Evolving Systems (2019) 10:425–451
https://doi.org/10.1007/s12530-017-9180-1

ORIGINAL PAPER

A C4.5 algorithm for english emotional classification

Phu Vo Ngoc[1] · Chau Vo Thi Ngoc[2] · Tran Vo Thi Ngoc[3] · Dat Nguyen Duy[4]

Received: 9 November 2016 / Accepted: 10 March 2017 / Published online: 8 April 2017

Abstract The solutions for processing sentiment analysis are very important and very helpful for many researchers, many applications, etc. This new model has been proposed in this paper, used in the English document-level sentiment classification. In this research, we propose a new model using C4.5 Algorithm of a decision tree to classify semantics (positive, negative, neutral) for the English documents. Our English training data set has 140,000 English sentences, including 70,000 English positive sentences and 70,000 English negative sentences. We use the C4.5 algorithm on the 70,000 English positive sentences to generate a decision tree and many association rules of the positive polarity are created by the decision tree. We also use the C4.5 algorithm on the 70,000 English negative sentences to generate a decision tree and many association rules of the negative polarity are created by the decision tree. Classifying sentiments of one English document is identified based on the association rules of the positive polarity and the negative polarity. Our English testing data set has 25,000 English documents, including 12,500 English positive reviews and 12,500 English negative reviews. We have tested our new model on our testing data set and we have achieved 60.3% accuracy of sentiment classification on this English testing data set.

Keywords Sentiment classification · English sentiment classification · English document opinion mining · C4.5 algorithm · c4.5 · CA · Decision tree

✉ Phu Vo Ngoc
vongocphu03hca@gmail.com; vongocphu@dtu.edu.vn

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

Tran Vo Thi Ngoc
vtntran@hcmut.edu.vn

Dat Nguyen Duy
duydatspk@gmail.com

1 Institute of Research and Development, Duy Tan University - DTU, Da Nang, Vietnam

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

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

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

1 Introduction

The solutions for processing the semantic analysis are very important and very helpful for many researchers, many applications, etc. Today there are many studies and many applications for sentiment classification in many languages.

In this work we propose a new model using a decision tree, specifically as C4.5 algorithm (CA), for English document-level emotional classification.

A decision tree is a decision support tool that uses a tree-like-graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.

The C4.5 algorithm is a famous algorithm of the decision tree which belongs to the data mining filed, but it has been used in many different fields for a long time. However, the C4.5 algorithm is not used in natural language processing (NLP), especially in sentiment classification. We thought that it can be used in the opinion analysis. Therefore, we try applying it into the semantic analysis. This is also very difficult for us to perform it into the sentiment analysis. This is very significantly important for the works and applications in the NLP. From the results which we got, it is true that the C4.5 algorithm is used in the NLP and also in the opinion classification. The aim of this research is to implement the C4.5 algorithm for the emotional analysis of the English documents based on the English sentences of the English training data set. We searched the surveys in the world, which is related to the decision tree, emotional classification. From the below proofs, we found that there is not any research in the world which is similar to this study. We looked for many methodologies to apply the C4.5 algorithm into the sentiment classification for the English documents and then, they are experimented on our data sets. Thus, this proposed model is the originality and novelty research and it also has many meanings in the data mining field, the NLP, the computer science field, etc.

We use the CA to classify semantics (positive, negative, neutral) of one English document in the English testing data set based on 140,000 English sentences of English testing data set which includes 70,000 English positive sentences and 70,000 English negative sentences.

We propose many basis principles to implement our new model as follows:

- Assuming that one English document in the English testing data set has n English sentences.
- Assuming that one English sentence in the English testing data set or in the English training data set has m English words (or English phrases).
- Assuming that there is one English sentence which has the longest length in both the English testing data set and the English training data set; and the longest length is m_max. It means that m_max is greater than m or m_max is as equal as m.
- We build a table of training data for the CA based on 140,000 English sentences of English testing data set as follows:
  - The table of training data has 140,000 records (or 140,000 rows) and (m_max + 1) columns.
  - Each column of the table from column 0 to column (m_max − 1) is one English word (or one English phrase) and value of each column is one English word (or one English phrase). If one English sentence has length m (m < m_max) then each column from m to (m_max − 1) is 0 (zero).
  - Column m_max in the table is polarity column. This column shows that the sentence belongs to positive in 70,000 English positive sentences or negative in 70,000 English negative sentences.
  - Example, we have three English sentences such as:
  - The film is very good ≥ the sentence belongs to the 70,000 English positive sentences.
  - The actor is very bad ≥ the sentence belongs to the 70,000 English negative sentences.
  - The film sounds good ≥ the sentence belongs to the 70,000 English positive sentences.
  - The table of training data is in the Table 1 below in the “Appendix”.
- When we use the IA on the Table 1, we get a decision tree to generate many association rules. The association rules have the format as “X ≥ positive” or “Y ≥ negative”. These rules are divided into two groups: the positive rule group and the negative rule group. The positive rule group contains all association rules having the format as “X ≥ positive”. The negative rule group contains all association rules having the format as “Y ≥ negative”.
- One English sentence of one English document in the English testing data set is the positive polarity if the sentence contains X fully. The English sentence is the negative polarity if the sentence contains Y fully. The English sentence is the neutral polarity if the sentence does not contain both X and Y fully.
- Assuming that we have some rules such as: “very good” ≥ positive; “very handsome” ≥ positive; “excellent” ≥ positive; “very bad” ≥ negative; “terrible” ≥ negative; we have three sentences such as “the film is very good”; “the actor is very bad”; and “he is drinking some beer”. With the first sentence “the film is very good”, the sentence only contains one rule “very good” ≥ positive,

Table 1 Training data set for a decision tree

Column 0	Column 1	Column 2	Column 3	Column 4	…	Column m_max
the	film	is	very	good	0	Positive
the	actor	is	very	bad	0	Negative
the	film	sounds	good	0	0	Positive
…	…	…	…	…	…	…

thus, the sentence is the positive polarity. With the second sentence “the actor is very bad”, the sentence only contains one rule “very bad”, therefore, the sentence is the negative polarity. With the third sentence “he is drinking some beer”, the sentence does not contain any rule in our rule set, so, the sentence is the neutral polarity.

- One English document in the English testing data set is the positive polarity if the number of the English sentence classified into the positive polarity is greater than the number of the English sentences classified into the negative polarity in the English document. The English document is the negative polarity if the number of the English sentences classified into the positive polarity is less than the number of the English sentences classified into the negative polarity in the document. The English document is the neutral polarity if the number of the English sentences classified into the positive polarity is as equal as the number of the English sentences classified into the negative polarity in the document.

In many researches related to the C4.5 algorithm (CA) in the world and in (Ruggieri 2002; Kretschmann et al. 2001; Quinlan 1996a, b; Xiaoliang et al. 2009, 2004; Korting 2006; Pan et al. 2003; Sornlertlamvanich et al. 2000, 2008; Steven 1994; Mazid et al. 2016; Muniyandi et al. 2012), there is not any CA—related work which is similar to our study.

In many studies related to the decision tree for sentiment classification (opinion analysis, semantic classification) in the world and in (Mita 2011; Taboada et al. 2008; Nizamani et al. 2012; Wan et al. 2015; Winkler et al. 2015, 20, 21; Vinodhini and Chandrasekaran 2013, 23, 24; Opinion 2015; Prasad et al. 2016, 27; Mugdha; Sharma 2014; Park et al. 2003; Loh and Mauricio 2003), there is not any CA—related research for semantic classification, which is similar to our work.

In many works related to the sentiment classification in the world and in (Manek et al. 2016; Agarwal and Mittal 2016a, b; Canuto et al. 2016, Kaur et al. 2016; Phu 2014; Tran et al. 2014; Li and Liu 2014), there is not any CA—related study for sentiment classification, which is similar to our model.

In many researches related to the unsupervised classification in the world and in (Turney 2002; Lee et al. 2002; Zyl 2002; Le Hegarat-Mascle et al. 2002; Ferro-Famil and Pottier 2002; Chaovalit and Zhou 2005; Te-Won; Lee and Lewicki 2002; Gllavata et al. 2004), there is not any CA—related study of unsupervised classification, which is similar to our work.

According to the CA in (Ruggieri 2002; Kretschmann et al. 2001; Quinlan 1996a, b; Xiaoliang et al. 2009, 2004; Korting 2006; Pan et al. 2003; Sornlertlamvanich et al. 2000; Rajeswari and Kannan 2008; Steven 1994; Mazid et al. 2016; Muniyandi et al. 2012), there are many advantages and disadvantages of the CA. Many advantages of the CA are as follows: builds models that can be easily interpreted; easy to implement; can use both categorical and continuous values; deals with noise. Many disadvantages of the CA are as follows: small variation in data can lead to different decision trees (especially when the variables are close to each other in value); does not work very well on a small training set.

Based on the works related to the C4.5 algorithm in (Ruggieri 2002; Kretschmann et al. 2001; Quinlan 1996a, b; Xiaoliang et al. 2009, 2004; Korting 2006; Pan et al. 2003; Sornlertlamvanich et al. 2000; Rajeswari and Kannan 2008; Steven 1994; Mazid et al. 2016; Muniyandi et al. 2012), we build the CA—related algorithms to perform our new model.

The motivation of the work is as follows: rule—based sentiment classification often has high accuracy and the rules are very popular in data mining. Researchers have sought to find many ways to use data mining rules in opinion analysis and to find the many different relationships between data mining and natural language processing. The C4.5 algorithm is a very popular and significant algorithm of the data mining, thus, the rules are generated by the C4.5 algorithm are very correct. This will result in many discoveries in scientific research, hence the motivation for this study.

The proposed approach is quite novel. The semantic analysis of an English document is based on many English sentences in the English training data set. The emotional classification of an English document is based on many association rules in the data mining field. Sentiment analysis is based on the FA algorithm. These principles are proposed to classify the semantics of an English document and data mining is used in natural language processing.

According to the researches in the world and in (Ruggieri 2002; Kretschmann et al. 2001; Quinlan 1996a, b; Xiaoliang et al. 2009, 2004; Korting 2006; Pan et al. 2003; Sornlertlamvanich et al. 2000; Rajeswari and Kannan 2008; Steven 1994; Mazid et al. 2016; Muniyandi et al. 2012; Mita 2011; Taboada et al. 2008; Nizamani et al. 2012; Wan et al. 2015; Winkler et al. 2015; Psomakelis et al. 2015; Shrivastava and Nair 2015; Vinodhini and Chandrasekaran 2013; Voll et al. 2007; Mandal et al. 2014; Kaur et al. 2015, 2016; Prasad et al. 2016, 27; Mugdha; Sharma 2014; Park et al. 2003; Loh and Mauricio 2003; Manek et al. 2016; Agarwal and Mittal 2016a, b; Canuto et al. 2016; Phu and Tuoi 2014; Tran et al. 2014; Li and Liu 2014; Turney 2002; Lee et al. 2002; Zyl 2002; Le Hegarat-Mascle et al. 2002; Ferro-Famil and Pottier 2002; Chaovalit and Zhou 2005; Lee and Lewicki 2002; Gllavata et al. 2004), to understand the significant contributions of this study, we present briefly as follows:

a. The C4.5 algorithm is a decision tree algorithm, but it is applied into the NLP.
b. It is not used in the sentiment classification, however, it is applied in the opinion analysis.
c. It is not used for the English document semantic analysis, whereas, it it applied in the emotional classification of the English documents.
d. From the results of this survey, it is widely applied in the different fields and the different applications.
e. This model can be applied into the other languages.
f. The C4.5—related algorithms are built in this search.
g. The rules are generated in this model.

Based on the above contributions, the model is clear superiority which is compared with the other methodologies and it is completely different from the other methods/models.

This study contains 6 sections: Sect. 1 is the introduction; Sect. 2 discusses the related works about the C4.5, etc., Sect. 3 is about the English data set of classifying sentences; Sect. 4 represents the methodology of our proposed model; Sect. 5 represents the experimental model and experimental results in this study; the conclusion of the proposed model is in Sect. 6. In addition, the References section displays many reference researches, and all the tables are shown in the Appendices section. Finally, all the codes of all algorithms in the Methodology are shown in the "Appendices of All Codes" section.

2 Related work

In this part, we summarize many studies related to our research, such as C4.5, sentiment analysis, etc.

