sections. Section 1introduces the study; Section 2 discusses the related works about

the Statistical Information Grid Algorithm (STING), etc.; Section 3 is about the English data set; Section 4 represents the methodology of our proposed model; Section 5 represents the experiment. Section 6 provides the conclusion. The References section comprises all the reference documents; all tables are shown in the Appendices section.

Related Work

In this section, we describe summaries of many studies related to a Statistical Information Grid Algorithm (STING), Vector Space Model (VSM), Hadoop, Cloudera, etc.

There are the works related to vector space modeling in (Carrera-Trejo *et al*., 2015; Amini *et al*., 2011; Soucy and Mineau, 2015). First, we transfer all English sentences into many vectors, which are used in the VSM algorithm. In this research Singh and Singh (2015), the authors will examine the vector space model, an information retrieval technique and its variation. The vector space model is an algebraic model used for information retrieval. It represents natural language documents in a formal manner using of vectors in a multi-dimensional space and allows decisions to be made as to which documents are similar to each other and to the queries fired. This research attempts to examine the vector space model, an information retrieval technique that is widely used today. It also explains the existing variations of VSM and proposes the new variation that should be considered. In text classification tasks, one of the main problems Carrera-Trejo *et al*. (2015) is to choose which features give the best results. Various features can be used like words, n-grams, syntactic n-grams of various types (POS tags, dependency relations, mixed, etc.); or a combination of these features can be considered. Also, algorithms for dimensionality reduction of these sets of features can be applied, such as Latent Dirichlet Allocation (LDA). In this research, 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. Amini *et al*. (2011) are two machine learning approaches to Text Categorization (TC) based on the vector space model. In this model, borrowed from information retrieval, documents are represented as a vector where each component is associated with a particular word from the vocabulary. Traditionally, each component value is assigned using the information retrieval TFIDF measure. While this weighting method seems very appropriate for IR, it is not clear that it is the best choice for TC problems. Actually, this weighting method does not leverage the information implicitly contained in the categorization task to represent documents. In this research, the authors 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 of the categorization problem considered get a very small weight. Extensive experiments reported in the research show that this new weighting method improves significantly the classification accuracy as measured on many categorization tasks.

The research projects related to implementing algorithms, applications, studies in parallel network environment in (Hadoop, 2017; Apache, 2017; Cloudera, 2017). In (Hadoop, 2017; Apache, 2017), Hadoop is an Apache-based framework used to handle large data sets on clusters consisting of multiple computers, using the Map and Reduce programming model. The two main projects of the Hadoop are Hadoop Distributed File System (HDFS) and Hadoop M/R (Hadoop Map/Reduce). Hadoop M/R allows engineers to program for writing applications for parallel processing of large data sets on clusters consisting of multiple computers. A M/R task has two main components: (1) Map and (2) Reduce. This framework splits inputting data into chunks which multiple Map tasks can handle with a separate data partition in parallel. The outputs of the map tasks are gathered and processed by the Reduce task ordered. The input and output of each M/R task are stored in HDFS because the Map tasks and the Reduce tasks perform on the pair (key, value) and formatted input and output formats will be the pair (key, value). Apache (2017), the global provider of the fastest, easiest and most secure data management and analytics platform built on Apache™ Hadoop® and the latest open source technologies, announced today that it will submit proposals for Impala and Kudu to join the Apache Software Foundation (ASF). By donating its leading analytic database and columnar storage projects to the ASF, Cloudera aims to accelerate the growth and diversity of their respective developer communities. Cloudera delivers the modern data management and analytics platform built on Apache Hadoop and the latest open source technologies. The world's leading organizations trust Cloudera to help solve their most challenging business problems with Cloudera Enterprise, the fastest, easiest and most secure data

platform available to the modern world. Cloudera's customers efficiently capture, store, process and analyze vast amounts of data, empowering them to use advanced analytics to drive business decisions quickly, flexibly and at lower cost than has been possible before. To ensure Cloudera's customers are successful, it offers comprehensive support, training and professional services.

There are the works related to the Statistical Information Grid Algorithm (STING) in (Wang *et al*., 1997; Pilevar and Sukumar, 2005; Lin *et al*., 2008; Halkidi *et al*., 2001; Joshi and Kaur, 2013; Madhulatha, 2012; Halkidi *et al*., 2001b; Murtagh and Contreras, 2011; Qian and Zhou, 2002; Amini *et al*., 2011; Wang *et al*., 1999; Yue, 2017; Lin *et al*., 2007a; Zhang *et al*., 2010).

The latest researches of the sentiment classification are (Agarwal and Mittal, 2016a; 2016b; Canuto *et al*., 2016; Ahmed and Danti, 2016; Phu and Tuoi, 2014; Tran *et al*., 2014; Phu *et al*., 2017a; Dat *et al*., 2017; Phu *et al*., 2016; 2017b; 2017c; 2017d; 2017e; 2017f; 2017g; 2017h; 2017i; 2017j). In the research Lin *et al*. (2007b), the authors present their machine learning experiments with regard to sentiment analysis in blog, review and forum texts found on the World Wide Web and written in English, Dutch and French. The survey in Agarwal and Mittal (2016a) discusses an approach where an exposed stream of tweets from the Twitter micro blogging site are preprocessed and classified based on their sentiments. In sentiment classification system the concept of opinion subjectivity has been accounted. In the study, the authors present opinion detection and organization subsystem, which have already been integrated into our larger question-answering system, etc.

Data Set

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

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

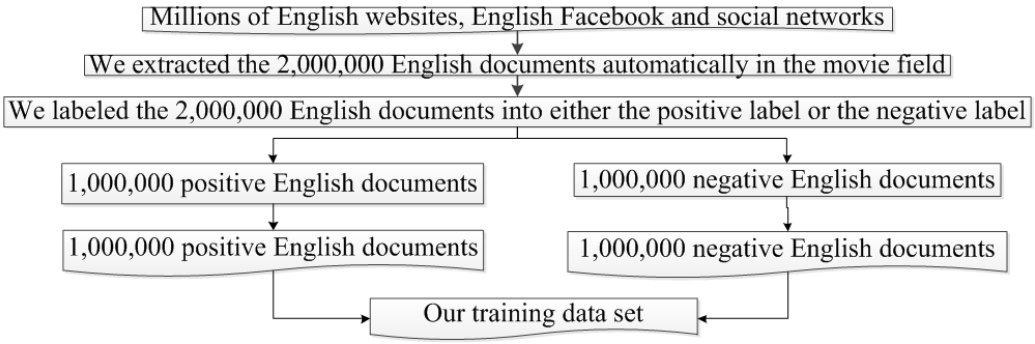

Fig. 1: Our training data set in English

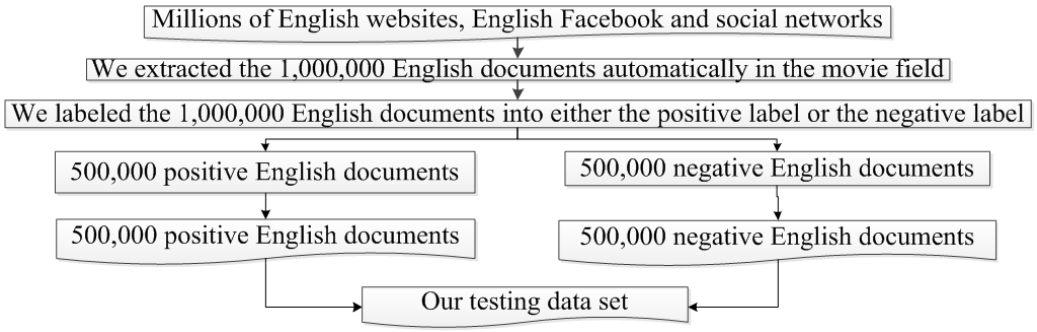

Fig. 2: Our testing data set in English

Methodology

This section has two parts as follows: Semantic classification for the documents of the testing in the sequential environment is presented in the first part. In the second part, sentiment classification for the reviews of the testing in the parallel network environment is displayed.

