Title page

Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms

Yang Liu

*Department of Management Science and Engineering, School of Business Administration,
Northeastern University, Shenyang 110167, China*

Liuy@mail.neu.edu.cn

Jian-Wu Bi

*Department of Management Science and Engineering, School of Business Administration,
Northeastern University, Shenyang 110167, China*

jianwubi@126.com

Zhi-Ping Fan

*Department of Management Science and Engineering, School of Business Administration,
Northeastern University, Shenyang 110167, China*

zpfan@mail.neu.edu.cn

Authors' Response to the Review of ESWA-D-16-03816R1

The authors want to thank the Editor-in-Chief and the anonymous reviewer giving us the second opportunity to improve our manuscript. According to your comment, we have made the revision. Our response to each comment is given below.

Editor-In-Chief's Comment: *Reviewer(s) have now commented on your paper. You will see that they are advising that you revise your manuscript. For your guidance, one reviewer's comments are appended below. If you decide to revise the work, please submit a list of changes or a rebuttal against each point which is being raised when you submit the revised manuscript. Please note that complete Point-to-Point responses MUST be inserted and clearly included in revised manuscript file itself. One of the best ways to place Point-to-Point responses in the revised manuscript is to insert Point-to-Point responses between the title page and Introduction section. The revised manuscript PDF file (a single file) you approve MUST clearly include complete Point-to-Point responses. We do NOT process any revision submission if Point-to-Point responses are incomplete or missing in the PDF file of the revised manuscript that you approve.*

Response: Thank you for giving us the second opportunity to improve our manuscript. According to the comment and suggestion given by the reviewer, we have revised our manuscript. Our point-to-point responses are given below.

Reviewer 1's Comments: *Complexity of a classification model is an important criterion used in model selection. In general, if two models are comparable in performance, a simpler model is preferable.*

In probabilistic models comparison of complexity is accomplished using AIC or BIC criteria, which are widely used. These criteria are not applicable in nonparametric models such as KNN or decision trees that you used in your analyses. For these models Kolmogorov complexity is perhaps the best approach to comparing their complexity.

My expectation was not that you would be the first in the field to include Kolmogorov

complexity measurements in your article, especially because Kolmogorov complexity is hard to compute.

But, let's ask who is your target audience? For instance, you feel you need to explain basics of SVM in your article so you expect that some readers will not be familiar with SVM. For those readers a sentence could and should be included pointing out that what models are more complex and require more learning parameters (like feature weights in SVM), or storage parameters (like points in Euclidean space in KNN), which are all indications of increasing Kolmogorov complexity.

I took notice of extensive and meticulous work you have done, but do you think my objection is unreasonable?

Response: Thank you for your insightful comment and valuable suggestion. According to your comment and suggestion, to illustrate the complexities of the algorithms used in our paper, we have made some revisions to Section 3.3, Section 4.3, Section 5.2 and Section 5.3, respectively. The brief illustration concerning the revision of each section is given below.

1) In Section 3.3, we have added the illustration to point out that the parameters will affect the complexities of the classification algorithms (please see page 14).

2) In Section 4.3, we have added Table 4 to illustrate the related parameters and the settings of the parameters in the experiment (please see Table 4 in page 21).

3) In Section 5.2, we have added Fig. 7 to show the computation results concerning average usages of memory and CPU of the four feature selection algorithms (please see Fig. 7 in page 30).

4) In Section 5.3, we have added Fig. 8 to show the computation results concerning average usages of memory and CPU of the five machine learning algorithms (please see Fig. 8 in page 32).

It is necessary to discuss with you that Kolmogorov complexity of algorithm is still not calculated in our paper. The main reasons are given below:

1) We have searched and studied the related literatures concerning Kolmogorov complexity, but we have not found the results of Kolmogorov complexity on measuring the complexities of the nine algorithms used in this paper. Thus, our study on Kolmogorov complexity of the algorithms lacks directly related references.

2) We have also consulted some related scholars to try to research Kolmogorov complexity, but we found that the calculation procedure of Kolmogorov complexity is very complicated. Till now, we have not known how to calculate the Kolmogorov complexity due to the limited time (only one month). Just as you said, Kolmogorov complexity is hard to compute.

Although the computation result of Kolmogorov complexity is not given in our paper, the execution times (time complexity), the average usages of memory and CPU (space complexity) and the description of main parameters are given to illustrate the complexities of the algorithms. We hope that you would think it is enough to compare the performances of different algorithms according to the above data.

Finally, we think that Kolmogorov complexity you mentioned is a valuable research topic. For this, we have added it as a future research direction in Section 6.

Furthermore, we have improved the linguistic quality of our paper. Some grammatical errors and improper expressions have been corrected.

We hope that the Editor and anonymous reviewer will find these revisions satisfactory.

Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms

Yang Liu ^a, Jian-Wu Bi ^{a, *}, Zhi-Ping Fan ^{a, b}

^a *Department of Management Science and Engineering, School of Business Administration, Northeastern University, Shenyang 110167, China*

^b *State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110819, China*

Abstract

Multi-class sentiment classification has extensive application backgrounds, whereas studies on this issue are still relatively scarce. In this paper, a framework for multi-class sentiment classification is proposed, which includes two parts: 1) selecting important features of texts using the feature selection algorithm, and 2) training multi-class sentiment classifier using the machine learning algorithm. Then, experiments are conducted for comparing the performances of four popular feature selection algorithms (document frequency, CHI statistics, information gain and gain ratio) and five popular machine learning algorithms (decision tree, naïve bayes, support vector machine, radial basis function neural network and K-nearest neighbor) in multi-class sentiment classification. The experiments are conducted on three public datasets which include twelve data subsets, and 10-fold cross validation is used to obtain the classification accuracy concerning each combination of feature selection algorithm, machine learning algorithm, feature set size and data subset. Based on the obtained 3600 classification accuracies (4 feature selection algorithms $\times$ 5 machine learning algorithms $\times$ 15 feature set sizes $\times$ 12 data subsets), the average classification accuracy of each algorithm is calculated, and the Wilcoxon test is used to verify the existence of significant difference between different algorithms in multi-class sentiment classification. The results show that, in terms of classification accuracy, gain ratio performs best among the four feature selection algorithms and support vector machine performs best among the five machine learning algorithms. In terms of execution time, the similar comparisons are also conducted. The

* Corresponding author. Tel.: +86-24-8368 7753; Fax: +86-24-2389 1569.
E-mail address: jianwubi@126.com (J.W. Bi).

obtained results would be valuable for further improving the existing multi-class sentiment classifiers and developing new multi-class sentiment classifiers.

