Journal of Computer Science

Original Research Paper

A CURE Algorithm for Vietnamese Sentiment Classification in a Parallel Environment

[1]Vo Ngoc Phu, [2]Vo Thi Ngoc Tran and [3]Jack Max

[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
[3]Department of Computer Science, Sumatra Univesity, Thailand

Article history
Received: 29-01-2018
Revised: 11-03-2018
Accepted: 19-04-2018

Corresponding Author:
Jack Max
Department of Computer Science, Sumatra Univesity, Thailand
Email: jack.max012018@gmail.com
Vo Ngoc Phu
Nguyen Tat Thanh University, 300A Nguyen Tat Thanh Street,
Ward 13, District 4, Ho Chi Minh City, 702000, Vietnam
Email: vongocphu03hca@gmail.com;vongocphu@ntt.edu.vn

Abstract: Solutions to process big data are imperative and beneficial for numerous fields of research and commercial applications. Thus, a new model has been proposed in this paper to be used for big data set sentiment classification in the Cloudera parallel network environment. Clustering Using Representatives (CURE), combined with Hadoop MAP (M)/REDUCE (R) in Cloudera – a parallel network system, was used for 20,000 documents in a Vietnamese testing data set. The testing data set included 10,000 positive Vietnamese documents and 10,000 negative ones. After testing our new model on the data set, a 62.92% accuracy rate of sentiment classification was achieved. Although our data set is small, this proposed model is able to process millions of Vietnamese documents, in addition to data in other languages, to shorten the execution time in the distributed environment.

Keywords: Sentiment Classification, Vietnamese Sentiment Classification, Vietnamese Sentence Sentiment Classification, Opinion Mining, Vietnamese Opinion Mining, Vietnamese Document Opinion Mining, Clustering Using Representatives, Cure, Cloudera, Parallel Environment, Parallel Network, Parallel Network Environment

Introduction

Solutions to process big data are imperative and beneficial for numerous fields of research and applications. Clustering can be considered the most significant unsupervised learning problem; similar to other problems of this kind, it deals with findings in a collection of unlabeled data. The clustering method refers to objects that are similar being clustered into the same group, whereas objects that are dissimilar will not be clustered into the same group (or the same cluster), as they will be in different groups (or different clusters) instead. A cluster only includes objects that share similar characteristics.

Clustering Using Representatives (CURE), which proposed the CURE algorithm – a hierarchical clustering algorithm (Guha *et al*., 1998), is an efficient data clustering algorithm for large databases. Therefore, the objective of this survey is to process numerous Vietnamese big data sets by using the CURE algorithm in the Cloudera distributed environment. The results of this study can be used to cross check sentiment classification for various fields of research and commercial applications.

In this work, each Vietnamese sentence was first transferred into a vector. The training data set had 40,000 Vietnamese sentences, which corresponded to 40,000 vectors that were divided into two groups: a positive vector group with 20,000 positive vectors, and a negative vector group with 20,000 negative vectors. In addition, we also transferred every sentence per document into our testing data set. A document consisted of n vectors if the document had n sentences. Therefore, if a document in the testing data set had n sentences, the document had a set of vectors, including n vectors. With 20,000 documents in the testing data set, we had 20,000 sets of vectors that corresponded to 20,000 documents.

Our new model has been proposed to classify the semantics (positive, negative, neutral) of each document in our testing data set as follows: in the Cloudera (2017) parallel network environment (Hadoop, 2017; Apache, 2017), Hadoop Map (M)/Reduce (R), we used the CURE algorithm to cluster each vector into the positive vector group or the negative vector group. Then, numerous vectors of each document were placed either in the positive vector group or the negative vector group: Documents were clustered into positive polarity if they

had more vectors in the positive vector group than in the negative vector group. Conversely, documents were clustered into negative polarity if they had more vectors in the positive vector group than in the negative vector group. Lastly, the remaining documents were clustered into neutral polarity if they had an equal number of vectors in both vector groups.

Our model is quite different from other studies related to Vietnamese vectors, Vietnamese segments, etc. (Hoang *et al*., 2007; Le *et al*., 2008; Nguyen *et al*., 2009), research related to the CURE algorithm (Guha *et al*., 1998; Yan-Hua *et al*., 2011; Nian-yun *et al*., 2009; Ertöz *et al*., 2002; Kaya and Alhajj, 2005; Ying *et al*., 2008; Rani *et al*., 2014), recent studies about Vietnamese sentiment classification (Ha *et al*., 2011; Bang *et al*., 2015; Kieu and Pham, 2010; Vu and Park, 2014; Nguyen *et al*., 2014; Hoanh-Su *et al*.; 2015; Anh and Dau, 2014; Phan and Cao, 2014; Nguyen *et al*., 2014; Duyen *et al*., 2014; Bach *et al*., 2015; Trinh *et al*., 2016), sentiment classification (Manek *et al*., 2016; Agarwal and Mittal, 2016a; 2016b; Canuto *et al*., 2016; Ahmed and Danti, 2016; Phu and Tuoi, 2014; Tran *et al*., 2014; Phu *et al*., 2016; Phu *et al*., 2017a; 2017b) and unsupervised classification (Turney, 2002; Lee *et al*., 2002a; van Zyl, 2002; Hegarat-Mascle *et al*., 2002; Ferro-Famil *et al*., 2002; Chaovalit and Zhou, 2005; Lee *et al*., 2002b; Gllavata *et al*., 2004).

The position of this study is used in a field of the Vietnamese sentiment classification for various Vietnamese surveys and commercial applications. The proposed model can also be applied to other languages.

The motivation of this new model is as follows: The emotional analysis of a document can be identified through its many sentences. Therefore, numerous algorithms in the data mining field can be applied to natural language processing and to semantic classification to process millions of documents.

The novelty of the proposed approach is as follows: A CURE algorithm in the data mining field is applied to sentiment analysis and to classify semantics of documents based on Vietnamese sentences. This algorithm has also been used to process and identify emotions for millions of English documents. The above principles are proposed to classify the semantics of Vietnamese documents, as data mining is used in natural language processing. Therefore, this research demonstrates that the proposed model can be successfully applied to numerous languages.

Our model has various significant contributions for countless research fields and commercial applications as follows:

1) The algorithm of data mining is applied to the semantic analysis of natural language processing
2) This study demonstrates that distinct fields of scientific research are interconnected
3) Emotional analysis has been successfully processed on millions of Vietnamese documents
4) Many studies and commercial applications can utilize the results of this survey
5) Semantic classification is implemented in the parallel network environment
6) Novel principles and algorithms are proposed in the research
7) The opinion classification of Vietnamese documents is performed on Vietnamese sentences
8) The Cloudera distributed system, Hadoop Map and Hadoop Reduce, is used in the proposed model
9) This model can be applied to various parallel network systems
10) This survey can be applied to parallel functions, such as Hadoop Map and Hadoop Reduce
11) The proposed model can be applied to several languages

This study contains 6 sections: Section 1 is the introduction section. Section 2 discusses related works about the Clustering Using Representatives (CURE). Section 3 concerns the Vietnamese data set to classify sentences. Section 4 represents the methodology of our proposed model. Section 5 represents the experimental model and results in this study. The conclusion of the proposed model is stated in section 6. In addition, the references section shows all the reference documents, and the tables are all shown in the appendices section.

Related Work

In this section, we display various summaries of the surveys related to our proposed model.

There are many studies that are related to Vietnamese vectors, Vietnamese segments, etc. (Hoang *et al*., 2007; Le *et al*., 2008; Nguyen *et al*., 2009). In a comparative study directed towards Vietnamese text classification methods, a modified version of the FCM algorithm was presented to address clusters with totally different geometrical properties. The proposed algorithm adopted a novel non-metric distance measure based on the idea of "point symmetry," and experimental results on several data sets were presented to illustrate its effectiveness. In a "hybrid approach to word segmentation of Vietnamese texts," the Bag Of Words (BOW) and Statistical N-Gram Language Modeling approaches achieved a high level of accuracy (Le *et al*., 2008). Additionally, the authors analyzed the advantages and disadvantages of each approach to find out the best method for specific circumstances. In a hybrid approach to automatically tokenize Vietnamese text, finite-state automata techniques, regular expression parsing, and the maximal-matching strategy were combined, which were augmented by statistical methods to resolve ambiguities of segmentation (Nguyen *et al*., 2009). The Vietnamese

lexicon in use was compactly represented by minimal finite-state automaton. A text to be tokenized is first parsed into lexical phrases and other patterns using pre-defined regular expressions. The automaton is then deployed to build linear graphs corresponding to the phrases to be segmented. Ultimately, the application of a maximum-matching strategy on a graph results in all candidate segmentations of a phrase.

Much research has been conducted related to implementing algorithms and applications in a parallel network environment (Hadoop, 2017; Apache, 2017; Cloudera, 2017). Hadoop is an Apache-based framework used to handle large data sets on clusters consisting of multiple computers and the Map and Reduce programming models (Hadoop, 2017; Apache, 2017). Its two main projects are the Hadoop Distributed File System (HDFS) and Hadoop M/R (Hadoop Map/Reduce). Hadoop M/R allows engineers to program writing applications for the parallel processing of large data sets of 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, as multiple Map tasks can handle a separate data partition in parallel. Then, the outputs of the map tasks are gathered and processed by the ordered Reduce tasks. The input and output of each M/R are stored in HDFS; as Map and Reduce tasks perform on the pair expressed as (key, value), the formatted input and output formats will also be expressed as (key, value). Cloudera, the global provider of the fastest, easiest, and most secure data management and analytics platform, which built upon the Apache™ Hadoop® and the latest open source technologies, announced that it will submit proposals for Impala and Kudu to join the Apache Software Foundation (ASF) (Cloudera, 2017). 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. The world's leading organizations trust Cloudera to help solve their most challenging business problems with Cloudera Enterprise. 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 costs.

Much research has been executed that is related to the CURE algorithm (Guha *et al*., 1998; Yan-Hua *et al*., 2011; Nian-Yun *et al*., 2009; Ertöz *et al*., 2002; Kaya and Alhajj, 2005; Ying *et al*., 2008; Rani *et al*., 2014). In data mining, clustering (Guha *et al*., 1998) has been useful for discovering groups and identifying interesting distributions in the underlying data. As traditional clustering algorithms either favor clusters with spherical shapes and similar sizes, or are very fragile in the presence of outliers, a new clustering algorithm called CURE was proposed. The CURE algorithm has also been used to analyze the behavior of users in large storage network databases (Yan-Hua *et al*., 2011). Experimental results showed that the improved algorithm is not only able to cluster, but also can distinguish between normal and abnormal behaviors; as the data was analyzed by the harmful behavior evaluation system, most of the abnormal behaviors observed were categorized under harmful behaviors (Yan-Hua *et al*., 2011). For increment data on real networks, the increment mining method was utilized, which is in accordance with the needs of real time network analysis. In another study, to inspect duplicated records, a new method of choosing representative records for a cluster was proposed, based on distance infection weight (Nian-Yun *et al*., 2009). Researchers have also proposed definitions of density and similarity that work well for high dimensional data (Ertöz *et al*., 2002), as well as an automated method for mining fuzzy association rules (Kaya and Alhajj, 2005), etc.