With the training data set, there are two groups. The first group includes the positive documents and the second group is the negative documents. The first group is called the positive cluster. The second group is called the negative cluster. All documents in both the first group and the second group go through the segmentation of words and stop-words removal; then, they are transferred into the multi-dimensional vectors (vector representation). The positive documents of the positive cluster are transferred into the positive multi-dimensional vectors which are called the positive vector group (or positive vector cluster). The negative documents of the negative cluster are transferred into the negative multi-dimensional vectors which are called the negative vector group (or negative vector cluster). Therefore, the training data set includes the positive vector group (or positive vector cluster) and the negative vector group (or negative vector cluster).

In (Carrera-Trejo *et al*., 2015; Amini *et al*., 2011; Soucy and Mineau, 2015), the VSM is an algebraic model used for information retrieval. It represents a natural language document in a formal manner by the use of vectors in a multidimensional space. The Vector Space Model (VSM) is a way of representing documents through the words they contain. The concepts behind vector space modeling are that by placing terms, documents and queries in a term-document space, it is possible to compute the similarities between queries and the terms or documents and allow the results of the computation to be ranked according to the similarity measure between them. The VSM allows decisions to be made about which documents are similar to each other and to queries.

We have transferred all English sentences into one-dimensional vectors similar to VSM (Carrera-Trejo *et al*., 2015; Amini *et al*., 2011; Soucy and Mineau, 2015).

A Statistical Information Grid Algorithm (STING) in A Sequential Environment

In Fig. 3, in the sequential environment, the documents of the English testing data set are transferred to the multi-dimensional vectors as follows: Each document of the testing data set is transferred to each multi-dimensional vector (each sentence of one document in the testing data set is transferred to the one-dimensional vector similar to VSM (Carrera-Trejo *et al*., 2015; Amini *et al*., 2011; Soucy and Mineau, 2015). The positive documents in the training data set are transferred to the positive multi-dimensional vectors, called the positive vector group in the sequential environment: Each document of the positive documents is transferred to each multi-dimensional vector (each sentence, of one document in the positive documents, is transferred to the one-dimensional vector similar to VSM (Carrera-Trejo *et al*., 2015; Amini *et al*., 2011; Soucy and Mineau, 2015) in the sequential environment). The negative documents in the training data set are transferred to the negative multi-dimensional vectors, called the negative vector group in the sequential environment: Each document of the negative documents is transferred to each multi-dimensional vector (each sentence, of one English document in the negative documents, is transferred to the one-dimensional vector similar to VSM (Carrera-Trejo *et al*., 2015; Amini *et al*., 2011; Soucy and Mineau, 2015) in the sequential environment).

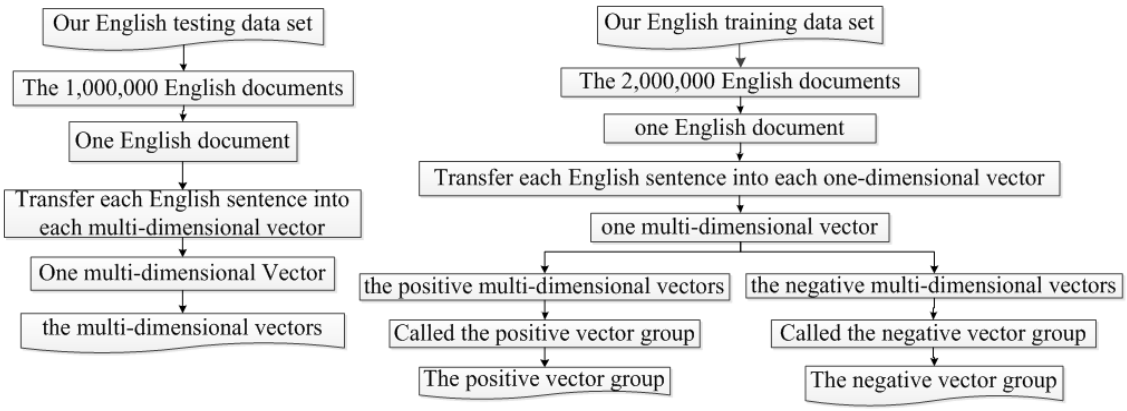

Fig. 3: Overview of transferring all documents into the multi-dimensional vectors

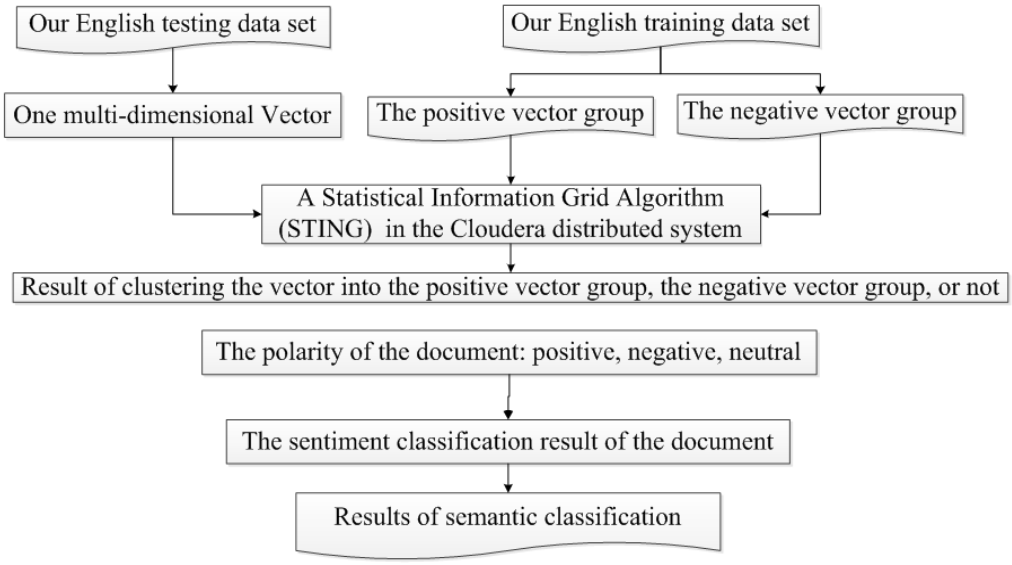

Fig. 4: A Statistical Information Grid Algorithm (STING) in the Sequential Environment

We perform this part in Fig. 4: In the sequential environment, the STING is implemented to cluster one multi-dimensional vector (called A) of the English testing data set to the positive vector group or the negative vector group. The document (corresponding to A) is the positive polarity if A is clustered to the positive vector group. The document (corresponding to A) is the negative polarity if A is clustered to the negative vector group. The document (corresponding to A) is the neutral polarity if A is not clustered to both the positive vector group and the negative vector group.