Keywords: Multi-class sentiment classification; Experimental comparison; Feature selection algorithms; Machine learning algorithms

1. Introduction

With the striking development of social media platforms, more and more people post online texts on different platforms to express their opinions on social issues and share their reviews on products and services (Ravi & Ravi, 2015; Zhang et al., 2012). These increasing numbers of online texts contain a large number of information, which is valuable for governments (Bertot, Jaeger, & Hansen, 2012), companies (Chen et al., 2015; Lee et al., 2016; Li et al., 2014) and consumers to make desirable decisions (Kang & Park, 2014; Liu, Bi, & Fan, 2017; Yang, Yang, & Wu, 2016). For this, many studies on opinion mining and sentiment analysis have been conducted, in which sentiment classification is regarded as an important research topic focusing on automatically identifying the sentiment classes of online texts (Zhang et al., 2012; Lee et al., 2016; Li et al., 2014; Chen et al., 2015; Yang, Yang, & Wu, 2016; Kang & Park, 2014). It is necessary to point out that, in most of the existing studies on sentiment classification, only two kinds of sentiment classes are identified, i.e., the sentiment class of each text is identified as either positive or negative sentiment orientation (Medhat, Hassan, & Korashy, 2014; Ravi & Ravi, 2015; Serrano-Guerrero et al., 2015; Balazs & Velásquez, 2016). Some scholars have pointed out that it is excessive simplicity that only positive and negative sentiment orientations of online texts can be identified (Tang, Tan, & Cheng, 2009; Serrano-Guerrero et al., 2015). If multiple sentiment classes of the online texts can be identified, they would be valuable for conducting more accurate and more effective analysis for supporting the decisions of governments, companies and consumers (Tang, Tan, & Cheng, 2009; Serrano-Guerrero et al., 2015).

Several studies on multi-class sentiment classification can be found, which will be introduced in Section 2. But, it is necessary to point out that the studies on multi-class sentiment classification are relatively scarce. Especially, the classification accuracies of the existing multi-class sentiment classifiers are relatively low, which restricts the applications of existing multi-class sentiment classifiers. Thus, it is necessary to conduct more studies on multi-class sentiment classification. According to the studies on binary sentiment

classification (Medhat, Hassan, & Korashy, 2014; Ravi & Ravi, 2015; Serrano-Guerrero et al., 2015; Balazs & Velásquez, 2016), it can be expected that multi-class sentiment classifiers could be developed by integrating the feature selection algorithm and the machine learning algorithm, i.e., important features of online texts could be selected using the feature selection algorithm and the multi-class sentiment classifiers could be trained using the machine learning algorithm. At present, there are multiple popular feature selection algorithms and multiple popular machine learning algorithms. The popular feature selection algorithms mainly include: document frequency (DF), CHI statistics (CHI), information gain (IG) and gain ratio (GR), while the popular machine learning algorithms mainly include: decision tree (DT), naïve bayes (NB), support vector machine (SVM), radial basis function neural network (RBFNN), and K-nearest neighbor (KNN). Which one among the popular feature selection algorithms and which one among the popular machine learning algorithms perform best in multi-class sentiment classification are the questions that need to be answered. The answers would be valuable for further improving the existing multi-class sentiment classifiers and developing new multi-class sentiment classifiers. With respect to this issue, only one study can be found. Zhang et al. (2010) conducted an experimental comparison on the performances of two feature selection algorithms (IG and CHI) and four machine learning algorithms (KNN, NB, SVM and DT) in multi-class sentiment classification of Chinese reviews. The experimental comparison is conducted based on a dataset with 493 Chinese reviews. The results have shown that multi-class sentiment classification problem is more difficult than binary sentiment classification problem. But, the significances of the experimental results are not verified since the limited Chinese reviews and the absence of statistical test.

In this paper, the experimental comparisons of the feature selection algorithms and the machine learning algorithms in multi-class sentiment classification are conducted. A framework of multi-class sentiment classification is first proposed, which includes two parts: 1) selecting important features of texts using the feature selection algorithm, and 2) training multi-class sentiment classifier using the machine learning algorithm. Then, experiments are conducted for assessing the performances of the four popular feature selection algorithms (DF, CHI, IG and GR) and the five popular machine learning algorithms (DT, NB, SVM, RBFNN and KNN) in multi-class sentiment classification. The experiments are conducted on three public datasets including twelve data subsets, in which a sentiment class using 3-class, 4-class or 5-class of sentiment classifications has been labeled to each one of the total 18,908 online texts. 10-fold cross validation is used to obtain the classification accuracy concerning each combination of feature selection algorithm, machine learning algorithm, feature set size and

data subset. Wilcoxon test is used to verify the existence of significant difference between each pair of algorithms in multi-class sentiment classification.

The rest of this paper is arranged as follows. Section 2 presents the related work on the experimental comparisons of feature selection and machine learning algorithms in multi-class sentiment classification. In Section 3, a framework for multi-class sentiment classification is described, and the related feature selection algorithms and machine learning algorithms are introduced. Section 4 presents the design and methodology used in the experiments. In Section 5, the experiment results are shown and analyzed. Finally, Section 6 summarizes and highlights the major contributions of this paper.

2. Related work

Although studies on experimental comparisons of feature selection and machine learning algorithms in multi-class sentiment classification are very limited, some related studies can be found, which can be classified into two categories, i.e., 1) studies on multi-class sentiment classifiers; 2) studies on experimental comparisons of feature selection and machine learning algorithms in binary sentiment classification. A brief literature review concerning each category will be given below.

2.1. Studies on multi-class sentiment classifiers

Although many studies on binary sentiment classifiers have been conducted, the studies on multi-class sentiment classifiers are relatively scarce. Only several studies on multi-class sentiment classifiers can be found.

Pang and Lee (2005) earlier paid attentions to the problem of multi-class sentiment classification, and proposed a meta-algorithm based on metric labeling. In their study, considering the persons' difference abilities on distinguishing the sentiment strengths of texts, two classification regimes are first constructed, i.e., both a sentiment class of 3-class sentiment classification and a sentiment class of 4-class sentiment classification are assigned to each online text. Then, based on the idea of metric labeling, the meta-algorithm for multi-class sentiment classification is proposed. Further, based on the labeled texts, the classification accuracy of the proposed meta-algorithm is measured, which is greater than those obtained by the one-vs-all SVM classification and SVM regression.

Goldberg and Zhu (2006) proposed a graph-based semi-supervised algorithm for solving the multi-class sentiment classification problem, in which the labeled texts are relatively scarce. In the algorithm, according to all the texts that have been labeled and have not been labeled sentiment classes, an undirected graph for multi-class sentiment classification is built. Then, according to the undirected graph, an optimization model is constructed, and a smooth rating function can be obtained by solving the model. Further, according to the smooth rating function, the sentiment class of each unlabeled text can be determined. Moreover, the proposed graph-based semi-supervised algorithm is used to identify the sentiment classes of the texts in the sentiment scale dataset built by Pang and Lee (2005).