For our data sets, we compared our results with those from numerous studies (Hoang *et al*., 2007; Le *et al*., 2008; Nguyen *et al*., 2009; Hadoop, 2017; Apache, 2017; Cloudera, 2017; Guha *et al*., 1998; Yan-Hua *et al*., 2011; Nian-Yun *et al*., 2009; Ertöz *et al*., 2002; Kaya and Alhajj, 2005; Ying *et al*., 2008; Rani *et al*., 2014; Ha *et al*., 2011; Bang *et al*., 2015; Kieu and Pham, 2010; Vu and Park, 2014; Nguyen *et al*., 2014; Hoanh-Su *et al*.; 2015; Anh and Dau, 2014; Phan and Cao, 2014; Nguyen *et al*., 2014; Duyen *et al*., 2014; Bach *et al*., 2015; Trinh *et al*., 2016; Manek *et al*., 2016; Agarwal and Mittal, 2016a; 2016b; Canuto *et al*., 2016; Ahmed and Danti, 2016; Phu and Tuoi, 2014; Tran *et al*., 2014; Turney, 2002; Lee *et al*., 2002a; van Zyl, 2002; Hegarat-Mascle *et al*., 2002; Ferro-Famil *et al*., 2002; Chaovalit and Zhou, 2005; Lee *et al*., 2002b; Gllavata *et al*., 2004; Phu *et al*., 2016; Phu *et al*., 2017a; 2017b; 2017c; 2017d; 2017e; 2017f; 2017g; 2017h; 2017i).

The sentiment analysis task classifies a sentence into one of the following predefined categories: positive, negative, or neutral. In order to analyze the sentiment, three different text categorization algorithms are often compared, including Decision Tree, Naive Bayes (NB) and Support Vector Machines (SVM).

Much recent research has addressed sentiment classification (Manek *et al*., 2016; Agarwal and Mittal, 2016a; 2016b; Canuto *et al*., 2016; Ahmed and Danti, 2016; Phu and Tuoi, 2014; Tran *et al*., 2014; Phu *et al*., 2016; Phu *et al*., 2017a; 2017b) such as a Gini-Index-based feature selection method with an SVM classifier for a large movie review data set (Manek *et al*., 2016), as well as a corpus-based semantic orientation approach for sentiment analysis (Agarwal and Mittal, 2016b). Furthermore, several studies have recently investigated Vietnamese sentiment classification (Ha *et al*., 2011; Bang *et al*., 2015; Kieu and Pham, 2010; Vu and Park, 2014; Nguyen *et al*., 2014; Hoanh-Su *et al*., 2015; Anh and Dau, 2014; Phan and Cao, 2014; Nguyen *et al*., 2014; Duyen *et al*., 2014; Bach *et al*., 2015; Trinh *et al*., 2016). Researchers have proposed various methods to

address such data sets, including an upgrading FOMS model on Vietnamese reviews on mobile phone products (Ha *et al*., 2011) and an improved technique to analyze sentiment for Vietnamese texts based on the term feature selection approach (Bang *et al*., 2015).

Unsupervised classification has been investigated through numerous means (Turney, 2002; Lee *et al*., 2002a; van Zyl, 2002; Hegarat-Mascle *et al*., 2002; Ferro-Famil *et al*., 2002; Chaovalit and Zhou, 2005; Lee *et al*., 2002b; Gllavata *et al*., 2004; Phu *et al*., 2016; Phu *et al*., 2017a; 2017b; 2017c; 2017d; 2017e; 2017f; 2017g; 2017h), such as a simple unsupervised learning algorithm to classify reviews as recommended or not recommended, which classified reviews according to the average semantic orientation of the phrases, specifically the adjectives and adverbs, used in the reviews (Hegarat-Mascle *et al*., 2002). In another study , the use of an imaging radar polarimeter data for unsupervised classification of scattering behavior was described by comparing the polarization properties of each pixel in an image to that of simple classes of scattering, such as even number of reflections, odd number of reflections, and diffuse scattering (van Zyl, 2002).

Data Set

Our new model was tested on our Vietnamese data set, which includes testing and training data sets.

Figure 1, the testing data set includes 20,000 Vietnamese documents, which contains 10,000 positive documents and 10,000 negative documents. All sentences and documents in our data set were automatically extracted from Vietnamese Facebook, websites, and social networks. Then, we labeled each sentence and document as either positive or negative.

Figure 2, the training data set includes 40,000 Vietnamese sentences, including 20,000 positive sentences and 20,000 negative sentences. All sentences and documents in our data set were automatically extracted from Vietnamese Facebook, websites, and social networks. Then, we labeled both the sentences and the documents as either positive or negative.

Methodology

In this section, we present how our new model is implemented in the Cloudera parallel network environment. The section has two main parts: the first part demonstrates how a Vietnamese sentence is transferred into a vector; and the second part displays how the CURE algorithm (CA) is performed. The second part includes two sub-sections: in the first sub-section, a document of our Vietnamese testing data set is classified into positive or negative polarity by using the CURE algorithm in the sequential environment; in the second sub-section, a document of our Vietnamese testing data set is classified into positive or negative polarity by using CA in the parallel network environment.

The methodology was executed as shown in Fig. 3.

The main ideas of the proposed model are as follows:

Step 1: Transfer all the Vietnamese sentences of the training data set into the vectors of the positive vector group and the negative vector group.
Step 2: Split each Vietnamese document of the testing data set into the Vietnamese sentences. Each Vietnamese sentence of this Vietnamese document is transferred into one vector.
Step 3: Use the CURE algorithm to cluster each vector of each Vietnamese document of the testing data set into the positive vector group or the negative vector group of the training data set.
Step 4: Identify the sentiment polarity of each Vietnamese document of the testing data set based on the classification results of clustering.
Step 5: Test the proposed model in the sequential system.
Step 6: Test this survey in the Cloudera parallel system – 2 nodes, the Cloudera parallel system – 3 nodes, and the Cloudera parallel system – 4 nodes

Transfer one Vietnamese Sentence into one Vector

Word Segmentation and Stop-Words Removal

As Vietnamese is an isolating language, the boundaries between words are not spaces as in English (Hoang *et al*., 2007; Nguyen *et al*., 2009). Therefore, we employed a Vietnamese word segmentation program (Le *et al*., 2008) in this work. All words and numbers are considered as features, usually referred to as tokens (Hoang *et al*., 2007). All documents are segmented into tokens, and the set of tokens is extracted by removing features that do not provide any information for document classification, such as numbers, dates, and function words (Hoang *et al*., 2007).

The main ideas of this segment are as follows:

Step 1: Word segmentation and stop-words removal (Hoang *et al*., 2007; Le *et al*., 2008; Nguyen *et al*., 2009) are applied to all the sentences in the training data set
Step 2: Each document of the testing data set is split into sentences. Word segmentation and stop-words removal are applied to every sentence of all the documents in the testing data set

Vector Representation

Each sentence of every document is a vector space model. The size of the vector comprises a maximum number of words/phrases; therefore, vector representation is executed via Word Segmentation and Stop-Words Removal. For example, if a Vietnamese sentence in the data set has 100 words/phrases, which is the maximum of all the sentences in the data set, then the size of the vector is 100. Each member of the vector has one value, which is calculated with Term Frequency-Inverse Document Frequency (TF-IDF).

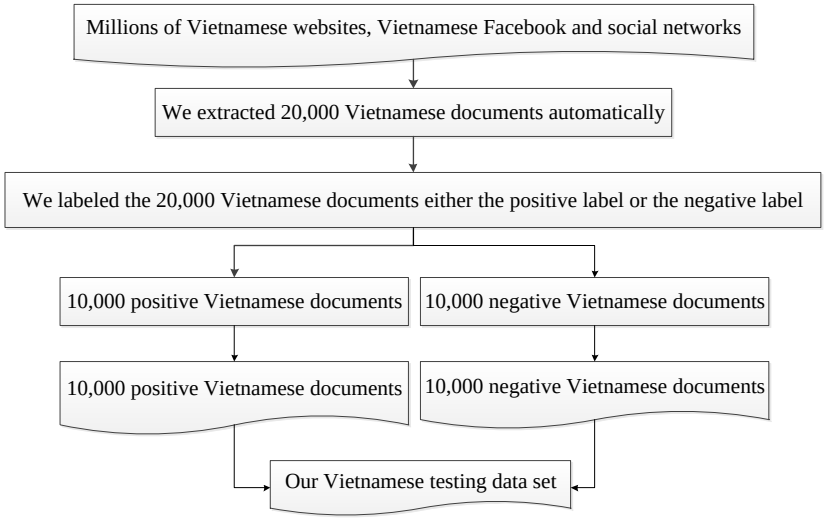

Fig. 1: Our Vietnamese data set

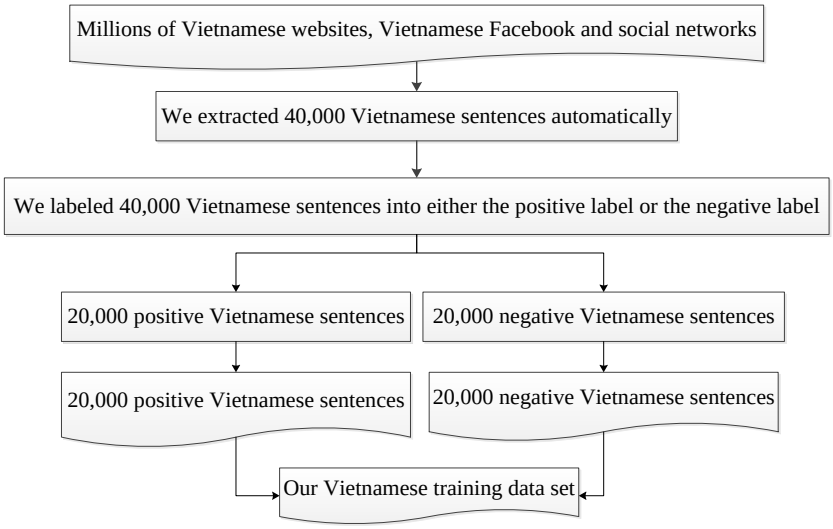

Fig. 2: Our Vietnamese data set

The overview process to transform each sentence into a vector in the parallel network environment is shown in Fig. 4.

In the Map (M) stage, the input of M is each sentence, after which M will undergo word segmentation and stop-words removal of said sentence. Then, M will transform this sentence into a vector. The output of M is the words/phrases of said sentence with their weight.

In the Reduce (R) stage, the input of R is the output of the M, and the output of R is one vector of each sentence.

After the implementation of word segmentation and stop-words removal, we used TF-IDF to transfer sentences in the training and testing data sets to the vector space model. The vector space model assigns weights to index terms. It is widely used in information retrieval to determine the relevance of a document for a given query. Both the document and the query are represented as weighted vectors of terms, and these weights are used to compute the degree of similarity between the query and the document.

TF-IDF is a measure that can also be applied as an algorithm to determine the ranking of a certain criterion of a word (phrase). The basic principle of this algorithm is that the significance of a word (phrase) will be proportional to its frequency of occurrence in a sentence and inversely proportional to its number of occurrences in other sentences in the dataset.

