Appl Intell (2017) 46:717–738
DOI 10.1007/s10489-016-0858-z

Fuzzy C-means for english sentiment classification in a distributed system

Vo Ngoc Phu[1,2] · **Nguyen Duy Dat**[3] · **Vo Thi Ngoc Tran**[4] · **Vo Thi Ngoc Chau**[5] · **Tuan A. Nguyen**[6]

Published online: 5 November 2016

Abstract Sentiment classification plays a significant role in everyday life, in political activities, in activities relating to commodity production, and commercial activities. Finding a solution for the accurate and timely classification of emotions is a challenging task. In this research, we propose a new model for big data sentiment classification in the parallel network environment. Our proposed model uses the Fuzzy C-Means (FCM) method for English sentiment classification with Hadoop MAP (M) /REDUCE (R) in Cloudera. Cloudera is a parallel network environment. Our proposed model can classify the sentiments of millions of English documents in the parallel network environment. We tested our model using the testing data set (which comprised 25,000 English reviews, 12,500 being positive and 12,500 negative) and achieved 60.2 % accuracy. Our English training data set has 60,000 English sentences, comprising 30,000 positive English sentences and 30,000 negative English sentences.

Keywords Sentiment classification · English sentiment classification · Opinion mining · English document opinion mining · Fuzzy C-Means · FCM · Cloudera · Parallel environment · Parallel network · Parallel network environment · Distributed system

✉ Vo Ngoc Phu
vongocphu@tdt.edu.vn

Nguyen Duy Dat
duydatspk@gmail.com

Vo Thi Ngoc Tran
vtntran@hcmut.edu.vn

Vo Thi Ngoc Chau
chauvtn@cse.hcmut.edu.vn; chauvtn@hcmut.edu.vn;
chauvtn2003@gmail.com

Tuan A. Nguyen
tuanna@uit.edu.vn

1 Division of Computational Mathematics and Engineering, Institute for Computational Science, Ton Duc Thang University, Ho Chi Minh City, Vietnam

2 Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam

3 Faculty of Information Technology, Ly Tu Trong Technical College, Ho Chi Minh City, Vietnam

4 School of Industrial Management (SIM), Ho Chi Minh City University of Technology, Vietnam National University, Ho Chi Minh City, Vietnam

5 Computer Science & Engineering (CSE), Ho Chi Minh City University of Technology, Vietnam National University, Ho Chi Minh City, Vietnam

6 Faculty of Computer Networks and Communications, University of Information Technology, Vietnam National University of Hochiminh City, Linh Trung Ward, Thu Duc District, Ho Chi Minh City, Vietnam

1 Introduction

Sentiment classification plays a significant role in everyday life, in political activities, in activities relating to commodity production, and commercial activities. Finding a solution for the accurate and timely classification of emotion is a challenging task.

Data clustering is the process of putting objects into classes where the objects are similar. A cluster is a set of data objects which are similar in scope, but are not similar to objects in other clusters. Number of data clusters are clustered which can be identified firstly following experience or can be automatically identified in the clustering method.

This technique clusters a set of n data object vectors X = {x1,x2, ... ,xn}⊂ Rs into c fuzzy clusters based on calculating the minimum objective function to measure the quality of clustering and find the cluster centers in each cluster to minimize the cost measurement function.A fuzzy set is a set in which each x basic member is assigned a $\mu(\xi)$ real value in [0, 1] to display the dependency measure of this member in the set. When the dependency measure equals 0, the basic member does not belong to the set; but if the dependency measure equals 1, the basic member belongs to the set completely. Therefore, a fuzzy set is a set of (x, μ(x)) pairs.

Fuzzy C-Means (FCM) is also a method of clustering which allows one element to belong to two or more clusters. It is often used to cluster the data but seldom used for data classification.

We suggest many basic principles of our model to classify the opinions (positive, negative, neutral) expressed in the English documents in the English testing data set, based on the large number of English sentences in the English training data set. The principle underpinning our proposed model, which uses clustering techniques to classify the semantics of the English documents are as follows: assuming that an English document contains n English sentences, the English document has a positive polarity if the number of English sentences clustered into the 30,000 positive English sentences of the training data set is greater than the number of English sentences clustered into the 30,000 negative English sentences of the training data set. Conversely, the English document has a negative polarity if the number of English sentences clustered into the 30,000 positive English sentences of the training data set is less than the number of English sentences clustered into the 30,000 negative English sentences of the training data set. Finally, the English document has a neutral polarity if the number of English sentences clustered into the 30000 positive English sentences of the training data set is equal to the number of English sentences clustered into the 30,000 negative English sentences of the training data set.

Based on these principles, we implement our proposed model in the Cloudera parallel network environment. Our model uses FCM combined with Hadoop Map (M)/Reduce (R) to classify the sentiments (positive, negative, neutral) of one English document in the English testing data set into either positive polarity, negative polarity or neutral polarity in the Cloudera parallel network environment.

To implement this study, we use the basis Fuzzy C-Means algorithm (the core basis of the Fuzzy C-Means algorithm) presented in [8–24]. There are also many studies which use the FCM in semantic classification (opinion mining, sentiment analysis) but there is not much work which uses FCM for sentiment analysis with the aforementioned principles of our proposed model.

FCM is a clustering technique in the data mining field and it has been applied in the natural language processing field where we have had many difficulties and it has taken a long time to implement this research. There are many advantages of FCM, such as: it is unsupervised, it always converges; it provides membership values which are useful for interpretation; it is flexible with respect to the distance used; and if some of the membership values are known, this can be incorporated into the numerical optimization. There are several disadvantages of FCM as follows: long computational time; sensitivity to the initial guess (speed, local minima); and sensitivity to noise - one expects low (or even no) membership degree for outliers (noisy points).

In addition, based on the work related to FCM and the sentiment analysis of big data in [8–24], there are not studies which use FCM for big data in sentiment classification. We use FCM in our model opinion mining in big data, although our English data set in this work is a small English testing data set with 25,000 English document in each testing data set.

In addition, based on many works related to FCM in the parallel system (or FCM in the distributed system) in [25–27], many studies relate to parallel systems or distributed systems, in [28–42], FCM used research for sentiment classification in [43–50], and many studies in the world, there is not any study related to FCM for sentiment classification in parallel system but our model uses FMC for semantic analysis in the distributed system.

Many studies, such as [2–56], use Hadoop Map (M)/Reduce (R), and Cloudera; Vector Space Models (VSM); FCM; FCM in parallel systems (distributed systems)sentiment classification and big data. However, to the best of our knowledge, no studies use all of them. Our proposed model uses all of these.

Finally, we build many FCM-related algorithms in our new model based on the basic FCM in the Cloudera distributed system with Hadoop Map (M) /Reduce (R) and these algorithms have not been used in any other study.

This study comprises six sections: Section 1 is the introduction; Section 2 discusses the related work on Fuzzy C-Means (FCM), Hadoop, Cloudera, etc.; Section 3 discusses the English data set; Section 4 overviews the methodology of our proposed model; Section 5 describes the experiment and Section 6 provides the conclusion.

2 Related work

In this section, we overview several studies related to Fuzzy C-Means (FCM), the Vector Space Model, Hadoop, Cloudera, etc.

There are many studies which are related to the Vector Space Model [2–4]. First of all, the authors of [2] transfer all English sentences into many factors which are used in VSM algorithm. In this research, the authors examine the Vector Space Model, an information retrieval technique and its variations. The rapid growth of the World Wide Web and the abundance of documents and different forms of information available on it, has resulted in the need for better information retrieval techniques. The Vector Space Model 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 multi-dimensional space, and allows decisions to be made as to which documents are similar to each other and to the queries fired. This work also explains the existing variations of the VSM and proposes a new variation that should be considered [3]. In the text classification task, one of the main problems is to choose which features give the best results. Various features can be used such as words, n-grams, syntactic n-grams of various types (POS tags, dependency relations, mixed, etc.), or a combination of these features. Also, algorithms to reduce the dimensionality of these sets of features can be applied, such as Latent Dirichlet Allocation (LDA). In this research, the authors consider the multi-label text classification task 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 the stop words. The authors also tried several combinations of features, like bi-grams and uni-grams. The authors also experimented by adding LDA results into Vector Space Models as new features. These latter experiments obtained the best results [4]. KNN and SVM 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 the statistical estimation of the importance of a word for a specific categorization problem. This method also has the benefit of making feature selection implicit, since useless features of the categorization problem considered are assigned a very small weight. Extensive experiments reported in the research show that this new weighting method significantly improves classification accuracy as measured on many categorization tasks.

Many studies such as [5–7] are related to the implementation of algorithms and applications in the parallel network environment. Hadoop is an Apache-based framework which is used to handle large data sets on clusters consisting of multiple computers, using the Map and Reduce programming model. The two main projects of Hadoop are the Hadoop Distributed File System (HDFS) and Hadoop M/R (Hadoop Map/Reduce). Hadoop M/R allows engineers to program for writing applications for the parallel processing of large datasets on clusters consisting of multiple computers. An M/R task has two main components: (1) Map and (2) Reduce. This framework splits the input data into chunks which multiple Map tasks can handle as a separate data partition in parallel. The outputs of the map tasks are gathered and processed by the Reduce task which is ordered. The inputs and outputs of each M/R are stored in HDFS because the Map tasks and the Reduce tasks are performed on the pair (key, value), and the formatted input and output formats will be the pair (key, value) [7]. Cloudera, 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 in November 2015 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 currently. Cloudera's customers are able to efficiently capture, store, process and analyze vast amounts of data, empowering them to use advanced analytics to drive business decisions quickly, flexibly and at a lower cost than has been possible before. To ensure Cloudera's customers are successful, it offers comprehensive support, training and professional services.

There are many studies, such as [8–24] which are related to the FCM algorithm.

Many studies are related to FCM in parallel systems (or FCM in distributed systems) such as the work in [25–27].

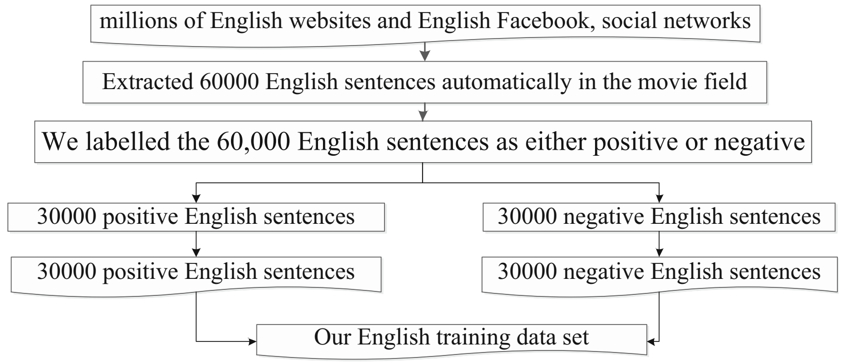


Fig. 1 Our english training data set

Many studies, such as [28–42] are related to parallel systems or distributed systems.