We built many algorithms to perform the STING in the sequential environment. We built the algorithm 1 to transfer one English document into one multi-dimensional vector. Each document is split into many sentences. Each sentence in each document is transferred to one one-dimensional vector based on VSM (Carrera-Trejo *et al*., 2015; Amini *et al*., 2011; Soucy and Mineau, 2015) in the sequential environment. We insert all the one-dimensional vectors of the sentences into one multi-dimensional vector of one document. The main ideas of the algorithm 1 are as follows:

Input: One English document
Output: One multi-dimensional vector
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 vector (one dimensiona) based on VSM (Carrera-Trejo *et al*., 2015; Amini *et al*., 2011; Soucy and Mineau, 2015)
Step 4: Add the transferred vector into one multi-dimensional vector
Step 5: End Repeat - End Step 2
Step 6: Return one multi-dimensional vector;

We built the algorithm 2 to create the positive vector group. Each document in the positive documents in 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 (Carrera-Trejo *et al*., 2015; Amini *et al*., 2011; Soucy and Mineau, 2015) in the sequential environment. We insert all the one-dimensional vectors of the sentences of the document into one one-dimensional vector of the document. Then, the positive documents in the English training data are transferred to the positive multi-dimensional vectors. 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 document of the training data set, do repeat:
Step 2: OneMultiDimensionalVector := Call Algorithm 1 with the positive English document in the English training data set;
Step 3: Add OneMultiDimensionalVector into PositiveVectorGroup;
Step 4: End Repeat – End Step 1
Step 5: Return PositiveVectorGroup;

We built the algorithm 3 to create the negative vector group. Each document in the negative documents in 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 (Carrera-Trejo *et al*., 2015; Amini *et al*., 2011; Soucy and Mineau, 2015) in the sequential environment. We insert all the one-dimensional vectors of the sentences of the document

into one one-dimensional vector of the document. Then, the negative documents in the English training data set are transferred to the negative multi-dimensional vectors. 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 PositiveVectorGroup
Step 1: Each document in the negative document of the training data set, do repeat:
Step 2: OneMultiDimensionalVector := Call Algorithm 1 with the negative English document in the English training data set;
Step 3: Add OneMultiDimensionalVector into NegativeVectorGroup;
Step 4: End Repeat - End Step 1
Step 5: Return Negative VectorGroup;

We built the algorithm 4 to cluster one multi-dimensional vector (corresponding to one document of the English testing data set) into the positive vector group PositiveVectorGroup, the negative vector group NegativeVectorGroup, or not. The main ideas of the algorithm 4 are as follows:

Input: One multi-dimensional vector A (corresponding to one English document of the English testing data set), the positive vector group PositiveVectorGroup, the negative vector group NegativeVectorGroup;
Output: Positive, negative, neutral;
Step 1: Implement the Statistical Information Grid Algorithm (STING) based on the Statistical Information Grid Algorithm in (Wang *et al*., 1997; Pilevar and Sukumar, 2005; Lin *et al*., 2008; Halkidi *et al*., 2001a; Joshi and Kaur, 2013; Madhulatha, 2012; Halkidi *et al*., 2001b; Murtagh and Contreras, 2011; Qian and Zhou, 2002; Amini *et al*., 2011; Wang *et al*., 1999; Yue, 2017; Lin *et al*., 2007a; Zhang *et al*., 2010) with input is one multi-dimensional vector (corresponding to one English document of the English testing data set), the positive vector group PositiveVectorGroup, the negative vector group NegativeVectorGroup;
Step 2: With the results of Step 1, If the vector is clustered into the positive vector group Then Return positive;
Step 3: Else If the vector is clustered into the negative vector group Then Return negative; End If - End Step 2
Step 4: Return neutral;

A Statistical Information Grid Algorithm (STING) in a Parallel Network Environment

In Fig. 5, all documents of both the English testing data set and the training data set are transferred into all the multi-dimensional vectors in the Cloudera parallel network environment. With the documents of the training data set, we transferred them into the multi-dimensional vectors by using Hadoop Map (M)/Reduce (R) in the Cloudera parallel network environment with the purpose of shortening the execution time of this task. The positive documents of the training data set are transferred into the positive vectors in the Cloudera parallel system and are called the positive vector group. The negative documents of the training data set are transferred into the negative vectors in the Cloudera parallel system and are called the negative vector group. Besides, the documents of the English testing data set are transferred to the multi-dimensional vectors by using Hadoop Map (M)/Reduce (R) in the Cloudera parallel network environment with the purpose of shortening the execution time of this task.

We perform this part in Fig. 6. In the Cloudera distributed network environment, by using the STING, one multi-dimensional vector (called A) of one document in the English testing data set is clustered into the positive vector group or the negative vector group. The document (corresponding to A) is the positive polarity if A is clustered into the positive vector group. The document (corresponding to A) is the negative polarity if A is clustered into the negative vector group. The document (corresponding to A) is the neutral polarity if A is not clustered into both the positive vector group and the negative vector group.

An overview of transferring each English sentence into one vector in the Cloudera network environment is follows in Fig 7.

In Fig. 7, transferring each English document into one vector in the Cloudera network environment includes two phases as follows: Map (M) phases and Reduce (R) phases. Input of the Map phase is one document and Output of the Map phase is many components of a vector which corresponds to the document. One document, input into Hadoop Map (M), is split into many sentences. Each sentence in the English document is transferred into one one-dimensional vector based on VSM (Carrera-Trejo *et al*., 2015; Amini *et al*., 2011; Soucy and Mineau, 2015). This is repeated for all the sentences of the document until all the sentences are transferred into all the one-dimensional vectors of the document. After finishing to transfer each sentence of the document into one one-dimensional vector, the Map phase of Cloudera automatically transfers the one-dimensional vector into the Reduce phase. The input of the Reduce phase is the output of the Map phase and this input comprises many components (many one-dimensional vectors) of a multi-dimensional vector. The output of the Reduce phase is a multi-dimensional vector which corresponds to the document. In the Reduce phase of Cloudera, those components of the vector are built into one multi-dimensional vector.

The documents of the testing data set are transferred into the multi-dimensional vectors based on Fig 7. The STING in the Cloudera parallel network environment has two main phases: The first main phase is Hadoop Map (M) phase in Cloudera and the second main phase is Hadoop Reduce (R) phase in Cloudera. In the Map phase of Cloudera, the input of the phase is the multi-dimensional vector of one English document (which is classified), the positive vector group,

DOI: 10.3844/ajeassp.2018.19.37

the negative vector group; and the output of the phase is the clustering results of the multi-dimensional vector of the document to the positive vector group or the negative vector group, or not.

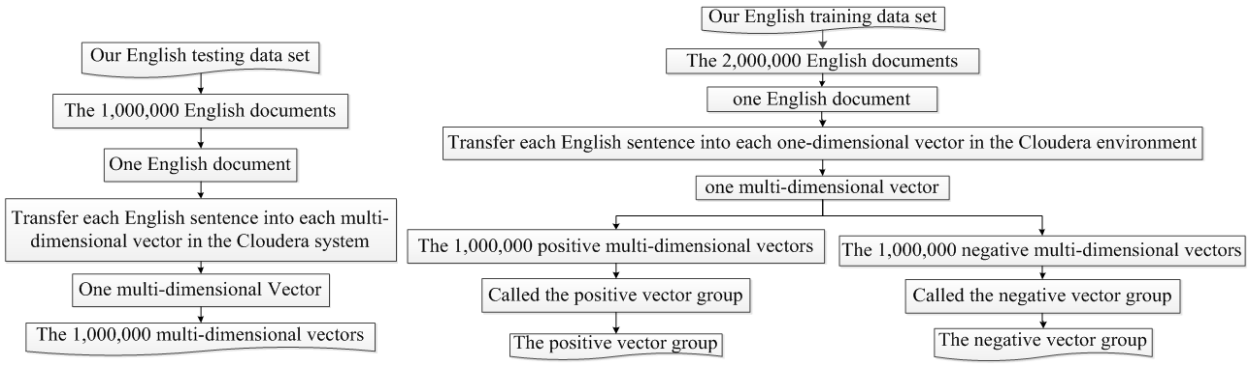

Fig. 5: Overview of transferring all English documents into the multi-dimensional vectors in the Cloudera distributed system