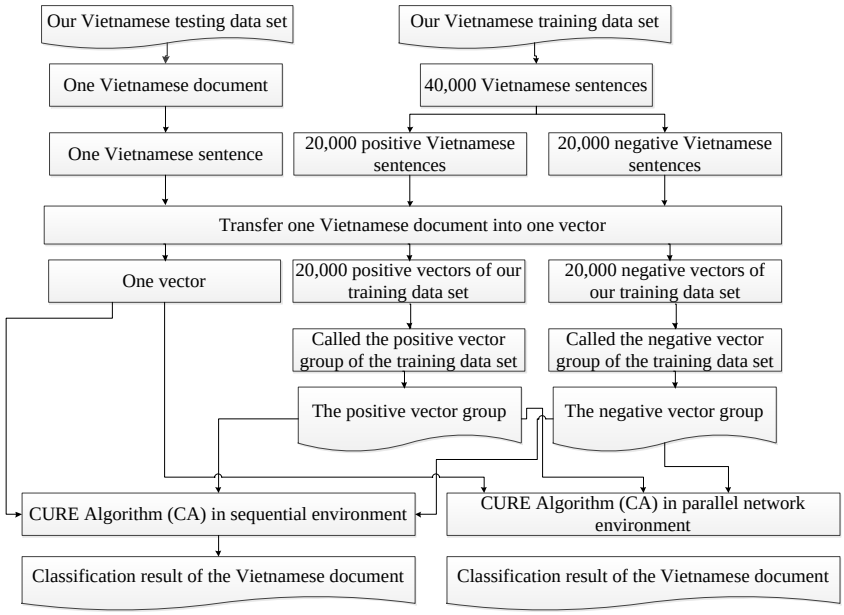

Fig. 3: Overview of our research

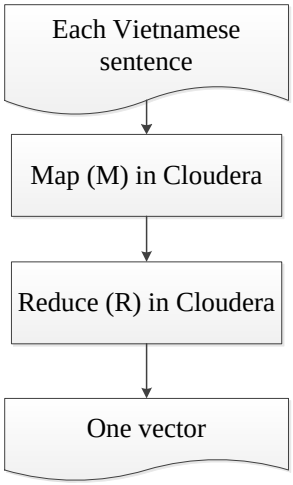

Fig. 4: Overview process of transforming each Vietnamese sentence into one vector in Cloudera

This algorithm combined with the model space vector is used widely in many fields, such as search engines and text mining. In this research, we only used the most common form of TF-IDF.

TF (Term Frequency)

TF is used to calculate the frequency of occurrence of the word (phrase) *t* in sentence *d*. If a word (phrase) appears more frequently, then TF is greater, and vice versa.

The simplest way to calculate *TF* of the word (phrase) t in sentence *d* is the frequency of occurrence of *t* in *d*:

$$Tf(t,d) = f(t,d) = \frac{Ns(t)}{W}$$

where, *Ns* (*t*) is the number of occurrences word (phrase) *t* in *d* and *W* is the total word (phrase) in *d*.

In addition to the above formula, there is another formula to calculate *TF*: A simple formula with enhanced frequency:

$$Tf(t,d) = \frac{f(t,d)}{\max\{f(w,d) : w \in d\}}$$

The numerator is the frequency of occurrence of the word (phrase) *t* in *d*. The denominator is the frequency of the word (phrase) that appears most in *d*.

TF is the only measure of the significance of a word (phrase) at the local (sentence) level. The significance level of word (phrase) in the entire data set is not shown, as numerous stop-words appeared several times. Therefore, we conducted the calculation of IDF to limit the significance of those words (phrases).

IDF (Inverse Document Frequency)

IDF is the inverse frequency of a word (phrase) in the data set. It shows the significance of a word (phrase) at the global level. IDF calculations reduce the value of popular words (phrases):

$$Idf(t,D) = \log\left(\frac{D}{d \in D : t \in d\}}\right)$$

with *D* is number of sentences in dataset and *d* is number of sentences in dataset which that sentences contain word (phrase) *t*.

In case if *t* does not appear in any sentence *d* of the *D* dataset then denominator equal to 0, the division is not valid, so it is often replaced by 1+ denominator ρ + (δ ∈ Δ: τ∈δ) that this does not affect the results of calculations.

We can notice that if a word (phrase) appears in the sentences of the data set more frequently, then its IDF value is smaller, and vice versa. However, a word (phrase) with a small IDF may be an important word (phrase), and a word (phrase) with a large IDF may be a popular word (phrase) and, thus, needs to be removed to avoid confounding results. Words (phrases) with small IDFs may be important words (phrases); this depends on the TF measure of that word (phrase), because words (phrases) that are rare may appear only in certain sentences of the dataset, and they are not useful in the classification process.

To identify important words (phrases), we conducted the TF-IDF calculation:

$$Tf - Idf(i,j) - Tf(i,j) * Idf(i)$$

If the TF-IDF measure is larger, then it is influential and will more greatly affect the classification. In the TF-IDF vector, if a word (phrase) t_i appears in d_j, then weight of the word (phrase) in the vector, which will be represented by the TF-IDF (t_i, d_j) value, is 0. Therefore, the following formula can be executed:

$$wij = \begin{cases} Tf - Idf(t_i, d_j), t_i \in d_j \\ 0, t_i \notin d_j \end{cases}$$

A CURE Algorithm

In the CURE algorithm, the CA in the sequential environment is implemented in the first sub-part, and the CA in the parallel network environment is performed in the second sub-part.

A CURE Algorithm in Sequential Environment

The CA in the sequential environment is executed as follows, as shown in Fig. 5.

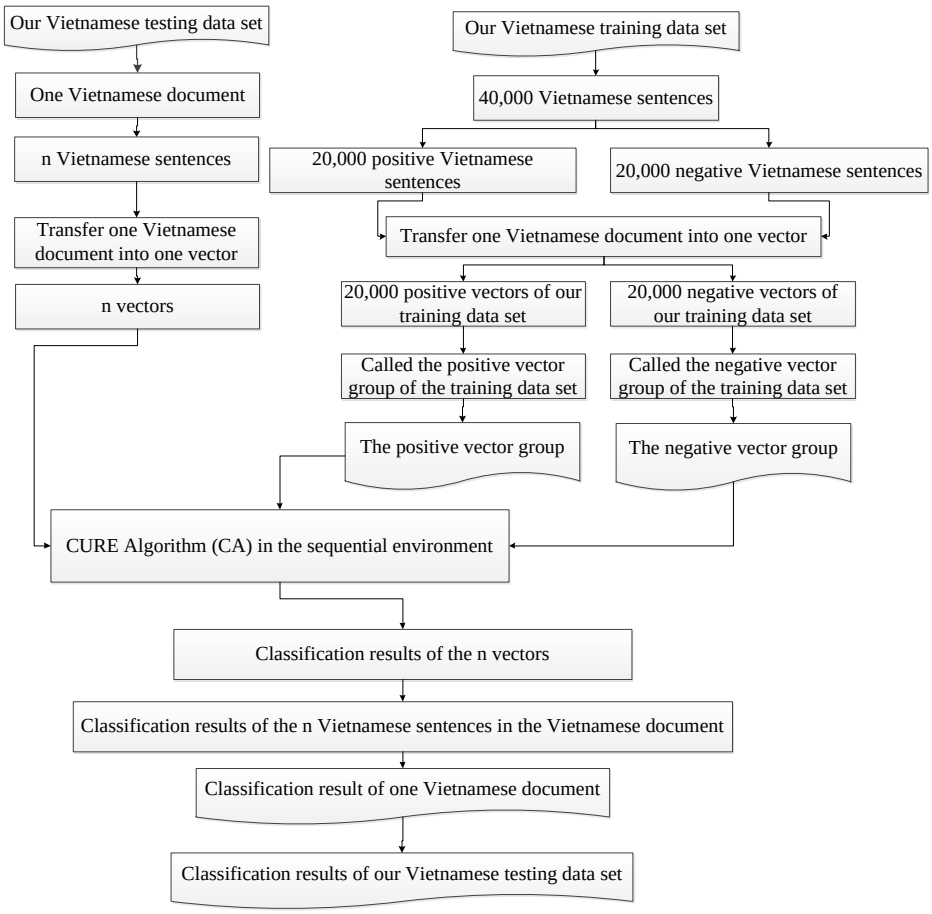

Fig. 5: A CURE algorithm in the sequential environment

The main ideas of this part are as follows:

Step 1: Transfer all the Vietnamese sentences of the training data set into the vectors of the positive vector group and the negative vector group in the sequential environment.
Step 2: Split each Vietnamese document of the testing data set into the Vietnamese sentences. Each Vietnamese sentence of this Vietnamese document is transferred into one vector in the sequential system.
Step 3: Use the CURE algorithm to cluster each vector of each Vietnamese document of the testing data set into the positive vector group or the negative vector group of the training data set in the sequential environment.
Step 4: Identify the sentiment polarity of each Vietnamese document of the testing data set based on the classification results of clustering in the sequential system.
Step 5: Classification results of the Vietnamese testing data set in the sequential environment.

CURE employs a novel hierarchical clustering algorithm that adopts a middle ground between the centroid and the all-point extremes (Guha *et al*., 1998). In CURE, first, a constant number c of well-scattered points in a cluster is chosen. The scattered points capture the shape and extent of the cluster. Next, the chosen scattered points are contracted towards the centroid of the cluster by a fraction cr. After shrinking, these scattered points are used as representations of the cluster. The clusters with the closest pair of representative points are merged at each step of CURE's hierarchical clustering algorithm.

CURE is robust to outliers and identifies clusters with non-spherical shapes and a wide variation in size. The CURE algorithm has many contributions: It can identify both spherical and non-spherical clusters and choose several well-scattered points as representatives of the cluster instead of a one-point centroid. It uses random sampling and partitioning to speed up clustering.

With the CURE algorithm, we follow the following steps.

For each cluster, *c* well-scattered points within the cluster are chosen and are contracted toward the mean of the cluster by a fraction α.

The distance between two clusters is equal to the distance between the closest pair of representative points from each cluster.

The c representative points attempt to capture the physical shape and geometry of the cluster. Shrinking the scattered points toward the mean removes surface abnormalities and decreases the effects of the outliers.

The CURE algorithm in the sequential environment is similar to algorithms used in numerous studies (Guha *et al*., 1998; Yan-Hua *et al*., 2011; Nian-Yun *et al*., 2009; Ertöz *et al*., 2002; Kaya and Alhajj, 2005; Ying *et al*., 2008; Rani *et al*., 2014): The input of the CURE algorithm is the positive vector group, the negative vector group and the n vectors of each document of our testing data set. The output of the CA in the sequential environment is the classification results of n vectors into the positive vector group or the negative vector group. After the CA in the sequential environment has been implemented, and the n vectors of the document have been classified, the document is deemed to exhibit a positive sentiment it has more vectors in the positive vector group than in the negative vector group, for the n vectors of the document. Conversely, the document exhibits negative semantics if it has fewer vectors in the positive vector group than in the negative vector group. Finally, the document is deemed to demonstrate neutral emotions if the number of the vectors in both the positive and negative vector groups are equal.

A CURE Algorithm in Parallel Network Environment

The CA in the Cloudera parallel network environment is executed as follows, as shown below in Fig. 6.