Research using FCM for sentiment classification can be found in [43–50].

The latest research on sentiment classification can be found in [51–54, 56].

3 Data set

The English training data set includes 60,000 English sentences in the movie field, of which 30,000 are positive English sentences and 30,000 are negative English sentences.

All English sentences in our English training data set have been automatically extracted from Facebook and websites in social networks, after which we labeled them as either positive or negative. Figure 1 is the English training data set of this model.

We used a publicly available large data set of movie reviews from the Internet Movie Database [1]. This English data set comprises a testing data setwhich we refer to as the first testing data set and also a training data setwhich we refer to as the second testing data set. Both our first testing data set and our second testing data set contain 25,000 English documents, each with 12,500 positive English movie reviews and 12,500 negative English movie

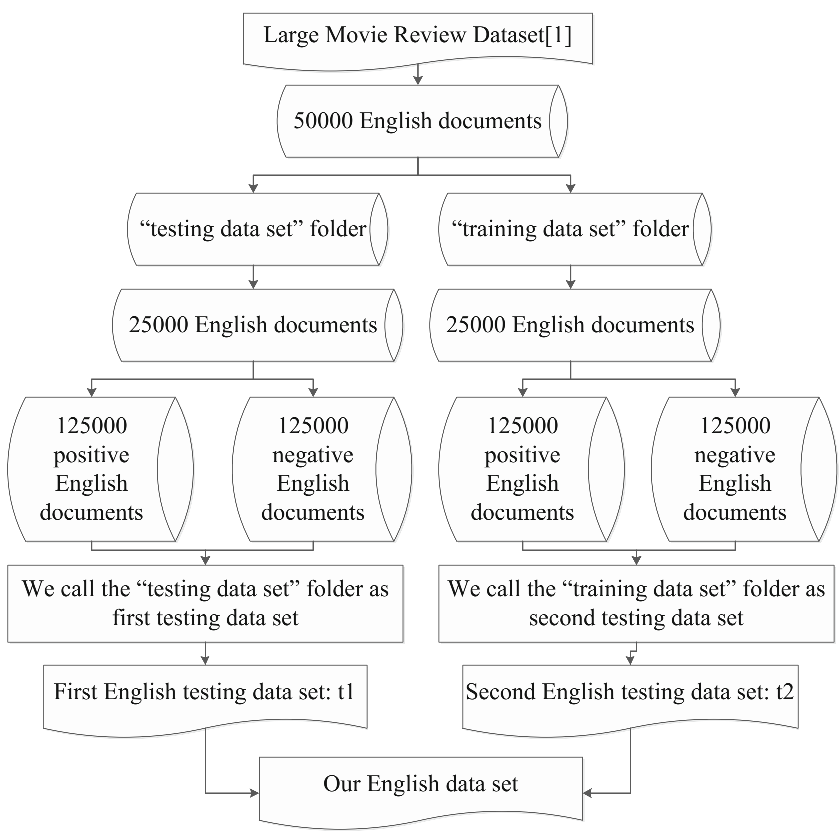


Fig. 2 Our english testing data set

reviews. Figure 2 is the English testing data set of this model.

4 Methodology

The methodology section comprises two parts: the semantic classification of the 25,000 English documents in the testing t1 and the 25,000 English documents of the testing t2 on the sequential environment is presented in the first part and the sentiment classification of the 25,000 English reviews of the testing t1 and the 25,000 English reviews of the testing t2 in the parallel network environment is presented in the second part.

In the English training data set, there are two clusters: the first, called the positive cluster, contains 30,000 positive English sentences and the second, called the negative cluster, contains 30,000 negative English sentences. All English sentences in both the first cluster and the second cluster have undergone word segmentation and stop-word removal after which they are transferred into vectors (vector representation). The 30,000 positive English sentences in the positive cluster are transferred into the 30,000 positive vectors, called the positive vector group (or the positive vector cluster). The 30,000 negative English sentences in the negative cluster are transferred into 30,000 negative vectors, called the negative vector group (or the negative vector cluster). Therefore, the English training data set includes the positive vector group (or the positive vector cluster) and the negative vector group (or the negative vector cluster) [2–4]. 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 VSM is a way of representing documents through the words that they contain. Vector space modeling places terms, documents, and queries in a term-document space so 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 transferred all the English sentences in the training data set into vectors similar to VSM [2–4].

4.1 Fuzzy C-means algorithm in the sequential environment

Figure 3 illustrates how sentiment classification is undertaken in the sequential environment.

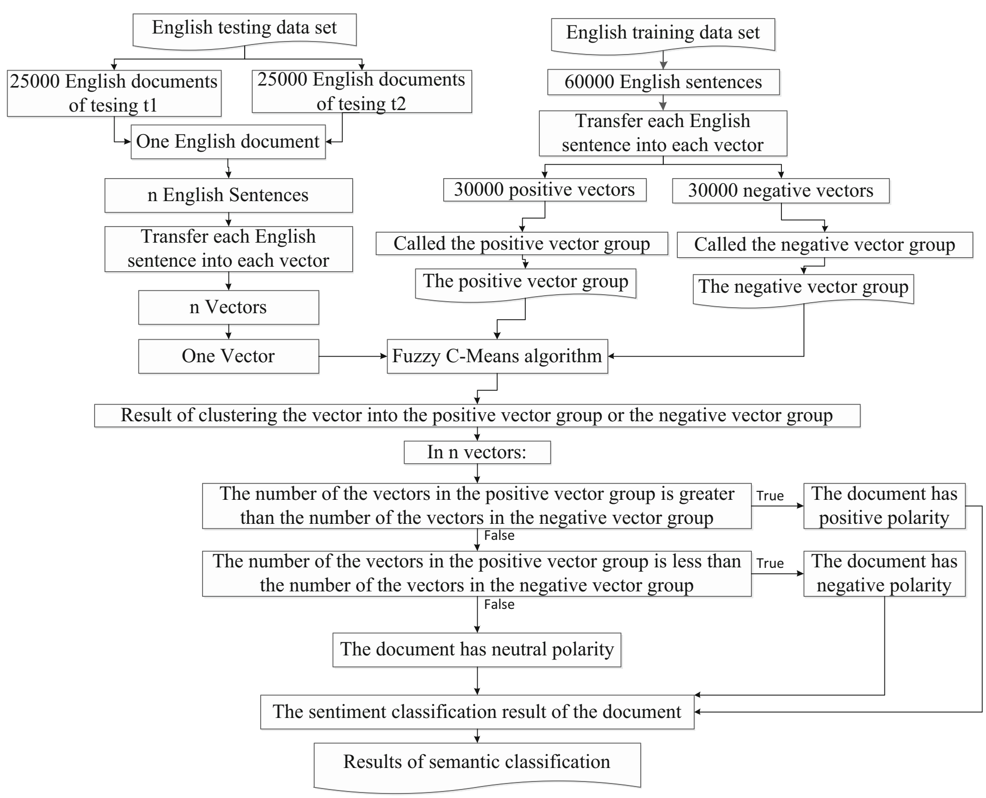


Fig. 3 Fuzzy c-means algorithm in the sequential environment

With each English document in the English testing data set, we assume that each English document has n English sentences and we transfer the n English sentences into n vectors similar to VSM [2–4]. Thus, the document has n vectors. For each vector of the n vectors, we use FCM to cluster the vector into the positive vector group or the negative vector group in the sequential environment. According to [8–17], we implement the FCM algorithm which is enhanced to be able to classify the sentiment of the English sentences.

The total all the fuzzy partitions which have c clusters of N objects in D is calculated as follows:

$$E_{fc} = \left\{ U \in R_{cN} | \underset{1 \le i \le c \wedge 1 \le k \le N}{\forall} u_{ik} \in [0,1], \sum_{i=1}^{c} u_{ik} = 1, 0 < \sum_{k=1}^{N} u_{ik} < N \right\}$$

Minimize the objective function:

$$J_m(U, V) = \sum_{i=1}^{c} \sum_{k=1}^{N} (u_{ik})^m d_{ik}^2$$

$$d_{ik}^2 = |x_k - v_i|_A$$

V = [v1, v2, ..., vc] is a matrix which represents the center object values of the cluster. A matrix is a positive finite. m is the exponent weight in [1, ∞).

The objective function reaches a minimum value if and only if:

$$\underset{1 \le k \le N}{\forall} I_k = \{i | 1 \le \mathrm{i} \le \ \mathrm{c};\ d_{\mathrm{ik}} = 0\}$$

$$\underset{1 \le i \le c \wedge 1 \le k \le N}{\forall} u_{ik} = \begin{cases} (d_{ik})^{\frac{2}{1-m}} \left[\sum_{j=1}^{c} (d_{ik})^{\frac{2}{1-m}} \right]^{-1} \\ \begin{cases} 0, i \notin \\ \sum_{i \in I_k} u_{ik} = 1, i \in I_k \end{cases} \end{cases} \tag{1}$$

$$\underset{1 \le i \le c}{\forall} v_i = \frac{\sum_{k=1}^{N} (u_{ik})^m x_k}{\sum_{k=1}^{N} (u_{ik})^m} \tag{2}$$

The FCM algorithm comprises the following steps:

Algorithm 1 Fuzzy c-means algorithm

Input: One vector of the document of the testing data set; the positive vector group of the training data set and the negative vector group of the training data set.
Output: The result of clustering the vector into the positive vector group or the negative vector group.
Begin
Step 1: Enter values for the two parameters: c (1 <c <N), m and initializing the sample matrix
Step 2: Repeat
 2.1 j=j+1;
 2.2 Calculating fuzzy partition matrix Uj following formula (1)
 2.3 Updating centers V(j) = [v_1(j), v_2(j), ..., v_c(j)] basing on (2) và U_j matrix;
Step 3: Untill ($||U_{(j+1)} - U_{(j)}||_F \le \xi$);
Step 4: Performing results of the clusters.
End;

with $||U||_F^2 = \sum_i \sum_k U_{ik}^2$

With the clustering results of the n vectors of the documents in the testing data set, the document has a positive sentiment if the number of vectors in the n vectors is greater than the number of vectors in the n vectors. The document has a negative sentiment if the number of vectors in the n vectors is less than the number of vectors in the n vectors. The document has a neutral sentiment if the number of vectors in the n vectors is equal to the number of vectors in the n vectors.

4.2 Fuzzy C-means (FCM) in the parallel network environment

Figure 4 illustrates how semantic classification is undertaken in a parallel network environment.

We transfer the 60,000 English sentences in the training data set into the 60,000 vectors using Hadoop Map (M)/Reduce (R) in the Cloudera parallel network environment to shorten the execution time of this task. Figure 5 overviews the process of transferring each English sentence into one vector in the Cloudera networkenvironment.