There are many works related to the C4.5 algorithm in (Ruggieri 2002; Kretschmann et al. 2001; Quinlan 1996a, b; Xiaoliang et al. 2009, 2004; Korting 2006; Pan et al. 2003; Sornlertlamvanich et al. 2000; Rajeswari and Kannan 2008; Steven 1994; Mazid et al. 2016; Muniyandi et al. 2012). (Ruggieri 2002) Authors present an analytic evaluation of the runtime behavior of the C4.5 algorithm which highlights some efficiency improvements. Based on the analytic evaluation, we have implemented a more efficient version of the algorithm, called EC4.5. It improves on C4.5 by adopting the best among the three strategies for computing the information gain of continuous attributes. All the strategies adopt a binary search of the threshold in the whole training set starting from the local threshold computed at a node. The first strategy computes the local threshold using the algorithm of C4.5, which, in particular, sorts cases by means of the quicksort method. The second strategy also uses the algorithm of C4.5, but adopts a counting sort method. The third strategy calculates the local threshold using a main-memory version of the RainForest algorithm, which does not need sorting. The authors' implementation computes the same decision trees as C4.5 with a performance gain of up to five times. (Kretschmann et al. 2001) The gap between the amount of newly submitted protein data and reliable functional annotation in public databases is growing. Traditional manual annotation by literature curation and sequence analysis tools without the use of automated annotation systems is not able to keep up with the ever increasing quantity of data that is submitted. Automated supplements to manually curated databases such as TrEMBL or GenPept cover raw data, but provide only limited annotation. To improve this situation automatic tools are needed that support manual annotation, automatically increase the amount of reliable information and help to detect inconsistencies in manually generated annotations. A standard data mining algorithm was successfully applied to gain knowledge about the Keyword annotation in SWISS-PROT. 11 306 rules were generated, which are provided in a database and can be applied to yet un-annotated protein sequences and viewed using a web browser. They rely on the taxonomy of the organism, in which the protein was found and on signature matches of its sequence. The statistical evaluation of the generated rules by cross-validation suggests that by applying them on arbitrary proteins 33% of their keyword annotation can be generated with an error rate of 1.5%. The coverage rate of the keyword annotation can be increased to 60% by tolerating a higher error rate of 5%, etc.

Then, we compare our proposed model's results with the surveys in (Ruggieri 2002; Kretschmann et al. 2001; Quinlan 1996a, b; Xiaoliang et al. 2009, 2004; Korting 2006; Pan et al. 2003; Sornlertlamvanich et al. 2000; Rajeswari and Kannan 2008; Steven 1994; Mazid et al. 2016; Muniyandi et al. 2012; Mita 2011; Taboada et al. 2008; Nizamani et al. 2012; Wan et al. 2015; Winkler et al. 2015; Psomakelis et al. 2015; Shrivastava and Nair 2015; Vinodhini and Chandrasekaran 2013; Voll et al. 2007; Mandal et al. 2014; Kaur et al. 2015, 2016; Prasad et al. 2016, 27; Mugdha; Sharma 2014; Park et al. 2003; Loh and Mauricio 2003, 31, 32, 33, 34; Phu and Tuoi 2014; Tran et al. 2014; Li and Liu 2014; Turney 2002; Lee et al. 2002; Zyl 2002; Le Hegarat-Mascle et al. 2002; Ferro-Famil and Pottier 2002; Chaovalit and Zhou 2005; Lee and Lewicki 2002; Gllavata et al. 2004; Phu et al. 2016, 2017a, b; Friedl and Brodley 1997; Freund and Mason 1999; Payne et al. 1978; Chang 1977; Mehta et al. 1995; Phu et al. 2017).

There are many researches related to a decision tree for sentiment classification in (Mita 2011; Taboada et al. 2008; Nizamani et al. 2012; Wan et al. 2015; Winkler et al. 2015;

Psomakelis et al. 2015; Vinodhini and Chandrasekaran 2013, 23; Mandal et al. 2014; Kaur et al. 2015; Prasad et al. 2016; Pong-Inwong et al. 2014; Mugdha; Sharma 2014; Park et al. 2003; Loh and Mauricio 2003). Automatic Text Classification (Mita 2011) is a semi-supervised machine learning task that automatically assigns a given document to a set of pre-defined categories based on its textual content and extracted features. Automatic Text Classification has important applications in content management, contextual search, opinion mining, analysis of product review, spam filtering and text sentiment mining. This survey (Mita 2011) explains the generic strategy for automatic text classification and surveys existing solutions. The authors in (Taboada et al. 2008) present an approach to extracting sentiment from texts that makes use of contextual information. Using two different approaches, the authors (Taboada et al. 2008) extract the most relevant sentences of a text, and calculate the semantic orientation weighing those more heavily, etc.

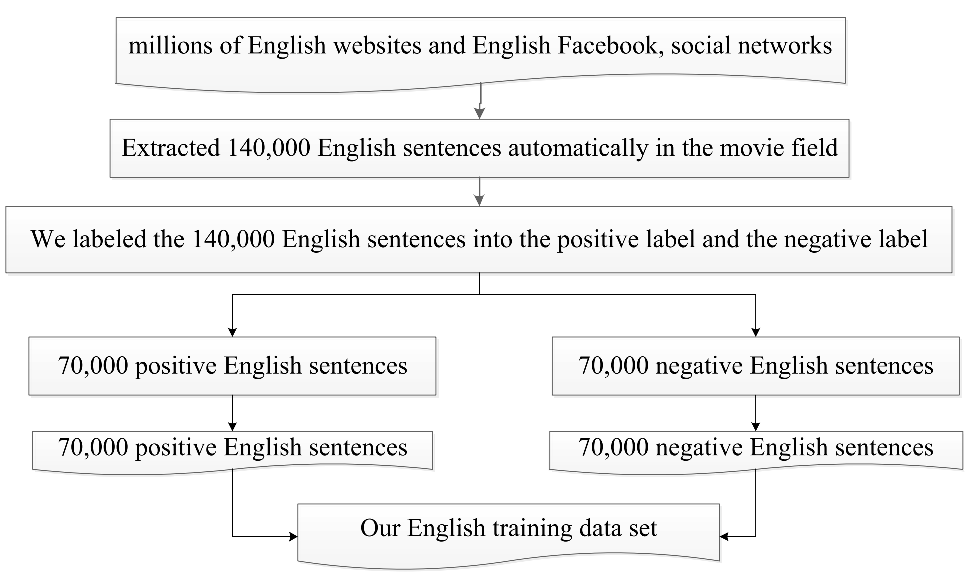
The latest researches of the sentiment classification are (Manek et al. 2016; Agarwal and Mittal 2016a, b, 34; Kaur et al.2016; Phu 2014; Tran et al. 2014; Li and Liu 2014; Phu et al. 2017a, b; Phu et al. 2017). With the rapid development of the World Wide Web in (Manek et al. 2016), electronic word-of-mouth 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 survey (Manek et al. 2016), a Gini Index based feature selection method with Support Vector Machine (SVM) classifier is proposed for sentiment classification for large movie review dataset. Opinion Mining or Sentiment Analysis in Agarwal an Mittal (2016a) is the 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 of availability and popularity of online review sites, blogs', forums', and social networking sites' necessity of analyzing and understanding these reviews has arisen. The main approaches for sentiment analysis can be categorized into semantic orientation-based approaches, knowledge-based, and machine-learning algorithms. This work (Agarwal an Mittal 2016a) surveys the machine learning approaches applied to sentiment analysis-based applications, etc.

The latest works of the unsupervised classification are (Turney 2002; Lee et al. 2002; Zyl 2002; Le Hegarat-Mascle et al. 2002; Ferro-Famil and Pottier 2002; Chaovalit and Zhou 2005; Lee and Lewicki 2002; Gllavata et al. 2004). This study in (Turney 2002) presents a simple unsupervised learning algorithm for classifying reviews as recommended (thumbs up) or not recommended (thumbs down). The authors in (Lee et al. 2002) propose a new method for unsupervised classification of terrain types and man-made objects using polarimetric synthetic aperture radar (SAR) data, etc.

3 Data set

In the Fig. 1, the English training data set includes 140,000 English sentences in the movie field, which contains 70,000 positive English sentences and 70,000 negative English sentences. All English sentences in our English

Fig. 1 Our English training data set



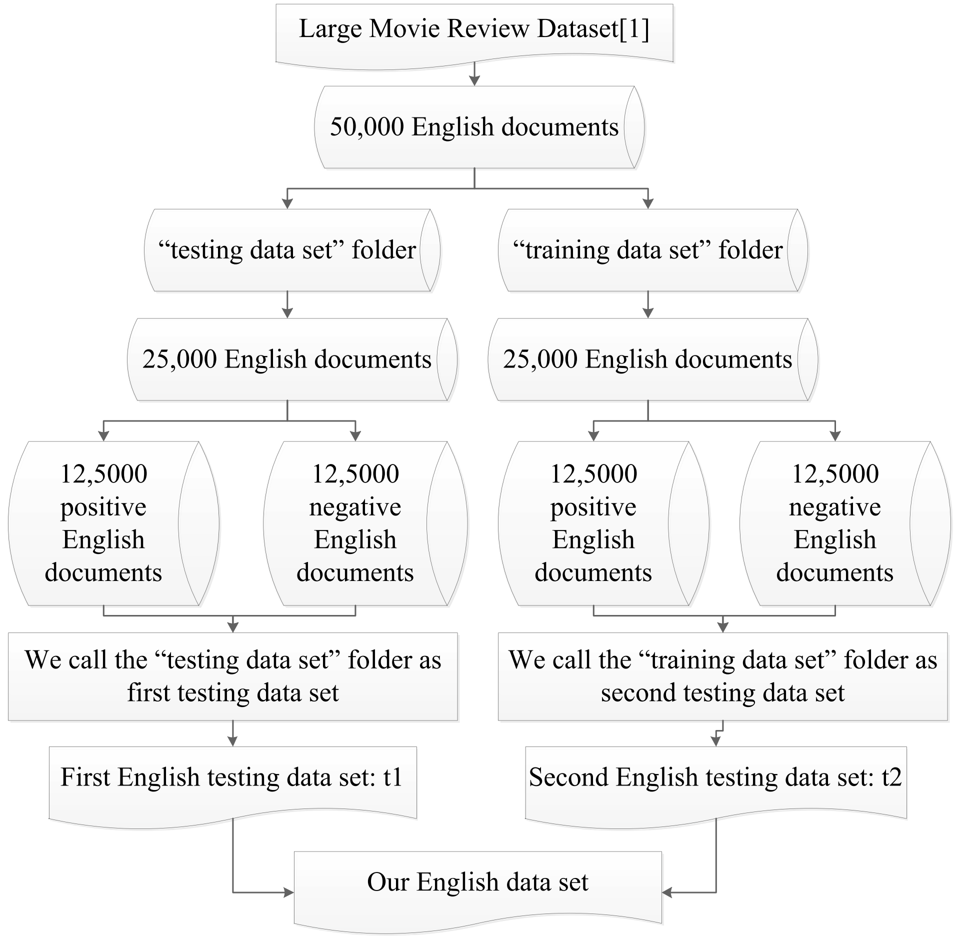


Fig. 2 Our English testing data set

training data set are automatically extracted from English Facebook, English websites; then we labeled positive and negative for them.

In Fig. 2, we use a public available large data set of classified movie reviews from the Internet Movie Database (IMDb) (Large 2016). This English data set includes two parts in two different folders. The first part is in the “testing data set” folder, it was named as the testing data set and we call it as the first testing data set; the second part is in the “training data set” folder, it was named as the training data set and we call it as the second testing data set. Both our first testing data set and our second testing data set have 25,000 English documents; and each the data set includes 12,500 positive English movie reviews and 12,500 negative English movie reviews.

4 Methodology

In this section, we present how our new model is implemented. First of all, the table of training dataset is created on the 70,000 positive sentences and the 70,000 negative sentences. Secondly, the C4.5 algorithm (CA) is applied to the table of the training dataset for generating the positive association rule set and the negative association rule set. Next, one English document of the English testing dataset is split into many English sentences. Then, the positive association rule set and the negative association rule set are applied to each English sentence of the English document, and the emotional classification of the English sentence is identified. Finally, the semantic classification of the English document is identified on its sentences.

In Fig. 3, this research is done as follows diagram below.