The main ideas of this part are as follows:

Step 1: Transfer all the Vietnamese sentences of the training data set into the vectors of the positive vector group and the negative vector group in the Cloudera parallel system – 2 nodes (the Cloudera parallel system – 3 nodes, and the Cloudera parallel system – 4 nodes).
Step 2: Split each Vietnamese document of the testing data set into the Vietnamese sentences. Each Vietnamese sentence of this Vietnamese document is transferred into one vector in the Cloudera parallel system – 2 nodes (the Cloudera parallel system – 3 nodes, and the Cloudera parallel system – 4 nodes).
Step 3: Use the CURE algorithm to cluster each vector of each Vietnamese document of the testing data set into the positive vector group or the negative vector group of the training data set in the Cloudera parallel system – 2 nodes (the Cloudera parallel system – 3 nodes, and the Cloudera parallel system – 4 nodes).
Step 4: Identify the sentiment polarity of each Vietnamese document of the testing data set based on the classification results of clustering in the Cloudera parallel system – 2 nodes (the Cloudera parallel system – 3 nodes, and the Cloudera parallel system – 4 nodes).
Step 5: Classification results of the Vietnamest testing data set in the Cloudera parallel system – 2 nodes (the Cloudera parallel system – 3 nodes, and the Cloudera parallel system – 4 nodes)

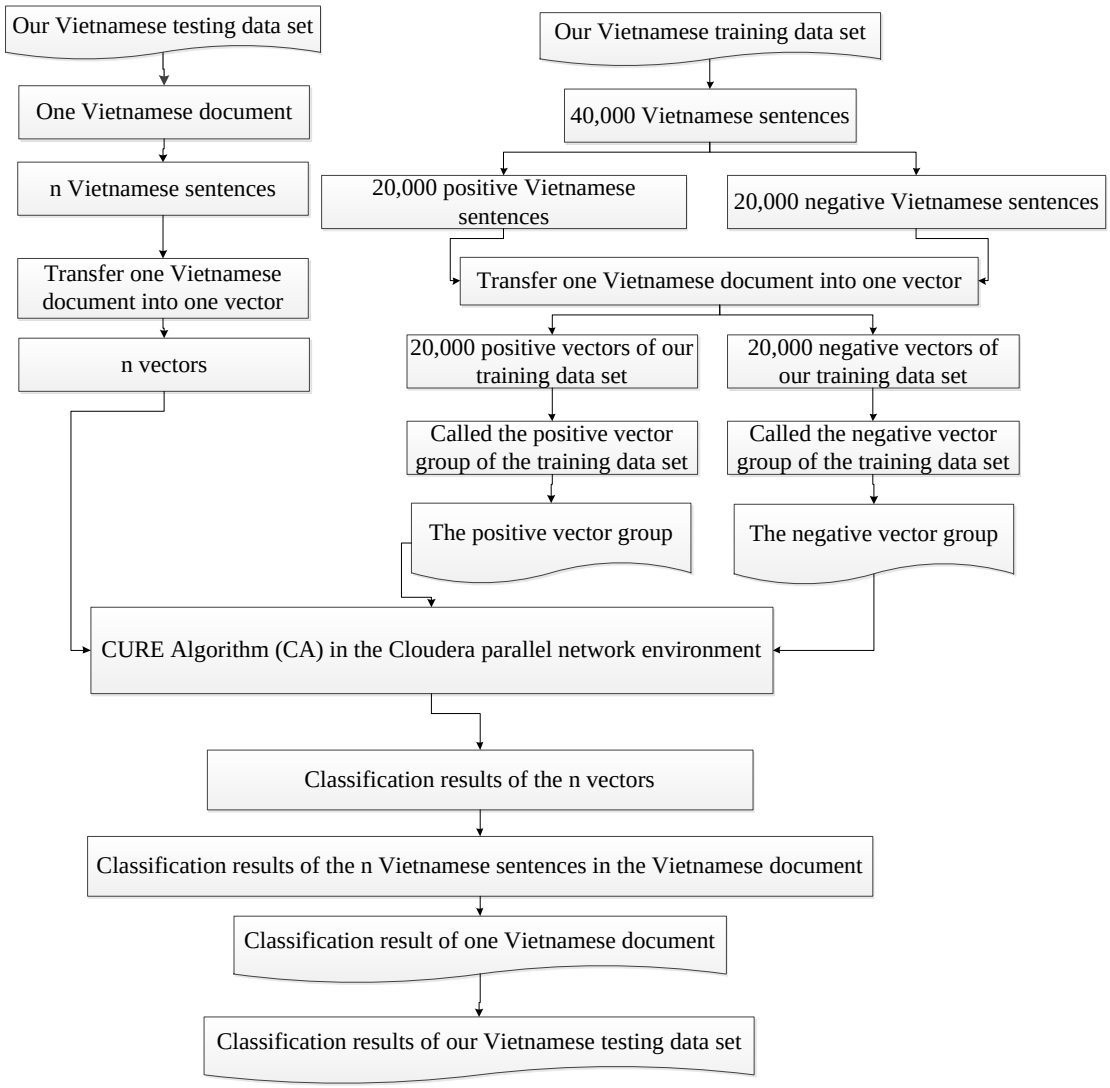

Fig. 6: A cure algorithm in the cloudera parallel network environment

Clustering each vector of each document of the testing data set into the positive vector group of the training data set, or the negative vector group of the training data set, is implemented by using the CURE algorithm with Hadoop Map (M)/Reduce (R) in the Cloudera parallel network environment.

The CURE algorithm in the Cloudera parallel network environment is similar to algorithms used in various works (Guha *et al*., 1998; Yan-Hua *et al*., 2011; Nian-Yun *et al*., 2009; Ertöz *et al*., 2002; Kaya and Alhajj, 2005; Ying *et al*., 2008; Rani *et al*., 2014). The input of the CURE algorithm is the positive vector group, the negative vector group, and the n vectors of each document of our testing data set. The output of the CA in the Cloudera environment is the classification results of n vectors to the positive vector group or the negative vector group. After the CA in the Cloudera environment is implemented and the n vectors of the document are classified, the document exhibits a positive sentiment if it has more vectors in the positive vector group than the negative vector group, for the n vectors of the document. The document exhibits negative semantics if it has fewer vectors in the positive vector group than the negative vector group. Finally, the document express neutral emotions if the number of the in both the positive and negative vector groups are equal.

This part includes two phases: the Hadoop Map (M) phase and the Hadoop Reduce (R) phase. The inputs of the Hadoop Map (M) phase in the Cloudera are the positive vector group of our training data set, the negative vector group of our training data set, and *n* vectors of each document of our testing data set.

The outputs of the Hadoop Map (M) phase in the Cloudera are the emotion classification results of the n vectors of each document of the testing data set. The inputs of the Hadoop Reduce (R) phase in Cloudera are the outputs of the Hadoop Map (M) phase in Cloudera and the semantic classification results of the n vectors of each document of the testing data set. The output of the Hadoop Reduce (R) phase in Cloudera is the sentiment classification result of the document of the testing data set; this document is classified into positive emotions, negative semantics, or neutral semantics.

The Hadoop Map (M) phase in the Cloudera parallel network environment is executed as follows, as shown in Fig. 7.

The main ideas of the Hadoop Map (M) phase are as follows:

Input: The *n* vectors of each Vietnamese document of the testing data set, the 20,000 vectors of the positive vector group, the 20,000 vector of the negative vector group.

Output: Results of clustering the *n* vectors of each Vietnamese document of the testing data set into the positive vector group or the negative vector group, The inputs of the Hadoop Reduce (R) phase.

Step 1: Each vector in the n vectors of the Vietnamese document, do repeat:
Step 2: Use the CURE algorithm to cluster this vector into the positive vector group or the negative vector group
Step 3: Get a result of this clustering.
Step 4: End Repeat – End Step 1.
Step 5: Results of clustering the n vectors of each Vietnamese document of the testing data set into the positive vector group or the negative vector group
Step 6: Transfer the results into the inputs of the Hadoop Reduce (R) phase.

The Hadoop Reduce (R) phase in the Cloudera parallel network environment is executed, as shown in Fig. 8.

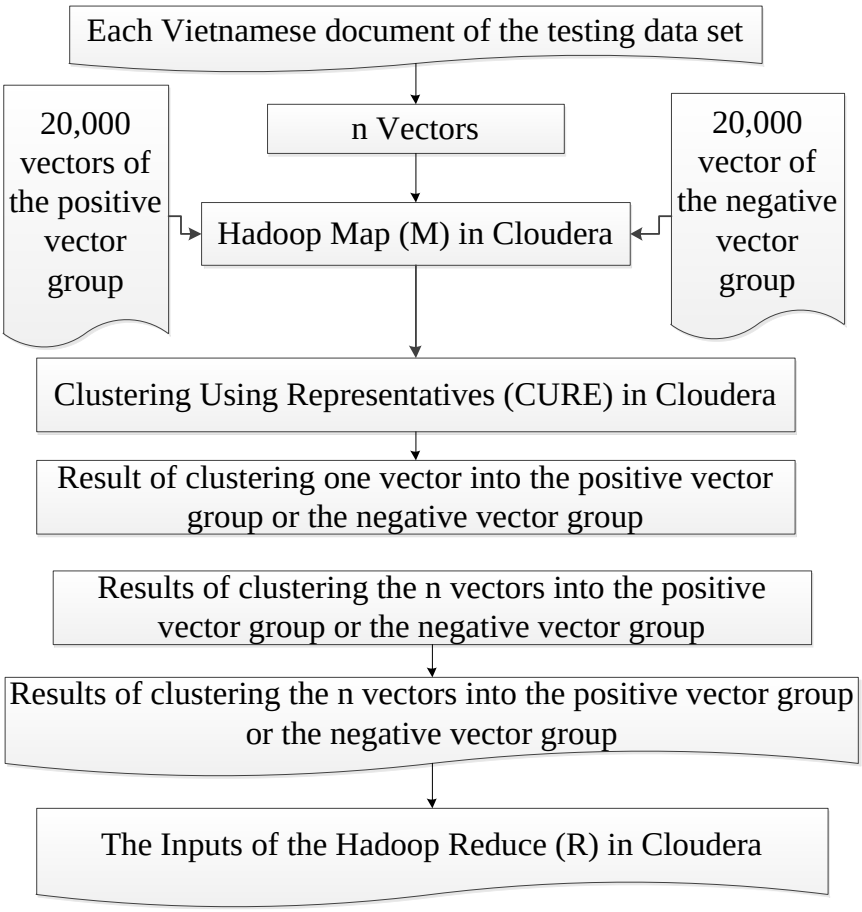

Fig. 7: Hadoop Map (M) phase in Cloudera

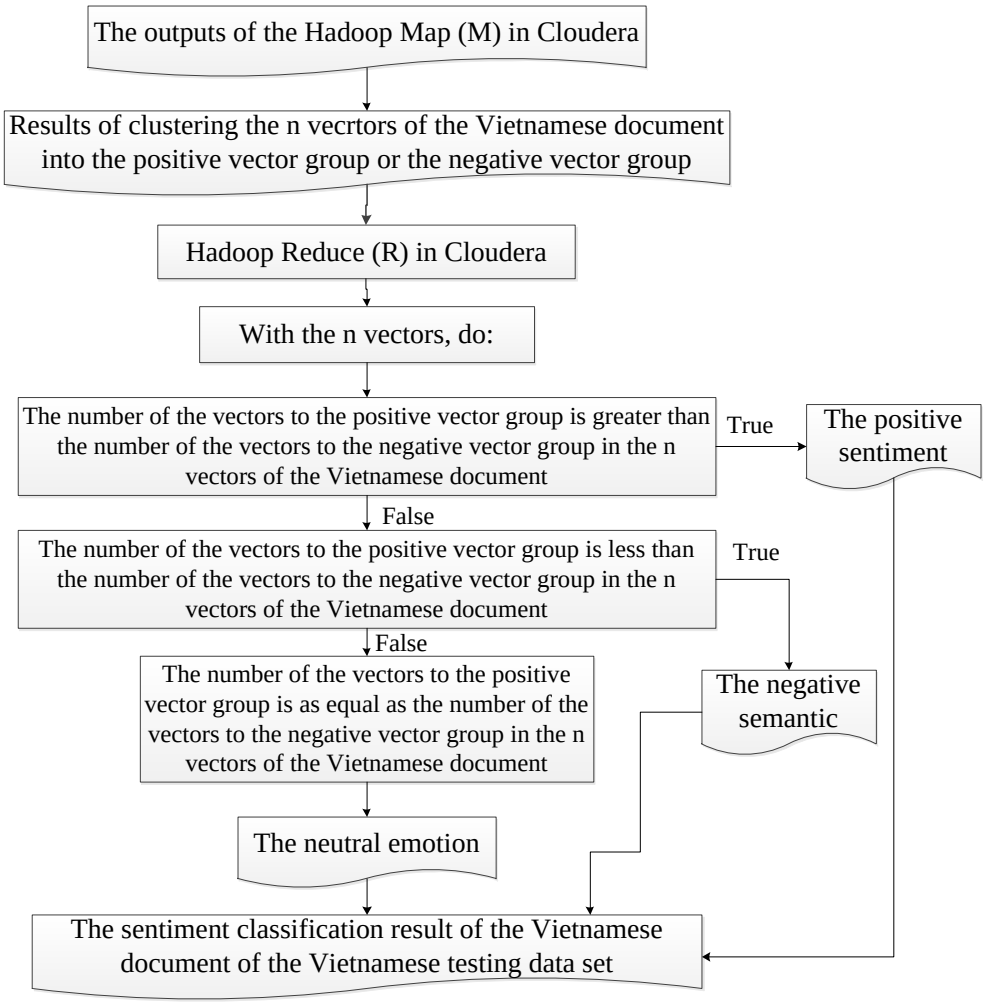

Fig. 8: Hadoop Reduce (R) phase in Cloudera

The main ideas of the Hadoop Reduce (R) phase are as follows:

Input: The outputs of the Hadoop Map (M) phase in the Cloudera - Results of clustering the n vectors of the Vietnamese document into the positive vector group or the negative vector group.