Transferring each English sentence into one vector in the Cloudera network environment involves two phases: Map (M) phases and Reduce (R) phases. The input of the Map phase is one English sentence and the output of the Map phase are the many components of a vector which correspond to the sentence. In the Map phase of Cloudera, we transfer the sentence into one vector similar to VSM [2–4]. The input of the Reduce phase is the output of the Map phase, which is many components of a vector. The output of the Reduce phase is a vector which corresponds to the

Fig. 4 Fuzzy c-means algorithm in the parallel network environment

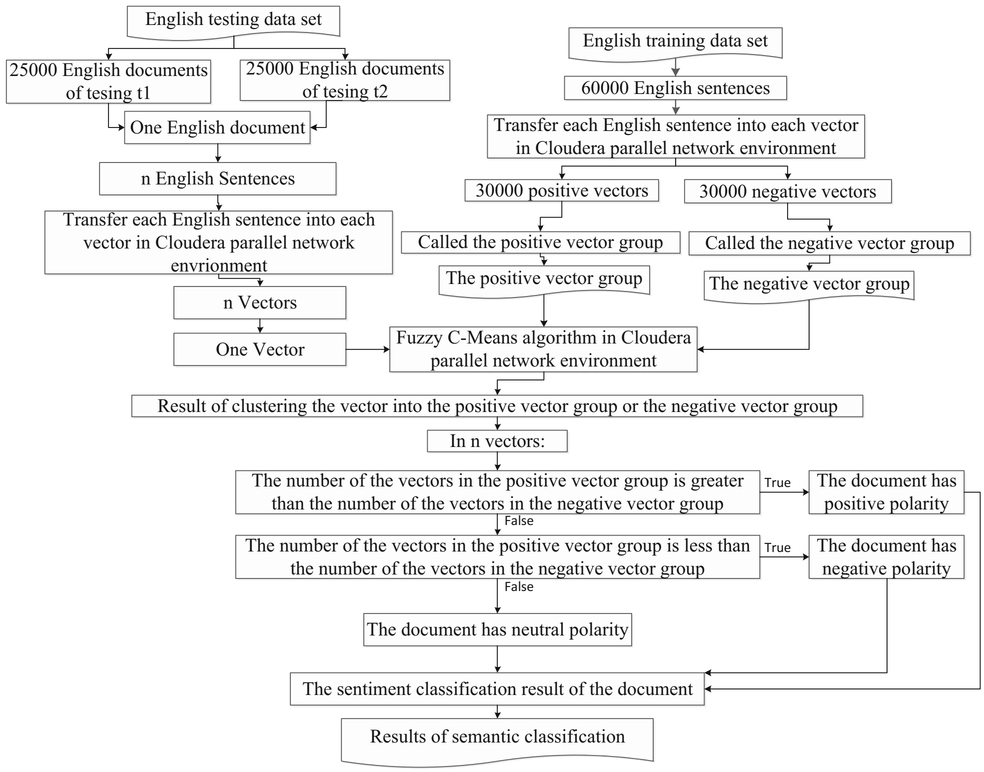


Fig. 5 Overview of the process of transforming each english sentence into one vector in Cloudera

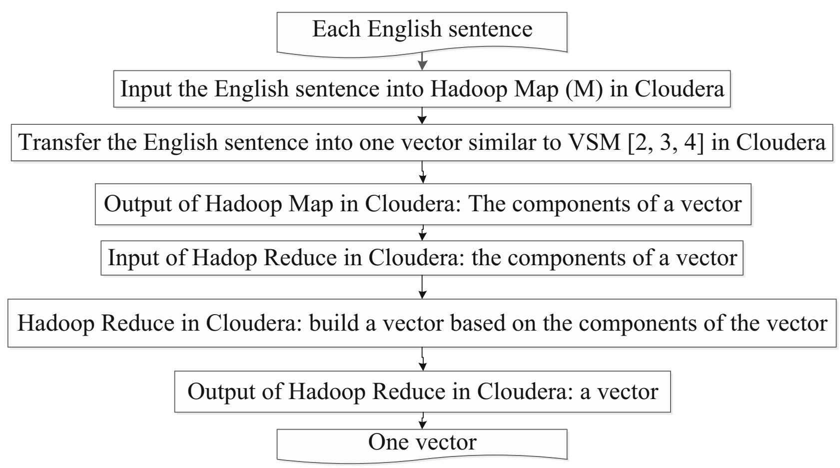
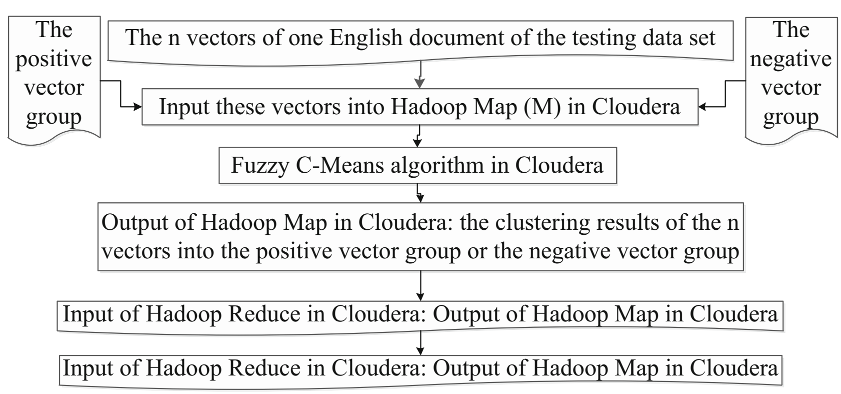


Fig. 6 Overview of fuzzy c-means in Hadoop map (M) in Cloudera



sentence. In the Reduce phase of Cloudera, these components of the vector are built into one vector.

Each English document in the testing data set contains n English sentences. We transfer each English sentence in the n English sentences into one vector similar to the process shown in Fig. 5. Hence, the document also has n vectors.

FCM in the Cloudera parallel network environment comprises two phases: the first phase is the Hadoop Map (M) phase in Cloudera and the second phase is the Hadoop Reduce (R) phase in Cloudera. In the Map phase, the input is the n vectors of one English document (which have been classified) into either the positive vector group or the negative vector group; and the output is the clustering results of the n vectors of the document into either the positive vector group or the negative vector group. In the Reduce phase, the input is the output of the Map phase and this input is the clustering results of the n vectors of the document into either the positive vector group or the negative vector group; and the output is the sentiment classification result of the document as either having positive polarity, negative polarity, or neutral polarity. In the Reduce phase, the English document is classified as having a positive sentiment if the number of vectors of the n vectors in the positive vector group is greater than the number of vectors of the n vectors in the negative vector group; the English document is classified as having a negative sentiment if the number of vectors of the n vectors in the positive vector group is less than the number of vectors of the n vectors in the negative vector group; and the English document is classified as having a neutral sentiment if the number of vectors of the n vectors in the positive vector group is equal to the number of vectors of the n vectors in the negative vector group.

4.2.1 Hadoop Map (M)

Figure 6 illustrates the Hadoop Map phase.

Similar to [7–17], we propose FCM as follows:

Algorithm 2 Fuzzy c-means algorithm

Input: The n vectors of the document of the testing data set; the positive vector group of the training data set and the negative vector group of the training data set.

Output: the result of clustering The n vectors into the positive vector group or the negative vector group.

Begin

Step 0: With each vector in the n vectors, repeat:

Step 1: Enter values for the two parameters: c ($1 < c < N$), m and initializing the sample matrix

Step 2: Repeat

  2.1 j=j+1;

  2.2 Calculating fuzzy partition matrix Uj following formula (1)

  2.3 Updating centers $V(j) = [v_1(j), v_2(j), \ldots, v_c(j)]$ basing on (2) và U_j matrix;

Step 3: Until ($\|U_{(j+1)} - U_{(j)}\|_F \leq \xi$);

Step 4: Performing results of the clusters.

Step 5: End Step 0;

End;

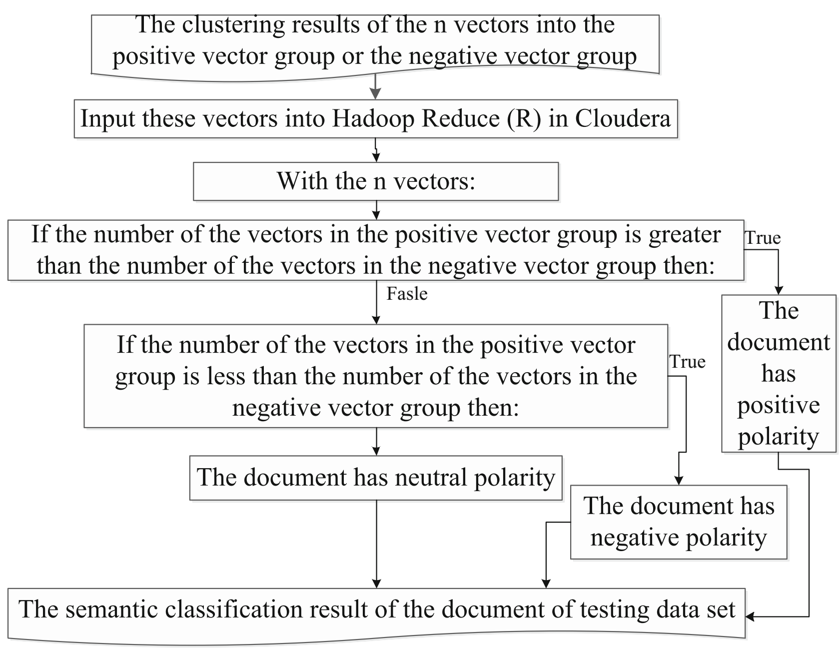


Fig. 7 Overview of Hadoop reduce (R) in Cloudera

Table 1 The results of the 25,000 english documents in testing data set t1

	Testing dataset t1	Correct classification	Incorrect classification
Negative	12500	7523	4977
Positive	12500	7527	4973
Summary	25000	15050	9950

4.2.2 Hadoop Reduce (R)

Figure 7 illustrates the Hadoop Reduce phase.

5 Experiment

We used measures such as accuracy (A) to calculate the accuracy of the results of sentiment classification.

The Java programming language was used to save the data sets in order to implement our proposed model to classify the 25,000 English documents in testing data set t1 and the 25,000 English documents of testing data set t2.

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

The sequential environment in this research comprises one node (one server). The Java language is used to program FCM. 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 implement FCM in the Cloudera parallel network environment - this Cloudera system comprises four nodes (four servers). The Java language is used to program the application of the FCM in 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 of the four nodes is Cloudera. All nodes have the same configuration.

The results of the sentiment classification of the 25,000 English documents in testing data set t1 are presented in Table 1.

The results of the sentiment classification of the 25,000 English documents in testing data set t2 are presented in Table 2.