Wilson et al. (2006) conducted the experimental studies for evaluating the performance of dependency trees in multi-class sentiment classification. In their study, a wide range of features (including new syntactic features developed for opinion recognition) and three different learning algorithms (including boosting, rule learning and support vector regression) are used. According to the experimental results, it is concluded that adding dependency trees to unigrams improved the performance of various classifiers compared to unigrams alone.

Bickerstaffe and Zukerman (2010) proposed a hierarchical classifier for multi-class sentiment classification. In their study, according to a shortest-spanning algorithm, the inter-class similarities are measured and the tree structures are first constructed, which can reduce the training times of binary classifiers. Then, according to the tree structures, multiple binary classifiers are combined to form the hierarchical classifier. The experiments are conducted on the sentiment scale dataset built by Pang and Lee (2005), and experimental results have shown that the proposed hierarchical classifier has a better performance on multi-class sentiment classification than the classifiers proposed by Pang and Lee (2005).

Cao and Zukerman (2012) suggested a probabilistic approach for multi-class sentiment classification. The approach combines the information obtained from a lexicon and the information obtained by the NB classifier. In the approach, to improve the accuracy of the sentiment classification, some characteristics of grammatical structure are further considered. The effectiveness of the proposed approach is illustrated based on an experimental study conducted on seven datasets of different text lengths and writing styles.

Thelwall et al. (2010) proposed a SentiStrength algorithm for classifying the sentiment

classes of short informal text. In their study, a sentiment word strength list is first constructed by the machine learning algorithm. Then, based on the sentiment word strength list, the SentiStrength algorithm is developed by considering multiple characteristics of short informal text, such as spelling correction, booster words, negating words list and repeated letters, etc. The experiments are conducted to measure the classification accuracies of the SentiStrength algorithm. The experimental results have shown that the SentiStrength algorithm is able to identify the sentiment classes of short informal texts with positive and negative sentiment orientations in the accuracies of 60.6% and 72.8%. On this basis, Thelwall et al. (2012) further proposed a revised version of SentiStrength (called SentiStrength 2) to overcome the shortage that the performance of the SentiStrength is relatively weak for identifying the sentiment classes of short informal texts with negative sentiment orientation. The main change of SentiStrength 2 is that the lexicon of negative words is significant extended by incorporating the negative General Inquirer terms into the original SentiStrength.

In the existing studies, different feature selection algorithms and machine learning algorithms have been employed for developing the multi-class sentiment classifiers. However, experimental comparisons of the performances of different feature selection algorithms and machine learning algorithms in multi-class sentiment classification are ignored. If suitable feature selection algorithms and machine learning algorithms are used, it can be expected that the performances of the existing multi-class sentiment classifiers could be further improved, and new multi-class sentiment classifiers could be developed.

2.2. Studies on experimental comparisons of feature selection and machine learning algorithms in binary sentiment classification

Although studies on experimental comparisons of feature selection and machine learning algorithms in multi-class sentiment classification are very limited, several experimental comparisons of feature selection and machine learning algorithms in binary sentiment classification can be found.

Tan and Zhang (2008) conducted experimental comparisons on the performances of feature selection and machine learning algorithms in binary sentiment classification of Chinese documents. In their study, four feature selection algorithms (mutual information, IG, CHI and

DF) and five machine learning algorithms (KNN, centroid classifier, NB, Winnow and SVM) are compared based on a dataset including 1021 Chinese documents. The experimental results have shown: 1) IG is the best one among the four feature selection algorithms in binary sentiment classification of Chinese documents; 2) SVM is the best one among the five machine learning algorithms in binary sentiment classification of Chinese documents; 3) the performances of sentiment classifiers are severely dependent on domains or topics.

Sharma and Dey (2012a) conducted the experimental comparisons on different feature selection and machine learning algorithms in binary sentiment classification of English documents. In their study, the performances of five feature selection algorithms (DF, IG, GR, CHI and Relief-F) and seven machine learning algorithms (NB, SVM, Maximum Entropy, DT, KNN, Winnow, Adaboost) are compared based on the online movie reviews dataset. The experimental results have shown: GR is the best one among the five feature selection algorithms and SVM is the best one among the seven machine learning algorithms.

Omar et al. (2014) conducted the experimental comparisons on different feature selection and machine learning algorithms in binary sentiment classification of Arabic documents. In their study, the performances of seven feature selection algorithms (IG, Principal Components Analysis, Relief-F, Gini Index, Uncertainty, CHI and SVMs) and three machine learning algorithms (NB, SVM and KNN) are compared based on an Arabic dataset. The experimental results have shown: 1) the use of feature selection algorithms is helpful for sentiment classification of documents in Arabic; 2) SVM-based feature selection algorithm is the best one among the seven feature selection algorithms; 3) SVM is the best one among the three machine learning algorithms.

Agarwal and Mittal (2016) conducted the experimental comparisons on prominent feature extraction for English review analysis. In their study, four types of basic features are first extracted, i.e., *unigram features*, *bi-gram features*, *bi-tagged features* and *dependency parsing tree-based features*. Then, IG and mRMR algorithms are respectively used to eliminate noisy and irrelevant features from the basic feature set. Further, based on the obtained basic feature set, linear SVM and Multinomial NB are used to identify the sentiment orientation of each review in four standard datasets. The results have shown that: 1) when the four types of basic features are considered, the mRMR is superior to IG in binary sentiment classification; 2)

multinomial NB performs better than SVM in the accuracy and execution time in binary sentiment classification.

In the existing studies, the performances of different feature selection and machine learning algorithms in binary sentiment classification of different languages have been compared. However, the performances of different feature selection and machine learning algorithms in multi-class sentiment classification are ignored. The performances of feature selection and machine learning algorithms in multi-class sentiment classification could be different with those in binary sentiment classification since the changes of multiple factors, such as the different datasets, different numbers of features, different numbers of sentiment classes, and so on. Thus, it is necessary to further conduct experimental comparisons of feature selection and machine learning algorithms in multi-class sentiment classification. The results of experimental comparisons would be valuable for further improving the existing multi-class sentiment classifiers and developing new multi-class sentiment classifiers.

3. A framework for multi-class sentiment classification and the related algorithms

In this section, a framework for multi-class sentiment classification is proposed, which includes two parts: 1) selecting important features of texts using the feature selection algorithm, and 2) training multi-class sentiment classifier using the machine learning algorithm. Then, a brief introduction of the four feature selection algorithms is given in Section 3.2, and a brief introduction of the five machine learning algorithms is given in Section 3.3.

3.1. A framework for multi-class sentiment classification

Fig. 1. A framework for multi-class sentiment classification.