The criteria of selection both positive and negative association rules are certainly dependent on the English training data set and the algorithm for generating them (in the paper, the algorithm is the C4.5 algorithm). The positive and negative association rules are very important for this model to identify the emotional polarities (positive, negative, neutral) of one English sentence. Then, the semantic classification of one English document is identified on its sentences.

We propose many algorithms to perform the model.

We build algorithm 1 to create the table of training data has 140,000 records (or 140,000 rows) and (m_max + 1) columns. Each English sentence in all the

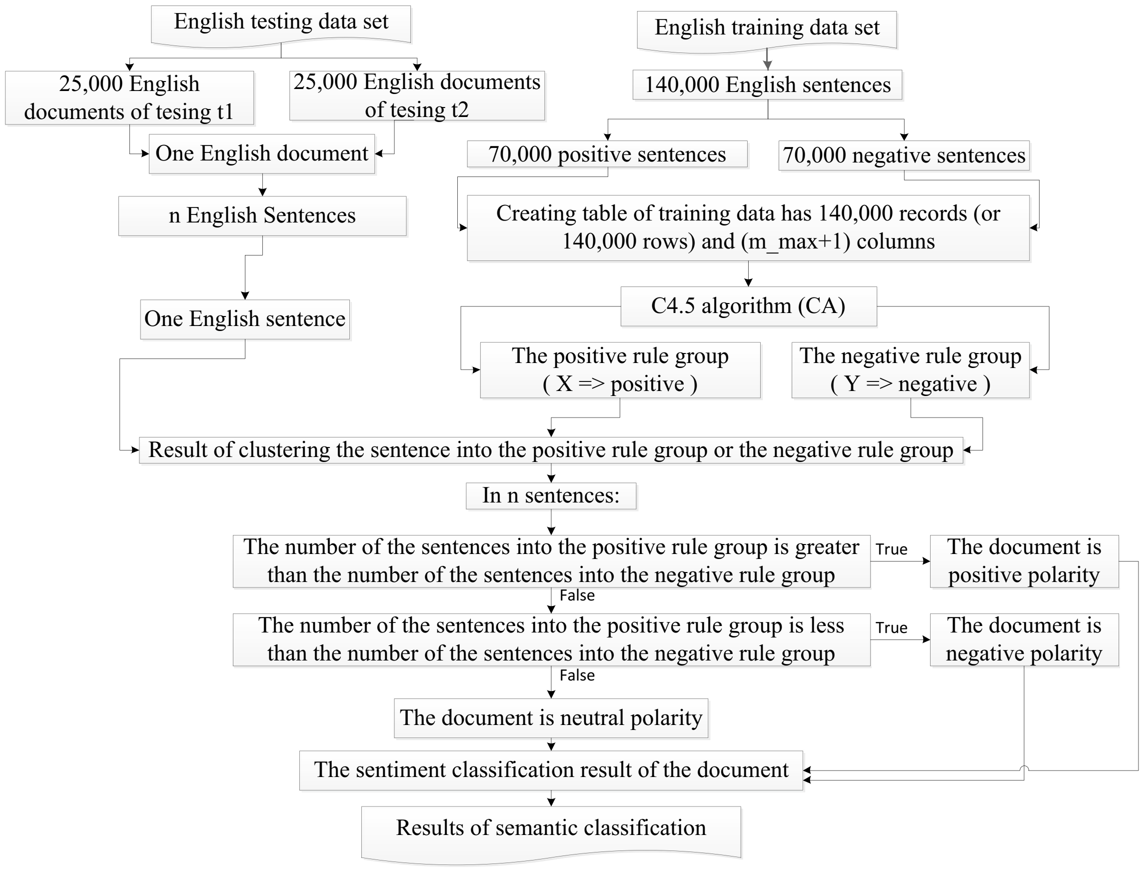


Fig. 3 Overview process of our new model

sentences of the training data set is split into the meaningful phrases (or the meaningful words). Each row of the table tableOfTrainingData is each English sentence. The columns of each row in the tableOfTrainingData are the meaningful phrases (or the meaningful words) of each English sentence in all the sentences of the English training data set.

The algorithm 1 is presented more detail in the Code 1 below. The main ideas of the algorithm 1 are as follows:

- Input: 140,000 English sentences of the English training data set including the 70,000 English positive sentences and the 70,000 English negative sentences
- Output: table of training data.
- Step 1: Create table tableOfTrainingData which has (m_max + 1) columns and 140,000 rows.
- Step 2: With each sentence (one sentence) in the 70,000 English positive sentences of the 140,000 sentences, do repeat:
- Step 3: Split this sentence into many words (or phrases) based on ‘ ’ or “ ”: arrayWords. Assuming that m is a number of words (or phraes) of this sentence which is split.
- Step 4: Create one new row in table tableOfTrainingData: NewRow
- Step 5: Do repeat i from 0 (the head of this sentence) to m-1 (the tail of this sentence):
- Step 6: NewRow.column[i] = arrayWords[i]
- Step 7: End of Step 5
- Step 8: If i is less than m_max Then: do repeat
- Step 9: NewRow.column[i] = 0 (or “ ”)
- Step 10: End of Step 8
- Step 11: NewRow.Column[m_max] = “positive”
- Step 12: End of Step 2
- Step 13: With each sentence (one sentence) in the 70,000 English negative sentences of the 140,000 sentences, do repeat:
- Step 14: Split this sentence into many words (or phrases) based on ‘ ’ or “ ”: arrayWords. Assuming that m is a number of words (or phraes) of this sentence which is split.
- Step 15: Create one new row in table tableOfTrainingData: NewRow
- Step 16: Do repeat i from 0 (the head of this sentence) to m-1 (the tail of this sentence):
- Step 17: NewRow.column[i] = arrayWords[i]

- Step 18: End of Step 16
- Step 19: If i is less than m_max Then: do repeat
- Step 20: NewRow.column[i] = 0 (or " ")
- Step 21: End of Step 19
- Step 22: NewRow.Column[m_max] = "negative"
- Step 23: End of Step 13
- Step 24: Return table tableOfTrainingData

According to the C4.5 algorithm in (Ruggieri 2002; Kretschmann et al. 2001; Quinlan 1996a, b; Xiaoliang et al. 2009, 2004; Korting 2006; Pan et al. 2003; Sornlertlamvanich et al. 2000; Rajeswari and Kannan 2008; Steven 1994; Mazid et al. 2016; Muniyandi et al. 2012), we build algorithm 2 to generate many association rules in the positive rule group and the negative rule group by using the C4.5 algorithm. The basic construction of C4.5 decision tree is

1. The root nodes are the top node of the tree. It considers all samples and selects the attributes that are most significant.
2. The sample information is passed to subsequent nodes, called 'branch nodes' which eventually terminate in leaf nodes that give decisions.
3. Rules are generated by illustrating the path from the root node to leaf node.

The algorithm 2 is presented more detail in the Code 2 below. The main ideas of the algorithm 2 are as follows:

Input:

- Table of training data tableOfTrainingData is the training examples.
- Attributes S is a list of other attributes that may be tested by the learned decision tree. (column from 0 to m_max − 1 of tableOfTrainingData)
- A decision tree (actually the root node of the tree) that correctly classifies the given Examples. This decision tree is divided into the positive rule group and the negative rule group.

Output: the positive rule group and the negative rule group.

From Step 1 to Step 26: Apply the C4.5 algorithm to the table tableOfTrainingData

- Step 27: Set positiveRuleGroup := null
- Step 28: Set negativeRuleGroup := null
- Step 29: Browse decision tree Tree, do:
- Step 30: If the rule is positive Then
- Step 31:positiveRuleGroup.Add (the rule);
- Step 32: Else If the rule is negative Then
- Step 33:negativeRuleGroup.Add (the rule);
- Step 34: End of Step 30
- Step 35: End of Step 29
- Step 36: Return positiveRuleGroup and negativeRuleGroup;

We build algorithm 3 to classify one English sentence into the positive polarity, the negative polarity or the neutral polarity. The positive association rule set in positiveRuleGroup and the negative association rule set in negativeRuleGroup are applied to one English sentence A. If the number of positive rules which A contains is greater than the number of negative rules which A contains, A is classified to the positive polarity. If the number of negative rules which A contains is less than the number of negative rules which A contains, A is classified to the negative polarity. If the number of positive rules which A contains is as equal as the number of negative rules which A contains, A is classified to the neutral polarity; or if A does not contain any positive rule and any negative rule, A is classified to the neutral polarity.

The algorithm 3 is presented more detail in the Code 3 below. The main ideas of the algorithm 3 are as follows:

- Input: one English sentence A, the positive rule group positiveRuleGroup and the negative rule group negativeRuleGroup
- Output: positive, negative, neutral of this sentence A.
- Step 1: With each rule (one rule) R in the positive rule group positiveRuleGroup, do repeat:
- Step 2: If the sentence A contains R Then
- Step 3: Set variable varibleOfPositive := varibleOfPositive + 1
- Step 4: End Of Step 2
- Step 5: End of Step 1
- Step 6: With each rule (one rule) R in the negative rule group negativeRuleGroup, do repeat:
- Step 7: If the sentence A contains R Then
- Step 8: Set variable varibleOfNegative := varibleOfNegative + 1
- Step 9: End Of Step 6
- Step 10: End of Step 7
- Step 11: If varibleOfPositive is greater than varibleOfNegative Then
- Step 12: Return positive
- Step 13: Else If varibleOfPositive is less than varibleOfNegative Then
- Step 14: Return negative
- Step 15: End If
- Step 16: Return neutral

We build algorithm 4 to classify one English document into the positive polarity, the negative polarity or the neutral polarity. The English document is classified to the positive polarity if the number of sentences classified to the positive polarity is greater than the number of sentences classified to the negative polarity in the document. The English document is classified to the negative polarity if the number of sentences classified to the positive polarity is less than the number of sentences classified to the negative polarity in the document. The English document is classified to the negative polarity if the number of sentences classified to the positive polarity is as equal as the number of sentences classified to the negative polarity in the document.

The algorithm 4 is presented more detail in the Code 4 below. The main ideas of the algorithm 4 are as follows:

- Input: one English document, including the n English sentences with the polarity result of each English sentence which is implemented by using the algorithm 3.
- Output: positive, negative, neutral of this English document
- Step 1: If the number of English sentences classified into the positive polarity is greater than the number of English sentences classified into the negative polarity in the document Then
- Step 2: Return positive;
- Step 3: End If
- Step 4: If the number of English sentences classified into the positive polarity is less than the number of English sentences classified into the negative polarity in the document Then
- Step 5: Return negative;
- Step 6: End If
- Step 7: Return neutral;

Or the main ideas of the algorithm 4 are as follows:

- Input: one English document A
- Output: positive, negative, neutral of this English document
- Step 1: Split this English document A into many English sentences: m sentences.
- Step 2: With each sentence (one sentence) i in m sentences, do repeat:
- Step 3: Run algorithm 3 with the sentence i
- Step 4: If the result is positive Then
- Step 5: Set variableOfPositive := variableOfPositive + 1
- Step 6: End of Step 4
- Step 7: If the result is negative Then
- Step 8: Set variableOfNegative := variableOfNegative + 1
- Step 9: End of Step 7
- Step 10: End of Step 2
- Step 11: If variableOfPositive is greater than variableOfNegative Then
- Step 12: Return positive
- Step 13: Else If variableOfPositive is less than variableOfNegative Then
- Step 14: Return negative
- Step 15: End of Step 11
- Step 16: Return neutral

5 Experiment

To implement the proposed model, we have already used Microsoft SQL Server 2008 R2 to save the English data sets and save the results of emotion classification.

Microsoft Visual Studio 2010 (C #) is used for programming to save data sets, implementing our proposed model to classify the 25,000 English documents of t1 and t2.

The experiment programs have been conducted on the Intel Dual laptop with Core i5 processor at 2.6 GHz Memory 8 GB; the operating system is Microsoft Windows 8.

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

The results of the 25,000 English documents of the testing data set t1 to test are presented in the Table 2 below in the Appendix.

The results of the 25,000 English documents of the testing data set t2 to test are presented in the Table 3 below in the "Appendix".

The accuracy of the 25,000 English documents in the testing dataset t1 is shown in the Table 4 below in the "Appendix".

The accuracy of the 25,000 English documents in the testing dataset t2 is shown in the Table 5 below in the "Appendix".

We also have the comparisons between our results with the surveys in the "Appendix".