The accuracy of the sentiment classification of the 25,000 English documents in testing dataset t1 is shown in Table 3.

The accuracy of the sentiment classification of the 25,000 English documents in testing dataset t2 is shown in Table 4.

6 Conclusion

Although our proposed model was tested on an English data set, it can also be applied to many other languages. In this paper, our model was tested on the 25,000 English documents in the testing data set t1 and the 25,000 English documents in the testing data set t2 which are small data sets. However, our model can be applied to a big data set containing millions of English documents in a very short time.

In this work, we proposed a new model to classify the sentiments of English documents using the Fuzzy C-Means Algorithm (FCM) with Hadoop Map (M) /Reduce (R) in the Cloudera parallel network environment. The experiment results show that our proposed model achieves 60.2 % and 59.8 % accuracy of the English documents. Currently, there is a paucity of research which shows that clustering methods can be used to classify data. Our research shows that

Table 2 The results of the 25,000 english documents in testing data set t2

	Testing dataset t2	Correct classification	Incorrect classification
Negative	12500	7437	5063
Positive	12500	7363	5137
Summary	25000	14800	10200

Table 3 The accuracy of our proposed model for the sentiment classification of the 25,000 english documents in testing data set t1

Model	Class	Accuracy	FCM Algorithm in the sequential environment	FCM Algorithm in the Cloudera distributed system
Our proposed model	Negative	60.2 %	Average time of the classification: 150,590 seconds/25,000 English documents	Average time of the classification: 37,659 seconds/25000 English documents
	Positive			

clustering methods are able to classify data and in particular, they are useful for sentiment classification for text.

As shown in Table 3, the average time taken for the sentiment classification of the 25,000 English documents in testing data set t1 using the FCM algorithm in the sequential environment is 150,590 seconds, which is greater than the average time taken for the sentiment classification of the 25,000 English documents using FCM in the Cloudera parallel network environment, which is 37,659 seconds.

As shown in Table 4, the average time taken for the sentiment classification of the 25,000 English documents in testing data set t2 using the FCM algorithm in the sequential environment is 151590 seconds, which is greater than the average time taken for the sentiment classification of the 25,000 English documents in testing data set t2 using FCM in the Cloudera parallel network environment, which is 37875 seconds.

The execution time of the FCM in Cloudera is dependent on the performance of the Cloudera parallel system and is also dependent on the performance of each server on the Cloudera system.

The principles underpinning our proposed model for classifying the sentiment (positive, negative, neutral) of the English documents in the English testing data set in the sequential environment, based on the numerous English sentences in the English training data set are similar to the principles underpinning our proposed model for classifying the sentiment (positive, negative, neutral) of the English documents in English testing data set in the distributed environment, based on the numerous English sentences in English training data set.

The FCM of our proposed model in the sequential environment is different from the FCM of our proposed model in the parallel environment. We built many algorithms related to the FCM to implement our model in the distributed system.

The execution time of our model in the parallel environment is less than the execution time of our model in the sequential environment. The execution of our model in the distributed system is shorter if the performance in the distributed system is longer.

In addition, the execution time of any model is also dependent on the algorithms. For example, using the same algorithms, different systems perform differently and have different execution times. Using the same system with the same performance, different algorithms may have different execution times.

Our survey has many advantages and disadvantages. The advantages are: it processes big data involving millions of English documents; the execution time of our model to conduct sentiment on big data is short, etc. However, the disadvantages are: it takes a long time to implement and it is costly to build the algorithms of the model in the distributed system.

Table 4 The accuracy of our proposed model for the sentiment classification of the 25,000 english documents in testing data set t2

Model	Class	Accuracy	FCM Algorithm in the sequential environment	FCM Algorithm in the Cloudera distributed system
Our proposed model	Negative	59.8 %	Average time of the classification: 151,590 seconds/25,000 English documents	Average time of the classification: 37,875 seconds/25,000 English documents
	Positive			

Table 5 Comparison of our model's results with the work in [2–4]

Studies	FCM	CT	SC	PNS	SD	DT	L	VSM
[2]	No	No	No	No	Yes	No	EL	Yes
Model/method in [2]	Examining the Vector Space Model, an information retrieval technique and its variations							
Summary of the work in [2]	In this work, the authors examine the Vector Space Model, an information retrieval technique and its variations. The rapid growth of World Wide Web and the abundance of documents and different forms of information available on it has resulted in the need for better information retrieval techniques. The Vector Space Model is an algebraic model used for information retrieval. It represents a natural language document in a formal manner with the use 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. It also explains the existing variations of the VSM and proposes a new variation that should be considered							
[3]	No	No	Yes	No	Yes	No	EL	Yes
Model/method in [3]	+Latent Dirichlet Allocation (LDA). +Multi-label text classification task and applies various feature sets. +Several combinations of features, like bi-grams and uni-grams.							
Summary of the work in [3]	+ The authors proposed the use of a VSM complement based on the LDA, adding the LDA results as features to Vector Space Models							
[4]	No	No	Yes	No	Yes	Yes	Yes	Yes
Model/method in [4]	KNN and SVM are two machine learning approaches to text categorization (TC) based on the Vector Space Model							
Summary of the work in [4]	+In this model, borrowed from information retrieval, documents are represented as vectors where each component is associated with a particular word from the vocabulary. Traditionally, each component value is assigned a weight 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 work, the authors introduce a new weighting method based on the statistical estimation of the importance of a word for a specific categorization problem. This method also has the benefit of making feature selection implicit, since the useless features of the categorization problem are assigned a very small weight. Extensive experiments reported in the work show that this new weighting method significantly improves classification accuracy as measured on many categorization tasks. + In this work, the authors presented a new method to weight features in the vector-space model for text categorization by leveraging the categorization task. The most commonly used method is TFIDF, which is unsupervised							
Our work	Yes	Yes	Yes	Yes	Yes	Yes	EL	Yes
Model/Method in our work	Fuzzy C-Means algorithm for English sentiment classification in the Cloudera distributed system							
Summary of our work	Firstly, we use the Fuzzy C-Means algorithm (FCM) to classify English documents as having either positive polarity, negative polarity or neutral polarity in the sequential environment. -Then, we use the Fuzzy C-Means algorithm (FCM) to classify the English documents as having either positive polarity, negative polarity or neutral polarity in the Cloudera distributed environment with the purpose of shortening the execution time							

To understand the scientific value of this research, we compare our model's results with the results of models used in other studies.

Table 5 compares our model's results with the studies in [2–4] as follows:

cluster technique: CT.
sentiment classification: SC (opinion mining, or semantic classification, or emotion classification).
parallel network system: PNS (distributed system).
special domain: SD.
dependence on the training data set: DT.
language: L
Vector Space Model: VSM
no mention: NM
English language: EL.
Fuzzy C-Means: FCM.

Table 6 Compares our model's results with the work related to the Fuzzy C-Means (FCM) algorithm in [8–24].

Table 7 compares our model's results with studies related to Fuzzy C-Means in the parallel system (or FCM in the distributed system) in [25–27].

Table 8 compares our model's results with studies related to FCM for sentiment classification in [43–50].

Table 9 compares our model's results with the latest research on sentiment classification (or sentiment analysis or opinion mining) in [51–56].

Table 6 Comparison of our model's results with the work related to the Fuzzy C-Means (FCM) algorithm in [8–24]

Work	FCM	CT	SC	PNS	SD	DT	L	VSM
[8]	Yes	Yes	No	No	Yes	Yes	NM	NM
Model/Method in [8]	Ambiguity-driven fuzzy C-means clustering							
Summary of the work in [8]	As a well-known clustering algorithm, Fuzzy C-Means (FCM) allows each input sample to belong to more than one cluster, providing more flexibility than non-fuzzy clustering methods. However, the accuracy of FCM is subject to false detections caused by noisy records, weak feature selection and low certainty of the algorithm in some cases. The false detections are very important in some decision-making application domains like network security and medical diagnosis, where weak decisions based on such false detections may lead to catastrophic outcomes. They mainly emerge from making decisions about a subset of records that do not provide sufficient evidence to make a good decision. In this survey, the authors propose a method for detecting such ambiguous records in FCM by introducing a certainty factor to decrease invalid detections. This approach enables us to send the detected ambiguous records to another discrimination method for a deeper investigation, thus increasing the accuracy by lowering the error rate. Most of the records are still processed quickly and with a low error rate preventing performance loss which is common in similar hybrid methods. Experimental results of applying the proposed method on several data sets from different domains show a significant decrease in error rate as well as improved sensitivity of the algorithm							
[9]	Yes	Yes	No	No	Yes	Yes	NM	NM
Model/Method in [9]	Generalized fuzzy c-means clustering strategies							
Summary of the work in [9]	Fuzzy c-means (FCM) is a useful clustering technique. Modifications of FCM using L_1 norm distances increase robustness to outliers. Object and relational data versions of FCM clustering are defined for the more general case where the Lp norm or semi-norm ($0<p<1$) is used as the measure of dissimilarity. The authors give simple (though computationally intensive) alternating optimization schemes for all object data cases of $p>0$ in order to facilitate the empirical examination of the object data models. Both object and relational approaches are included in a numerical study							
[10]	Yes	Yes	No	No	Yes	Yes	NM	NM
Model/Method in [10]	Fuzzy Kohonen clustering networks							
Summary of the work in [10]	Kohonen networks are well known for cluster analysis (unsupervised learning). This class of algorithms is a set of heuristic procedures that suffers from several major problems (e.g. neither termination or convergence is guaranteed, no model is optimized by the learning strategy, and the output is often dependent on the sequence of data). A fuzzy Kohonen clustering network is proposed which integrates the Fuzzy c-Means (FCM) model into the learning rate and updating the strategies of the Kohonen network. This yields an optimization problem related to FCM, and the numerical results show improved convergence as well as reduced labeling errors. It is proved that the proposed scheme is equivalent to the c-Means algorithms. The new method can be viewed as a Kohonen type of FCM, but is "self-organizing" since the "size" of the update neighborhood and learning rate in the competitive layer are automatically adjusted during learning. Anderson's IRIS data is used to illustrate this method; and results are compared with the standard Kohonen approach							
[11]	Yes	Yes	No	No	Yes	Yes	NM	NM
Model/Method in [11]	Fuzzy c-means clustering of incomplete data							
Summary of the work in [11]	The problem of clustering a real s-dimensional data set X={x(1),,,,,x(n)}subset R(s) is considered. Usually, each observation (or datum) consists of numerical values for all s features (such as height, length, etc.), but sometimes data sets can contain vectors that are missing one or more of the feature values. For example, a particular datum x(k) might be incomplete, having the form x(k)=(254.3, ?, 333.2, 47.45, ?)(T), where the second and fifth feature values are missing. The fuzzy c-means (FCM) algorithm is a useful tool for clustering real s-dimensional data, but it is not directly applicable to the case of incomplete data. Four strategies for doing FCM clustering of incomplete data sets are given, three of which involve the modified versions of the FCM algorithm. Numerical convergence properties of the new algorithms are discussed, and all approaches are tested using real and artificially generated incomplete data sets							
[12]	Yes	Yes	No	No	Yes	Yes	NM	NM
Model/Method in [12]	The color image segmentation algorithm based on the thresholding and the fuzzy c-means techniques							
Summary of the work in [12]	In this research, a segmentation algorithm for color images based on the thresholding and the fuzzy c-means (FCM) techniques is presented. The scale-space filter is used as a tool for analyzing the histograms of three color components. The methodology uses a coarse-fine concept to reduce the computational burden required for the FCM. The coarse segmentation attempts to segment coarsely using the thresholding technique, while the fine segmentation assigns the pixels, which remain unclassified after the coarse segmentation, to the closest class using the FCM. Attempts also have been made to compare the performance of the proposed algorithm with other existing algorithms—Ohlander's, Rosenfeld's, and Bezdek's. Intensive computer simulation has been performed and the results are discussed in this paper. The simulation results indicate that the proposed algorithm yields the most accurate segmented image on the color coordinate proposed by Ohta et al., while requiring a reasonable amount of computational effort							