Output: The results of the sentiment classification of the Vietnamese document of the Vietnamese testing data set.

Step 1: If the number of the vectors of the positive vector group is greater than the number of the vectors of the negative vector group in the n vectors of the Vietnamese document Then return positive

Step 2: Else If the number of the vectors of the positive vector group is less than the number of the vectors of the negative vector group in the n vectors of the Vietnamese document Then return negative

Step 3: Else return neutral

Step 4: Get the polarity of the Vietnamese document of the testing data set.

Step 5: The results of the sentiment classification of the Vietnamese document of the Vietnamese testing data set.

Experiment

We have used the measure such as Accuracy (A) to calculate the accuracy of the results of emotion classification.

The Java programming language was used to save our data sets and implement our proposed model to classify the 20,000 Vietnamese documents.

To implement the proposed model, we used the Java programming language to save the training data set, testing data set, and the results of emotion classification.

The sequential environment in this research includes one node (one server). The Java language was used to program CA. 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 performed CA in the Cloudera parallel network environment. This Cloudera system includes three nodes (three servers). The Java language is used in programming the application of the CURE 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 four servers is Cloudera. All four nodes have the same configuration information.

The results of the 20,000 Vietnamese documents tested are presented in Table 1.

The accuracy of the 20,000 Vietnamese documents in the testing dataset is presented in Table 2.

We also tested the 20,000 Vietnamese documents in the sequential environment, the Cloudera system with two nodes, the Cloudera system with three nodes, and the Cloudera system with four nodes in Table 3.

Table 1: The results of the 20,000 Vietnamese documents in the testing data set

	Testing dataset	Correct classification	Incorrect classification
Negative	10,000	62,89	3,711
Positive	10,000	6,295	3,705
Summary	20,000	12,584	7,416

Table 2: The accuracy of our new model for the 20,000 Vietnamese documents in the testing data set

Proposed model	Class	Accuracy
Our new model	Negative Positive	62.92%

Table 3: The execution time of our new model for the 20,000 Vietnamese documents in the testing data set

	Average execution time
CURE algorithm in the sequential environment	21,600 sec/20,000 Vietnamese documents
CURE algorithm in the cloudera parallel network environment-2 nodes	9,495 sec/20,000 Vietnamese documents
parallel cure algorithm in the cloudera network environment-3 nodes	7,198 sec/20,000 Vienamese documents
parallel cure algorithm in the cloudera network environment-4 nodes	4,689 sec/20,000 Vietnamese documents

Results and Discussion

With our proposed new model, we achieved 62.92% accuracy for the Vietnamese documents in Table 2. In Table 3, the average time of the semantic classification of the CURE algorithm in the sequential environment is 21,600 seconds/20,000 documents. This rate is greater than the average time of the emotion classification of the CA in the Cloudera parallel network environment with three nodes, which is 7,198 seconds/20,000 documents. The average execution time of the sentiment classification of the CA in the Cloudera parallel network environment with four nodes demonstrated the fastest time, which is 4,689 seconds/20,000 documents. The average execution time of the sentiment classification of the CA in the Cloudera parallel network environment with two nodes is faster than the average execution time of the sentiment classification of the CA in the Cloudera parallel network environment with three nodes.

Conclusion

Although our new model was tested on a Vietnamese data set, it can be applied to other languages. In this paper, our model was tested on 20,000 documents, which is a small data set. However, our model can be applied to data sets for big data, with millions of Vietnamese documents.

In this work, we proposed a new model to classify sentiments for the Vietnamese documents by Clustering Using Representatives (CURE) with Hadoop Map (M) /Reduce (R) in Cloudera parallel network environment. At the time of the execution of this study, not much research has been published that has shown the use of clustering methods to classify data. Our research shows that clustering methods can be used to classify data and, in particular, to classify emotion for text.

The accuracy of the proposed model depends on many factors:

1) The CURE-related algorithms: To increase accuracy, the CURE-related algorithms can be improved or replaced with another algorithm.
2) The positive vector group: The positive vector group depends on the 20,000 positive Vietnamese sentences and the algorithms that are used to transfer the sentences to the vectors. To obtain an increased rate of accuracy, the algorithms can be further developed.
3) The negative vector group: The negative vector group depends on the 20,000 negative Vietnamese

sentences and the algorithms that are used to transfer the sentences to the vectors. To obtain an increased rate of accuracy, the algorithms can be further developed.

4) The 20,000 positive sentences of the training data set: To increase the accuracy, we can increase the number of the positive sentences in the training data set.
5) The 20,000 negative sentences of the training data set: To increase the accuracy, we can increase the number of the negative sentences in the training data set.
6) The testing data set: The training data set must be similar to the testing data set.
7) The testing and training data sets: To improve the accuracy of the data sets, the selected sentences could be chosen with regard to a specific domain (such as textbooks, cartoons, etc.). This selection could also be used to standardize sentences and documents on the Internet, as well.

The execution time of the proposed model depends on many factors:

1) The performance of the distributed environment: The performance of the parallel system depends on the Cloudera system, Hadoop Map /Reduce, the algorithms, and the performance of the nodes
2) The Cloudera system and Hadoop Map/Reduce: To increase the execution time, the Cloudera system and Hadoop Map/Reduce can be improved or replaced with other distributed environments and parallel functions
3) The CURE-related algorithms: To increase the execution time, the CURE-related algorithms can be improved or replaced with other algorithms
4) The performance of the nodes: The performance of the nodes depends on the number of servers and the performance of each server. To increase the execution time, the number of the servers can be increased, or the performance of each server can be improved

The proposed model has many advantages and limitations. It uses the CURE algorithm to classify semantics of Vietnamese documents based on Vietnamese sentences. The proposed model can process millions of Vietnamese documents in the shortest time. This study can be performed in the distributed systems. It can be applied to other languages. However, low accuracy and a relatively high investment of financial costs and time are also associated with the implementation of this proposed model.

To understand the scientific values of this research, we conduct to compare our model's results with many studies, as shown in the tables.

Our model's results are compared with the works in Tables 4 and 5 (Hoang *et al*., 2007; Le *et al*., 2008; Nguyen *et al*., 2009).

In Tables 6 and 7, our model's results are compared with the research related to the CURE algorithm (Guha *et al*., 1998; Yan-Hua *et al*., 2011; Nian-Yun *et al*., 2009; Ertöz *et al*., 2002; Kaya and Alhajj, 2005; Ying *et al*., 2008; Rani *et al*., 2014).

Tables 8 and 9, our model's results are compared with the latest research on Vietnamese sentiment classification, Vietnamese sentiment analysis, and Vietnamese opinion mining.

Tables 10 and 11, our model's results are compared with the latest researches of the sentiment classification, sentiment analysis, and opinion mining.

Tables 12 and 13, our model's results are compared with the latest works on unsupervised classification (Turney, 2002; Lee *et al*., 2002a; van Zyl, 2002; Hegarat-Mascle *et al*., 2002; Ferro-Famil *et al*., 2002; Chaovalit and Zhou, 2005; Lee *et al*., 2002b; Gllavata *et al*., 2004).

Table 4: Comparisons of our model's results with the works in (Hoang *et al*., 2007; Le *et al*., 2008; Nguyen *et al*., 2009)

Studies	CA	SC	L	SD	DT	PNE	Approach
(Hoang *et al*., 2007)	No	Yes	VL	Yes	Yes	No	Two different approaches for the Vietnamese text classification problem: The Bag of Words - BOW and Statistical N-Gram Language Modeling - N-Gram approaches
(Le *et al*., 2008)	No	Yes	VL	Yes	Yes	No	A hybrid approach to automatically tokenize Vietnamese text.
(Nguyen *et al*., 2009)	No	Yes	VL	Yes	Yes	No	A new syllable-based document representation at the morphological level of the language for efficient classification.
Our work	Yes	Yes	VL	No	Yes	Yes	A new model for Vietnamese document-level sentiment classification using the CURE Algorithm (CA) in the Cloudera parallel network environment.

CURE Algorithm: CA; Sentiment Classification: SC; Language: L; Special Domain: SD; Depending on the training data set: DT; Parallel network environment: PNE; Vietnamese language: VL; English language: EL; No mention: NM

Table 5: Comparisons of our model's advantages and disadvantages with the works in (Hoang *et al*., 2007; Le *et al*., 2008; Nguyen *et al*., 2009)

Works	Advantages	Disadvantages
(Hoang *et al*., 2007)	With the differences between Vietnamese and English, finding a feasible approach for Vietnamese TC is very interesting. With the authors' experiments, they prove that both SVM with average accuracy 96.21% and N-Gram with average accuracy 95.58% absolutely suitable to use for Vietnamese TC. Especially, the N-Gram model seems to be preferable to SVM for the following reasons: The higher classification speed, avoidance of the word segmentation and explicit feature selection procedure and giving the equivalent F1-score result.	The authors also recognize that these approaches for Vietnamese TC occur some errors such as: (1) The limitations from tokenizer (word segmentation tool) affects to classification performance (in BOW approach) (2) The documents have the ambiguities between two or many topics because these documents have too many tokens or phrases which both express the content of the topic.
(Le *et al*., 2008)	The authors present in this survey a hybrid approach to automatically tokenize Vietnamese text. The application of a maximal matching strategy on a graph results in all candidate segmentations of a phrase. It is the responsibility of an ambiguity resolver, which uses a smoothed bi-gram language model, to choose the most probable segmentation of the phrase. The hybrid approach is implemented to create vnTokenizer, a highly accurate tokenizer for Vietnamese texts	The authors found that the majority of errors of segmentation is due to the presence in the texts of compounds absent from the lexicon. Unknown compounds are a much greater source of segmenting errors than segmentation ambiguities. Future efforts should therefore be geared in priority towards the automatic detection of new compounds, which can be performed by means either statistical in a large corpus or rule-based using linguistic knowledge about word composition.
(Nguyen *et al*., 2009)	This work introduces a new syllable-based document representation at the morphological level of the language for efficient classification. The authors tested the representation on their corpus with different classification tasks using six classification algorithms and two feature selection techniques. The authors' experiments show that the new representation is effective for Vietnamese categorization, and suggest that the best performance can be achieved using syllable-pair document representation, an SVM with a polynomial kernel as the learning algorithm, and using Information gain and an external dictionary for feature selection.	No mention
Our study	The advantages and disadvantages of the proposed model are given in the Conclusion section.	