A framework for multi-class sentiment classification is shown in Fig. 1. It can be seen from

Fig. 1 that the process can be divided into two parts: 1) selecting important features of texts using the feature selection algorithm, and 2) training multi-class sentiment classifier using the machine learning algorithm. In the first part, the labeled texts are converted into vectors concerning text features, where each text feature represents a term in the texts; then, to reduce the number of text features and select the important features, feature selection algorithm is used for selecting a pre-defined number of text features. In the second part, based on the selected text features, the multi-class sentiment classifier is trained using the machine learning algorithm. The obtained classifier can be used to identify the different sentiment classes of unlabeled texts.

3.2. A brief introduction of the four feature selection algorithms

In this section, a brief introduction of each one of the four feature selection algorithms (DF, CHI, IG and GR) is given. The main ideas of the four feature selection algorithms are similar. In each algorithm, a score of each feature is first calculated, then the important features are selected according to a pre-defined number of features or a pre-defined threshold of the feature scores. The introduction of each feature selection algorithm is given below.

1) Document frequency (DF)

DF denotes the frequency that each term occurs in all texts (document) (Yang & Pedersen, 1997; Tan & Zhang, 2008). If the frequency of a term is too low, then it is regarded that the term is not typical enough to represent the texts; while if the frequency of a term is too high, then it is regarded that the term is not significant enough to distinguish the texts. Thus, in the DF algorithm, two pre-defined thresholds are used, i.e., a smaller pre-defined threshold and a bigger pre-defined threshold. The term will be removed if its frequency is less than the smaller pre-defined threshold or greater than the bigger pre-defined threshold.

2) CHI statistics (CHI)

CHI is one of the most used feature selection algorithms in sentiment classification, which is based on the statistical test. CHI measures the divergence of the observed data from the expected distribution based on the assumption that the occurrence of text feature is perfectly independent of the sentiment class (Galavotti, Sebastiani, & Simi, 2000; Tan & Zhang, 2008). CHI value represents the association between the text feature and its associated sentiment

class. The larger CHI value of a text feature is, the greater ability of the text feature for distinguishing sentiment classes of the texts will be. The CHI value is defined as

$$\text{CHI}(t, C_i) = \frac{n \times (ad - bc)^2}{(a + c) \times (b + d) \times (a + b) \times (c + d)} \quad (1)$$

$$\text{CHI}_{\max}(t) = \max_i (\text{CHI}(t, C_i)) \quad (2)$$

where t denotes the text feature, C_i denotes the i th sentiment class, a denotes the number of times that t and C_i co-occur; b denotes the number of times that t occurs without C_i ; c denotes the number of times that C_i occurs without t ; d is the number of times that neither C_i nor t occurs; n denotes the total number of texts.

3) Information gain (IG)

IG is also one of the most used feature selection algorithms in sentiment classification (Yang & Pedersen, 1997; Tan & Zhang, 2008). In IG algorithm, it is regarded that the larger uncertainty of a text is, the greater uncertainty of the sentiment class of the text will be. Thus, in IG algorithm, the importance of each text feature is measured according to the reduction of uncertainty if the value of the text feature is known. For this, the entropy is used to measure the uncertainty of a text. The IG value of a text feature is measured by the difference of the uncertainty of the text with and without the text feature. The IG value of a text feature can be calculated as

$$\text{IG}(t) = -\sum_{i=1}^m P(C_i) \log P(C_i) + P(t) \sum_{i=1}^m P(C_i | t) \log P(C_i | t) + P(\bar{t}) \sum_{i=1}^m P(C_i | \bar{t}) \log P(C_i | \bar{t}) \quad (3)$$

where $P(C_i)$ denotes the probability that sentiment class C_i occurs; $P(t)$ denotes the probability that text feature t occurs; $P(\bar{t})$ denotes the probability that text feature t does not occur; m denotes the total number of sentiment classes.

4) Gain ratio (GR)

GR is introduced in decision tree (C4.5) algorithms and is a type of feature selection algorithm based on the principle of IG (Quinlan, 1993; Sharma & Dey, 2012b). The GR value of a text feature is calculated by normalizing the IG value of the text feature. The high GR value indicates that the text feature will be useful for classification. GR is an iterative process for selecting the subset of text features by utilizing GR value. The iterations terminate when

the pre-defined number of features remain. The split information for text feature t is generated by splitting the data set D into v separations, where v is the outcome of a test on the feature t . The split information of text feature t is defined as:

$$\text{SplitInfo}(t) = - \sum_{j=1}^v \frac{|D_j|}{|D|} \times \log_2 \frac{|D_j|}{|D|} \quad (4)$$

where $|D_j|$ represents the number of the texts belonged to D_j . The GR value is defined as

$$\text{GR}(t) = \frac{\text{IG}(t)}{\text{SplitInfo}(t)} \quad (5)$$

where $\text{IG}(t)$ denotes IG value of the text feature t .

3.3. A brief introduction of the five machine learning algorithms

In this section, a brief introduction of each one of the five machine learning algorithms (DT, NB, SVM, RBFNN and KNN) is given as follows.

1) Decision tree (DT)

DT is an algorithm commonly used in data mining (Quinlan, 1993; Chang & Chang, 2012; Sharma & Dey, 2012a). The goal of DT is to construct a tree structure that predicts the values of test samples based on the training samples. A tree structure is a graphical representation of training samples, where leaves represent class labels of training samples and branches represent conjunctions of features that lead to the class labels. When DT is used to multi-class sentiment classification of texts, the leaves represent the multiple sentiment classes of training texts and branches represent terms occurring in the texts. There are many specific algorithms that can be used to generate the tree structure, such as ID3 (Quinlan, 1986), C4.5 (Quinlan, 1993) and CART (Breiman, 1984) etc., in which C4.5 is regarded as the most effective one (Quinlan, 1993). Thus, in this study, we use C4.5 to generate the tree structure for multi-class sentiment classification. In C4.5, the idea of GR is employed, which has been described in Section 3.2.

2) Naïve bayes (NB)

NB is a popular machine learning algorithm performing well in many domains, and it is one of the most used algorithms in binary sentiment classification (McCallum & Nigam, 1998; Domingos & Pazzani, 1997; Tan & Zhang, 2008). The NB algorithm is based on the

assumption that the probability of each term in a text is independence in the term's context and position in the text. Then, the probability of each text belonging to each sentiment class is estimated according to the joint probabilities of terms and sentiment classes. The sentiment class of a text is determined according to the highest probability among the probabilities of the text belonging to different sentiment classes.