Table 6 (continued)

Work	FCM	CT	SC	PNS	SD	DT	L	VSM
[13]	Yes	Yes	NM	NM	NM	NM	NM	NM
Model/Method in [13]	A FORTRAN-IV coding of the fuzzy c-means (FCM) clustering program.							
Summary of the work in [13]	This paper develop a FORTRAN-IV coding of the fuzzy c-means (FCM) clustering program. The FCM program is applicable to a wide variety of Geo-statistical data analysis problems. This program generates fuzzy partitions and prototypes for any set of numerical data. These partitions are useful for corroborating known substructures or suggesting substructures in unexplored data. The clustering criterion used to aggregate subsets is a generalized least-squares objective function. Features of this program include a choice of three norms (Euclidean, Diagonal, or Mahalonobis), an adjustable weighting factor that essentially controls sensitivity to noise, acceptance of variable numbers of clusters, and outputs that include several measures of cluster validity							
[14]	Yes	Yes	NM	NM	NM	NM	NM	NM
Model/Method of [14]	Many functions have been proposed for validation of partitions of object data produced by the fuzzy c-means (FCM) clustering algorithm							
Summary of [14]	Many functions have been proposed for validation of partitions of object data produced by the fuzzy c-means (FCM) clustering algorithm. The authors examine the role a subtle but important parameter-the weighting exponent m of the FCM model-plays in determining the validity of FCM partitions. The functionals considered are the partition coefficient and entropy indexes of Bezdek, the Xie-Beni (1991), and extended Xie-Beni indexes, and the Fukuyama-Sugeno index (1989). Limit analysis indicates, and numerical experiments confirm, that the Fukuyama-Sugeno index is sensitive to both high and low values of m and may be unreliable because of this. Of the indexes tested, the Xie-Beni index provided the best response over a wide range of choices for the number of clusters, (2–10), and for m from 1.01-7. The authors' calculations suggest that the best choice form is probably in the interval [1.5, 2.5], whose mean and midpoint, m=2, have often been the preferred choice for many users of FCM							
[15]	Yes	Yes	NM	NM	NM	NM	NM	NM
Model/Method in [15]	A new model called possibilistic-fuzzy c-means (PFCM) model							
Summary of the work in [15]	The authors proposed the fuzzy-possibilistic c-means (FPCM) model and algorithm that generates both membership and typicality values when clustering unlabeled data. FPCM constrains the typicality values so that the sum over all data points of typicalities to a cluster is one. The row sum constraint produces unrealistic typicality values for large data sets. The authors proposed a new model called possibilistic-fuzzy c-means (PFCM) model. PFCM produces memberships and possibilities simultaneously, along with the usual point prototypes or cluster centers for each cluster. PFCM is a hybridization of possibilistic c-means (PCM) and fuzzy c-means (FCM) that often avoids the various problems of PCM, FCM and FPCM. PFCM solves the noise sensitivity defect of FCM, overcomes the coincident clusters problem of PCM and eliminates the row sum constraints of FPCM. The authors derive the first-order necessary conditions for extrema of the PFCM objective function, and use them as the basis for a standard alternating optimization approach to finding local minima of the PFCM objective functional. Several numerical examples are given that compare FCM and PCM to PFCM							
[16]	Yes	Yes	NM	NM	NM	NM	NM	NM
Model/Method in [16]	A novel algorithm for fuzzy segmentation of magnetic resonance imaging (MRI) data and estimation of intensity inhomogeneities using fuzzy logic							
Summary of the work in [16]	The authors present a novel algorithm for the fuzzy segmentation of magnetic resonance imaging (MRI) data and the estimation of intensity inhomogeneities using fuzzy logic. MRI intensity inhomogeneities can be attributed to imperfections in the radio-frequency coils or to problems associated with the acquisition sequences. The result is a slowly varying shading artifact over the image that can produce errors with conventional intensity-based classification. The authors' algorithm is formulated by modifying the objective function of the standard fuzzy c-means (FCM) algorithm to compensate for such in homogeneities and to allow the labeling of a pixel (voxel) to be influenced by the labels in its immediate neighborhood. The neighborhood effect acts as a regularizer and biases the solution toward piecewise-homogeneous labelings. Such a regularization is useful in segmenting scans corrupted by salt and pepper noise							
[17]	Yes	Yes	NM	NM	NM	NM	NM	NM
Model/Method in [17]	An approximate fuzzy c-means (AFCM) implementation based upon replacing the necessary "exact"' variates in the FCM equation with integer-valued or real-valued estimates							
Summary of the work in [17]	This research reports the results of a numerical comparison of two versions of the fuzzy c-means (FCM) clustering algorithms. In particular, the authors propose and exemplify an approximate fuzzy c-means (AFCM) implementation based upon replacing the necessary "exact" variates in the FCM equation with integer-valued or real-valued estimates. This approximation enables AFCM to exploit a lookup table approach for computing Euclidean distances and for exponentiation. The net effect of the proposed implementation is that CPU time during each iteration is reduced to approximately one sixth of the time required for a literal implementation of the algorithm, while apparently preserving the overall quality of terminal clusters produced. The two implementations are tested numerically on a nine-band digital image, and a pseudo-code subroutine is given for the convenience of applications oriented readers. Our results suggest that AFCM may be used to accelerate FCM processing whenever the feature space is comprised of tuples having a finite number of integer-valued coordinates							

Table 6 (continued)

Work	FCM	CT	SC	PNS	SD	DT	L	VSM
[18]	Yes	Yes	NM	NM	NM	NM	NM	NM
Model/Method in [18]	This counterexample establishes the existence of saddle points of the FCM objective function at locations other than the geometric centroid of fuzzy c-partition space							
Summary of the work in [18]	A counterexample to the original incorrect convergence theorem for the fuzzy c-means (FCM) clustering algorithms (see J.C. Bezdak, IEEE Trans. Pattern Anal. and Math. Intell., vol.PAMI-2, no.1, pp.1-8, 1980) is provided. This counterexample establishes the existence of saddle points of the FCM objective function at locations other than the geometric centroid of fuzzy c-partition space. Counterexamples previously discussed by W.T. Tucker (1987) are summarized. The correct theorem is stated without proof: every FCM iterate sequence converges, at least along a subsequence, to either a local minimum or saddle point of the FCM objective function. Although Tucker's counterexamples and the corrected theory appear elsewhere, they are restated as a caution not to further propagate the original incorrect convergence statement							
[19]	Yes	Yes	NM	NM	NM	NM	NM	NM
Model/Method in [19]	The authors substantially improve RFCM by generalizing it to the case of arbitrary (symmetric) dissimilarity data							
Summary of the work in [19]	The relational fuzzy c-means (RFCM) algorithm can be used to cluster a set of n objects described by pair-wise dissimilarity values if (and only if) there exist n points in Rn − 1 whose squared Euclidean distances precisely match the given dissimilarity data. This strong restriction on the dissimilarity data renders RFCM inapplicable to most relational clustering problems. This work substantially improves RFCM by generalizing it to the case of arbitrary (symmetric) dissimilarity data. The generalization is obtained using a computationally efficient modification of the existing algorithm that is equivalent to applying a "spreading" transformation to the dissimilarity data. While the method given applies specifically to dissimilarity data, a simple transformation can be used to convert similarity relations into dissimilarity data, so the method is applicable to any numerical relational data that are positive, reflexive (or antireflexive) and symmetric. Numerical examples illustrate and compare the present approach to problems that can be studied with alternatives such as the linkage algorithms							
[20]	Yes	Yes	NM	NM	NM	NM	NM	NM
Model/Method in [20]	A novel algorithm for the fuzzy segmentation of magnetic resonance imaging (MRI) data							
Summary of the work in [20]	Image segmentation plays a crucial role in many medical imaging applications. In this work, the authors present a novel algorithm for fuzzy segmentation of magnetic resonance imaging (MRI) data. The algorithm is realized by modifying the objective function in the conventional fuzzy C-means (FCM) algorithm using a kernel-induced distance metric and a spatial penalty on the membership functions. Firstly, the original Euclidean distance in the FCM is replaced by a kernel-induced distance, and thus the corresponding algorithm is derived and called as the kernelized fuzzy C-means (KFCM) algorithm, which is shown to be more robust than FCM. Then a spatial penalty is added to the objective function in KFCM to compensate for the intensity inhomogeneities of MR image and to allow the labeling of a pixel to be influenced by its neighbors in the image. The penalty term acts as a regularizer and has a coefficient ranging from zero to one. Experimental results on both synthetic and real MR images show that the proposed algorithms have better performance when noise and other artifacts are present than the standard algorithms							
[21]	Yes	Yes	NM	NM	NM	NM	NM	NM
Model/Method in [21]	New relational versions of the hard and fuzzy c-means algorithms							
Summary of the work in [21]	The hard and fuzzy c-means algorithms are widely used, effective tools for the problem of clustering n objects into (hard or fuzzy) groups of similar individuals when the data is available as object data, consisting of a set of n feature vectors in R. However, object data algorithms are not directly applicable when the n objects are implicitly described in terms of relational data, which consists of a set of n measurements of relations between each of the pairs of objects. New relational versions of the hard and fuzzy c-means algorithms are presented here for the case when the relational data can reasonably be viewed as some measure of distance. Some convergence properties of the algorithms are given along with a numerical example							
[22]	Yes	Yes	NM	NM	NM	NM	NM	NM
Model/Method in [22]	A fuzzy c-means (FCM) algorithm that incorporates spatial information into the membership function for clustering							
Summary of the work in [22]	A conventional FCM algorithm does not fully utilize the spatial information in the image. In this work, the authors present a fuzzy c-means (FCM) algorithm that incorporates spatial information into the membership function for clustering. The spatial function is the summation of the membership function in the neighborhood of each pixel under consideration. The advantages of the new method are the following: (1) it yields regions more homogeneous than those of other methods, (2) it reduces the spurious blobs, (3) it removes noisy spots, and (4) it is less sensitive to noise than other techniques. This technique is a powerful method for noisy image segmentation and works for both single and multiple-feature data with spatial information							