Table 6: Comparisons of our model's results with the researches related to CURE algorithm in (Guha *et al*., 1998; Yan-Hua *et al*., 2011; Nian-Yun *et al*., 2009; Levent Ertöz *et al*., 2002; Kaya and Alhajj, 2005; Ying *et al*., 2008; Rani *et al*., 2014)

Works	CA	SC	L	SD	DT	PNE	Model/method
(Guha *et al*., 1998)	Yes	No	NM	NM	NM	No	A new clustering algorithm called CURE that is more robust to outliers, and identifies clusters having non-spherical shapes and wide variations in size
(Yan-Hua *et al*., 2011)	Yes	No	NM	NM	NM	No	A network user behavior data model based on Netflow statistics.
(Nian-Yun *et al*., 2009)	Yes	No	NM	NM	NM	No	A new method of choosing representative records for a cluster.
(Ertöz *et al*., 2002)	Yes	No	NM	NM	NM	No	A novel method of implementing a density based approach over a Produs algorithm to cluster even small data points.
(Kaya and Alhajj, 2005)	Yes	NM	NM	NM	NM	No	An automated method for mining fuzzy association rules
(Ying *et al*., 2008)	Yes	NM	NM	NM	NM	No	An appropriate CURE algorithm and C4.5 decision tree method are adopted to establish a new costume sales forecasting model
(Rani *et al*., 2014)	Yes	NM	NM	NM	NM	No	A comparative analysis of these two algorithms, namely BIRCH and CURE
Our study	Yes	Yes	VL	No	Yes	Yes	A new model for Vietnamese document-level sentiment classification using the CURE algorithm (CA) in the Cloudera parallel network environment.

Table 7: Comparisons of our model's merits and demerits with the researches related to CURE algorithm in (Guha *et al*., 1998; SUN Yan-Hua *et al*., 2011; Nian-Yun *et al*., 2009; Ertöz *et al*., 2002; Kaya and Alhajj, 2005; Ying *et al*., 2008; Rani *et al*., 2014)

Works	Merits	Demerits
(Guha *et al*., 1998)	The authors' experimental results confirm that the quality of clusters produced by CURE is much better than those found by existing algorithms. Furthermore, they demonstrate that random sampling and partitioning enable CURE to not only outperform existing algorithms, but also to scale well for large databases without sacrificing clustering quality.	No mention
(Yan-Hua *et al*., 2011)	Experiment results show that the improved algorithm is not only able to cluster, but also can distinguish the normal and abnormal behaviors. Analyzed by harm behavior evaluating system, most of the abnormal behaviors belong to harm behaviors. For increment data on the real net, it also gives the method of increment mining, which accords with the need of real time network analyzing.	No mention
(Nian-Yun *et al*., 2009)	To inspect duplicated records, the Clustering Using Representatives (CURE) algorithm is ameliorated. The definition of pre-sampling is put forward, which can find the distribution of duplicated records so as to improve exactness of random sampling in record sets. A new method of choosing representative records for a cluster is proposed, which is based on distance infection weight. With this method, representative points are selected not only according to the density of the clusters, but also according to the importance of points including some isolated points. This method can make selecting representative points suitable. Both theory and experiment show that it is an effective approach to detect the similar duplicated records	No mention
(Ertöz *et al*., 2002)	This approach handles many problems that traditionally plague clustering algorithms, e.g., finding clusters in the presence of noise and the outliers and finding clusters in data that has clusters of different shapes, sizes, and density. The authors have used their clustering algorithm on a variety of high and low dimensional data sets with good results, but in this work, they present only a couple of examples involving high dimensional data sets: Word clustering and time series derived from NASA Earth science data.	No mention
(Kaya and Alhajj, 2005)	The authors compared the proposed GA-based approach with other approaches from the literature. Experiments conducted on 100K transactions from the US census in the year 2000 show that the proposed method exhibits a good performance in terms of execution time and interesting fuzzy association rules.	No mention
(Ying *et al*., 2008)	The CURE algorithm carries out grouping of similar items in terms of sales prospect and the C4.5 decision tree finds understandable links between these clusters and selected descriptive criteria. Based on the test of 568 historical sales data and 326 new data, the performance efficiency of forecasting model is analyzed. Finally, aiming at the classification error of forecasting model, a further improvement method is given.	No mention
(Rani *et al*., 2014)	These are not pure hierarchical clustering algorithm, some other clustering algorithm techniques are merged into hierarchical clustering in order to improve cluster quality and also to perform multiple phase clustering. This study presents a comparative analysis of these two algorithms namely BIRCH and CURE by applying Weka 3.6.9 data mining tool on Iris Plant dataset.	No mention
Our research	Our model's merits and demerits are illustrated in the Conclusion section.	

Table 8: Comparisons of our model with the latest Vietnamese sentiment classification models (or the latest Vietnamese sentiment classification methods) in (Ha *et al*., 2011; Bang *et al*., 2015; Kieu and Pham, 2010; Xuan-Son Vu and Park, 2014; Nguyen *et al*., 2014; Le *et al*., 2015; Trinh and Dau, 2014; Hoanh-Su *et al*., 2015; Phan and Cao, 2014; Nguyen *et al*., 2014; Duyen *et al*., 2014; Bach *et al*., 2015; Son Trinh *et al*., 2016)

Studies	CA	SC	L	SD	DT	PNE	Approach
(Ha *et al*., 2011)	No	Yes	VL	Yes	Yes	No	+HAC clustering +Semi-supervised SVM-kNN classification
(Bang *et al*., 2015)	No	Yes	VL	Yes	Yes	No	+Decision Tree +Naive Bayes (NB) +Support Vector Machines (SVM +Feature selection technique, $\chi 2$ (CHI).
(Kieu and Pham, 2010)	No	Yes	VL	Yes	Yes	No	A rule-based system using the Gate framework
(Vu and Park, 2014)	No	Yes	VL	No	No	No	A method to construct VSWN from a Vietnamese dictionary, not from WordNet.
(Nguyen *et al*., 2014)	No	Yes	EL	Yes	Yes	No	A supervised machine learning approach to handle the task of document-level sentiment polarity classification
(Le *et al*., 2015)	No	Yes	VL	Yes	Yes	No	+An approach to extracting and classifying aspect-terms for Vietnamese language. +Semi-supervised learning GK-LDA
(Trinh *et al*., , 2016)	No	Yes	EL	Yes	Yes	No	A crossed-domain sentiment analysis system

Table 8: Continue

(Hoanh-Su *et al*., 2015)	No	Yes	VL	Yes	Yes	No	+Naive-Bayes for Vietnamese text. +Decision trees for Vietnamese text +Support Vector Machine (SVM) for Vietnamese text
(Phan and Cao, 2014)	No	Yes	VL	Yes	Yes	No	+Skip-gram based model +A machine learning based classification, SVM
(Nguyen *et al*., 2014)	No	Yes	VL	Yes	Yes	No	+A domain specific sentiment dictionary +Statistical methods for a specific domain
(Duyen *et al*., 2014)	No	Yes	VL	Yes	Yes	No	Machine learning
(Bach *et al*., 2015)	No	Yes	VL	Yes	Yes	No	A general framework for mining Vietnamese comparative sentences
(Trinh *et al*., 2016)	No	Yes	VL	No	No	No	A lexicon based method
This study	Yes	Yes	VL	No	Yes	Yes	A new model for Vietnamese document-level sentiment classification using the CURE algorithm (CA) in the Cloudera parallel network environment.

Table 9: Comparisons of our model's benefits and drawbacks with the latest Vietnamese sentiment classification models (or the latest Vietnamese sentiment classification methods) in (Ha *et al*., 2011; Bang *et al*., 2015; Kieu and Pham, 2010; Vu and Park, 2014; Nguyen *et al*., 2014; Le *et al*., 2015; Anh and Dau, 2014; Hoanh-Su *et al*., 2015; Phan and Cao, 2014; Nguyen *et al*., 2014; Duyen *et al*., 2014; Bach *et al*., 2015; Trinh *et al*., 2016)

Studies	Benefits	Drawbacks
(Ha *et al*., 2011)	In this study, an upgrading FOMS model on Vietnamese reviews on mobile phone products is described. Feature words and opinion words were extracted based on some Vietnamese syntactic rules. Extracted feature words were grouped by using HAC clustering and semi-supervised SVM-KNN classification. Customers' opinion orientation and summarization on features was determined by using a VietSentiWordNet, which had been extended from an initial VietSentiWordNet. Experiments on feature extraction and opinion summarization on features are shown.	No mention
(Bang *et al*., 2015)	In order to analyze the sentiment, the authors compare three different Text categorization algorithms, including Decision Tree, Naive Bayes (NB) and Support Vector Machines (SVM). Furthermore, the authors enhance the efficiency of the text categorization by applying feature selection technique, $\chi 2$ (CHI). The evaluation was conducted on 1,650 hotel reviews written in Vietnamese languages. The experimental results showed that applying term feature selection could significantly improve the performance of the sentiment analysis.	No mention
(Kieu and Pham, 2010)	In this work, the authors address this problem at the sentence level and rule-based system using the Gate framework. build a Experimental results on a corpus of computer product reviews are very promising. To the best of the authors' knowledge, this is the first work that analyzes the sentiment at the sentence level in Vietnamese.	In the future, the authors plan to collect a larger data set with more diverse domains and combine the authors' system with machine learning approaches
(Vu and Park, 2014)	The authors propose a method to construct VSWN from a Vietnamese dictionary, not from WordNet. The authors show the effectiveness of the proposed method by generating a VSWN with 39,561 synsets automatically. The method is experimentally tested with 266 synsets with an aspect of positivity and negativity. It attains a competitive result compared with English SentiWordNet that is 0.066 and 0.052 differences for positivity and negativity sets respectively.	No mention
(Nguyen *et al*., 2014)	The authors reach state-of-the-art accuracies at 91.6 and 89.87% on the dataset PL04 and IMDB11 respectively. Furthermore, by analyzing the effects of rating-based feature of the accurate performance, the authors show that the rating-based feature is very efficient to sentiment classification on polarity reviews. And adding bigram and trigram features also enhances accuracy performance. Furthermore, the authors get an accuracy of 93.24% on the dataset SAR14, and they also share	No mention
(Le *et al*., 2015)	In this work, the authors propose an approach to extracting and classifying aspect-terms in Vietnamese language. The semi- supervised learning GK-LDA is proved to have better performance than the traditional topic modeling LDA. In the aspect inference, the authors use dictionary-based method which can extract noun-phrases for obtaining better performance than just extract word seeds or use a complete sentence to infer aspects. The authors' experimental results show that their proposed method can effectively perform the aspect extraction and classification task. task. Even though the authors' approach is initially proposed for handling Vietnamese text, the authors believe that it is also applicable to other languages.	No mention