Let $D=\{d_1, d_2, \dots, d_n\}$ denote a set of n texts, where d_l is the l th text, $l=1, 2, \dots, n$. Let $X_l=(x_{1l}, x_{2l}, \dots, x_{el})$ denote the feature vector of text d_l concerning the feature set $T=\{t_1, t_2, \dots, t_e\}$, where t_k denotes the k th feature (term) in the set T , $l=1, 2, \dots, n$, $k=1, 2, \dots, e$. Let $C=\{C_1, C_2, \dots, C_m\}$ denote a set of m sentiment classes, where C_i is the i th sentiment class of the texts, $i=1, 2, \dots, m$. The probability of text d_l belonging to sentiment class C_i can be estimated by

$$p(d_l | C_i) = p(C_i) \prod_{k=1}^e p(t_k | C_i), \quad l=1, 2, \dots, n, \quad i=1, 2, \dots, m \quad (6)$$

where $p(t_k | C_i)$ is the conditional probability of term t_k occurring in a text that belongs to sentiment class C_i and $p(C_i)$ is the prior probability of a text that belongs to sentiment class C_i . $p(t_k | C_i)$ and $p(C_i)$ can be estimated according to the training texts.

Then, the sentiment class of a test text is determined according to the highest probability among the probabilities of the text belonging to the different sentiment classes, i.e.,

$$C^* = \arg \max_{i \in \{1, 2, \dots, m\}} p(d_l | C_i), \quad l=1, 2, \dots, n \quad (7)$$

3) Support vector machine (SVM)

SVM is a popular machine learning algorithm with a solid theoretical foundation (Vapnik, 2000). SVM has been widely used in binary sentiment classification, and studies have shown that SVM is one of the most accurate algorithms in binary sentiment classification (Isa et al., 2008; Tsytarau & Palpanas, 2012). Based on the structural risk minimization principle, the SVM seeks a hyperplane that separates the feature vectors of training texts into different sentiment classes with the maximum margin. The problem of finding this hyperplane can be transformed into a constrained optimization problem. It is necessary to point out that different kernels of SVM have been developed, such as linear kernel, RBF kernel, sigmoid kernel (Vapnik, 2000). To save the space of this paper, only the linear kernel SVM is introduced,

which is regarded as the basic form of SVM.

Let C_i and $C_{i'}$ denote two different sentiment classes of training texts. Let $y_l = +1$ if text d_l belongs to C_i , and let $y_l = -1$ if text d_l belongs to $C_{i'}$. Let $\vec{w}_{ii'}$ be a vector representing the hyperplane that separates the feature vectors of training texts belonging the two sentiment classes C_i and $C_{i'}$. The solution of the optimization problem, i.e., $\vec{w}_{ii'}$, can be represented by

$$\vec{w}_{ii'} = \sum_{l=1}^{n_{ii'}} \alpha_l^* y_l \vec{d}_l, \quad \alpha_l \geq 0 \quad (8)$$

where α_l^* is obtained by solving a dual optimization problem; $n_{ii'}$ is the number of feature vectors of training texts, i.e., $n_{ii'} = n_i + n_{i'}$, n_i and $n_{i'}$ are the numbers of training texts belonging to the sentiment classes C_i and $C_{i'}$, respectively; $\vec{d}_l$ is a support vector of the hyperplane $\vec{w}_{ii'}$ if and only if the α_l is greater than zero. It is necessary to point out that SVM cannot be directly used to train a multi-class sentiment classifier since SVM is a binary classifier. Thus, to train the multi-class sentiment classifiers based on the SVM, one-vs-one strategy needs to be used. Several existing one-vs-one strategies can be considered, such as weighted voting (Hüllermeier & Vanderlooy, 2010), classification by pairwise coupling (Hastie & Tibshirani, 1998), non-dominance criterion (Fernández et al., 2010), probability estimates method (Wu, Lin, & Weng, 2004), binary-pair partitioning method (Zahorian & Nossair 1999) and bayes covariant method (Šuch & Barreda, 2016), etc.

4) Radial basis function neural network (RBFNN)

RBFNN is an artificial neural network that uses radial basis functions as activation functions (Musavi et al., 1992). RBFNN has strong nonlinear fitting ability and high reliability, and it performs well in many domains (Chen, Cowan, & Grant, 1991). RBFNN is a three-layer (input layer, hidden layer and output layer) feedforward network, where the movement of signal from input layer to hidden layer is nonlinear and that from hidden layer to output layer is linear. The output is a linear combination of radial basis functions of the inputs and neuron parameters. When RBFNN is used to identify the sentiment classes of test texts, the classifier is trained using the feature vectors of labeled training texts, then the sentiment class of each test text can be obtained by inputting the feature vector of the test text into the trained classifier (Ravi & Ravi, 2016).

Let $X_l = (x_{l1}, x_{l2}, \dots, x_{el})$ denote the feature vector of text d_l concerning the set of terms

$T = \{t_1, t_2, \dots, t_e\}$, where t_k denotes the k th term in the set T , $k = 1, 2, \dots, e$. The output of RBFNN concerning d_l can be calculated as

$$y_l = \sum_{u=1}^h w_u \exp\left(-\frac{1}{2\delta_u^2} \|X_l - Center_u\|^2\right) \quad (9)$$

where h represents the number of neurons in hidden layer, w_u is the connecting weights from u th hidden layer to output layer, $Center_u$ is the center of u th activation function, $\|\cdot\|$ is Euclid norm, and δ_u^2 is the width of the receptive field. According to the output y_l , the classification result of text d_l can be determined (Schwenker, Kestler & Palm, 2001).

5) *K-nearest neighbor (KNN)*

KNN is a popular instance-based learning algorithm (Yang & Liu, 1999), which has been shown as one of the most effective algorithms in sentiment classification (Lam & Han, 2003; Wu et al., 2008). In KNN, to determine the sentiment class of a test text d_l , K nearest neighbors are first selected from the training texts, which are denoted as $td_1, td_2, \dots, td_K$. Let $C = \{C_1, C_2, \dots, C_m\}$ denote a set of m sentiment classes, where C_i is the i th sentiment class of the texts, $i = 1, 2, \dots, m$. According to the similarities of d_l between the K nearest neighbors and the different sentiment classes of the K nearest neighbors, a KNN score of text d_l concerning sentiment class C_i can be calculated by

$$\text{Score}(d_l, C_i) = \sum_{k=1}^K \text{sim}(d_l, td_k) \delta(td_k, C_i) \quad (10)$$

$$\delta(td_k, C_i) = \begin{cases} 1, & \text{if sentiment class of } td_k \text{ is } C_i, \\ 0, & \text{if sentiment class of } td_k \text{ is not } C_i, \end{cases} \quad (11)$$

where $\text{sim}(d_l, td_k)$ is the similarity between text d_l and training text td_k ; $\delta(td_k, C_i)$ is the indicator of the training text td_k with respect to sentiment class C_i . According to the obtained KNN scores, the sentiment class of the test text is determined, i.e.,

$$f(d_l) = \arg \max_{i \in \{1, 2, \dots, m\}} \text{Score}(d_l, C_i) \quad (12)$$

When the above machine learning algorithms are used in multi-class sentiment classification, some parameters need to be defined beforehand, such as the number of nearest neighbors (K) in KNN, kernel type in SVM and so on. The parameters are regarded as indications of increasing Kolmogorov complexity (Watanabe, 1992). For simplicity of

description, detailed description of these parameters will be given in Section 4.3 along with the parameter settings in the experiment.