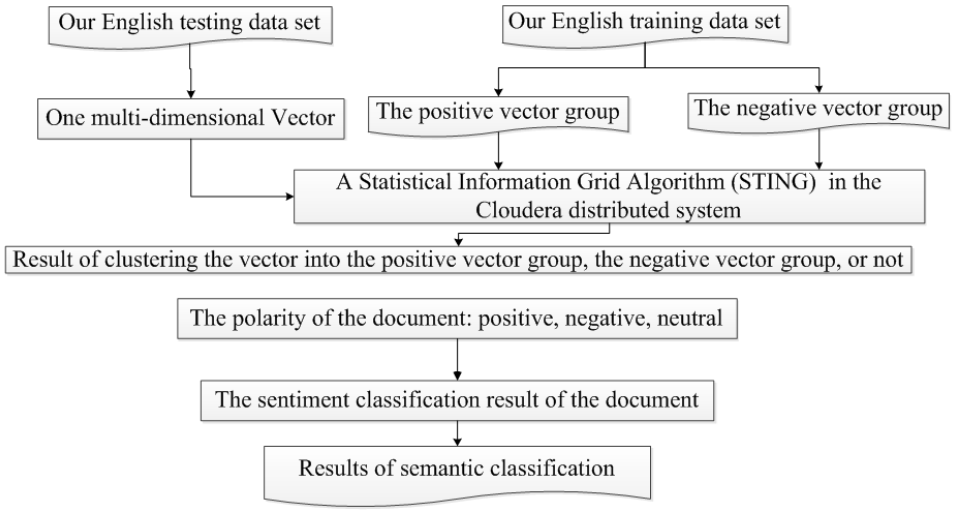

Fig. 6: A Statistical Information Grid Algorithm (STING) in the Parallel Network Environment

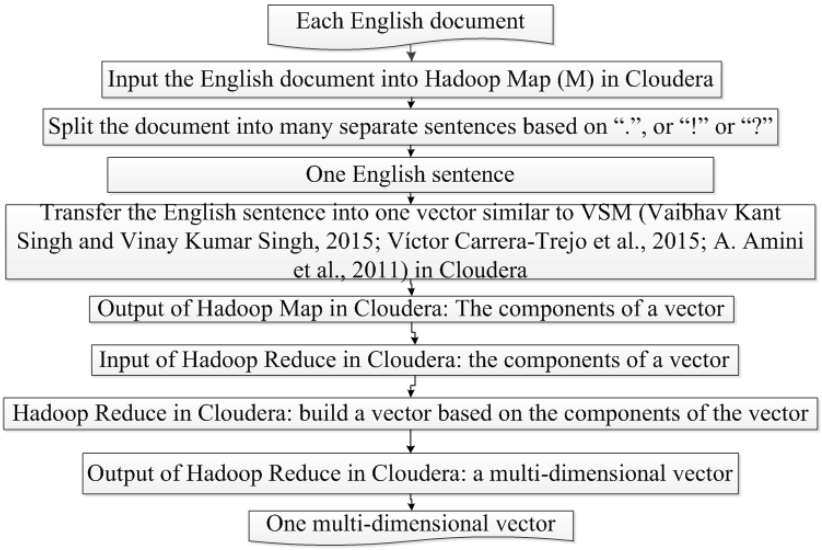

Fig. 7: Overview of transforming each English sentence into one vector in Cloudera

With the Reduce phase of the Cloudera, the input of the phase is the output of the Map phase of the Cloudera and this input is the clustering results of the multi-dimensional vector of the document to the positive vector group or the negative vector group or not; and the output of the phase is the sentiment classification result of the document into the positive polarity, the negative polarity, or the neutral polarity. In the Reduce phase, the document is classified as the positive emotion if the multi-dimensional vector is clustered into the positive vector group; the document is classified as the negative semantic if the multi-dimensional vector into the negative vector group; and the document is classified as the neutral sentiment if the multi-dimensional vector is not clustered into the positive vector group, or the negative vector group, or not.

Hadoop Map (M)

The template is used to format your paper and style the text. All margins, column widths, line spaces and text fonts are prescribed

We implement this phase in Fig. 8. The Statistical Information Grid Algorithm (STING) in Cloudera is based on the Statistical Information Grid Algorithm in (Wang *et al*., 1997; Pilevar and Sukumar, 2005; Lin *et al*., 2008; Halkidi *et al*., 2001a; Joshi and Kaur, 2013; Madhulatha, 2012; Halkidi *et al*., 2001b; Murtagh and Contreras, 2011; Qian and Zhou, 2002; Amini *et al*., 2011; Wang *et al*., 1999; Yue, 2017; Lin *et al*., 2007a; Zhang *et al*., 2010). The input is one multi-dimensional vector in the English testing data set, the positive vector group and the negative vector group of the English training data set. The output of the STING is the clustering results of the multi-dimensional vector into the positive vector group or the negative vector group, or not. The main ideas of the STING are as follows:

- Determine a layer to begin with
- For each cell of this layer, we calculate the confidence interval (or estimated range) of probability that this cell is relevant to the query
- From the interval calculated above, we label the cell as relevant or not relevant
- If this layer is the bottom layer, go to (6); otherwise, go to (5)
- We go down the hierarchy structure by one level. Go to (2) for those cells that form the relevant cells of the higher level layer
- If the specification of the query is met, go to (8); otherwise, go to (7)
- Retrieve those data fall into the relevant cells and do further processing
- Return the result that meets the requirement of the query. Go to (9)
- Find the regions of relevant cells. Return those regions that meet the requirement of the query. Go to (9)
- Stop

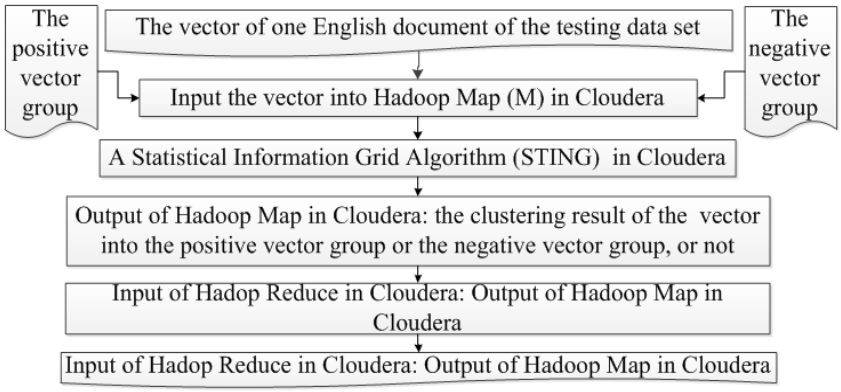

Fig. 8: Overview of the STING in Hadoop Map (M) in Cloudera

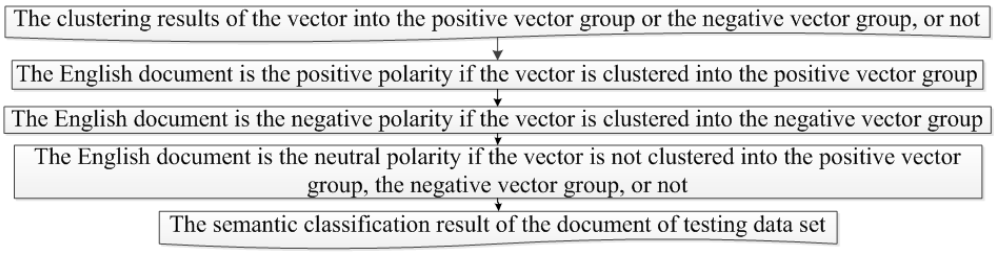

Fig. 9: Overview of Hadoop Reduce (R) in Cloudera

After finishing to cluster the multi-dimensional vector into the positive vector group, or the negative vector group, or not, Hadoop Map transfers this results into Hadoop Reduce in the Cloudera system.