Table 9: Continue

(Anh and Dau, 2014)	In this study, the authors proposed a crossed-domain sentiment analysis system for the discovery of current careers from social networks. The proposed system can capture the sentiment of career-related messages from two famous social networks, including Twitter and Facebook. The experimental results clearly pointed out that the most favorite careers which enjoy the highest positive sentiment and the least favorite careers that have the highest negative sentiment. The performance results of the proposed system are promising for cross-domain sentiment analysis, with the precision of over 85% and the recall of over 90%.	No mention
(Hoanh-Su *et al*., 2015)	This research applies machine learning with several algorithms such Naive-Bayes, decision trees and Support Vector Machine (SVM) for Vietnamese text data collected from fast-food industry on Facebook. The experiment results show that machine learning methods are able to classify Vietnamese sentiment text with the accuracy over 70%. Thus we proposed several recommendations for mining Vietnamese social text data.	No mention
(Phan and Cao, 2014)	The core of this research focuses on contributing an approach for word representation that reflects Vietnamese semantic information contexts for analyzing as an input of a machine learning based classification, SVM. The application of this research is applied to the STAAR project's opinion mining system.	No mention
(Nguyen *et al*., 2014)	This research proposes an approach to mining public opinions from Vietnamese text using a domain specific sentiment dictionary in order to improve the accuracy. The sentiment dictionary is built incrementally using statistical methods for a specific domain. The efficiency of the approach is demonstrated through an application which is built to extract public opinions on online products and services. Even though this approach is designed initially for Vietnamese text, we believe that it is also applicable to other languages.	No mention
(Duyen *et al*., 2014)	This work presents an empirical study on machine learning based sentiment analysis for Vietnamese, in which we focus on the task of sentiment classification. The study provides useful information for further research as well as for building a real sentiment analysis system for Vietnamese.	No mention
(Bach *et al*., 2015)	The authors present a general framework for mining Vietnamese comparative sentences in which we formulate the first subtask, i.e., Identifying comparative sentences, as a classification problem and the second subtask, i.e., Recognition of relations, as a sequence learning problem. The authors introduce a new corpus for the task in Vietnamese and conduct a series of experiments on that corpus to investigate the task in both linguistic and modeling aspects. The authors' work provides promising results for further research on this interesting task.	No mention
(Trinh *et al*., 2016)	In this survey, the authors propose a lexicon based method for sentiment analysis with Facebook data for Vietnamese language by focus on two core component in a sentiment system. That is to build a Vietnamese Emotional Dictionary (VED) including 5 sub-dictionaries: Noun, verb, adjective and adverb and propose features which based-on the English emotional analysis method and adaptive with traditional Vietnamese language and then support vector machine classification method to be used to identify the emotional of the user's message. The experiment shows that our system has very good performance.	No mention
This study	Our model's benefits and drawbacks are shown in the Conclusion section.	

Table 10: Comparisons of our model with the latest sentiment classification models (or the latest sentiment classification methods) in (Manek *et al*., 2016; Agarwal and Mittal, 2016a; 2016b; Canuto *et al*., 2016; Ahmed and Danti, 2016; Phu and Tuoi, 2014; Tran *et al*., 2014)

Works	CA	SC	L	SD	DT	PNE	Model/Method
(Manek *et al*., 2016)	No	Yes	EL	Yes	Yes	No	+A Gini Index based feature selection method +Support Vector Machine (SVM) classifier
(Agarwal and Mittal, 2016a)	No	Yes	EL	Yes	Yes	No	Machine Learning Approach
(Agarwal and Mittal, 2016b)	No	Yes	EL	Yes	No	No	The corpus-based semantic orientation approach for sentiment analysis.

Table 10: Continue

(Canuto *et al*., 2016)	No	Yes	EL	Yes	Yes	No	New meta-level features, especially designed for the sentiment analysis of short messages
(Ahmed and Danti, 2016)	No	Yes	EL	Yes	Yes	No	SentiWordNet that generates score count words into one of the seven categories like strong-positive, positive, weak-positive, neutral, weak-negative, negative and strong-negative words.
(Phu and Tuoi, 2014)	No	Yes	EL	No	No	No	+Terms-Counting method. +Contextual Valence Shifters method
(Tran *et al*., 2014)	No	Yes	EL	Yes	Yes	No	+Naïve Bayes. +N-Gram +Chi-Square, etc.
This work	Yes	Yes	VL	No	No	Yes	A new model for Vietnamese document-level sentiment classification using the CURE algorithm (CA) in the Cloudera parallel network environment.

Table 11: Comparisons of our model's benefits and drawbacks with the latest sentiment classification models (or the latest sentiment classification methods) in (Manek *et al*., 2016; Agarwal and Mittal, 2016a; 2016b; Canuto *et al*., 2016; Ahmed and Danti, 2016; Phu and Tuoi, 2014; Tran *et al*., 2014)

Works	Benefits	Drawbacks
(Manek *et al*., 2016)	In this work, a Gini Index based feature selection method with Support Vector Machine (SVM) classifier is proposed for sentiment classification for large movie review dataset. The results show that our Gini Index method has better classification performance in terms of reduced error rate and accuracy.	No mention
(Agarwal and Mittal, 2016a)	The main emphasis of this chapter 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) features weighting schemes, (3) feature selection, and (4) machine-learning methods. This survey also discusses the standard free benchmark datasets and evaluation methods for sentiment analysis. The authors conclude the chapter with a comparative study of some state-methods for sentiment analysis and some possible future research directions in opinion mining and sentiment analysis.	No mention
(Agarwal and Mittal, 2016b)	This approach initially mines sentiment-bearing terms from 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 the multi- word features.	No mention
(Canuto *et al*., 2016)	This study provides useful insights into how to enhance the performance of sentiment analysis by improving the representation schemes for instances, categories and their relationships.	A line of future research would be to explore the authors' meta features with other classification Lgorithms and feature selection techniques in different sentiment analysis tasks, such as scoring movies or products according to their related reviews.
(Ahmed and Danti, 2016)	The proposed approach is experimented with online books and political reviews and demonstrates the efficacy through Kappa measures, which has a higher accuracy of 97.4 % and lower error rate. 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 11: Continue

(Phu and Tuoi, 2014)	The authors combine five dictionaries into the new one with 21137 entries. No Mention The new dictionary has many verbs, adverbs, phrases and idioms are not in five ones before. +The work 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.
(Tran *et al*., 2014)	+The authors have explored Naive Bayes model with N-GRAM method, No Mention Negation Handling method, Chi- Square method and Good-Turing discounting with selecting different thresholds of Good-Turing discounting method and and different minimum frequencies of Chi-Square method to improve the accuracy of sentiment classification.
This work	The proposed model's benefits and drawbacks are shown in the Conclusion section.

Table 12: Comparisons of our model's results with the latest works of the unsupervised classification in [(Turney, 2002; Lee *et al*., 2002; van Zyl, 2002; Hegarat-Mascle *et al*., 2002; Ferro-Famil *et al*., 2002; Chaovalit and Zhou, 2005; Lee *et al*., 2002; Gllavata *et al*., 2004) Unsupervised Classification: UC

Studies	CA	SC	L	SD	DT	PNE	Approach	UC
(Turney, 2002)	No	Yes	EL	Yes	Yes	No	A simple unsupervised learning algorithm for classifying reviews as recommended (thumbs up) or not recomm- ended (thumbs down).	Yes
(Lee *et al*., 2002)	No	No	NM	NM	NM	No	A new method for unsupervised classification of terrain types and man-made objects using polarimetric Synthetic Aperture Radar (SAR) data	Yes
(van Zyl, 2002)	No	NM	NM	NM	NM	No	The use of an imaging radar polarimeter data for Unsupervised classification of scattering behavior	Yes
(Hegarat-Mascle *et al*., 2002)	No	NM	NM	NM	NM	No	Dempster-Shafer evidence theory may be successfully applied to unsupervised classification in multisource remote sensing.	Yes
(Ferro-Famil *et al*., 2002)	No	NM	NM	NM	NM	No	A new classification scheme for dual frequency polarimetric SAR data sets. A (6×6) polarimetric coherency matrix is defined to simultaneously take into account the full polarimetric information from both images	Yes
(Chaovalit and Zhou, 2005)	No	Yes	EL	Yes	Yes	No	+Machine learning +Semantic orientation	Yes
(Lee *et al*., 2002)	No	No	NM	NM	NM	No	The algorithm estimates the density of each class and is able to model class distributions with non-Gaussian structure	Yes
(Gllavata *et al*., 2004)	No	NM	NM	NM	NM	No	A robust text localization approach.	Yes
This study	Yes	Yes	VL	No	No		A new model for Vietnamese document-level sentiment classification using the CURE Algorithm (CA) in the Cloudera parallel network environment.	Yes

Table 13: Comparisons of our model's positives and negatives with the latest works of the unsupervised classification in [(Turney, 2002; Lee *et al*., 2002; van Zyl, 2002; Hegarat-Mascle *et al*., 2002; Ferro-Famil *et al*., 2002; Chaovalit and Zhou, 2005; Lee *et al*., 2002; Gllavata *et al*., 2004)

Works	Positives	Negatives
(Peter D. Turney, 2002)	In experiments with 410 reviews from Epinions, the algorithm attains an average accuracy of 74%. It appears that movie reviews are difficult to classify, because the whole is not necessarily the sum of the parts; thus the accuracy on movie reviews is about 66%.On the other hand, for banks and automobiles, it seems that the whole is the sum of the parts and the accuracy is 80 to 84%. Travel reviews are an intermediate case.	The limitations of this work include the time required for queries and, for some applications, the level of accuracy that was achieved. The former difficulty will be eliminated by progress in hardware. The latter difficulty might be addressed by using semantic orientation combined with other features in a supervised classification algorithm
(Lee *et al*., 2002)	Significant improvement has been observed in the iteration. The iteration ends when the number of pixels switching classes becomes smaller than a predetermined number or when other criteria are met. The authors observed that the class centers in the entropy-alpha plane are shifted by each iteration. The final class centers in the entropy-alpha plane are useful for class identification of the scattering mechanism associated with each zone. The advantages of this method are the automated	No mention

Table 13: Continue

	classification, and the interpretation of each class based on scattering mechanism. The effectiveness of this algorithm is demonstrated using a JPL/AIRSAR polarimetric SAR image.	
(van Zyl, 2002)	When this algorithm is applied to data acquired over the San Francisco Bay area in California, it classifies scattering by the ocean as being similar to that predicted by the class of odd number of reflections, scattering by the urban area as being similar to that predicted by the class of even number of reflections, and scattering by the Golden Gate Park as being similar to that predicted by the diffuse scattering class. It also classifies the scattering by a lighthouse in the ocean and boats on the ocean surface as being similar to that predicted by the even number of reflections class, making it easy to identify these objects against the background of the surrounding ocean.	No mention
(Hegarat-Mascle *et al*., 2002)	The performance of classification is studied in terms of identification of the different land cover types. Comparing Dempster–Shafer data fusion to two other simple data fusion methods (the concatenated vector and the class subdivision approaches), the authors show that the former generally performs better (e.g., a 20% improvement in the identification rates for corn using L and C band data). On the MAC-Europe campaign data, the best two-data-set fusion results are obtained either using optical and L band SAR images, or multi-band L and C SAR images.	No mention
(Ferro-Famil *et al*., 2002)	The data sets are then classified by an iterative algorithm based on a complex Wishart density function of the 6/spl times/6 matrix. A class number reduction technique is then applied on the 64 resulting clusters to improve the efficiency of the interpretation and representation of each class. An alternative technique is also proposed which introduces the x alternative technique is also information to refine the results of classification to a small number of clusters using the correlation conditional probability of the cross- matrix. These classification schemes are applied to full polarimetric P, L, and C-band SAR images of the Nezer Forest, France, acquired by the NASA/JPL AIRSAR sensor in 1989.	No mention
Chaovalit and Zhou, 2005)	This research investigates movie review mining using two approaches: Machine learning and semantic orientation. The approaches are adapted to the movie review domain for comparison. The results show that the authors' results are comparable to or even better than previous findings. The authors also find that movie review mining is a more challenging application than many other types of review mining. The challenges of the movie review mining lie in that factual information is always mixed with real-life review data and ironic words are used in writing movie reviews.	Future work for improving existing approaches is also suggested.
(Lee *et al*., 2002)	The new algorithm can improve classification accuracy compared with standard Gaussian mixture models. When applied to blind source separation in nonstationary environments, the method can switch automatically between classes, which correspond to contexts with different mixing properties. The algorithm can learn efficient codes for images containing both natural scenes and text. This method shows promise for modeling non-Gaussian structure in high-dimensional data and has many potential applications.	No mention
(Gllavata *et al*., 2004)	In this study, a robust text localization approach is presented, which can automatically detect horizontally aligned text with different sizes, fonts, colors and languages. First, a wavelet transform is applied to the image and the distribution of high-frequency wavelet coefficients is considered to statistically characterize text and non-text areas. Then, the k-means algorithm is used to classify text areas in the image. The detected text areas undergo a projection analysis in order to refine their localization. Finally, a binary segmented text image is generated, to be used as input to an OCR engine. The detection performance of the authors' approach is demonstrated by presenting experimental results for a set of video frames taken from the MPEG-7 video test set.	No mention
This study	The positives and negatives of this survey are illustrated in the Conclusion section.	