4. Experimental design

The experimental datasets used in this study are introduced in Section 4.1. Then, the experimental procedure is described in Section 4.2, and the experimental setup is illustrated in Section 4.3.

4.1. Experimental datasets

In this study, we conduct the experiments on three public datasets, which can be further divided into twelve data subsets. The first dataset (Dataset 1) is the sentiment scale datasets in the Movie Review Data (<http://www.cs.cornell.edu/people/pabo/movie-review-data/>), which includes four corpora of movie reviews. In each of the four corpora, both a sentiment class of 3-class sentiment classification and a sentiment class of 4-class sentiment classification are assigned to each movie review (Pang, & Lee, 2005). Thus, Dataset 1 is further divided into eight data subsets, denoted as Data subset 1, Data subset 2,, Data subset 8. The second dataset (Dataset 2) and the third dataset (Dataset 3) respectively consist of 1041 and 3407 short informal texts, which are provided by Thelwall et al. (2012). In Dataset 2 and Dataset 3, both a positive sentiment strength (from 1-5, where 1 denotes no positive sentiment, 5 denotes very strong positive sentiment) and a negative sentiment strength (from 1-5, where 1 denotes no negative sentiment, 5 denotes very strong negative sentiment) are assigned to each short informal text according to the ratings of different coders. Two sentiment strengths are assigned to a short text because both positive and negative sentiments can be simultaneously contained in the short text (Thelwall et al., 2012). Thus, Dataset 2 and Dataset 3 are respectively divided into two data subsets, i.e., Dataset 2 is further divided into Data subset 9 and Data subset 10, Dataset 3 is further divided into Data subset 11 and Data subset 12. The related information of the experimental datasets, i.e., the three datasets and the twelve data subsets, is shown in Table 1. In Table 1, #Cl. denotes the number of sentiment classes that have been labeled to the texts; #T denotes the number of texts; #F denotes the number of text

features; C_1 , C_2 , C_3 , C_4 and C_5 denote the numbers of texts belonging to the different sentiment classes. For example, in Data subset 1, there are 1027 texts with 19237 text features, where three kinds of sentiment classes C_1 , C_2 and C_3 have been labeled to the 1027 texts. The numbers of texts that belong to the sentiment classes C_1 , C_2 and C_3 are 359, 427 and 241, respectively.

Table 1

The related information of the experimental datasets.

Dataset	Data subset	#Cl.	#T	#F	C_1	C_2	C_3	C_4	C_5
Dataset 1	Data subset 1	3	1027	19237	359	427	241		
	Data subset 2	3	1307	31188	186	491	630		
	Data subset 3	3	902	25897	239	349	314		
	Data subset 4	3	1770	26439	413	648	709		
	Data subset 5	4	1027	19237	171	440	302	114	
	Data subset 6	4	1307	31188	138	292	596	281	
	Data subset 7	4	902	25897	115	295	334	158	
	Data subset 8	4	1770	26439	191	526	766	287	
Dataset 2	Data subset 9	5	1041	4249	728	215	71	23	4
	Data subset 10	5	1041	4249	196	492	291	54	8
Dataset 3	Data subset 11	5	3407	12376	1708	1049	384	203	63
	Data subset 12	5	3407	12376	924	1056	893	431	103

Notes: #Cl. indicates the number of sentiment classes that have been labeled to the texts; #T indicates the number of texts; #F indicates the number of text features.

4.2. Experimental procedure

The experimental procedure is shown in Fig. 2. It can be seen from Fig. 2 that the experimental procedure can be divided into three stages, i.e., 1) selecting important features of texts using the feature selection algorithm, 2) training multi-class sentiment classifier using the machine learning algorithm, and 3) evaluating the performances of the different algorithms. A brief description of each stage is given below.

1) Selecting important features of texts using the feature selection algorithm

Since the terms in the texts can not be directly identified by machine learning algorithms, the texts are usually converted into feature vectors concerning the terms. Bag-of-words (BOW) model is the most used approach for converting the texts into feature vectors. By the BOW model, multiple texts can be converted into the feature vectors concerning terms as shown in Table 2. In Table 2, n denotes the number of texts, e denotes the number of terms (features), w_{ik} denotes the weight of term t_k for representing the sentiment class of the text

d_l . The results of existing studies have shown that term frequency-inverse document frequency (TF-IDF) is an efficient approach for determining the weight w_{lk} , $l=1,2,\dots,n$, $k=1,2,\dots,e$. According to TF-IDF, the formula for calculating the value of w_{lk} can be represented by

$$w_{lk} = tf_{lk} \times \log\left(\frac{n}{n_k}\right), \quad l=1,2,\dots,n, \quad k=1,2,\dots,e. \quad (13)$$

where tf_{lk} is the frequency of term t_k in text d_l , n is the number of texts, n_k is number of the texts that include term t_k .

Fig. 2. Experimental procedure on comparing different feature selection algorithms and machine learning algorithms

Table 2

The feature vectors concerning terms.

Texts	t_1	t_2	...	t_e
d_1	w_{11}	w_{12}	...	w_{1e}
d_2	w_{21}	w_{22}	...	w_{2e}
...	...	...	...	...
d_n	w_{n1}	w_{n2}	...	w_{ne}

Although the texts can be converted into feature vectors by BOW model, the dimensionalities of the vectors could be huge since the number of different terms (features) in the texts could be huge. In the huge number of features, there would be not only effective features but also noisy and irrelevant features. Thus, to improve the efficiency and accuracy of sentiment classification, the feature selection algorithm is usually used to reduce the dimensionalities of the feature vectors, i.e., partial important features (terms) are selected using the feature selection algorithm. In this paper, the DF, CHI, IG and GR are respectively used to select important text features, and fifteen pre-defined feature set sizes are used, which are 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1500, 2000, 2500, 3000 and all (all features).

2) Training multi-class sentiment classifier using the machine learning algorithm

To train the multi-class sentiment classifiers and evaluate the performances of the different classifiers, it is necessary to divide the texts in each data subset into the training set and the test set. In this paper, to minimize the influence of variability of the training set on the experimental results, the approach of 10-fold cross validation is used. Each data subset shown in Table 1 is further divided into 10 subsets with similar sizes and distributions. Then, the union of 9 subsets is used as the training set while the remaining subset is used as the test set. The process is repeated ten times, such that every subset is used as the test set once. The average test result of the ten times is regarded as the result of the 10-fold cross validation. In this study, using the above 4 feature selection algorithms (DF, CHI, IG and GR), 720 groups of features are selected (4 feature selection algorithms $\times$ 12 data subsets $\times$ 15 feature set sizes). Further, each of the above 5 machine learning algorithms (DT, NB, SVM, RBFNN and KNN) is trained based on the 720 groups of features, then 3600 average test results concerning the performances of different classifiers can be obtained by the 10-fold cross

validation.