Hadoop Reduce (R)

The template is used to format your paper and style the text. All margins, column widths, line spaces and text fonts are prescribed.

We perform this phase in Fig. 9 as follows: After receiving the clustering result of Hadoop Map, Hadoop Reduce labels the semantics polarity for the multi-dimensional vector which is classified. Then, the output of Hadoop Reduce will return the semantics polarity of one document (corresponding to the multi-dimensional vector) in the English testing data set. One document is the positive polarity if the multi-dimensional vector is clustered into the positive vector group. One document is the negative polarity if the multi-dimensional vector is clustered into the negative vector group. One document is the neutral polarity if the multi-dimensional vector is not clustered into both the positive vector group and the negative vector group.

Experiment

We have measured an Accuracy (A) to calculate the accuracy of the results of the emotion classification. A Java programming language is used for programming to save data sets, implementing our proposed model to classify the 1,000,000 documents of the testing data set. To implement the proposed model, we have already used the Java programming language to save the English training data set and 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 the STING. The configuration of the server in the sequential environment is: Intel® Server Board S1200V3RPS, Intel® Pentium® Processor G3220 (3M Cache, 3.00 GHz), 2GB PC3-10600 ECC 1333 MHz LP Unbuffered DIMMs. The operating system of the server is: Cloudera. We perform the STING 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 STING 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 PC3-10600 ECC 1333 MHz LP Unbuffered DIMMs. The operating system of each server in the 9 servers is: Cloudera. All 9 nodes have the same configuration information.

Results and Discussion

In this section, we show the results of this survey in the tables as follows: Table 1 to 3.

We show the results of the documents in the testing data set in Table 1.

The accuracy of the sentiment classification of the documents of the testing data set is presented in Table 2.

We display the average execution times of the classification of our novel model for the documents of the testing data set in Table 3.

In Table 1, we have had the 839,200 documents of the correct classification of the testing data set comprising the 420,559 negative documents and the 418,641 positive documents. We have also had the 160,800 documents of the incorrect classification of the testing data set including the 79,441 negative documents and the 81,359 positive documents in this table.

In Table 2, we had achieved 83.92% accuracy of the testing data set.

Table 1: The results of the 1,000,000 English documents in the testing data set

	Testing dataset	Correct classification	Incorrect classification
Negative	500,000	420,559	79,441
Positive	500,000	418,641	81,359
Summary	1,000,000	839,200	160,800

Table 2: The accuracy of our new model for the 1,000,000 English documents in the testing data set

Proposed Model	Class	Accuracy
Our novel model	Negative Positive	83.92%

Table 3: Average time of the classification of our new model for the 1,000,000 English documents in testing data set

	Average time of the classification/ 1,000,000 documents
The novel model in the sequential environment	5,214,291 sec
The novel model in the Cloudera distributed system with 3 nodes	1,738,317 sec
The novel model in the Cloudera distributed system with 6 nodes	849,399 sec
The novel model in the Cloudera distributed system with 9 nodes	596,565 sec

In Table 3, the average time of the semantic classification of the STING algorithm in the sequential environment is 5,214,291 sec/1,000,000 documents and it is greater than the average time of the emotion classification of the STING in the Cloudera parallel network environment with 3 nodes which is 1,738,317 sec/1,000,000 documents. The average time of the emotion classification of the STING in the Cloudera parallel network environment with 9 nodes, which is 596,565 sec/1,000,000 documents, is the shortest time. Besides, the average time of the emotion classification of the STING in the Cloudera parallel network environment with 6 nodes is 849,399 sec/1,000,000 documents.

Conclusion

Although our new model has been tested on our English data set, it can be applied to many other languages. In this study, our model has been tested on the 1,000,000 English documents of the testing data set in which the data sets are small. However, our model can be applied to larger data sets with millions of English documents in the shortest time.

Table 4: Comparisons of our model's results with the works in (Carrera-Trejo *et al*., 2015; Amini *et al*., 2011)

Studies	STING	CT	Sentiment classification	PNS	SD	DT	Language	VSM
Singh and Singh (2015)	No	No	No	No	Yes	No	EL	Yes
Carrera-Trejo *et al*. (2015)	No	No	Yes	No	Yes	No	EL	Yes
Amini *et al*. (2011)	No	No	Yes	No	Yes	Yes	EL	Yes
Our work	Yes	Yes	Yes	Yes	Yes	Yes	EL	Yes

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

Table 5: Comparisons of our model's advantages and disadvantages with the works in (Carrera-Trejo *et al*., 2015; Amini *et al*., 2011)

Researches	Approach	Advantages	Disadvantages
Singh and Singh (2015)	Examining the vector space model, an information retrieval technique and its variation	In this study, the authors have given an insider to the working of vector space model techniques used for efficient retrieval techniques. It is the bare fact that each system has its own strengths and weaknesses. What we have sorted out in the authors' work for vector space modeling is that the model is easy to understand and cheaper to implement, considering the fact that the system should be cost effective(i.e., should follow the space/time constraint. It is also very popular. Although the system has all these properties, it is facing some major drawbacks.	The drawbacks are that the system yields no theoretical findings. Weights associated with the vectors are very arbitrary and this system is an independent system, thus requiring separate attention. Though it is a promising technique, the current level of success of the vector space model techniques used for information retrieval are not able to satisfy user needs and need extensive attention.
Carrera-Trejo *et al*. (2015)	+Latent Dirichlet allocation (LDA). +Multi-label text classification tasks and apply various feature sets. +Several combinations of features, like bi-grams and uni-grams.	In this study, the authors consider multi-label text classification tasks and apply various feature sets. The authors consider a subset of multi- labeled files of the Reuters-21578 corpus. The authors use traditional TF-IDF values of the features and tried both considering and ignoring stop words. The authors also tried several combinations of features, like bi-grams and uni-grams. The authors also experimented with adding LDA results into vector space models as new features. These last experiments obtained the best results.	No mention
Amini *et al*. (2011)	The K-Nearest Neighbors algorithm for English sentiment classification in the Cloudera distributed system.	In this study, the authors introduce a new weighting method based on statistical estimation of the importance of a word for a specific categorization problem. One benefit of this method is that it can make feature selection implicit, since useless features of the categorization problem considered get a very small weight. Extensive experiments reported in the work show that this new weighting method improves significantly the classification accuracy as measured on many categorization tasks.	Despite positive results in some settings, GainRatio failed to show that supervised weighting methods are generally higher than unsupervised ones. The authors believe that ConfWeight is a promising supervised weighting technique that behaves gracefully both with and without feature selection. Therefore, the authors advocate its use in further experiments.
Our work	The K-Nearest Neighbors algorithm for English sentiment classification in the Cloudera distributed system. The advantages and disadvantages of the proposed model are shown in the Conclusion section.		

In this study, we have proposed a new model to classify sentiment of English documents using the Statistical Information Grid Algorithm (STING) with Hadoop Map (M)/Reduce (R) in the Cloudera parallel network environment. With our proposed new model, we have achieved 83.92% accuracy of the testing data set. Until now, not many studies have shown that the clustering methods can be used to classify data. Our research shows that clustering methods are used to classify data and, in particular, can be used to classify emotion in text.