Table 2 The results of the 25,000 English documents in testing data set t1

	Testing dataset t1	Correct classification	Incorrect classification
Negative	12,500	7,533	4,967
Positive	12,500	7,542	4,958
Summary	25,000	15,075	9,925

Table 3 The results of the 25,000 English documents in testing data set t2

	Testing dataset t2	Correct classification	Incorrect classification
Negative	12,500	7,584	4,916
Positive	12,500	7,591	4,909
Summary	25,000	15,175	9,825

Table 4 The accuracy of our new model for the 25,000 English documents in testing data set t1

Proposed Model	Class	Accuracy
This survey	Negative Positive	60.3%

Table 5 The accuracy of our new model for the 25,000 English documents in testing data set t2

Proposed Model	Class	Accuracy
This research	Negative Positive	60.7%

6 Conclusion

Classification result of 25,000 English documents of t1 data set by using our model has achieved accuracy 60.3 and 60.7% of t2 data set.

With the same of the English training data set, the classification results of the different English testing data sets are very different from each others. The classification results are depending on the association rules of the positive rule group and the negative rule group. The association rules of the positive rule group and the negative rule group are depending on the algorithms and the English training data sets.

With the same of the English training data set, the association rules of the positive rule group and the negative rule group are very different from each others by using the different algorithms. Thus, the classification results are very different from each others.

With the same of the algorithms, the association rules of the positive rule group and the negative rule group are very different from each others by using the different data sets. Thus, the classification results are very different from each others.

To increase the accuracy of the classification results significantly, we can increase the association rules of the positive rule group and the negative rule group certainly.

To increase the association rules of the positive rule group and the negative rule group significantly, we can improve the algorithms, or the English training data sets, or both the algorithms and the English training data sets.

Although our model's accuracy is not high, our model is a new contribution to English sentiment classification and sentiment classification of other languages.

Based on the basis the C4.5 algorithm, we build the algorithms related to the CA for performing our new model.

This model also has many benefits and drawbacks. The benefits of the model are as follows: the document-level emotional analysis is based on the English sentences. The rules are generated by the C4.5 algorithm are high correct. The rules are used in many researches and commercial applications. The drawbacks of the model are as follows: The accuracy of the model is low, because the rule-based sentiment classification often has better accuracy. It takes too much time to generate the rules.

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

In the Table 6 below, we compare our model's results with many researches related to the C4.5 algorithm in (Ruggieri 2002; Kretschmann et al. 2001; Quinlan 1996a, b; Xiaoliang et al. 2009, 2004; Korting 2006; Pan et al. 2003; Sornlertlamvanich et al. 2000; Rajeswari and Kannan 2008; Steven 1994; Mazid et al. 2016; Muniyandi et al. 2012).

In the Table 7 below, we compare our model's results with many researches related to the decision tree for sentiment classification in (Mita 2011; Taboada et al. 2008; Nizamani et al. 2012; Wan et al. 2015; Winkler et al. 2015; Vinodhini and Chandrasekaran 2013, 2007, 2014; Kaur et al. 2015; Prasad et al. 2016, 2014; Sharma 2014).

In the Table 8 below, we compare our model's results with the latest researches of the sentiment classification in (2016, Kaur et al. 2016; Phu 2014; Tran et al. 2014).

In the Table 9 below, we compare our model's results with the latest works of the unsupervised classification in (Turney 2002; Lee et al. 2002; Zyl 2002; Le Hegarat-Mascle et al. 2002; Ferro-Famil and Pottier 2002; Chaovalit and Zhou 2005; Lee and Lewicki 2002; Gllavata et al. 2004).

We compare our model with many algorithms for the decision tree in (Friedl and Brodley 1997; Freund and Mason 1999; Payne et al. 1978; Chang 1977; Mehta et al. 1995) in the Table 10.

Appendix

See Tables (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)

Table 6 Comparison our model's results with many researches related to the C4.5 algorithm in (Ruggieri 2002; Kretschmann et al. 2001; Quinlan 1996a, b; Xiaoliang et al. 2009, 2004; Korting 2006; Pan et al. 2003; Sornlertlamvanich et al. 2000; Rajeswari and kannan 2008; Steven 1994; Mazid et al. 2016; Muniyandi et al. 2012)

Works	SC	Language	SD	DT	c4.5 algorithm	Decision tree
Ruggieri (2002)	No	NM	Yes	Yes	Yes	Yes
Model/method of Ruggieri (2002)	Efficient C4.5 [classification algorithm]					
Summary of Ruggieri (2002)	The authors present an analytic evaluation of the runtime behavior of the C4.5 algorithm which highlights some efficiency improvements. Based on the analytic evaluation, the authors have implemented a more efficient version of the algorithm, called EC4.5. It improves on C4.5 by adopting the best among the three strategies for computing the information gain of continuous attributes. All the strategies adopt a binary search of the threshold in the whole training set starting from the local threshold computed at a node. The first strategy computes the local threshold using the algorithm of C4.5, which, in particular, sorts cases by means of the quicksort method. The second strategy also uses the algorithm of C4.5, but adopts a counting sort method. The third strategy calculates the local threshold using a main-memory version of the RainForest algorithm, which does not need sorting. The authors' implementation computes the same decision trees as C4.5 with a performance gain of up to five times					
Kretschmann (2001)	No	NM	Yes	Yes	Yes	Yes
Model/method of Kretschmann (2001)	Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT					
Summary of Kretschmann (2001)	The gap between the amount of newly submitted protein data and reliable functional annotation in public databases is growing. Traditional manual annotation by literature curation and sequence analysis tools without the use of automated annotation systems is not able to keep up with the ever increasing quantity of data that is submitted. Automated supplements to manually curated databases such as TrEMBL or GenPept cover raw data, but provide only limited annotation. To improve this situation automatic tools are needed that support manual annotation, automatically increase the amount of reliable information and help to detect inconsistencies in manually generated annotations. A standard data mining algorithm was successfully applied to gain knowledge about the Keyword annotation in SWISS-PROT. 11,306 rules were generated, which are provided in a database and can be applied to yet un-annotated protein sequences and viewed using a web browser. They rely on the taxonomy of the organism, in which the protein was found and on signature matches of its sequence. The statistical evaluation of the generated rules by cross-validation suggests that by applying them on arbitrary proteins 33% of their keyword annotation can be generated with an error rate of 1.5%. The coverage rate of the keyword annotation can be increased to 60% by tolerating a higher error rate of 5%					
Quinlan (1996a)	No	NM	Yes	Yes	Yes	Yes
Model/method of Quinlan (1996a)	Improved use of continuous attributes in C4.5					
Summary of Quinlan (1996a)	A reported weakness of C4.5 in domains with continuous attributes is addressed by modifying the formation and evaluation of tests on continuous attributes. An MDL-inspired penalty is applied to such tests, eliminating some of them from consideration and altering the relative desirability of all tests. Empirical trials sEnglish document in the English testing data set based onhow that the modifications lead to smaller decision trees with higher predictive accuracies. Results also confirm that a new version of C4.5 incorporating these changes is superior to recent approaches that use global discretization and that construct small trees with multi-interval splits					
Xiaoliang et al. (2009)	No	NM	Yes	Yes	Yes	Yes
Model/Method of Xiaoliang et al. (2009)	Research and application of the improved algorithm C4.5 on Decision tree					
Summary of Xiaoliang et al. (5)	The algorithm on the Decision tree is the most widely used method of inductive inference, and it is a simple method of knowledge representation, Different examples can be divided into representative categories, such as a classifier and prediction models. This work introduces the basic concepts of a classifier, the principle of the decision tree and algorithm ID3, analyses the algorithm C4.5 and gives further research to improve it, and the trials show that the improved algorithm has the reliable results and high efficiency					
Zhou and Jiang (2004)	No	NM	Yes	Yes	Yes	Yes
Model/method of Zhou and Jiang (2004)	NeC4.5: neural ensemble based C4.5					

Table 6 (continued)

Works	SC	Language	SD	DT	c4.5 algorithm	Decision tree
Summary of Zhou and Jiang (2004)	The decision tree is with good comprehensibility while neural network ensemble is with strong generalization ability. These merits are integrated into a novel decision tree algorithm NeC4.5. This algorithm trains a neural network ensemble at first. Then, the trained ensemble is employed to generate a new training set through replacing the desired class labels of the original training examples with those outputs from the trained ensemble. Some extra training examples are also generated from the trained ensemble and added to the new training set. Finally, a C4.5 decision tree is grown from the new training set. Since its learning results are decision trees, the comprehensibility of NeC4.5 is better than that of te neural network ensemble. Moreover, experiments show that the generalization ability of NeC4.5 decision trees can be better than that of C4.5 decision trees					
Korting (2006)	No	NM	Yes	Yes	Yes	Yes
Model/method of Korting (2006)	C4.5 algorithm and multivariate decision trees					
Summary of Korting (2006)	The aim of this article is to show a brief description about the C4.5 algorithm, used to create Univariate Decision Trees. The authors also talk about Multivariate Decision Trees, their process to classify instances using more than one attribute per node in the tree. The authors try to discuss how they work, and how to implement the algorithms that build such trees, including examples of Univariate and Multivariate results					
Pan et al. (2003)	No	NM	Yes	Yes	Yes	Yes
Model/method of Pan et al. (2003)	Hybrid neural network and C4.5 for misuse detection					
Summary of Pan et al. (2003)	Intrusion detection technology is an effective approach to dealing with the problems of network security. In this study, the authors present an intrusion detection model based on hybrid neural network and C4.5. The key idea is to take advantage of different classification abilities of neural network and the C4.5 algorithm for different attacks. What is more, the model could also be updated by the C4.5 rules mined from the dataset after the event (intrusion). The authors employ data from the third international knowledge discovery and data mining tools competition (KDDcup '99) to train and test feasibility of the authors' proposed model. From the authors' experimental results with different network data, the authors' model achieves more than 85 percent detection rate on average, and less than 19.7 percent false alarm rate for five typical types of attacks. Through the analysis after-the-event module, the average detection rate of 93.28 percent and false positive rate of 0.2 percent can respectively be obtained					
Sornlertlamvanich et al. (2000)	No	NM	Yes	Yes	Yes	Yes
Model/method of Sornlertlamvanich et al. (2000)	Automatic corpus-based Thai word extraction with the c4.5 learning algorithm					
Summary of Sornlertlamvanich et al. (2000)	"Word" is difficult to define in the languages that do not exhibit explicit word boundary, such as Thai. Traditional methods on defining words for this kind of languages have to depend on human judgement which bases on unclear criteria or procedures, and have several limitations. This research proposes an algorithm for word extraction from Thai texts without borrowing a hand from word segmentation. The authors employ the c4.5 learning algorithm for this task. Several attributes such as string length, frequency, mutual information and entropy are chosen for word/non-word determination. The authors' experiment yields high precision results about 85% in both training and test corpus					
Quinlan (1996b)	No	NM	Yes	Yes	Yes	Yes
Model/method of Quinlan (1996b)	Bagging, boosting, and C4.5					
Summary of Quinlan (1996b)	Breiman's bagging and Freund and Schapire's boosting are recent methods for improving the predictive power of classifier learning systems. Both form a set of classifiers that are combined by voting, bagging by generating replicated bootstrap samples of the data, and boosting by adjusting the weights of training instances. This study reports results of applying both techniques to a system that learns decision trees and testing on a representative collection of datasets. While both approaches substantially improve predictive accuracy, boosting shows the greater benefit. On the other hand, boosting also produces severe degradation on some data sets. A small change to the way that boosting combines the votes of learned classifiers reduces this downside and also leads to slightly better results on most of the datasets considered					

Table 6 (continued)