Table 6 (continued)

Work	FCM	CT	SC	PNS	SD	DT	L	VSM
[23]	Yes	Yes	No	No	NM	Yes	NM	No
Model/Method in [23]	A new weighted and constrained possibilistic C-means clustering algorithm							
Summary of the work in [23]	In this work, a new weighted and constrained possibilistic C-means clustering algorithm is proposed for process fault detection and diagnosis (FDI) in offline and online modes for both already known and novel faults. A possibilistic clustering based approach is utilized here to address some of the deficiencies of the fuzzy C-means (FCM) algorithm leading to more consistent results in the context of the FDI tasks by relaxing the probabilistic condition in FCM cost function. The proposed algorithm clusters the historical data set into C different dense regions without having precise knowledge about the number of the faults in the data set. The algorithm incorporates simultaneously possibilistic algorithm and local attribute weighting for time-series segmentation. This allows different weights to be allocated to different features responsible for the distinguished process faults which is an essential characteristic of proper FDI operations. A set of comparative studies have been carried out on the large-scale Tennessee Eastman industrial challenge problem and the DAMADICS actuator benchmark to demonstrate the superiority of the proposed algorithm in process FDI applications with respect to some available alternative approaches							
[24]	Yes	Yes	No	No	NM	Yes	NM	No
Model/Method in [24]	Clustering Incomplete Data Using Kernel-Based Fuzzy C-means Algorithm							
Summary of the work in [24]	There is a recent trend in recent machine learning community to construct a nonlinear version of a linear algorithm using the ’kernel method’, e.g. support vector machines (SVMs), kernel principal component analysis, kernel fisher discriminant analysis and the recent kernel clustering algorithms. In unsupervised clustering algorithms using kernel method, typically, a nonlinear mapping is used first to map the data into a potentially much higher feature space, where clustering is then performed. A drawback of these kernel clustering algorithms is that the clustering prototypes lie in high dimensional feature space and hence lack clear and intuitive descriptions unless using additional projection approximation from the feature to the data space as done in the existing literatures. In this work, a novel clustering algorithm using the ’kernel method’ based on the classical fuzzy clustering algorithm (FCM) is proposed and called the kernel fuzzy c-means algorithm (KFCM). KFCM adopts a new kernel-induced metric in the data space to replace the original Euclidean norm metric in FCM and the clustered prototypes still lie in the data space so that the clustering results can be reformulated and interpreted in the original space. Authors’ analysis shows that KFCM is robust to noise and outliers and also tolerates unequal sized clusters. Finally this property is utilized to cluster incomplete data. Experiments on two artificial and one real datasets show that KFCM has better clustering performance and more robust than several modifications of FCM for incomplete data clustering							
Our work	Yes	Yes	Yes	Yes	Yes	Yes	English	Yes
Model/Method in our work	Fuzzy C-Means algorithm for English sentiment classification in the Cloudera distributed system							
Summary of our work	Firstly, we use Fuzzy C-Means algorithm (FCM) to classify the English documents as having either positive polarity, negative polarity, or neutral polarity in the sequential environment. Then, we use the Fuzzy C-Means algorithm (FCM) to classify the English documents as having either positive polarity, negative polarity, or neutral polarity in the Cloudera distributed environment with the purpose of shortening the execution time							

Table 7 Comparison of our model's results with studies related to Fuzzy C-Means in the parallel system (or FCM in the distributed system) in [25–27]

Studies	FCM	CT	SC	PNS	SD	DT	L	VSM
[25]	Yes	Yes	NM	NM	NM	NM	NM	NM
Model/Method in [25]	The literal and approximate fuzzy c-means unsupervised clustering algorithms, and a supervised computational neural network							
Summary of the work in [25]	Magnetic resonance (MR) brain section images are segmented and then synthetically colored to give visual representations of the original data with three approaches: the literal and approximate fuzzy c-means unsupervised clustering algorithms, and a supervised computational neural network. The initial clinical results are presented on normal volunteers and selected patients with brain tumors surrounded by edema. Supervised and unsupervised segmentation techniques provide broadly similar results. Unsupervised fuzzy algorithms were visually observed to show better segmentation when compared with raw image data for volunteer studies. For a more complex segmentation problem with tumor/edema or cerebrospinal fluid boundary, where the tissues have similar MR relaxation behavior, inconsistency in rating among experts was observed, with fuzz-c-means approaches being slightly preferred over feedforward cascade correlation results. Various facets of both approaches, such as supervised versus unsupervised learning, time complexity, and utility for the diagnostic process, are compared							
[26]	Yes	Yes	NM	NM	NM	NM	NM	NM
Model/Method in [26]	Three clustering methods: (1) the conventional two-stage method, (2) the self-organizing feature maps, and (3) the authors' proposed two-stage method, via both simulated and real-world data							
Summary of the work in [26]	Cluster analysis is a common tool for market segmentation. Conventional research usually employs the multivariate analysis procedures. In recent years, due to their high performance in engineering, artificial neural networks have also been applied in the area of management. Thus, this study aims to compare three clustering methods: (1) the conventional two-stage method, (2) the self-organizing feature maps and (3) the authors' proposed two-stage method, via both simulated and real-world data. The proposed two-stage method is a combination of the self-organizing feature maps and the K-means method. The simulation results indicate that the proposed scheme is slightly better than the conventional two-stage method with respect to the rate of misclassification, and the real-world data on the basis of Wilk's Lambda and discriminant analysis							
[27]	Yes	Yes	NM	Yes	NM	NM	NM	NM
Model/Method in [27]	The parallel fuzzy c-means (PFCM) algorithm for clustering large data sets							
Summary of the work in [27]	The parallel fuzzy c-means (PFCM) algorithm for clustering large data sets is proposed in this survey. The proposed algorithm is designed to run on parallel computers of the Single Program Multiple Data (SPMD) model type with the Message Passing Interface (MPI). A comparison is made between PFCM and an existing parallel k-means (PKM) algorithm in terms of their parallelisation capability and scalability. In an implementation of PFCM to cluster a large data set from an insurance company, the proposed algorithm is demonstrated to have almost ideal speedups as well as an excellent scaleup with respect to the size of the data sets							
Our work	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Model/Method in our work	Fuzzy C-Means algorithm for English sentiment classification in the Cloudera distributed system							
Summary of our work	Firstly, we use the Fuzzy C-Means algorithm (FCM) to classify the English documents as having either positive polarity, negative polarity or neutral polarity in the sequential environment							
	Then, we use the Fuzzy C-Means algorithm (FCM) to classify the English documents as having either positive polarity, negative polarity or neutral polarity in the Cloudera distributed environment with the purpose of shortening the execution time							

Table 8 Comparison of our model's results with the FCM used for sentiment classification in [43–50]

Work	FCM	CT	SC	PNS	SD	DT	L	VSM
[43]	Yes	Yes	Yes	NM	NM	NM	NM	NM
Model/Method in [43]	EmoSenticSpace, a new framework for affective common-sense reasoning that extends WordNet-Affect and SenticNet by providing both emotion labels and polarity scores for a large set of natural language concepts							
Summary of the work in [43]	Emotions play a key role in natural language understanding and sensemaking. Pure machine learning usually fails to recognize and interpret emotions in text accurately. The need for knowledge bases that give access to semantics and sentics (the conceptual and affective information) associated with natural language is growing exponentially in the context of big social data analysis. To this end, this survey proposes EmoSenticSpace, a new framework for affective common-sense reasoning that extends WordNet-Affect and SenticNet by providing both emotion labels and polarity scores for a large set of natural language concepts. The framework is built by means of fuzzy c-means clustering and support-vector-machine classification, and takes into account a number of similarity measures, including point-wise mutual information and emotional affinity. EmoSenticSpace was tested on three emotion-related natural language processing tasks, namely sentiment analysis, emotion recognition, and personality detection. In all cases, the proposed framework outperforms the state-of-the-art. In particular, the direct evaluation of EmoSenticSpace against psychological features provided in the benchmark ISEAR dataset shows a 92.15 % agreement							
[44]	Yes	Yes	Yes	NM	NM	NM	NM	NM
Model/Method in [44]	The task of semi-supervised classification: extending category labels from a small dataset of labeled examples to a much larger set							
Summary of the work in [44]	The authors consider the task of semi-supervised classification: extending category labels from a small dataset of labeled examples to a much larger set. The authors show that, at least in their case study task, unsupervised fuzzy clustering of the unlabeled examples helps in obtaining the hard clusters. Namely, the authors used the membership values obtained with fuzzy clustering as additional features for hard clustering. The authors also used these membership values to reduce the confusion set for the hard clustering. As a case study, the authors use applied the proposed method to the task of constructing a large emotion lexicon by extending the emotion labels from the WordNet Affect lexicon using various features of words. Some of the features were extracted from the emotional statements of the freely available ISEAR dataset; other features were WordNet distance and the similarity measured via the polarity scores in the SenticNet resource. The proposed method classified words by emotion labels with high accuracy							
[45]	Yes	Yes	Yes	NM	NM	NM	NM	NM
Model/Method in [45]	Techniques that have been used for the task of opinion mining							
Summary of the work in [45]	As individuals impart their sentiments on the Web on products and services they have used, it has become important to formulate methods to automatically classify and judge them. The task of examining such data, collectively called client feedback data, is known as opinion mining. Opinion mining consists of several steps, and different techniques have been proposed for different steps. This survey basically explains such techniques that have been used for the implementation of task of opinion mining. On the basis of this analysis the authors provide an overall system design for the development of opinion mining approach							
[46]	Yes	Yes	Yes	NM	NM	NM	NM	NM
Model/Method in [46]	Core concepts and techniques in the large subset of cluster analysis.							
Summary of the work in [46]	The fast retrieval of relevant information from databases has always been a significant issue. Many techniques have been developed for this purpose, of which data clustering is one of the major techniques. The process of creating vital information from the huge amount of data is learning, which can be classified as either supervised learning or unsupervised learning. Clustering is a kind of unsupervised data mining technique. It describes the general working behavior, the methodologies followed by these approaches and the parameters which affect the performance of these algorithms. In classifying web pages, the similarity between web pages is a very important feature. The main objective of this survey is to gather more core concepts and techniques in the large subset of cluster analysis							
[47]	Yes	Yes	NM	NM	NM	NM	NM	NM
Model/Method in [47]	A novel combination of fuzzy inference system and Dempster–Shafer Theory							
Summary of the work in [47]	Brain Magnetic Resonance Imaging (MRI) segmentation is a challenging task due to the complex anatomical structure of brain tissues as well as intensity non-uniformity, partial volume effects and noise. Segmentation methods based on fuzzy approaches have been developed to overcome the uncertainty caused by these effects. In this study, a novel combination of fuzzy inference system and Dempster–Shafer Theory is applied to brain MRI for the purpose of segmentation where the pixel intensity and the spatial information are used as features. In the proposed modeling, the consequent part of rules is a Dempster–Shafer belief structure. The novelty aspect of this work is that the rules are paraphrased as evidences. The results show that the proposed algorithm, called FDSIS has satisfactory outputs on both simulated and real brain MRI datasets							