The execution time of the STING 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 Statistical Information Grid Algorithm 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. It can be applied to other languages. Its negatives are as follows: It has a low rate of accuracy. It costs too much and takes too much time to implement this proposed model.

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

In Table 4 and 5 below, we compare our model's results with the studies in (Singh and Singh, 2015; Carrera-Trejo *et al*., 2015; Amini *et al*., 2011; Soucy and Mineau, 2015).

In Table 6 and 7 below, we compare our model's results with the works related to the Statistical Information Grid Algorithm (STING) in (Wang *et al*., 1997; Pilevar and Sukumar, 2005; Lin *et al*., 2008; Halkidi *et al*., 2001a; Joshi and Kaur, 2013; Madhulatha, 2012; Halkidi *et al*., 2001b; Murtagh and Contreras, 2011; Qian and Zhou, 2002; Amini *et al*., 2011; Wang *et al*., 1999; Yue, 2017; Lin *et al*., 2007a; Zhang *et al*., 2010).

In Table 8 and 9 below, we compare our model's results with the latest research on sentiment classification (or sentiment analysis or opinion mining) in (Agarwal and Mittal, 2016a; 2016b; Canuto *et al*., 2016; Ahmed and Danti, 2016; Phu and Tuoi, 2014).

Table 6: Comparisons of our model's results with the works related to the Statistical Information Grid Algorithm (STING) in (Wang *et al*., 1997; Pilevar and Sukumar, 2005; Lin *et al*., 2008; Halkidi *et al*., 2001a; Joshi and Kaur, 2013; Madhulatha, 2012; Halkidi *et al*., 2001b; Murtagh and Contreras, 2011; Qian and Zhou, 2002; Amini *et al*., 2011; Wang *et al*., 1999; Yue, 2017; Lin *et al*., 2007a; Zhang *et al*., 2010)

Surveys	STING	CT	Sentiment classification	PNS	SD	DT	Language	VSM
Cloudera (2017)	Yes	Yes	NM	NM	NM	NM	NM	NM
Wang *et al*. (1997)	Yes	Yes	NM	NM	NM	NM	NM	NM
Pilevar and Sukumar (2005)	Yes	Yes	NM	NM	NM	NM	NM	NM
Lin *et al*. (2008)	Yes	Yes	NM	NM	NM	NM	NM	NM
Halkidi *et al*. (2001a)	Yes	Yes	NM	NM	NM	NM	NM	NM
Joshi and Kaur (2013)	Yes	Yes	NM	NM	NM	NM	NM	NM
Madhulatha (2012)	Yes	Yes	NM	NM	NM	NM	NM	NM
Halkidi *et al*. (2001b)	Yes	Yes	NM	NM	NM	NM	NM	NM
Murtagh and Contreras (2011)	Yes	Yes	NM	NM	NM	NM	NM	NM
Qian and Zhou (2002)	Yes	Yes	NM	NM	NM	NM	NM	NM
Amini *et al*. (2011)	Yes	Yes	NM	NM	NM	NM	NM	NM
Wang *et al*. (1999)	Yes	Yes	NM	NM	NM	NM	NM	NM
Yue (2017)	Yes	Yes	NM	NM	NM	NM	NM	NM
Lin *et al*. (2007a)	Yes	Yes	NM	NM	NM	NM	NM	NM
Zhang *et al*. (2010)	Yes	Yes	NM	NM	NM	NM	NM	NM
Our work	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 7: Comparisons of our model's merits and demerits with the works related to the Statistical Information Grid Algorithm (STING) in (Wang *et al*., 1997; Pilevar and Sukumar, 2005; Lin *et al*., 2008; Halkidi *et al*., 2001a; Joshi and Kaur, 2013; Madhulatha, 2012; Halkidi *et al*., 2001b; Murtagh and Contreras, 2011; Qian and Zhou, 2002; Amini *et al*., 2011; Wang *et al*., 1999; Yue, 2017; Lin *et al*., 2007a; Zhang *et al*., 2010)

Works	Approach	Merits	Demerits
Cloudera (2017)	STING: A Statistical Information Grid Approach to Spatial Data Mining	In this survey, the authors propose a hierarchical statistical information grid based approach for spatial data mining to reduce the cost further. The idea is to capture statistical Information associated with spatial cells in such a manner that whole classes of queries and clustering problems can be answered without recourse to the individual objects. In theory and confirmed by empirical studies, this approach outperforms the best previous method by at least an order of magnitude, especially when the data set is very large.	No mention

Table 7: Continue

Wang *et al*. (1997)	GCHL: A grid-clustering algorithm for high-dimensional very large spatial data bases	The algorithm work as well in the feature space of any data set. The method operates on a limited memory buffer and requires at most a single scan through the data. The authors demonstrate the high quality of the obtained clustering solutions, capability of discovering concave/ deeper and convex/higher regions, their robustness to outlier and noise and GCHL excellent scalability.	No mention
Pilevar and Sukumar (2005)	A Deflected Grid-based Algorithm for Clustering Analysis	The main idea of DGD algorithm is to deflect the original grid structure in each dimension of the data space after the clusters generated from this original structure have been obtained. The deflected grid structure can be considered a dynamic adjustment of the size of the original cells and thus, the clusters generated from this deflected grid structure can be used to revise the originally obtained clusters. The experimental results verify that, indeed, the effect of DGD algorithm is less influenced by the size of the cells than other grid-based ones.	No mention
Lin *et al*. (2008)	On Clustering Validation Techniques	This paper introduces the fundamental concepts of clustering while it surveys the widely known clustering algorithms in a comparative way. Moreover, it addresses an important issue of clustering process regarding the quality assessment of the clustering results. This is also related to the inherent features of the data set under concern. A review of clustering validity measures and approaches available in the literature is presented. Furthermore, the paper illustrates the issues that are under-addressed by the recent algorithms and gives the trends in clustering process.	No mention
Halkidi *et al*. (2001a)	A Review: Comparative Study of Various Clustering Techniques in Data Mining	Many c lustering algorithms have been developed and are categorized from several aspects such as partitioning methods, hierarchical methods, density based methods and grid-based methods. Further data set can be numeric or categorical. Inherent geometric properties of numeric data can be exploited to naturally define distance function between data points. Whereas categorical data can be derived from either quantitative or qualitative data where observations are directly observed from counts.	No mention
Joshi and Kaur (2013)	An Overview on Clustering Methods	The survey provides an overview of the different representative clustering methods. In addition, several clustering validations indices are shown. Furthermore, approaches to automatically determine the number of clusters are presented. Finally, application of different heuristic approaches to the clustering problem is also investigated.	No mention
Madhulatha (2012)	Clustering algorithms and validity measures	The authors surveys clustering methods and approaches available in the literature in a comparative way. It also presents the basic concepts, principles and assumptions upon which the clustering algorithms are based. Another important issue is the validity of the clustering schemes resulting from applying algorithms. This is also related to the inherent features of the data set under concern. The authors review and compare clustering validity measures available in the literature. Furthermore, the authors illustrate the issues that are under-addressed by the recent algorithms and we address new research directions	No mention
Halkidi *et al*. (2001b)	Algorithms for hierarchical clustering: an overview	The authors look at hierarchical self-organizing maps and mixture models. The authors review grid-based clustering, focusing on hierarchical density-based approaches. Finally, the authors describe a recently developed very efficient(linear time) hierarchical clustering algorithm, which can also be viewed as a hierarchical grid-based algorithm	No mention
Murtagh and Contreras (2011)	Analyzing PopularClustering Algorithms from Different Viewpoints	Different clustering methods use different similarity definition and techniques. Several popular clustering algorithms are analyzed from three different viewpoints: (1) clustering criteria, (2) cluster representation and (3) algorithm framework. Furthermore, some new built algorithms, which mix or generalize some other algorithms, are introduced. Since the analysis is from several viewpoints, it can cover and distinguish most of the existing algorithms. It is the basis of the research of self-tuning algorithm and clustering benchmark.	No mention
Qian and Zhou (2002)	A study of density-grid based clustering algorithms on data streams	In this survey, the authors review the grid based clustering algorithms that use density-based algorithms or density concept for the clustering. The authors called them density-grid clustering algorithms. The authors explore the algorithms in details and the merits and limitations of them. The algorithms are also summarized in a table based on the important features.Besides that, the authors discuss about how well the algorithms address the challenging issues in the clustering data streams.	No mention