Works	SC	Language	SD	DT	c4.5 algorithm	Decision tree
Rajeswari and Kannan (2008)	No	NM	Yes	Yes	Yes	Yes
Model/method of Rajeswari and Kannan (2008)	An active rule approach for network intrusion detection with enhanced C4.5 algorithm					
Summary of Rajeswari and Kannan (2008)	Intrusion detection systems provide additional defense capacity to a networked information system in addition to the security measures provided by the firewalls. This work proposes an active rule based enhancement to the C4.5 algorithm for network intrusion detection in order to detect misuse behaviors of internal attackers through effective classification and decision making in computer networks. This enhanced C4.5 algorithm derives a set of classification rules from network audit data and then the generated rules are used to detect network intrusions in a real-time environment. Unlike most existing decision trees based approaches, the spawned rules generated and fired in this work are more effective because the information-theoretic approach minimizes the expected number of tests needed to classify an object and guarantees that a simple (but not necessarily the simplest) tree is found. The main advantage of this proposed algorithm is that the generalization ability of enhanced C4.5 decision trees is better than that of C4.5 decision trees. The authors have employed data from the third international knowledge discovery and data mining tools competition (KDDcup'99) to train and test the feasibility of this proposed model. By applying the enhanced C4.5 algorithm an average detection rate of 93.28 percent and a false positive rate of 0.7 percent have respectively been obtained in this work					
Salzberg (1994)	No	NM	Yes	Yes	Yes	Yes
Model/method of Salzberg (1994)	C4.5: programs for machine learning					
Summary of Salzberg (1994)	Algorithms for constructing decision trees are among the most well-known and widely used of all machine learning methods. Among decision tree algorithms, J. Ross Quinlan'sID3 and its successor, C4.5, are probably the most popular in the machine learning community. These algorithms and variations on them have been the subject of numerous research works since Quinlan introduced ID3. Until recently, most researchers looking for an introduction to decision trees turned to Quinlan's seminal 1986 Machine Learning journal article [Quinlan, 1986]. In his new work, C4.5: Programs for Machine Learning, Quinlan has put together a definitive, much needed description of his complete system, including the latest developments. As such, this study will be a welcome addition to the library of many researchers and students					
Mazid et al. (2016)	No	NM	Yes	Yes	Yes	Yes
Model/Method of Mazid et al. (2016)	Improved C4.5 algorithm for rule based classification					
Summary of Mazid et al. (2016)	C4.5 is one of the most popular algorithms for rule base classification. There are many empirical features in this algorithm such as continuous number categorization, missing value handling, etc. However, in many cases it takes more processing time and provides less accuracy rate for correctly classified instances. On the other hand, a large dataset might contain hundreds of attributes. Authors need to choose most related attributes among them to perform higher accuracy using C4.5. It is also a difficult task to choose a proper algorithm to perform efficient and perfect classification. With the authors' proposed method, we select the most relevant attributes from a dataset by reducing input space and simultaneously improve the performance of this algorithm. The improved performance is measured based on better accuracy and less computational complexity. The authors' measure Entropy of Information Theory to identify the central attribute for a dataset. Then apply correlation coefficient measure, namely, Pearson's, Spearman, Kendall correlation utilizing the central attribute of the same data set. The authors conduct a comparative study using these three most popular correlation coefficient measures to choose the best method on eight well known data mining problem from UCI (University of California Irvine) data repository. The authors use box plot to compare experimental results. The authors' proposed method shows better performance in most of the individual experiments					
Muniyandi et al. (2012)	No	NM	Yes	Yes	Yes	Yes
Model/method of Muniyandi et al. (2012)	Network anomaly detection by cascading K-means clustering and C4.5 decision tree algorithm					

Table 6 (continued)

Works	SC	Language	SD	DT	c4.5 algorithm	Decision tree
The summary of Muniyandi et al. (2012)	Intrusions pose a serious securing risk in a network environment. Network intrusion detection system aims to identify attacks or malicious activity in a network with a high detection rate while maintaining a low false alarm rate. Anomaly detection systems (ADS) monitor the behavior of a system and flag significant deviations from the normal activity as anomalies. In this work, the authors propose an anomaly detection method using "K-Means + C4.5", a method to cascade k-Means clustering and the C4.5 decision tree methods for classifying anomalous and normal activities in a computer network. The k-Means clustering method is first used to partition the training instances into k clusters using Euclidean distance similarity. On each cluster, representing a dense region of normal or anomaly instances, the authors build decision trees using C4.5 decision tree algorithm. The decision tree on each cluster refines the decision boundaries by learning the subgroups within the cluster. To obtain a final conclusion the authors exploit the results derived from the decision tree on each cluster					
Our study	Yes	English	Yes	Yes	Yes	Yes
Model/method of our study	C4.5 Algorithm for English sentiment classification					
The summary of our study	We use the C4.5 algorithm to classify semantics (positive, negative, neutral) of one English document in the English testing data set based on 140,000 English sentences of English training data set which includes 70,000 English positive sentences and 70,000 English negative sentences					

SC sentiment classification, *SD* special domain, *DT* depending on the training data set, *VL* Vietnamese language, *EL* English language, *NM* no mention

Table 7 Comparison our model's results with many researches related to the decision tree for sentiment classification in (Mita 2011; Taboada et al. 2008; Nizamani et al. 2012; Wan et al. 2015; Winkler et al. 2015; Vinodhini and Chandrasekaran 2013, 2007, 2014; Kaur et al. 2015; Prasad et al. 2016, 2014; Sharma 2014)

Works	SC	Language	SD	DT	C4.5 Algorithm	Decision Tree
Dalal and Zaveri (2011)	Yes	English	Yes	Yes	Yes	Yes
Model/method of Dalal and Zaveri (2011)	Automatic text classification: a technical review					
Summary of Dalal and Zaveri (2011)	Automatic Text Classification is a semi-supervised machine learning task that automatically assigns a given document to a set of pre-defined categories based on its textual content and extracted features. Automatic Text Classification has important applications in content management, contextual search, opinion mining, product review analysis, spam filtering and text sentiment mining. This work explains the generic strategy for automatic text classification and surveys existing solutions to major issues such as dealing with unstructured text, handling large number of attributes and selecting a machine learning technique appropriate to the text-classification application					
Taboada et al. (2008)	Yes	English	Yes	Yes	Yes	Yes
Model/method of Taboada et al. (2008)	Extracting sentiment as a function of discourse structure and topicality					
Summary of Taboada et al. (2008)	Authors present an approach to extracting sentiment from texts that makes use of contextual information. Using two different approaches, we extract the most relevant sentences of a text, and calculate the semantic orientation weighing those more heavily. The first approach makes use of discourse structure via Rhetorical Structure Theory, and extracts nuclei as the relevant parts; the second approach uses a topic classifier built using support vector machines, which extracts topic sentences from texts. The use of weights on relevant sentences shows an improvement over word-based methods that consider the entire text equally. In the study, the authors also describe an enhancement of our previous word-based methods in the treatment of intensifiers and negation, and the addition of other parts of speech beyond adjectives					
Nizamani et al. (2013)	Yes	English	Yes	Yes	Yes	Yes
Model/method of Nizamani et al. (2013)	Modeling suspicious email detection using enhanced feature selection					

Table 7 (continued)

Works	SC	Language	SD	DT	C4.5 Algorithm	Decision Tree
Summary of Nizamani et al. (2013)	The research presents a suspicious email detection model which incorporates an enhanced feature selection. In the work authors proposed the use of feature selection strategies along with classification techniques for terrorists email detection. The presented model focuses on the evaluation of machine learning algorithms such as decision tree (ID3), logistic regression, Naive Bayes (NB), and Support Vector Machine (SVM) for detecting emails containing suspicious content. In the literature, various algorithms achieved good accuracy for the desired task. However, the results achieved by those algorithms can be further improved by using appropriate feature selection mechanisms. The authors have identified the use of a specific feature selection scheme that improves the performance of the existing algorithms					
Wan and Gao (2015]	Yes	English	Yes	Yes	Yes	Yes
Model/method of Wan and Gao (2015]	An ensemble sentiment classification system of twitter data for airline services analysis					
Summary of Wan and Gao (2015]	In airline service industry, it is difficult to collect data about customers' feedback by questionnaires, but Twitter provides a sound data source for them to do customer sentiment analysis. However, little research has been done in the domain of Twitter sentiment classification about airline services. In this study, an ensemble sentiment classification strategy was applied based on Majority Vote principle of multiple classification methods, including Naive Bayes, SVM, Bayesian Network, C4.5 Decision Tree and Random Forest algorithms. In the authors' experiments, six individual classification approaches, and the proposed ensemble approach were all trained and tested using the same data set of 12,864 tweets, in which 10-fold evaluation is used to validate the classifiers. The results show that the proposed ensemble approach outperforms these individual classifiers in this airline service Twitter dataset. Based on the authors' observations, the ensemble approach could improve the overall accuracy in twitter sentiment classification for other services as well					
Winkler et al. (2015)	Yes	German	Yes	Yes	Yes	Yes
Model/Method of Winkler et al. (2015)	Data-based prediction of sentiments using heterogeneous model ensembles					
Summary of Winkler et al. (2015)	In this work, the authors present an ensemble modeling approach for sentiment analysis using machine learning algorithms. The main goal of sentiment analysis is to develop estimators that are able to identify the sentiment orientation (positive, negative, or neutral) of sentences found in any arbitrary source. The novel approach presented here relies on the analysis of the words found in sentences and the formation of large sets of heterogeneous models, i.e., binary as well as multi-class classification models that are calculated by various different machine learning methods; these models shall represent the relationship between the presence of given words (or combination of words) and sentiments. All models trained during the learning phase are applied during the test phase and the final sentiment assessment is annotated with a confidence value that specifies, how reliable the models are regarding the presented decision. In the empirical part of this study, the authors show results achieved using a German corpus of Amazon recensions and a set of machine learning methods (decision trees and adaptive boosting, Gaussian processes, random forests, k-nearest neighbor classification, support vector machines and artificial neural networks with evolutionary feature and parameter optimization, and genetic programming). Using a heterogeneous model ensemble learning approach that combines multi-class classifiers as well as binary classifiers, the classification accuracy can be increased significantly and the ratio of totally wrongly classified samples (i.e., those that are assigned to the completely opposite sentiment orientation) can be decreased significantly					
Psomakelis et al. (2015)	Yes	English	Yes	Yes	Yes	Yes
Model/Method of Psomakelis et al. (2015)	Comparing methods for twitter sentiment analysis					

Table 7 (continued)

Works	SC	Language	SD	DT	C4.5 Algorithm	Decision Tree
Summary of Psomakelis et al. (2015)	This work extends the set of works which deal with the popular problem of sentiment analysis in Twitter. It investigates the most popular document ("tweet") representation methods which feed sentiment evaluation mechanisms. In particular, the authors study the bag-of-words, n-grams and n-gram graphs approaches and for each of them the authors evaluate the performance of a lexicon-based and 7 learning-based classification algorithms (namely SVM, Naive Bayesian Networks, Logistic Regression, Multilayer Perceptrons, Best-First Trees, Functional Trees and C4.5) as well as their combinations, using a set of 4451 manually annotated tweets. The results demonstrate the superiority of learning-based methods and in particular of n-gram graphs approaches for predicting the sentiment of tweets. They also show that the combinatory approach has impressive effects on n-grams, raising the confidence up to 83.15% on the 5-Grams, using majority vote and a balanced dataset (equal number of positive, negative and neutral tweets for training). In the n-gram graph cases the improvement was small to none, reaching 94.52% on the 4 g graphs, using Orthodromic distance and a threshold of 0.001					
Shrivastava and Nair (2015)	Yes	English	Yes	Yes	Yes	Yes
Model/method of Shrivastava and Nair (2015)	Mood prediction on tweets using classification algorithm					
Summary of Shrivastava and Nair (2015)	Data mining is a technique which offers the computer algorithm to compute patterns and find the category of data using classification and clustering. In data mining classification is performed with supervised learning and unsupervised learning. Selection of algorithm depends upon the type and behavior of data. The data can be as structured and unstructured. Structured data is that which reside in fixed field. It is first depends on creating data model. Unstructured data refers to information that does not have a predefined data model or not organized in a predefined manner. In data mining text mining has become an important research area. Text mining I is a discovery of new, previously unknown information by automatically extracting information from different resources. The various applications in text mining are information retrieval, machine learning, data mining, and statics and computation semantics. In form of text data most of the information is stored. Now a days in a direction of multiple language support most of the research is progressing. This system is capable to group the similar data from different kinds of language source according to their original semantic and also being able to gain information across language. In the presented work the identified twitter data set isused to perform text analysis. Therefore the entire input data samples are required to classify in two classes namely positive and negative. Therefore a binary classifier namely ID3 decision tree and their improved variant is utilized for analysis and performing the classification task. Before classification of text data there is need to improve the quality of data. Therefore the raw text data is first pre-processed then tagged according to the lexical means. After tagging on the original text data the classification algorithms are trained and make use to classify the text according to their sentiments. The implementation of the improved ID3text classification technique and their performance is evaluated in terms of their accuracy and the error rate. These parameters show how accurately the text patterns are identified using the data mining technique. Additionally for finding their performance in terms of their efficiency the time and space complexity is also measured that shows the effective classification with less consumption					
Vinodhini and Chandrasekaran (2013)	Yes	English	Yes	Yes	Yes	Yes
Model/method of Vinodhini and Chandrasekaran (22)	The performance of sentiment mining classifiers for problems of unbalanced and balanced large data sets for three different products					