3) Evaluating the performances of different algorithms

To evaluate the performances of different algorithms in multi-class sentiment classification, the classification accuracy of each classifier needs to be calculated and the existence of significant difference between each pair of algorithms needs to be verified. For this, a brief introduction of classification accuracy in multi-class sentiment classification and a brief description of Wilcoxon test for comparing different algorithms are given below.

i) Classification accuracy in multi-class sentiment classification

Similar with the studies of Thelwall et al. (2010) and Wang et al. (2014), the classification accuracy is used to evaluate the performances of different algorithms in multi-class sentiment classification. To calculate the classification accuracy, a confusion matrix is constructed as shown in Table 3. In Table 3, $C_1, C_2, \dots, C_m$ denote the m sentiment classes, $c_{ii'}$ denotes the number of texts whose actual sentiment class are $C_{i'}$ and are classified into sentiment class C_i , $i, i' = 1, 2, \dots, m$.

Table 3

Confusion matrix for multi-class sentiment classification.

		Actual sentiment class			
		C_1	C_2	$\dots$	C_m
Sentiment classification result	C_1	c_{11}	c_{12}	$\dots$	c_{1m}
	C_2	c_{21}	c_{22}	$\dots$	c_{2m}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	C_m	c_{m1}	c_{m2}	$\dots$	c_{mm}

According to Table 3, classification accuracy of the multi-class sentiment classifier can be defined as

$$\text{Classification accuracy} = \frac{\sum_{i=1}^m c_{ii}}{\sum_{i=1}^m \sum_{i'=1}^m c_{ii'}}. \quad (14)$$

ii) Wilcoxon test for comparing different algorithms

Further, to verify the significance of the experimental results obtained by this study, the non-parametric tests are used. According to the recommendations of the study of Demšar (2012), Wilcoxon test (Wilcoxon, 1945) is used in this paper. The process of Wilcoxon test is described in the following. To save the space of this paper, we only take the process of

Wilcoxon test for comparing feature selection algorithms as an example. The process for comparing machine learning algorithms can be conducted by the similar way.

Let α_{ij}^k denote the classification accuracy of the multi-class sentiment classifier that combines the i th feature selection algorithm and j th machine learning algorithm on identifying sentiment classes of texts in the k th data subset, $i=1,2,3,4$, $j=1,2,\dots,5$, $k=1,2,\dots,12$.

Let $d_{ii'}^{kj}$ denote the difference between α_{ij}^k and $\alpha_{i'j}^k$, i.e.,

$$d_{ii'}^{kj} = \alpha_{ij}^k - \alpha_{i'j}^k, \quad i, i' = 1, 2, 3, 4, \quad i \neq i', \quad j = 1, 2, \dots, 5, \quad k = 1, 2, \dots, 12. \quad (15)$$

Since there are 12 data subsets and 5 machine learning algorithms, sixty $d_{ii'}^{kj}$ with respect to the i th and i' th feature selection algorithms can be obtained, which are denoted as $d_{ii'}^{11}, d_{ii'}^{12}, \dots, d_{ii'}^{15}, \dots, d_{ii'}^{11,5}, d_{ii'}^{12,1}, \dots, d_{ii'}^{12,5}$. Let $|d_{ii'}^{(1)}| \leq |d_{ii'}^{(2)}| \leq \dots \leq |d_{ii'}^{(60)}|$ denote a ranking of the absolute values of the sixty $d_{ii'}^{kj}$ from the smallest to the greatest. Let $r_{ii'}^{kj}$ denote the ranking position of $d_{ii'}^{kj}$ in the ranking $|d_{ii'}^{(1)}| \leq |d_{ii'}^{(2)}| \leq \dots \leq |d_{ii'}^{(60)}|$ such that $r_{ii'}^{kj} \in \{1, 2, \dots, 60\}$ and $r_{ii'}^{kj} \neq r_{ii'}^{k'j'}$, if $j \neq j'$ or $k \neq k'$, $i, i' = 1, 2, 3, 4$, $i \neq i'$, $j = 1, 2, \dots, 5$, $k = 1, 2, \dots, 12$. Let $I_{ii'}^{kj}$ denote an indication variable concerning $d_{ii'}^{kj}$, the value of $I_{ii'}^{kj}$ can be represented by the following Eq. (16), i.e.,

$$I_{ii'}^{kj} = \begin{cases} 1, & \text{if } d_{ii'}^{kj} > 0, \\ 0.5, & \text{if } d_{ii'}^{kj} = 0, \\ 0, & \text{if } d_{ii'}^{kj} < 0. \end{cases} \quad i, i' = 1, 2, 3, 4, \quad i \neq i', \quad j = 1, 2, \dots, 5, \quad k = 1, 2, \dots, 12 \quad (16)$$

Let $R_{ii'}^+$ be a score indicating the degree that the i th feature selection algorithm outperforms the i' th feature selection algorithm, and $R_{ii'}^-$ denote a score indicating the degree that the i' th feature selection algorithm outperforms i th feature selection algorithm. According to the study of Demšar (2012), $R_{ii'}^+$ and $R_{ii'}^-$ can be respectively calculated by

$$R_{ii'}^+ = \sum_{j=1}^5 \sum_{k=1}^{12} r_{ii'}^{kj} I_{ii'}^{kj}, \quad i, i' = 1, 2, 3, 4, \quad i \neq i', \quad (17)$$

$$R_{ii'}^- = \sum_{j=1}^5 \sum_{k=1}^{12} r_{ii'}^{kj} (1 - I_{ii'}^{kj}), \quad i, i' = 1, 2, 3, 4, \quad i \neq i'. \quad (18)$$

Let $T_{ii'} = \min\{R_{ii'}^+, R_{ii'}^-\}$ denote the value for verifying significant of the difference between

the i th and i' th feature selection algorithms. Let q denote the number of samples for Wilcoxon test. In the Wilcoxon test for comparing feature selection algorithms, $q=60$; in the Wilcoxon test for comparing machine learning algorithms, $q=48$ since $i=1,2,3,4$, $j=1,2,\dots,5$, $k=1,2,\dots,12$. According to $T_{ii'}$ and q , the significant of the difference between the i th and i' th feature selection algorithms can be verified by looking up the Wilcoxon rank sum table.

4.3. Experimental setup

We conduct the experiments on a PC using Windows 7 operating system. In our experiments, the data mining toolkit WEKA (Waikato Environment for Knowledge Analysis) version 3.7.0 (<http://www.cs.waikato.ac.nz/ml/weka/>) is used, which is a collection of various feature selection algorithms and machine learning algorithms for data mining tasks (Hall, Witten, & Frank, 2011).