Table 7: Continue

Amini *et al*. (2011)	STING+: An approach to active spatial data mining	The authors employ a hierarchical structure with associated statistical information at the various levels of the hierarchy and decompose the user-defined trigger into a set of sub- triggers associated with cells in the hierarchy. Updates are suspended in the hierarchy until their cumulative effect might cause the trigger to fire. It is shown that this approach achieves three orders of magnitude improvement over the naïve approach that re-evaluates the condition over the database for each update, while both approaches produce the same result without any delay. Moreover this scheme can support incremental query processing as well.	No mention
Wang *et al*. (1999)	Research on the clustering analysis algorithm for data mining	This research analyzes the clustering with the focus on the data mining methodologies of the listed to capture the systematic combination. (1) Discriminant analysis, to establish one or more of the discriminant function and to determine the discriminant criteria. (2) Neural network method. At present, in data mining, it is the most commonly used neural network BP and RBF network. (3) Correlation analysis and the regression analysis, correlation analysis is made of the relevance of the correlation coefficient between measured variables. Besides this, the authors propose the new clustering algorithm with the integration of the related methods and the experiment result shows the effectiveness of the method. The method outperforms compared with the other state-of-the-art methodologies	No mention
Yue (2017)	An Adaptive Crossover-Imaged Clustering Algorithm	The experimental results verify that, indeed, the effect of ACICA algorithm is less influenced by the size of the cells than other grid-based algorithms. Finally, the authors will verify by experiment that the results of ourproposed ACICA algorithm outperforms than others.	No mention
Lin *et al*. (2007a)	Analyze the Wild Birds' Migration Tracks by MPI-Based Parallel Clustering Algorithm	In this study, parallel STING (statistical information grid) algorithm is designed and implemented based on Message Passing Interface (MPI) for spatial clustering. By using parallel STING algorithm, it only takes several seconds to get the result.	No mention
Zhang *et al*. (2010)	An Adaptable Deflect and Conquer Clustering Algorithm	The idea of ADCC is to utilize the predefined grids and predefined threshold to identify the significant cells, by which nearby cells that are also significant can be merged to develop a cluster in the first place. Next, the modified grids which are deflected to half size of the grid are used to identify the significant cells again. Finally, the new generated significant cells and the initial significant cells are merged so as to offset the round-off error and improve the precision of clustering task and the authors verify by experiment that the performance of our new grid-based clustering algorithm, ADCC, is good.	No mention
Our work	The Statistical Information Grid Algorithm (STING) for English sentiment classification in the Cloudera distributed system. Our research's merits and demerits are shown in the Conclusion section.		

Table 8: Comparisons of our model with the latest sentiment classification models (or the latest sentiment classification methods) in (Agarwal and Mittal, 2016a; 2016b; Canuto *et al*., 2016; Ahmed and Danti, 2016; Phu and Tuoi, 2014)

Studies	STING	CT	Sentiment classification	PNS	SD	DT	Language	VSM
Lin *et al*. (2007b)	No	No	Yes	NM	Yes	Yes	Yes	vector
Agarwal and Mittal (2016a)	No	No	Yes	NM	Yes	Yes	NM	NM
Agarwal and Mittal (2016b)	No	No	Yes	NM	Yes	Yes	EL	NM
Canuto *et al*. (2016)	No	No	Yes	NM	Yes	Yes	NM	NM
Ahmed and Danti (2016)	No	No	Yes	No	No	No	EL	No
Phu and Tuoi (2014)	No	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 latest sentiment classification models (or the latest sentiment classification methods) in (Agarwal and Mittal, 2016a; 2016b; Canuto *et al*., 2016; Ahmed and Danti, 2016; Phu and Tuoi, 2014)

Studies	Approach	Positives	Negatives
Lin *et al*. (2007b)	The Machine Learning Approaches Applied to Sentiment Analysis-Based Applications	The main emphasis of this survey is to discuss the research involved in applying machine learning methods, mostly for sentiment classification at document level. Machine learning-based approaches work in the following phases, which are discussed in detail in this study for sentiment classification: (1) feature extraction, (2) feature weighting schemes, (3) feature selection and (4) machine-learning methods. This study also discusses the standard free benchmark datasets and evaluation methods for sentiment analysis. The authors conclude the research with a comparative study of some state-of-the-art methods for sentiment analysis and some possible future research directions in opinion mining and sentiment analysis.	No mention
Agarwal and Mittal (2016a)	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
Agarwal and Mittal (2016b)	Exploiting New Sentiment-Based Meta-Level Features for Effective Sentiment Analysis	Experiments performed with a substantial number of datasets (nineteen) demonstrate that the effectiveness of the proposed sentiment-based meta-level features is not only superior to the traditional bag-of-words representation (by up to 16%) but also is also superior in most cases to state-of-art meta-level features previously proposed in the literature for text classification tasks that do not take into account any idiosyncrasies of sentiment analysis. The authors' proposal is also largely superior to the best lexicon-based methods as well as to supervised combinations of them. In fact, the proposed approach is the only one to produce the best results in all tested datasets in all scenarios.	A line of future research would be to explore the authors' meta features with other classification algorithms and feature selection techniques in different sentiment analysis tasks such as scoring movies or products according to their related reviews.
Canuto *et al*. (2016)	Rule-Based Machine Learning Algorithms	The proposed approach is tested by experimenting with online books and political reviews and demonstrates the efficacy through Kappa measures, which have a higher accuracy of 97.4% and a lower error rate. The weighted average of different accuracy measures like Precision, Recall and TP-Rate depicts higher efficiency rate and lower FP-Rate. Comparative experiments on various rule-based machine learning algorithms have been performed through a ten-fold cross validation training model for sentiment classification.	No mention

Table 9: Continue

Ahmed and Danti (2016)	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
Phu and Tuoi (2014)	Naive Bayes Model with N-GRAM Method, Negation Handling Method,Chi-Square Method and Good-Turing Discounting, etc.	The authors have explored the Naïve Bayes model with N-GRAM method, Negation Handling method, Chi-Square method and Good-Turing Discounting by selecting different thresholds of Good-Turing Discounting method and different minimum frequencies of Chi-Square method to improve the accuracy of sentiment classification.	No Mention
Our work	The Statistical Information Grid Algorithm for English sentiment classification in the Cloudera distributed system. The positives and negatives of the proposed model are given in the Conclusion section.		

Author's Contributions

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

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

Conflict of Interest

The authors declare that they have no conflict of interest.