Table 7 (continued)

Works	SC	Language	SD	DT	C4.5 Algorithm	Decision Tree
Summary of Vinodhini and Chandrasekaran (22)	The transition from Web 2.0 to Web 3.0 has resulted in creating the dissemination of social communication without limits in space and time. Sentiment analysis has really come into its own in the past couple of years. It's been a part of text mining technology for some time, but with the rise in social media popularity, the amount of unstructured textual data that can be used as a machine learning data source, is enormous. Marketers use this data as an intelligent indicator for customer preferences. This work aims to evaluate the performance of sentiment mining classifiers for problems of unbalanced and balanced large data sets for three different products. The classifiers used for sentiment mining in this paper are Support Vector Machine (SVM), Naïve bayes and C5.The results shows that the performance of the classifiers depends on the class distribution in the dataset. Also balanced data sets achieve better results than unbalanced datasets in terms of overall misclassification rate					
Voll and Taboada (2007)	Yes	English	Yes	Yes	Yes	Yes
Model/method of Voll and Taboada (2007)	Not all words are created equal: extracting semantic orientation as a function of adjective relevance					
Summary of Voll and Taboada (2007)	Semantic orientation (SO) for texts is often determined on the basis of the positive or negative polarity, or sentiment, found in the text. Polarity is typically extracted using the positive and negative words in the text, with a particular focus on adjectives, since they convey a high degree of opinion. Not all adjectives are created equal, however. Adjectives found in certain parts of the text, and adjectives that refer to particular aspects of what is being evaluated have more significance for the overall sentiment of the text. To capitalize upon this, we weigh adjectives according to their relevance and create three measures of SO: a baseline SO using all adjectives (no restriction); SO using adjectives found in on-topic sentences as determined by a decision-tree classifier; and SO using adjectives in the nuclei of sentences extracted from a high-level discourse parse of the text. In both cases of restricting adjectives based on relevance, performance is comparable to current results in automated SO extraction. Improvements in the decision classifier and discourse parser will likely cause this result to surpass current benchmarks					
Mandal and Sen (2014)	Yes	English Bangla	Yes	Yes	Yes	Yes
Model/method of Mandal and Sen (2014)	Supervised learning Methods for Bangla Web Document Categorization					
Summary of Mandal and Sen (2014)	This study explores the use of machine learning approaches, or more specifically, four supervised learning Methods, namely Decision Tree(C 4.5), K-Nearest Neighbour (KNN), Naive Bays (NB), and Support Vector Machine (SVM) for categorization of Bangla web documents. This is a task of automatically sorting a set of documents into categories from a predefined set. Whereas a wide range of methods have been applied to English text categorization, relatively few studies have been conducted on Bangla language text categorization. Hence, the authors attempt to analyze the efficiency of those four methods for categorization of Bangla documents. In order to validate, Bangla corpus from various websites has been developed and used as examples for the experiment. For Bangla, empirical results support that all four methods produce satisfactory performance with SVM attaining good results in terms of high dimensional and relatively noisy document feature vectors					
Kaur et al. (2014)	Yes	English	Yes	Yes	Yes	Yes
Model/method of Kaur et al. (2014)	Presenting the various existing techniques and work done for sentiment analysis					

Table 7 (continued)

Works	SC	Language	SD	DT	C4.5 Algorithm	Decision Tree
Summary of Kaur et al. (2014)	Sentiment analysis or opinion mining is the computational study of people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes. This is an Information Extraction task which is technically very challenging, but also practically very useful. With the advent of web 2.0, huge volumes of opinionated text is available on the web. To extract sentiment about an object from this huge web, automated opinion mining systems are thus needed. The existing techniques for sentiment analysis includes machine learning and lexical-based approaches. This work aims at presenting the various existing techniques and work done for sentiment analysis till date with issues pertaining to this field and future research prospects in this area					
Prasad et al. (2016)	Yes	Indian	Yes	Yes	Yes	Yes
Model/method of Prasad et al. (2016)	Sentiment classification: an approach for indian language tweets using decision tree					
Summary of Prasad et al. (2016)	This study describes the system we used for Shared Task on Sentiment Analysis in Indian Languages (SAIL) Tweets, at MIKE-2015. Twitter is one of the most popular platform which allows users to share their opinion in the form of tweets. Since it restricts the users with 140 characters, the tweets are actually very short to carry opinions and sentiments to analyze. The authors take the help of a twitter training dataset in Indian Language (Hindi) and apply data mining approaches for analyzing the sentiments. We used a state-of-the-art Data Mining tool Weka to automatically classify the sentiment of Hindi tweets into positive, negative or neutral					
Pong-Inwong and Rungworawut (2014)	Yes	English	Yes	Yes	Yes	Yes
Model/method of Pong-Inwong and Rungworawut (2014)	Teaching senti-lexicon for automated sentiment polarity definition in teaching evaluation					
Summary of Pong-Inwong and Rungworawut (2014)	This research significantly achieved the construction of a teaching evaluation sentiment lexicon and an automated sentiment orientation polarity definition in teaching evaluation. The Teaching Senti-lexicon will compute the weights of terms and phrases obtained from student opinions, which are stored in teaching evaluation suggestions in the form of open-ended questions. This Teaching Senti-lexicon consists of three main attributes, including: teaching corpus, category and sentiment weight score. The sentiment orientation polarity was computed with its meaning function being sentiment class definitions. A number of 175 instances were randomized using teaching feedback responses which were posted by students studying at Loei Raja hat University. The contributions of this work propose an effective teaching sentiment analysis method, especially for teaching evaluation. In this study, the experimented model employed SVM, ID3 and Naïve Bayes algorithms, which were implemented in order to analyze sentiment classifications with a 97% highest accuracy of SVM. This model is also applied to improve upon their teaching as well					
Sharma (2014)	Yes	English	Yes	Yes	Yes	Yes
Model/method of Sharma (2014)	Z-CRIME: a data mining tool for the detection of suspicious criminal activities based on decision tree					
Summary of Sharma (2014)	Data mining is the extraction of knowledge from large databases. One of the popular data mining techniques is Classification in which different objects are classified into different classes depending on the common properties among them. Decision Trees are widely used in the Classification. This study proposes a tool which applies an enhanced Decision Tree Algorithm to detect the suspicious e-mails about the criminal activities. An improved ID3 Algorithm with enhanced feature selection method and attribute- importance factor is applied to generate a better and faster Decision Tree. The objective is to detect the suspicious criminal activities and minimize them. That's why the tool is named as "Z-Crime" depicting the "Zero Crime" in the society. This paper aims at highlighting the importance of data mining technology to design a proactive application to detect the suspicious criminal activities					

Table 7 (continued)

Works	SC	Language	SD	DT	C4.5 Algorithm	Decision Tree
Our study	Yes	English	Yes	Yes	Yes	Yes
Model/method of our study	C4.5 algorithm for English sentiment classification					
The summary of our study	We use the C4.5 algorithm to classify semantics (positive, negative, neutral) of one English document in the English testing data set based on 140,000 English sentences of English training data set which comprises 70,000 English positive sentences and 70,000 English negative sentences					

Table 8 Comparison our model with the latest sentiment classification models in (2016, Kaur et al. 2016; Phu and Tuoi2014; Tran et al. 2014)

Works	SC	Language	SD	DT	C4.5 Algorithm	Decision Tree
Manek et al. (2016)	Yes	English	Yes	Yes	No	No
Model/method of Manek et al. (2016)	+A Gini Index based feature selection method +Support Vector Machine (SVM) classifier					
Summary of Manek et al. (2016)	+A Gini Index based feature selection method with Support Vector Machine (SVM) classifier is proposed for sentiment classification for large movie review dataset					
Agarwal and Mittal (2016a]	Yes	English	Yes	Yes	No	No
Model/method of Agarwal and Mittal (2016a)	Machine learning approach					
Summary of Agarwal and Mittal (2016a)	+Machine Learning Approach, which uses the bag-of-words (BoW) with the help of feature selection techniques which selects only important features by eliminating the noise and irrelevant features					
Agarwal and Mittal (2016b]	Yes	English	Yes	No	No	No
Model/method of Agarwal and Mittal (2016b)	The corpus-based semantic orientation approach for sentiment analysis					
Summary of Agarwal and Mittal (2016b)	+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					
Canuto et al. (2016)	Yes	English	Yes	Yes	No	No
Model/method of Canuto et al. (2016)	New meta-level features, especially designed for the sentiment analysis of short messages					
Summary of Canuto et al. (2016)	+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					
Ahmed and Danti (2016)	Yes	English	Yes	Yes	No	No
Model/method of Ahmed and Danti (2016)	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					
Summary of Ahmed and Danti (2016)	The focus is to perform effectively Sentimental analysis and Opinion mining of Web reviews using various rules based machine learning algorithms The study uses 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 + Comparative experiments on various rules based machine learning algorithms have been performed through a ten-fold cross validation training model for sentiment classification					
Phu and Tuoi (2014)	Yes	English	No	No	No	No
Model/method of Phu and Tuoi (2014)	Terms-Counting method Contextual Valence Shifters method					

Table 8 (continued)

Works	SC	Language	SD	DT	C4.5 Algorithm	Decision Tree
Summary of Phu and Tuoi (2014)	The authors combine five dictionaries into the new one with 21,137 entries. 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)	Yes	English	Yes	Yes	No	No
Model/method of Tran et al. (2014)	+Naïve Bayes +N-Gram +Chi-Square, etc					
Summary of Tran et al. (2014)	+The authors have explored Naive Bayes model with N-GRAM method, Negation Handling method, Chi-Square method and Good-Turing Discounting with selecting different thresholds of Good-Turing Discounting method and different minimum frequencies of Chi-Square method to improve the accuracy of sentiment classification					
This work	Yes	English	Yes	Yes	Yes	Yes
Model/method of this work	C4.5 Algorithm for English sentiment classification					
The summary of this works	We use the C4.5 algorithm to classify semantics (positive, negative, neutral) of one English document in the English testing data set based on 140,000 English sentences of English training data set which includes 70,000 English positive sentences and 70,000 English negative sentences					

Table 9 Comparison our model with the latest unsupervised classification works in (Turney 2002; Lee et al. 2002; Zyl 2002; Le Hegarat-Mascle et al. 2002; Ferro-Famil and Pottier 2002; Chaovalit and Zhou 2005; Lee and Lewicki 2002; Gllavata et al. 2004)

Studies	SC	L	SD	DT	C4.5 Algorithm	Decision Tree	Unsupervised Classification
Turney (2002)	Yes	EL	Yes	Yes	No	No	Yes
Model/method of Turney (2002)	A simple unsupervised learning algorithm for classifying reviews as recommended (thumbs up) or not recommended (thumbs down)						
Summary of Turney (2002)	The work presents a simple unsupervised learning algorithm for classifying reviews as recommended (thumbs up) or not recommended (thumbs down). The classification of a review is predicted by the average semantic orientation of the phrases in the review that contain adjectives or adverbs. A phrase has a positive semantic orientation when it has good associations (e.g., "subtle nuances") and a negative semantic orientation when it has bad associations (e.g., "very cavalier"). In the research, the semantic orientation of a phrase is calculated as the mutual information between the given phrase and the word "excellent" minus the mutual information between the given phrase and the word "poor". A review is classified as recommended if the average semantic orientation of its phrases is positive. The algorithm achieves an average accuracy of 74% when evaluated on 410 reviews from Epinions, sampled from four different domains (reviews of automobiles, banks, movies, and travel destinations)						
Lee et al. (2002)	No	NM	NM	NM	No	No	Yes
Model/method of Lee et al. (2002)	A new method for unsupervised classification of terrain types and man-made objects using polarimetric synthetic aperture radar (SAR) data						

Table 9 (continued)