Table 8 (continued)

Work	FCM	CT	SC	PNS	SD	DT	L	VSM
[48]	Yes	Yes	Yes	NM	Yes	Yes	EL	NM
Model/Method in [48]	This survey aims to add some additional features for improving the classification method							
Summary of the work in [48]	Sentiment classification aims to detect information such as opinions, explicit, implicit feelings expressed in text. Most of the existing approaches are able to detect either explicit expressions or implicit expressions of sentiments in the text separately. In this proposed framework, it will detect both implicit and explicit expressions available in the meeting transcripts. It will classify the positive, negative and neutral words and also identifies the topic of the particular meeting transcripts by using fuzzy logic. This work aims to add some additional features for improving the classification method. The quality of the sentiment classification is improved using proposed fuzzy logic framework. This fuzzy logic includes features like fuzzy rules and fuzzy c-means algorithm. The quality of the output is evaluated using parameters such as precision, recall, f-measure. Here, Fuzzy C-means Clustering technique is measured in terms of purity and entropy							
[49]	Yes	Yes	Yes	NM	Yes	Yes	Yes	Yes
Model/Method in [49]	Clustering documents with labeled and unlabeled documents using fuzzy semi-Kmeans							
Summary of the work in [49]	While focusing on document clustering, this work presents a fuzzy semi-supervised clustering algorithm called fuzzy semi-Kmeans. The fuzzy semi-Kmeans is an extension of K-means clustering model, and it is inspired by an EM algorithm and a Gaussian mixture model. Additionally, the fuzzy semi-Kmeans provides the flexibility to employ different fuzzy membership functions to measure the distance between data. This work employs Gaussian weighting function to conduct experiments, but cosine similarity function can be used as well. This work conducts experiments on three data sets and compares fuzzy semi-Kmeans with several methods. The experimental results indicate that fuzzy semi-Kmeans can generally outperform the other methods							
[50]	Yes	Yes	Yes	NM	NM	NM	NM	NM
Model/Method in [50]	Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier							
Summary of the work in [50]	With the rapid development of the World Wide Web, electronic word-of-mouth interaction has made consumers active participants. Nowadays, a large number of reviews posted by the consumers on the Web provide valuable information to other consumers. Such information is highly essential for decision making and hence popular among the internet users. This information is very valuable not only for prospective consumers to make decisions, but also for businesses in predicting the success and sustainability. In this study, a Gini Index based feature selection method with Support Vector Machine (SVM) classifier is proposed for sentiment classification for large movie review data. The results show that our Gini Index method has better classification performance in terms of reduced error rate and accuracy							
Our study	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Model/Method in our study	Fuzzy C-Means algorithm for English sentiment classification in the Cloudera distributed system							
Summary of our study	Firstly, we use the Fuzzy C-Means algorithm (FCM) to classify the English documents as having either a positive polarity, negative polarity or neutral polarity in the sequential environment Then, we use the Fuzzy C-Means algorithm (FCM) to classify the English documents as having either positive polarity, negative polarity neutral polarity in the Cloudera distributed environment with the purpose of shortening the execution time							

Table 9 Comparison of the proposed model with the latest sentiment classification models (or the latest sentiment classification methods) in [51–56]

Studies	FCM	CT	SC	PNS	SD	DT	L	VSM
[51]	No	No	Yes	NM	Yes	Yes	Yes	vector
Model/Method in [51]	The machine learning approaches applied to sentiment analysis-based applications							
Summary of the work in [51]	Opinion mining or sentiment analysis is a study that analyzes people's opinions or sentiments from the text towards entities such as products and services. It has always been important to know what other people think. With the rapid growth in the availability and popularity of online review sites, blogs, forums, and social networking sites, it has become necessary to analyze and understand these reviews. The main approaches to sentiment analysis can be categorized into semantic orientation-based approaches, knowledge-based, and machine-learning algorithms. The authors survey the machine learning approaches applied to sentiment analysis-based applications. The main emphasis of this work is to discuss the research involved in applying machine learning methods mostly for sentiment classification at the document level. Machine learning-based approaches to sentiment classification work in the following phases: (1) feature extraction, (2) feature weighting schemes, (3) feature selection, and (4) machine-learning methods. The study also discusses the standard free benchmark datasets and evaluation methods for sentiment analysis. The authors conclude the work 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							
[52]	No	No	Yes	NM	Yes	Yes	NM	NM
Model/Method in [52]	Two types of techniques have been used in the literature for semantic orientation-based approach for sentiment analysis, viz., (i) corpus based and (ii) dictionary or lexicon or knowledge based							
Summary of the work in [52]	Two types of techniques have been used in the literature for semantic orientation-based approach for sentiment analysis, viz., (i) corpus based and (ii) dictionary or lexicon or knowledge based. In this work, the authors explore the corpus-based semantic orientation approach for sentiment analysis. Corpus-based semantic orientation approach requires large dataset to detect the polarity of the terms and therefore the sentiment of the text. The main problem with this approach is that it relies on the polarity of the terms that have appeared in the training corpus since polarity is computed for the terms that are in the corpus. 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. The performance of the semantic orientation-based approach has been limited in the literature due to inadequate coverage of the multi-word features							
[53]	No	No	Yes	NM	Yes	Yes	EL	NM
Model/Method in [53]	New meta-level features especially designed for the sentiment analysis of short messages such as: (i) information derived from the sentiment distribution among the k nearest neighbors of a given short test document x, (ii) the distribution of distances of x to their neighbors and (iii) the document polarity of these neighbors given by unsupervised lexical-based methods							
Summary of the work in [53]	In this work, the authors address the problem of automatically learning to classify the sentiment of short messages/reviews by exploiting information derived from meta-level features i.e., features derived primarily from the original bag-of-words representation. The authors propose new meta-level features especially designed for the sentiment analysis of short messages such as: (i) information derived from the sentiment distribution among the k nearest neighbors of a given short test document x, (ii) the distribution of distances of x to their neighbors and (iii) the document polarity of these neighbors given by unsupervised lexical-based methods. The authors' approach is also capable of exploiting information from the neighborhood of document x regarding (highly noisy) data obtained from 1.6 million Twitter messages with emoticons. The set of proposed features is capable of transforming the original feature space into a new one, potentially smaller and more informed. 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-word representation (by up to 16 %) but 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 some 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							
[54]	No	No	Yes	NM	Yes	Yes	NM	NM
Model/Method in [54]	Rule based machine learning algorithms							
Summary of the work in [54]	Sentiment analysis is becoming a promising topic with the strengthening of social media such as blogs, networking sites etc. where people exhibit their views on various topics. In this work, the focus is to perform effective sentiment analysis and opinion mining of Web reviews using various rule-based machine learning algorithms. The authors use SentiWordNet which generates score count words into one of seven categories, strong-positive, positive, weak-positive, neutral, weak-negative, negative and strong-negative words. The proposed approach is tested on online books and political reviews and demonstrates efficacy through Kappa measures, which has a higher accuracy of 97.4 % and lower error rate. The weighted average of different accuracy measures like Precision, Recall, and TP-Rate depicts a 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							

Table 9 (continued)

Studies	FCM	CT	SC	PNS	SD	DT	L	VSM
[55]	No	No	Yes	No	No	No	EL	No
Model/Method in [55]	A combination of the term-counting method and enhanced contextual valence shifters method							
Summary of the work in [55]	The authors 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, which are not in the five previous before. The study shows that the authors' proposed method based on the combination of the term-counting method and the enhanced contextual valence shifters method has improved the accuracy of sentiment classification. The combined method has an accuracy of 68.984 % of the testing dataset, and 69.224 % on the training data set. All these methods are implemented to classify the reviews based on our new dictionary and the Internet Movie data set							
[56]	No	No	Yes	No	No	No	EL	No
Model/Method in [56]	Naive Bayes model with N-GRAM method, Negation Handling method, Chi-Square method and Good-Turing Discounting, etc							
Summary of the work in [56]	The authors explored the Naive Bayes model with N-GRAM method, Negation Handling method, Chi-Square method and Good-Turing Discounting by selecting different thresholds of Good-Turing Discounting method and different minimum frequencies of Chi-Square method to improve the accuracy of sentiment classification							
Our work	Yes	Yes	Yes	Yes	Yes	Yes	EL	Yes
Model/Method in our work	Fuzzy C-Means algorithm for English sentiment classification in the Cloudera distributed system							
Summary of our work	Firstly, we use the Fuzzy C-Means algorithm (FCM) to classify the English documents as having either positive polarity, negative polarity or neutral polarity in the sequential environment							
	Then, we use the Fuzzy C-Means algorithm (FCM) to classify the English documents as having either positive polarity, negative polarity or neutral polarity in the Cloudera distributed environment with the purpose of shortening the execution time							