References

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

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

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

Amini, A. T.Y. Wah, M.R. Saybani and S.R.A.S. Yazdi, 2011. A study of density-grid based clustering algorithms on data streams. Proceedings of the 8th International Conference on Fuzzy Systems and Knowledge Discovery, Jul. 26-28, IEEE Xplore Press, Shanghai, China, pp: 1652-1656. DOI: 10.1109/FSKD.2011.6019867

Apache, 2017. http://apache.org

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

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

Cloudera, 2017. http://www.cloudera.com

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

Hadoop, 2017. http://hadoop.apache.org

Halkidi, M., Y. Batistakis and M. Vazirgiannis, 2001a. On clustering validation techniques. J. Intell. Inform. Syst., 17: 107-145. DOI: 10.1023/A:1012801612483

Halkidi, M., Y. Batistakis and M. Vazirgiannis, 2001b. Clustering algorithms and validity measures. Proceedings of the 13th International Conference on Scientific and Statistical Database Management, Jul. 18-20, IEEE Xplore Press, Fairfax, VA, USA, pp: 3-22. DOI: 10.1109/SSDM.2001.938534

Joshi, A. and R. Kaur, 2013. A review: Comparative study of various clustering techniques in data mining. Int. J. Adv. Res. Comput. Sci. Software Eng., 3: 55-57.

Lin, N.P., C.I. Chang and C.L. Pan, 2007a. An adaptable deflect and conquer clustering algorithm. Proceedings of the 6th WSEAS International Conference on Applied Computer Science, Apr. 15-17, World Scientific and Engineering Academy and Society (WSEAS) Hangzhou, China, pp: 155-159.

Lin, N.P., C.I. Chang, H.E. Chueh, H.J. Chen and W.H. Hao, 2007b. An adaptive crossover-imaged clustering algorithm. Proceedings of the 7th WSEAS International Conference on Simulation, Modelling and Optimization, (SMO' 07), Beijing, China.

Lin, N.P., C.I. Chang, H.E. Chueh, H.J. Chen and W.H. Hao, 2008. A deflected grid-based algorithm for clustering analysis. WSEAS Trans. Comput., 7: 125-132.

Madhulatha, T.S., 2012. An overview on clustering methods. IOSR J. Eng., 2: 719-725.

Murtagh, F. and P. Contreras, 2011. Algorithms for hierarchical clustering: An overview. WIRES Data Min. Knowl. Discovery, 2: 86-97. DOI: 10.1002/widm.53

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

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

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

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

Phu, V.N., V.T.N. Chau, V.T.N. Tran, N.D. Dat and T.A. Nguyen. 2017c. STING algorithm used english sentiment classification in a parallel environment. Int. J. Patt. Recognit. Artificial Intell. DOI: 10.1142/S0218001417500215

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

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

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

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

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

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

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

Pilevar, A.H. and M. Sukumar, 2005. GCHL: A grid-clustering algorithm for high-dimensional very large spatial data bases. Patt. Recognit. Lett., 26: 999-1010. DOI: 10.1016/j.patrec.2004.09.052

Qian, W.N. and A.Y. Zhou, 2002. Analyzing popular clustering algorithms from different viewpoints. J. Software, 13: 1383-1394.

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

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

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

Wang, W., J. Yang and R. Muntz, 1997. STING: A statistical information grid approach to spatial data mining. Proceedings of the 23rd International Conference on Very Large Data Bases, Aug. 25-29, Morgan Kaufmann Publishers Inc., Athens, Greece, pp: 186-195.

Wang, W., J. Yang and R. Muntz, 1999. STING+: An approach to active spatial data mining. Proceedings of the 15th International Conference on Data Engineering, Mar. 23-26, IEEE Xplore Press, Sydney, NSW, Australia, pp: 116-125. DOI: 10.1109/ICDE.1999.754914

Yue, W., 2017. Research on the clustering analysis algorithm for data mining. RISTI [Revista Iberica de Sistemas e Tecnologias de Informacao].

Zhang, H.M., Y.C. Zhou, J.H. Li, X.Z. Wang and B.P. Yan, 2010. Analyze the wild birds' migration tracks by MPI-based parallel clustering algorithm. Proceedings of the 6th international conference on Advanced data mining and applications: Part I, Nov. 19-21, Springer-Verlag, Chongqing, China, pp: 383-393.

Appendix

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

Input: one English document
Output: one multi-dimensional vector
Begin
Step 0: Set LengthOfMultiDimensionalVector := 0;
Step 1: Set MultiDimensionalVector := {}{} with n_max rows and m_max columns.
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 vector (one dimensiona) based on VSM (Singh and Singh, 2015; Carrera-Trejo *et al*., 2015; Amini *et al*., 2011; Soucy and Mineau, 2015)
Step 5: If OneDimensionalVector.length is less than m_max Then
Step 6: For j = OneDimensionalVector.length; j< m_max; j++; do:
Step 7: OneDimensionalVector[j] := 0;
Step 8: End For;
Step 9: End If;
Step 10: MultiDimensionalVector.AddOneDimensionalVector (OneDimensionalVector);
Step 11: LengthOfMultiDimensionalVector = LengthOfMultiDimensionalVector + 1;
Step 12: End For;
Step 13: If LengthOfMultiDimensionalVector is less than n_max Then
Step 14: For k = LengthOfMultiDimensionalVector; k < n_max; k++; do:
Step 15: MultiDimensionalVector.AddOneDimensionalVector (zero vector – one dimension);
Step 16: End For;
Step 17: Return MultiDimensionalVector;
End;

Algorithm 2: Creating the Positive Vector Group

Input: the 1,000,000 positive English documents of the English training data set.
Output: the positive vector group PositiveVectorGroup
Begin
Step 0: Set PositiveVectorGroup := {};
Step 1: For i = 0; i < 1,000,000; i++; do:
Step 2: Set OneMultiDimensionalVector := Call Algorithm 1 with the positive English document i in the English training data set;
Step 3: PositiveVectorGroup.AddMultiDimensionalVector(OneMultiDimensionalVector);
Step 4: End For;
Step 5: Return PositiveVectorGroup;
End;

Algorithm 3: Creating the Negative Vector Group

Input: the 1,000,000 negative English documents of the English training data set.
Output: the negative vector group NegativeVectorGroup
Begin
Step 0: Set NegativeVectorGroup := {};
Step 1: For i = 0; i < 1,000,000; i++; do:
Step 2: Set OneMultiDimensionalVector := Call Algorithm 1 with the negative English document i in the English training data set;
Step 3: NegativeVectorGroup.AddMultiDimensionalVector(OneMultiDimensionalVector);
Step 4: End For;
Step 5: Return NegativeVectorGroup;
End;

Algorithm 4: Clustering one Multi-Dimensional Vector (Corresponding to One English Document of the English Testing Data Set) into the Positive Vector Group PositiveVectorGroup, the Negative Vector Group NegativeVectorGroup, or Not

Input: one multi-dimensional vector A (corresponding to one English document of the English testing data set), the positive vector group PositiveVectorGroup, the negative vector group NegativeVectorGroup;
Output: positive, negative, neutral;
Begin
Step 1: Implement the Statistical Information Grid Algorithm (STING) based on the Statistical Information Grid Algorithm in (Wang *et al*. 1997; Pilevar and Sukumar, 2005; Lin *et al*., 2008; Halkidi *et al*., 2001a; Joshi and Kaur, 2013; Madhulatha, 2012; Halkidi *et al*., 2001b; Murtagh and Contreras, 2011; Qian and Zhou, 2002; Amini *et al*., 2011; Wang *et al*., 1999; Yue, 2017; Lin *et al*., 2007a; Zhang *et al*., 2010) with input is one multi-dimensional vector (corresponding to one English document of the English testing data set), the positive vector group PositiveVectorGroup, the negative vector group NegativeVectorGroup;
Step 2: With the results of Step 1, If the vector is clustered into the positive vector group Then
Step 3: Return positive;
Step 4: Else If the vector is clustered into the negative vector group Then
Step 5: Return negative;
Step 6: Else
Step 7: Return neutral;
Step 8: End If;
Step 9: Return neutral;
End;