Studies	SC	L	SD	DT	C4.5 Algorithm	Decision Tree	Unsupervised Classification
Summary of Lee et al. (2002)	The authors propose a new method for unsupervised classification of terrain types and man-made objects using polarimetric synthetic aperture radar (SAR) data. This technique is a combination of the unsupervised classification based on polarimetric target decomposition, S.R. Cloude et al. (1997), and the maximum likelihood classifier based on the complex Wishart distribution in the polarimetric covariance matrix, J.S. Lee et al. (1994). The authors use Cloude and Pottier's method to initially classify the polarimetric SAR image. The initial classification map defines training sets for classification based on the Wishart distribution. The classified results are then used to define training sets for the next iteration. 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 by the scattering mechanism associated with each zone. The advantages of this method are the automated 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						
Zyl (2002)	NM	NM	NM	NM	No	No	Yes
Model/method of Zyl (2002)	The use of an imaging radar polarimeter data for unsupervised classification of scattering behavior						
Summary of Zyl (2002)	The use of an imaging radar polarimeter data for unsupervised classification of scattering behavior is described by comparing the polarization properties of each pixel in an image to that of simple classes of scattering such as an even number of reflections, odd number of reflections, and diffuse scattering. For example, 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 an odd number of reflections, scattering by the urban area as being similar to that predicted by the class of an 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 reflection class, making it easy to identify these objects against the background of the surrounding ocean						
Le Hegarat-Mascle et al. (2002)	NM	NM	NM	NM	No	No	Yes
Model/Method of Le Hegarat-Mascle et al. (2002)	Dempster-Shafer evidence theory may be successfully applied to unsupervised classification in multi-source remote sensing						
Summary of Le Hegarat-Mascle et al. (2002)	The aim of the work is to show that Dempster-Shafer evidence theory may be successfully applied to unsupervised classification in multi-source remote sensing. Dempster-Shafer formulation allows for consideration of unions of classes, and to represent both imprecision and uncertainty, through the definition of belief and plausibility functions. These two functions, derived from mass function, are generally chosen in a supervised way. In this work the authors describe an unsupervised method, based on the comparison of mono-source classification results, to select the classes necessary for Dempster-Shafer evidence combination and to define their mass functions. Data fusion is then performed, discarding invalid clusters (e.g.corresponding to conflicting information) thank to an iterative process. The unsupervised multi-source classification algorithm is applied to MAC-Europe'91 multi-sensor airborne campaign data collected over the Orgeval French site. Classification results using different combinations of sensors (TMS and AirSAR) or wavelengths (L- and C-bands) are compared. Performance of data fusion is evaluated in terms of identification of land cover types. The best results are obtained when all three data sets are used						
Ferro-Famil et al. (2002)	NM	NM	NM	NM	No	No	Yes
Model/method of Ferro-Famil et al. (2002)	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						

Table 9 (continued)

Studies	SC	L	SD	DT	C4.5 Algorithm	Decision Tree	Unsupervised Classification
Summary of Ferro-Famil et al. (2002)	Introduces 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. This matrix is composed of the two coherency matrices and their cross-correlation. A decomposition theorem is applied to both images to obtain 64 initial clusters based on their scattering characteristics. The data sets are then classified by an iterative algorithm based on a complex Wishart density function of the 6×6 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 polarimetric cross-correlation information to refine the results of classification to a small number of clusters using the conditional probability of the cross-correlation 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						
Chaovalit and Zhou (2005)	Yes	EL	Yes	Yes	No	No	Yes
Model/method of Chaovalit and Zhou (2005)	+Machine learning +Semantic orientation						
Summary of Chaovalit and Zhou (2005)	Web content mining is intended to help people discover valuable information from the large amount of un-structured data on the web. Movie review mining classifies movie reviews into two polarities: positive and negative. As a type of sentiment-based classification, movie review mining is different from other topic-based classifications. Few empirical studies have been conducted in this domain. This work 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 are always mixed with real-life review data and ironic words are used in writing movie reviews						
Lee et al. (2002a, b)	No	NM	NM	NM	No	No	Yes
Model/method of Lee et al. (2002a, b)	The algorithm estimates the density of each class and is able to model class distributions with non-Gaussian structure						
Summary of Lee et al. (2002a, b)	An unsupervised classification algorithm is derived by modeling observed data as a mixture of several mutually exclusive classes that are each described by linear combinations of independent, non-Gaussian densities. The algorithm estimates the density of each class and is able to model class distributions with non-Gaussian structure. The new algorithm can improve classification accuracy compared with standard Gaussian mixture models. When applied to blind source separation in non-stationary 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						
Gllavata (2004)	NM	NM	NM	NM	No	No	Yes
Model/method of Gllavata (2004)	A robust text localization approach						
Summary of Gllavata (2004)	Text localization and recognition in images is important for searching information in digital photo archives, video databases and Web sites. However, since the text is often printed against a complex background, it is often difficult to detect. In the work, 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. 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						

Table 9 (continued)

Studies	SC	L	SD	DT	C4.5 Algorithm	Decision Tree	Unsupervised Classification
This study	Yes	EL	Yes	Yes	Yes	Yes	No
Model/method of this study	C4.5 Algorithm for English sentiment classification						
The summary of this study	We use the C4.5 algorithm to classify semantics (positive, negative, neutral) of one English document in the English testing data set based on 140,000 English sentences of English training data set which comprises 70,000 English positive sentences and 70,000 English negative sentences						

Table 10 Comparison our model with many algorithms for the decision tree in (Friedl and Brodley 1997; Freund and Mason 1999; Payne et al. 1978; Chang 1977; Mehta et al. 1995)

Researches	SC	L	SD	DT	C4.5 Algorithm	Decision Tree	Unsupervised Classification
Friedl and Brodley (1997)	No	NM	Yes	Yes	No	Yes	No
Model/method of Friedl and Brodley (1997)	Decision tree classification of land cover from remotely sensed data						
Summary of Friedl and Brodley (1997)	Decision tree classification algorithms have significant potential for land cover mapping problems and have not been tested in detail by the remote sensing community relative to more conventional pattern recognition techniques such as maximum likelihood classification. In this survey, the authors present several types of decision tree classification algorithms arid evaluate them on three different remote sensing data sets. The decision tree classification algorithms tested include an univariate decision tree, a multivariate decision tree, and a hybrid decision tree capable of including several different types of classification algorithms within a single decision tree structure. Classification accuracies produced by each of these decision tree algorithms are compared with both maximum likelihood and linear discriminant function classifiers. Results from this analysis show that the decision tree algorithms consistently outperform the maximum likelihood and linear discriminant function classifiers in regard to classification accuracy. In particular, the hybrid tree consistently produced the highest classification accuracies for the data sets tested. More generally, the results from this work show that decision trees have several advantages for remote sensing applications by virtue of their relatively simple, explicit, and intuitive classification structure. Further, decision tree algorithms are strictly nonparametric and, therefore, make no assumptions regarding the distribution of input data, and are flexible and robust with respect to nonlinear and noisy relations among input features and class labels						
Freund and Mason (1999)	No	NM	Yes	Yes	No	Yes	No
Model/method of Freund and Mason (1999)	The Alternating Decision Tree Learning Algorithm						
Summary of Freund and Mason (1999)	The application of boosting procedures to decision tree algorithms has been shown to produce very accurate classifiers. These classifiers are in the form of a majority vote over a number of decision trees. Unfortunately, these classifiers are often large, complex and difficult to interpret. This stud describes a new type of classification rule, the alternating decision tree, which is a generalization of decision trees, voted decision trees and voted decision stumps. At the same time classifiers of this type are relatively easy to interpret						
Payne and Tignor (1978)	No	NM	Yes	Yes	No	Yes	No
Model/method of Payne and Tignor (1978)	Freeway incident-detection algorithms based on decision trees with states						
Summary of Payne and Tignor (1978)	Incident-detection algorithms are a part of an overall freeway-traffic management system. These algorithms provide indications of the probable presence of freeway incidents by processing electronic surveillance data. In this survey, a class of algorithms that are designed to discriminate patterns in the data peculiar to incidents are described. The generic structure of these algorithms is the decision tree with states, the states corresponding to distinct traffic conditions. Ways to calibrate algorithm thresholds are described and applied to the algorithms. Performance evaluations based on traffic data from the Los Angeles system are presented						
Chang and Pavlidis (1977)	No	NM	Yes	Yes	No	Yes	No
Model/method of Chang and Pavlidis (1977)	Fuzzy decision tree algorithms						

Table 10 (continued)

Researches	SC	L	SD	DT	C4.5 Algorithm	Decision Tree	Unsupervised Classification
Summary of Chang and Pavlidis (1977)	Certain theoretical aspects of fuzzy decision trees and their applications are discussed. The main result is a branchbound-backtrack algorithm which, by means of pruning subtrees unlikely to be traversed and installing tree-traversal pointers, has an effective backtracking mechanism leading to the optimal solution while still requiring usually only O (log n) time, where n is the number of decision classes						
Mehta et al. (1995)	No	NM	Yes	Yes	No	Yes	No
Model/method of Mehta et al. (1995)	MDL-based decision tree pruning						
Summary of Mehta et al. (1995)	This paper explores the application of the Minimum Description Length principle for pruning decision trees. The authors present a new algorithm that intuitively captures the primary goal of reducing the mis-classification error. An experimental comparison is presented with three other pruning algorithms. The results show that the MDL pruning algorithm achieves good accuracy, small trees, and fast execution times						
This study	Yes	EL	Yes	Yes	Yes	Yes	No
Model/method of this study	C4.5 Algorithm for English sentiment classification						
The summary of this study	We use the C4.5 algorithm to classify semantics (positive, negative, neutral) of one English document in the English testing data set based on 140,000 English sentences of English training data set which comprises 70,000 English positive sentences and 70,000 English negative sentences						

Appendices of all codes

CODE 1: Creating table of training data

Input: 115,000 English sentences of the English training data set including the 57,500 English positive sentences and the 57,500 English negative sentences

Output: table of training data.

```
Begin
Step 1: Set tableOfTrainingData := create Table (m_max + 1 columns) (115,000 rows);
Step 2: For each sentence in the 57,500 English positive sentences, do:
Step 3:     Set arrayWords := Split the sentence based on ' ' or " "
Step 4:     tableOfTrainingData.Rows.Add ( new Rows());
Step 5:     For i = 0; i < arrayWords.length; i++, do:
Step 6:             tableOfTrainingData.Column[i].Add(arrayWords[i]  );
Step 7:     End For;
Step 8:     For j = i; j < m_max; j++, do:
Step 9:             tableOfTrainingData.Column[j].Add(0);
Step 10:    End For;
Step 11:    tableOfTrainingData.Column[m_max].Add(positive);
Step 12:End For;
Step 13: For each sentence in the  57,500 English negative sentences, do:
Step 14:    Set arrayWords := Split the sentence based on ' ' or " "
Step 15:    tableOfTrainingData.Rows.Add ( new Rows());
Step 16:    For i = 0; i < arrayWords.length; i++, do:
Step 17:            tableOfTrainingData.Column[i].Add(arrayWords[i]  );
Step 18:    End For;
Step 19:    For j = i; j < m_max; j++, do:
Step 20:            tableOfTrainingData.Column[j].Add(0);
Step 21:    End For;
Step 22:    tableOfTrainingData.Column[m_max].Add(negative);
Step 23:End For;
Step 24: Return tableOfTrainingData;
End;
```

CODE 2: Generating many association rules in the positive rule group and the negative rule group