References

1. Large movie review dataset (2016) http://ai.stanford.edu/~amaas/data/sentiment/
2. Singh VK, Singh VK (2015) Vector space model: an information retrieval system. International Journal of Advanced Engineering Research and Studies
3. Carrera-Trejo V, Sidorov G, Miranda-Jiménez S, Moreno Ibarra M, Cadena Martínez R (2015) Latent Dirichlet allocation complement in the vector space model for multi-label text classification. International Journal of Combinatorial Optimization Problems and Informatics 6(1):7–19
4. Soucy P, Mineau GW (2005) Beyond TFIDF weighting for text categorization in the vector space model. In: Proceedings of the 19th international joint conference on Artificial intelligence, USA, pp 1130–1135
5. Hadoop (2016). http://hadoop.apache.org
6. Apache (2016). http://apache.org
7. Cloudera (2016). http://www.cloudera.com
8. Ghaffari M, Ghadiri N (2016) Ambiguity-driven fuzzy C-means clustering: how to detect uncertain clustered records. Applied Intelligence (APIN):1–12
9. RJ Hathaway JC, Bezdek YHu (2000) Generalized fuzzy c-means clustering strategies using L/sub p/ norm distances. IEEE Trans Fuzzy Syst 8(5):576–582
10. Tsao EC-K, Bezdek JC, Pal NR (1994) Fuzzy Kohonen clustering networks. Pattern Recogn 27(5):757–764
11. Hathaway RJ, Bezdek JC (2001) Fuzzy c-means clustering of incomplete data. IEEE Trans Syst Man Cybern B (Cybern) 31(5):735–744
12. Lim YW, Lee SU (1990) On the color image segmentation algorithm based on the thresholding and the fuzzy c-means techniques. Pattern Recogn 23(9):935–952
13. Bezdek JC, Ehrlich R, Full W (1984) FCM: the fuzzy c-means clustering algorithm. Comput Geosci 10(2–3):191–203
14. Pal NR, Bezdek JC (2002) On cluster validity for the fuzzy c-means model. IEEE Trans Fuzzy Syst 3(3):370–379
15. Pal NR, Pal K, Keller JM, Bezdek JC (2005) A possibilistic fuzzy c-means clustering algorithm. IEEE Trans Fuzzy Syst 13(4):517–530
16. Ahmed MN, Yamany SM, Mohamed N, Farag AA (2002) A modified fuzzy c-means algorithm for bias field estimation and segmentation of MRI data. IEEE Trans Med Imaging 21(3):193–199
17. Cannon RL, Dave JV, Bezdek JC (2009) Efficient implementation of the fuzzy c-means clustering algorithms. IEEE Trans Pattern Anal Mach Intell 8(2):248–255
18. Bezdek JC, Hathaway RJ, Sabin MJ, Tucker WT (1987) Convergence theory for fuzzy c-means: Counterexamples and repairs. IEEE Trans Syst Man Cybern 17(5):873–877
19. Hathaway RJ, Bezdek JC (1994) Nerf c-means: non-euclidean relational fuzzy clustering. Pattern Recogn 27(3):429–437
20. D-Q Zhang S-C, Chen A (2004) Novel kernelized fuzzy C-means algorithm with application in medical image segmentation. Artif Intell Med 32(1):37–50
21. Hathaway RJ, Davenport JW, Bezdek JC (1989) Relational duals of the c-means clustering algorithms. Pattern Recogn 22(2):205–212
22. Chuang K-S, Tzeng H-L, Chena S, Wu J, Chen T-J (2006) Fuzzy c-means clustering with spatial information for image segmentation. Comput Med Imaging Graph 30(1):9–15
23. Bahrampour S, Moshiri B, Salahshoor K (2011) Weighted and constrained possibilistic C-means clustering for online fault detection and isolation. Appl Intell (APIN) 35(2):269–284
24. Zhang D-Q, Chen S-C (2003) Clustering incomplete data using kernel-based fuzzy c-means algorithm. Neural Process Lett 18(3):155–162

25. Hall LO, Bensaid AM, Clarke LP, Velthuizen RP (2002) A comparison of neural network and fuzzy clustering techniques in segmenting magnetic resonance images of the brain. IEEE Trans Neural Netw 3(5):672–682
26. Kuo RJ, Ho LM, Hu CM (2002) Integration of self-organizing feature map and K-means algorithm for market segmentation. Comput Oper Res 29(11):1475–1493
27. Kwok T, Smith K, Lozano S, Taniar D (2002) Parallel Fuzzy c-Means Clustering for Large Data Sets, Euro-Par 2002 Parallel Processing, Volume 2400 of the series Lecture Notes in Computer Science, pp 365–374
28. Xylogiannopoulos KF, Karampelas P, Alhajj R (2016) Repeated patterns detection in big data using classification and parallelism on LERP Reduced Suffix Arrays. Appl Intell (APIN):1–31
29. Carns PH, Ligon III WB, Ross RB, Thakur R (2000) PVFS: A parallel file system for linux clusters. In: Proceedings of the extreme linux track: 4th annual linux showcase and conference
30. Moyer SA, Sunderam VS (1994) PIOUS: a scalable parallel I/o system for distributed computing environments. In: Proceedings of the scalable high-performance computing conference
31. Shirazi BA, Kavi KM, Hurson AR (1995) Scheduling and load balancing in parallel and distributed systems, scheduling and load balancing in parallel and distributed systems, USA
32. Andrews GR (1999) Foundations of parallel and distributed programming. In: Foundations of parallel and distributed programming 1st, USA
33. Gropp W, Lusk E, Doss N, Skjellum A (1996) A high-performance, portable implementation of the MPI message passing interface standard. Parallel Comput 22(6):789–828
34. Yu Y, Isard M, Fetterly D, Budiu M, Erlingsson Ú, Gunda PK, Currey J (2008) dryadLINQ: a system for general-purpose distributed data-parallel computing using a high-level language symposium on operating system design and implementation (OSDI)
35. Shvachko K, Kuang H, Radia S, Chansler R (2010) The hadoop distributed file system
36. Guerrero JM, Matas J, Garcia de Vicuna L, Castilla M, Miret J (2007) Decentralized control for parallel operation of distributed generation inverters using resistive output impedance. IEEE Trans Ind Electron 54:2
37. van Steen M, Homburg P, Tanenbaum AS (1999) Globe: a wide-area distributed system. IEEE Concurr 7(1):70–78
38. Shende SS, Malony AD (2006) The tau parallel performance system. Int J High Perform Comput Appl 20(2):287–311
39. Bagrodia R, Meyer R, Takai M, Chen Y-A, Zeng X, Martin J, Song HY (1998) Parsec: a parallel simulation environment for complex systems. Computer 31(10):77–85
40. RumelHart DE, Hinton GE, McClelland JL (1986) A general framework for parallel distributed processing. In: Parallel distributed processing: explorations in the microstructure of cognition, USA, vol 1, pp 45–76
41. Ikudome K, Fox GC, Kolawa A, Flower JW (1990) An automatic and symbolic parallelization system for distributed memory parallel computers. In: Proceedings of the fifth distributed memory computing conference
42. Wang HO, Tanaka K, Griffin M (1995) Parallel distributed compensation of nonlinear systems by Takagi-Sugeno fuzzy model
43. Poria S, Gelbukh A, Cambria E, Hussain A, Huang G-B (2014) EmoSenticSpace: a novel framework for affective common-sense reasoning. Knowl-Based Syst 69:108–123
44. Poria S, Gelbukh A, Das D, Bandyopadhyay S (2013) Fuzzy clustering for semi-supervised learning – case study: construction of an emotion lexicon. In: Advances in artificial intelligence, volume 7629 of the series lecture notes in computer science, pp 73–86
45. Vinchurkar SV, Nirkhi SM (2012) feature extraction of product from customer feedback through blog. International Journal of Emerging Technology and Advanced Engineering 2(1):2250–2459
46. IndiraPriya P, Ghosh DK (2013) A Survey on Different Clustering Algorithms in Data Mining Technique. International Journal of Modern Engineering Research (IJMER) 3(1):267–274
47. Ghasemi J, Ghaderi R, Karami Mollaei MR, Hojjatoleslami SA (2013) A novel fuzzy Dempster–Shafer inference system for brain MRI segmentation. Inf Sci 223:205–220
48. Sheeba JI, Vivekanandan K (2014) A fuzzy logic based on sentiment classification. International Journal of Data Mining & Knowledge Management Process (IJDKP) 4(4)
49. Liu C-L, Chang T-H, Li H-H (2013) Clustering documents with labeled and unlabeled documents using fuzzy semi-Kmeans. Fuzzy Sets Syst 221:48–64
50. Manek AS, Deepa Shenoy P, Chandra Mohan M, Venugopal KR (2016) Aspect term extraction for sentiment analysis in large movie reviews using gini index feature selection method and SVM classifier. World wide web, 1–20. doi:10.1007/s11280-015-0381-x. Print ISSN1386-145x, US
51. Agarwal B, Mittal N (2016) Machine learning approach for sentiment analysis. Prominent feature extraction for sentiment analysis, 21–45. doi:10.1007/978-3-319-25343-5_3. Print ISBN 978-3-319-25341-1
52. Agarwal B, Mittal N (2016) Semantic orientation-based approach for sentiment analysis. Prominent feature extraction for sentiment analysis, 77–88. doi:10.1007/978-3-319-25343-5_6. Print ISBN 978-3-319-25341-1
53. Canuto S, André M, Gonçalves FB (2016) Exploiting new sentiment-based meta-level features for effective sentiment analysis. In: Proceedings of the ninth ACM international conference on web search and data mining (WSDM '16), New York, USA, pp 53–62
54. Ahmed S, Danti A (2016) Effective sentimental analysis and opinion mining of web reviews using rule based classifiers. Computational Intelligence in Data Mining 1:171–179. doi:10.1007/978-81-322-2734-2_18. Print ISBN 978-81-322-2732-8, India
55. Phu VN, Tuoi PT (2014) Sentiment classification using enhanced contextual valence shifters. In: International Conference on Asian Language Processing (IALP), pp 224–229
56. Tran VTN, Phu VN, Tuoi PT (2014) Learning more chi square feature selection to improve the fastest and most accurate sentiment classification. In: The third asian conference on information systems (ACIS 2014)

Vo Ngoc Phu received B. Sc Degree in Computer Science from Ho Chi Minh City University of Science, Ho Chi Minh City, Viet Nam in 2000, M. Sc Degree in Computer Science from Ho Chi Minh City University of Technology, Ho Chi Minh City, Viet Nam in 2014. He is currently a lecturer and a researcher in Ton Duc Thang University, Ho Chi Minh City, Vietnam and Duy Tan University, Da Nang City, Vietnam.

Nguyen Duy Dat received B. Sc Degree from Ho Chi Minh City University of Technology and Education, Ho Chi Minh City, Vietnam in 2013; M. Sc Degree from Ho Chi Minh City University of Technology, Ho Chi Minh City, Viet Nam in 2014. He is currently a lecturer in Cao Thang College, Ho Chi Minh City, Vietnam.

Vo Thi Ngoc Tran received her B.Sc Degree and M. Sc Degree in Computer Science from Ho Chi Minh City University of Technology, Vietnam National University, HoChi Minh City, Vietnam in 1989 and 1990.

Vo Thi Ngoc Chau received her B.Sc Degree, M. Sc Degree, and Dr.Sc Degree in Computer Science from Ho Chi Minh City University of Technology, Vietnam National University, Ho Chi Minh City, Vietnam in 1989, 1990 and 1995, respectively. She is currently a lecturer at Ho Chi Minh City University of Technology, Vietnam National University, Ho Chi Minh City, Vietnam.

Tuan A. Nguyen received his B.Sc Degree in Computer Science from Ho Chi Minh City University of Science, Ho Chi Minh City, Viet Nam in 1989; M. Sc. Degree and Dr. Sc. Degree in Computer Science from La Trobe University, Australia in 1990 and in 1995, respectively. He is currently a lecturer at University of Information Technology, Vietnam National University of Hochiminh City.