Acknowledgement

This survey is funded by Nguyen Tat Thanh University, 300A Nguyen Tat Thanh Street, Ward 13, District 4, Ho Chi Minh City, 702000, Vietnam (Vo Ngoc Phu, vongocphu03hca@gmail.com and vongocphu@ntt.edu.vn)

Author's Contributions

Dr. Vo Thi Ngoc Tran: Built our data sets and "Vo Ngoc Phu" checked them finally and Wrote the draft document of our manuscript.

Dr. Vo Ngoc Phu: Implemented this survey and "Dr.Vo Thi Ngoc Tran" helps "Dr.Vo Ngoc Phu" a lot to perform this study and Checked, fixed and wrote it finally.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

Hoang, V.C.D., D. Dinh, N.L. Nguyen and H.Q. Ngo, 2007. A comparative study on Vietnamese text classification methods. Proceedings of the IEEE International Conference on Computer Science - Research, Innovation and Vision for the Future, (IVF' 07), pp: 267-273.

Agarwal, B. and N. Mittal, 2016a. Machine learning approach for sentiment analysis. Prominent Feature Extraction for Sentiment Analysis.

Agarwal, B. and N. Mittal, 2016b. Semantic orientation-based approach for sentiment analysis. Prominent Feature Extraction for Sentiment Analysis.

Ahmed, S. and A. Danti, 2016. Effective sentimental analysis and opinion mining of web reviews using rule based classifiers. Computat. Intell. Data Min, 1: 171-179. DOI: 10.1007/978-81-322-2734-2_18

Anh, T.T.V. and H.X. Dau, 2014. A crossed-domain sentiment analysis system for the discovery of current careers from social networks. Proceedings of the 5th Symposium on Information and Communication Technology, Dec. 04-05, Hanoi, pp: 226-231. DOI: 10.1145/2676585.2676614

Apache, 2017. http://apache.org

Bach, N.X., P.D. Van, N.D. Tai and T.M. Phuong, 2015. Mining vietnamese comparative sentences for sentiment analysis. Proceedings of the 7th International Conference on Knowledge and Systems Engineering, Oct. 8-10, IEEE Xplore Press, Ho Chi Minh City, Vietnam, pp: 162-167. DOI: 10.1109/KSE.2015.36

Bang, T.S., C. Haruechaiyasak and V. Sornlertlamvanich, 2015. Vietnamese sentiment analysis based on term feature selection approach. Proceedings of The Tenth International Conference on Knowledge, Information and Creativity Support Systems (CSS' 15), Phuket, Thailand.

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

Chaovalit, P. and L. Zhou, 2005. Movie review mining: A comparison between supervised and unsupervised classification approaches. Proceedings of the 38th Annual Hawaii International Conference on System Sciences, Jan. 6-6, IEEE Xplore Press, Big Island, pp: 112-112. DOI: 10.1109/HICSS.2005.445

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

Duyen, N.T., N.X. Bach and T.M. Phuong, 2014. An empirical study on sentiment analysis for Vietnamese. Proceedings of the International Conference on Advanced Technologies for Communications, Oct. 15-17, IEEE Xplore Press, Hanoi, Vietnam, pp: 309-314. DOI: 10.1109/ATC.2014.7043403

Ertöz, L., M. Steinbach and V. Kumar, 2002. A new shared nearest neighbor clustering algorithm and its applications. Proceedings of the Workshop on Clustering High Dimensional Data and its Applications at 2nd SIAM International Conference on Data Mining, (CDM' 02), pp: 105-15.

Ferro-Famil, L., E. Pottier and J.S. Lee, 2002. Unsupervised classification of multifrequency and fully polarimetric SAR images based on the H/A/Alpha-Wishart classifier. IEEE Trans. Geosci. Remote Sens., 39: 2332-2342. DOI: 10.1109/36.964969

Gllavata, J., R. Ewerth and B. Freisleben, 2004. Text detection in images based on unsupervised classification of high-frequency wavelet coefficients. Proceedings of the 17th International Conference on Pattern Recognition, Aug. 26-26, IEEE Xplore Press, Cambridge, pp: 425-428. DOI: 10.1109/ICPR.2004.1334146

Guha, S., R. Rastogi and K. Shim, 1998. CURE: An efficient clustering algorithm for large databases. Proceedings of the International Conference on Management of Data, Jun. 01-04, Seattle, Washington, pp: 73-84. DOI: 10.1145/276304.276312

Ha, Q.T., T.T. Vu, H.T. Pham and C.T. Luu, 2011. An upgrading feature-based opinion mining model on vietnamese product reviews. Proceedings of the 7th International Conference on Active Media Technology, Sept. 7-9, Lanzhou, China, pp: 173-185. DOI: 10.1007/978-3-642-23620-4_21

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

Hegarat-Mascle, S.L. I. Bloch and D. Vidal-Madjar, 2002. Application of dempster-shafer evidence theory to unsupervised classification in multisource remote sensing. IEEE Trans. Geosci. Remote Sens., 35: 1018-1031. DOI: 10.1109/36.602544

Kaya, M. and R. Alhajj, 2005. Genetic algorithm based framework for mining fuzzy association rules. Fuzzy Sets Syst., 152: 587-601.

Kieu, B.T. and S.B. Pham, 2010. Sentiment Analysis for Vietnamese. Proceedings of the 2nd International Conference on Knowledge and Systems Engineering, Oct. 7-9, IEEE Xplore Press, Hanoi, Vietnam, pp: 152-157. DOI: 10.1109/KSE.2010.33

Hoanh-Su, L., H.S., J.H. Lee and H.K. Lee, 2015. Applying machine learning to classify sentiment text for vietnamese language on social network data. Korea Society Manage. Inform. Syst.

Le, H.S., T.V. Le and T.V. Pham, 2015. Aspect analysis for opinion mining of vietnamese text. Proceedings of the International Conference on Advanced Computing and Applications, Nov. 23-25, IEEE Xplore Press, Ho Chi Minh City. DOI: 10.1109/ACOMP.2015.21

Le, P.H., H.M.T. Nguyen, A. Roussanaly and V.T. Ho, 2008. A hybrid approach to word segmentation of Vietnamese texts. Proceedings of the 2nd International Conference on Language and Automata Theory and Applications, Mar. 13-19, Tarragona, Spain, pp: 240-249. DOI: 10.1007/978-3-540-88282-4_23

Lee, J.S., M.R. Grunes, T.L. Ainsworth and L.J. Du, 2002a. Unsupervised classification using polarimetric decomposition and the complex Wishart classifier. IEEE Trans. Geosci. Remote Sens., 37: 2249-2258.

Lee, T.W., M.S. Lewicki and T.J. Sejnowski, 2002b. ICA mixture models for unsupervised classification of non-Gaussian classes and automatic context switching in blind signal separation. IEEE Trans. Pattern Anal. Machine Intelligence, 22: 1078-1089.

Manek, A.S., P.D. Shenoy, MC. Mohan and K.R. Venugopal, 2016. Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. World Wide Web, Print.

Nguyen, D.Q., T. Vu and S.B. Pham, 2014. Sentiment classification on polarity reviews: An empirical study using rating-based features. Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, (SMA' 14), pp: 128-135.

Nguyen, G.S., X. Gao and P. Andreae, 2009. Vietnamese document representation and classification. Proceedings of the 22nd Australasian Joint Conference on Advances in Artificial Intelligence, Dec. 01-04, Melbourne, pp: 577-586. DOI: 10.1007/978-3-642-10439-8_58

Nguyen, H.N., T.V. Le, H.S. Le and T.V. Pham, 2014. Domain specific sentiment dictionary for opinion mining of vietnamese text. Proceedings of the International Workshop on Multi-disciplinary Trends in Artificial Intelligence, Dec. 08-10, Bangalore, India, pp: 136-148. DOI: 10.1007/978-3-319-13365-2_13

Nian-Yun, S., Z. Jin-Ming and C. Xi, 2009. CURE algorithm-based inspection of duplicated records. Comput. Eng., 35: 56-58.

Phan, D.H. and T.D. Cao, 2014. Applying skip-gram word estimation and SVM-based classification for opinion mining Vietnamese food places text reviews. Proceedings of the 5th Symposium on Information and Communication Technology, Dec. 04-05, Hanoi, pp: 232-239. DOI: 10.1145/2676585.2676606

Phu, V.N. and P.T. Tuoi, 2014. Sentiment classification using Enhanced Contextual Valence Shifters. Proceedings of the International Conference onAsian Language Processing, Oct. 20-22, IEEE Xplore Press, Kuching, Malaysia, pp: 224-229. DOI: 10.1109/IALP.2014.6973485

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

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

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. Int. J. Artificial Intell. Rev. DOI: 10.1007/s10462-017-9538-6, pp 1-67

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

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

Phu, V.N., V.T.N. Chau, V.T.N. Tran, N.D. Dat and T.A. Nguyen, 2017b. STING algorithm used English sentiment classification in a parallel environment. Int. J. Pattern Recognition Artificial Intell., 31: 30- 30.
DOI: 10.1142/S0218001417500215

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

Rani, Y., Manju and H. Rohil, 2014. Comparative Analysis of BIRCH and CURE hierarchical clustering algorithm using weka 3.6.9. SIJ Tran. Comput. Sci. Eng. Applic.

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

Trinh, S., L. Nguyen, M. Vo and P. Do, 2016. Lexicon-based sentiment analysis of facebook comments in vietnamese language. Recent Developments in Intelligent Information and Database Systems, 263-276.

Turney, P.D., 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, (ACL' 02), pp: 417-424.

van Zyl, J.J., 2002. Unsupervised classification of scattering behavior using radar polarimetry data. IEEE Trans. Geosci. Remote Sens., 27: 36-45.
DOI: 10.1109/36.20273

Vu, X.S. and S.B. Park, 2014. Construction of vietnamese sentiwordnet by using vietnamese dictionary. Proceedings of the 40th Conference of the Korea Information Processing Society, (IPS' 14), South Korea, pp: 745-748.

Yan-Hua, S., L. Jie and L. Jian, 2011. Network users behavior analysis based on CURE algorithm. J. Comput. Technol. Dev.

Ying, W., L. Renwang, L. Bin and Z. Zhile, 2008. Costume sales forecasting model based on CURE algorithm and C4.5 decision tree. J. Textile Res.