Input:
Table of training data tableOfTrainingData is the training examples.
Attributes S is a list of other attributes that may be tested by the learned decision tree. (column from 0 to m_max -1 of tableOfTrainingData)
A decision tree (actually the root node of the tree) that correctly classifies the given Examples. This decision tree is divided into the positive rule group and the negative rule group.
Output: the positive rule group and the negative rule group
Begin

```
Step 1: If T is null Then
Step 2:     Return failure
Step 3: End If
Step 4: If S is null Then
Step 5:     Return Tree as a single node with most frequent class label in tableOfTrainingData
Step 6: End If
Step 7: If |S| = 1 Then
Step 8:     Return Tree as a single node S
Step 9: End If
Step 10: set Tree = {}
Step 11: for a ∈ S do:
Step 12:        set Info(a, tableOfTrainingData) = 0, and SplitInfo(a, tableOfTrainingData) = 0
Step 13:        comput Entroby(a)
Step 14:        for v ∈ values(a, tableOfTrainingData) do:
Step 15:            set tableOfTrainingData_{a, v} as the subset of tableOfTrainingData with attribute a=v
Step 16:            Info(a, tableOfTrainingData) += (|tableOfTrainingData_{a,v}|/| tableOfTrainingData_a|)Entroby(a_v)
Step 17:            SplitInfor(a, tableOfTrainingData) += - (|tableOfTrainingData_{a,v}|/|
     tableOfTrainingData_a|)log(|tableOfTrainingData_{a,v}|/| tableOfTrainingData_a|)
Step 18:        End For;
Step 19:        Gain(a, tableOfTrainingData) = Entropy(a) - Info(a, tableOfTrainingData)
Step 20:        GainRatio (a, tableOfTrainingData) = Gain(a, tableOfTrainingData)/SplitInfor(a, tableOfTrainingData)
Step 21: End For;
Step 22: set a_best = argmax{GainRatio (a, tableOfTrainingData)}
Step 23: attach a_best t into Tree
Step 24: for v ∈ values(a_best, tableOfTrainingData) do
Step 25:        call Algorithm 2 (tableOfTrainingData_{a,v})
Step 26: end for;
Step 27: Set positiveRuleGroup := {}
Step 28: Set negativeRuleGroup := {}
Step 29: Browse decision tree Tree, do:
Step 30:   If the rule is positive Then
Step 31:        positiveRuleGroup.Add (the rule);
Step 32:   Else If the rule is negative Then
Step 33:        negativeRuleGroup.Add (the rule);
Step 34:   End If
Step 35: End Browse;
Step 36: Return positiveRuleGroup and negativeRuleGroup;
```

End;

CODE 3: Classifying one English sentence into the positive polarity, the negative polarity or the neutral polarity

Input: one English sentence A, the positive rule group positiveRuleGroup and the negative rule group negativeRuleGroup
Output: positive, negative, neutral
Begin

```
Step 1: For each rule in positiveRuleGroup (X => positive), do:
Step 2:     If (the sentence contains X fully) = = True Then
Step 3:         Return positive;
Step 4:     End If
Step 5: End For;
Step 6: For each rule in negativeRuleGroup (Y => negative), do:
Step 7:     If (the sentence contains Y fully) = = True Then
Step 8:         Return positive;
Step 9:     End If
Step 10: End For;
Step 11: Return neutral;
```

End;

CODE 4: Classifying one English document into the positive polarity, the negative polarity or the neutral polarity

Input: one English document, including the n English sentences with the polarity result of each English sentence
Output: positive, negative, neutral
Begin
Step 1: If the number of English sentences classified into the positive polarity is greater than the number of English sentences classified into the negative polarity in the document Then
Step 2: Return positive;
Step 3: End If
Step 4: If the number of English sentences classified into the positive polarity is less than the number of English sentences classified into the negative polarity in the document Then
Step 5: Return negative;
Step 6: End If
Step 7: Return neutral;
End;

References

Agarwal B, Mittal N (2016a) Semantic orientation-based approach for sentiment analysis. Promin Feature Extr Sentim Anal doi:10.1007/978-3-319-25343-5_6 **(ISBN 978-3-319-25341-1)**

Agarwal B, Mittal N (2016b) Machine Learning Approach for Sentiment Analysis. Promin Feature Extr Sentim Anal doi:10.1007/978-3-319-25343-5_3 **(ISBN 978-3-319-25341-1)**

Ahmed S, Danti A (2016) Effective sentimental analysis and opinion mining of web reviews using rule based classifiers. Comput Intell Data Mining 1:171–179, doi:10.1007/978-81-322-2734-2$418, **(India, Print ISBN 978-81-322-2732-8)**

Canuto S, Gonçalves AM, Benevenuto F (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

Chang RL, Pavlidis T (1977) Fuzzy decision tree algorithms. IEEE Trans Syst Man Cybern 7:28–35

Chaovalit P, Zhou L (2005) Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches, Proceedings of the 38th Annual Hawaii International Conference on System Sciences, pp 112

Dalal MK, Zaveri M (2011) Automatic text classification: a technical review. Int J Comput Appl 28(2):0975–8887

Ferro-Famil L, Pottier E, Lee J-S (2002) Unsupervised classification of multifrequency and fully polarimetric SAR images based on the H/A/Alpha-Wishart classifier. IEEE Trans Geosci Remote Sens 39(11):2332–2342

Freund Y, Mason L (1999) The alternating decision tree learning algorithm, ICML '99 Proceedings of the Sixteenth International Conference on Machine Learning, pp 124–133

Friedl MA, Brodley CE (1997) Decision tree classification of land cover from remotely sensed data. Remote Sens Environ 61(3):399–409

Gllavata J, Ewerth R, Freisleben B (2004) Text detection in images based on unsupervised classification of high-frequency wavelet coefficients. In: Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004) 1:425–428

Kaur A, Duhan N (2015) A survey on sentiment analysis and opinion mining. Int J Innov Adv Comput Sci (IJIACS) **(ISSN 2347–8616, Volume 4, Special Issue)**

Korting TS (2006) C4.5 algorithm and Multivariate Decision Trees. National Institute for Space Research–INPE, SP Brazil

Kretschmann E, Fleischmann W, Apweiler R (2001) Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. Bioinformatics 17(10):920–926

Large Movie Review Dataset (2016) http://ai.stanford.edu/~amaas/data/sentiment/. Accessed Jun 2011

Le Hegarat-Mascle S, Bloch I, Vidal-Madjar D (2002) Application of Dempster–Shafer evidence theory to unsupervised classification in multisource remote sensing. IEEE Trans Geosci Remote Sens 35(4):1018–1031

Lee T-W, Lewicki MS, Sejnowski TJ (2002a) ICA mixture models for unsupervised classification of non-Gaussian classes and automatic context switching in blind signal separation, IEEE Trans Pattern Anal Mach Intell 22(10):1078–1089

Lee J-S, Grunes MR, Ainsworth TL, Du L-J (2002b) Unsupervised classification using polarimetric decomposition and the complex Wishart classifier. IEEE Trans Geosci Remote Sens 37(5):2249–2258

Li G, Liu F (2014) Sentiment analysis based on clustering: a framework in improving accuracy and recognizing neutral opinions, Appl Intell (APIN) 40(3):441–452

Loh S, de Oliveira JPM, Gameiro MA (2003) Gameiro, knowledge discovery in texts for constructing decision support systems. Appl Intell (APIN) 18(3):357–366

Mandal AK, Sen R (2014) Supervised learning Methods for Bangla Web Document Categorization. Int J Artif Intell Appl (IJAIA) 5(5)

Manek AS, Shenoy PD, Mohan MC, 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. doi:10.1007/s11280-015-0381-x **(ISSN1386-145X)**

Mazid MM, Ali ABMS, Tickle KS (2016) Improved C4.5 algorithm for rule based classification. In: AIKED'10 proceedings of the 9th WSEAS international conference on artificial intelligence, knowledge engineering and data bases, UK, pp 296–301

Mehta M, Rissanen J, Agrawal R (1995) MDL-based Decision Tree Pruning KDD-95Proceedings

Muniyandi AP, Rajeswari R, Rajaram R (2012) network anomaly detection by cascading K-means clustering and C4.5 decision tree algorithm. Procedia Eng (International Conference on Communication Technology System Design 2011) 30:174–182

Nizamani S, Memon N, Wiil UK, Karampelas P (2012) Modeling suspicious email detection using enhanced feature selection. IJMO 2(4):371–377 **(ISSN: 2010–3697, 2013)**

Pan Z-S, Chen S-C, Hu G-B, Zhang D-Q (2003) Hybrid neural network and C4.5 for misuse detection. Int Conf Mach Learn Cybern 4:2463–2467

Park S-B, Zhang B-T, Kim YT (2003) Word sense disambiguation by learning decision trees from unlabeled data, Appl Intell (APIN) 19(1):27–38

Payne HJ, Tignor SC (1978) Freeway incident-detection algorithms based on decision trees with states. 57th Annual Meeting of the Transportation Research Board, pp 30–37

Phu VN, Tuoi PT (2014) Sentiment classification using Enhanced Contextual Valence Shifters. International Conference on Asian Language Processing (IALP), pp 224–229

Phu VN, Dat ND, Tran VTN, Chau VTN, Nguyen TA (2016) Fuzzy C-means for english sentiment classification in a distributed system. Int J Appl Intell (APIN), pp 1–22

Phu VN, Chau VTN, Tran VTN, Dat ND (2017a) A Vietnamese adjective emotion dictionary based on exploitation of Vietnamese language characteristics, Int J Artif Intell Rev (AIR). doi:10.1007/s10462-017-9538-6

Phu VN, Chau VTN, Tran VTN, Dat ND, Nguyen TA (2017b) STING algorithm used english sentiment classification in a parallel environment. Int J Pattern Recognit Artif Intell. doi:10.1142/S0218001417500215

Pong-Inwong C, Rungworawut WS (2014) Teaching senti-lexicon for automated sentiment polarity definition in teaching evaluation. 10th International Conference on Semantics, Knowledge and Grids (SKG), pp 84–91

Prasad SS, Kumar J, Prabhakar DK, Pal S (2016) Sentiment classification: an approach for indian language tweets using decision tree. Mining Intelligence and Knowledge Exploration. In: Lecture Notes in Computer Science, Vol 9468, pp 656–663

Psomakelis E, Tserpes K, Anagnostopoulos D, Varvarigou T (2015) Comparing methods for Twitter Sentiment Analysis, arXiv:1505.02973 [cs.CL], 2015

Quinlan JR (1996a) Improved use of continuous attributes in C4.5. J Artif Intell Res 4(1):77–90

Quinlan JR (1996b) Bagging, Boosting, and C4.5 In: Proceedings of the thirteenth national conference on Artificial intelligence (AAAI'96) 1:725–730

Rajeswari LP, Arputharaj K (2008) An active rule approach for network intrusion detection with enhanced C4.5 algorithm. Int J Commun Netw Syst Sci 1:314–321

Ruggieri S (2002) Efficient C4.5 [classification algorithm]. IEEE Trans Knowl Data Eng 14(2):438–444

Sharma M (2014) Z-CRIME: a data mining tool for the detection of suspicious criminal activities based on decision tree. International Conference on Data Mining and Intelligent Computing (ICDMIC), pp 1–6

Shrivastava S, Nair PS (2015) Mood prediction on tweets using classification algorithm. Int J Sci Res (IJSR) 4(11):295–299

Sornlertlamvanich V, Potipiti T, Charoenporn T (2000) Automatic corpus-based Thai word extraction with the c4.5 learning algorithm. In: Proceedings of the 18th conference on Computational linguistics (COLING'00), Vol 2, pp 802–807, USA

Steven L (1994) Salzberg, C4.5: programs for machine learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993, Mach Learn 16(3):235–240

Taboada M, Voll K, Brooke J (2008) Extracting sentiment as a function of discourse structure and topicality, Technical Report 2008-20, School of Computing Science, Simon Fraser University, Burnaby

Tran VTN, Phu VN, Tuoi PT (2014) Learning More Chi Square Feature Selection to Improve the Fastest and Most Accurate Sentiment Classification, The Third Asian Conference on Information Systems, ACIS

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

van Zyl JJ (2002) Unsupervised classification of scattering behavior using radar polarimetry data. IEEE Trans Geosci Remote Sens 27(1):36–45

Vinodhini G, Chandrasekaran RM (2013) Performance evaluation of sentiment mining classifiers on balanced and imbalanced dataset. Int J Comput Sci Bus Inform 6(1)

Voll K, Taboada M (2007) Not all words are created equal: extracting semantic orientation as a function of adjective relevance, AI 2007: advances in artificial intelligence. In: Lecture notes in computer science. vol 4830, pp 337–346

Wan Y, Gao Q (2015) An ensemble sentiment classification system of twitter data for airline services analysis. IEEE International Conference on Data Mining Workshop (ICDMW), pp 1318–1325

Winkler S, Schaller S, Dorfer V, Affenzeller M, Petz G, Karpowicz M (2015) Data-based prediction of sentiments using heterogeneous model ensembles, Soft Comput 19(12):3401–3412

Xiaoliang Z, Hongcan Y, Jian W, Shangzhuo W (2009) Research and application of the improved algorithm C4.5 on Decision tree. Int Conf Test Meas 2:184–187

Zhou Z-H, Jiang Y (2004) NeC4.5: neural ensemble based C4.5. IEEE Trans Knowl Data Eng 16(6):770–773