Table 4

The description of main parameters and their settings in the experiment.

Algorithm	Parameter settings	Description of these parameters
DT	Confidence Factor = 0.25	The confidence factor used for pruning.
	Min NumObj = 2	The minimum number of instances per leaf.
NB	—	—
SVM	Cost = 1	The cost parameter C.
	Degree = 3	The degree of the kernel.
	Eps = 0.001	The tolerance of the termination criterion.
	Gamma = 0	The parameter of kernels ('RBF', 'Poly', and 'Sigmoid').
	Kernel Type = RBF'	The type of kernel to use.
	Probability Estimates = true	Whether to generate probability estimates instead of -1/+1 for classification problems.
RBFNN	Clustering Seed = 1	The random seed to pass on to K-means.
	Min Std Dev = 0.1	Sets the minimum standard deviation for the clusters.
	Num Clusters = 2	The number of clusters for K-Means to generate.
KNN	KNN = 5	The number of neighbors to use.
	Distance Metric = Euclidean Distance	The distance function to use for finding neighbors.

In our study, we conduct the experimental comparisons of the performances of the four feature selection algorithms (DF, CHI, IG and GR) and the five machine learning algorithms

(DT, NB, SVM, RBFNN and KNN) in multi-class sentiment classification. The four feature selection algorithms and the five machine learning algorithms are implemented by corresponding modules of WEKA, respectively. In the LIBSVM module, the default one-vs-one strategy is used, which is the second method proposed by Wu, Lin, & Weng (2004). The description of main parameters and the settings of the parameters in the experiment are shown in Table 4. The parameters are regarded as indications of increasing Kolmogorov complexity (Watanabe, 1992).

5. Experimental results and analysis

According to the experimental design in Section 4, the experiments are conducted. Basic experimental results are first shown in Section 5.1. Then, based on the basic experimental results, comparisons of different feature selection algorithms are presented in Section 5.2, comparisons of different machine learning algorithms are presented in Section 5.3, and comparisons between classification accuracies obtained in the existing studies and those obtained in this study are given in Section 5.4.

5.1. Basic experimental results

The experimental results are shown in Fig. 3- Fig. 6. To illustrate the meanings of Fig. 3- Fig. 6, we take Fig. 3 as an example. Fig. 3 shows the accuracy curves of the multi-class sentiment classifiers that combine DF with the five machine learning algorithms (DT, NB, SVM, RBFNN and KNN) concerning the twelve data subsets, where Fig. 3(a) - Fig. 3(l) show the accuracy curves concerning the Data subset 1, Data subset 2,, Data subset 12, respectively. The horizontal axis of each one of Fig. 3(a) - Fig. 3(l) represents the fifteen pre-defined feature set sizes, i.e., 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1500, 2000, 2500, 3000 and all (all features), while the vertical axis of each one of Fig. 3(a) - Fig. 3(l) represents the classification accuracy. It can be seen from Fig. 3 that most of the classification accuracies of the multi-class sentiment classifiers that combine DF with different machine learning algorithms present the increasing tendencies with the increasing numbers of features. It can be seen from Fig. 4- Fig. 6 that, with the increasing numbers of features, most of the classification accuracies of the multi-class sentiment classifiers that

combine CHI, IG or GR with different machine learning algorithms present the tendencies that increase at first and then decrease at some points.

Fig. 3. The accuracy curves of the multi-class sentiment classifiers that combine DF with the five machine learning algorithms.

Fig. 4. The accuracy curves of the multi-class sentiment classifiers that combine CHI with the five machine learning algorithms.

Fig. 5. The accuracy curves of the multi-class sentiment classifiers that combine IG with the five machine learning algorithms.

Fig. 6. The accuracy curves of the multi-class sentiment classifiers that combine GR with the five machine learning algorithms.

Table 5 shows the best classification accuracies concerning the twelve data subsets using the multi-class sentiment classifiers that combine different feature selection algorithms and machine learning algorithms. For example, 48.491 on the top left corner of Table 5 denotes the best one among the fifteen classification accuracies concerning Data subset 1 using the multi-class sentiment classifier that combines DF and DT, when the fifteen pre-defined numbers of features are used. Concretely, 48.491 on the top left corner of Table 5 is corresponding to the peak of curve of DT (—) in Fig. 3(a). Table 6 shows the execution time of each multi-class sentiment classifier corresponding to the best classification accuracy shown in Table 5. For example, the 0.62 on the top left corner of Table 6 denotes the execution time corresponding to the classification accuracy 48.491 obtained by the classifier that combines DF and DT. It can be seen from Table 5 that, in terms of classification accuracy, among the four feature selection algorithms, GR performs best on Data subsets 1-6, 8, 11 and 12; CHI performs best on Data subset 7; IG performs best on Data subsets 9 and 10. On the other hand, among the five machine learning algorithms, SVM performs best on all data subsets. It can be seen from Table 6 that, in terms of execution time, among the four feature selection algorithms DF performs best on Data subsets 1, 3 and 11; IG performs best on Data subsets 2, 4-6 and 8-10; GR performs best on Data subsets 7 and 12. On the other hand, among the five machine learning algorithms, KNN performs best on all data subsets.

Table 5

The best classification accuracies concerning the twelve data subsets using the multi-class sentiment classifiers that combine different feature selection and machine learning algorithms.

Data subset 1						Data subset 2					
	DF	CHI	IG	GR	Average		DF	CHI	IG	GR	Average
DT	48.491	52.191	52.191	54.236	51.777	DT	57.427	62.328	60.873	64.089	61.179
NB	57.546	65.336	65.336	65.239	63.364	NB	71.286	77.412	77.489	77.335	75.881
SVM	63.973	73.807	73.418	73.710	71.227	SVM	74.349	85.299	84.839	85.528	82.504
RBFNN	41.577	73.418	72.931	72.639	65.141	RBFNN	49.311	87.519	87.366	87.596	77.948
KNN	43.233	45.180	44.791	63.291	49.124	KNN	51.225	58.806	59.571	61.103	57.676
Average	50.964	61.986	61.733	65.823		Average	60.720	74.273	74.028	75.130	
Data subset 3						Data subset 4					
	DF	CHI	IG	GR	Average		DF	CHI	IG	GR	Average
DT	43.063	52.830	52.497	50.167	49.639	DT	52.346	57.038	57.434	56.473	55.823
NB	58.935	70.255	70.255	71.587	67.758	NB	65.630	71.792	71.849	73.149	70.605
SVM	61.709	79.356	79.023	80.466	75.139	SVM	68.570	77.954	77.897	79.367	75.947
RBFNN	40.289	84.351	84.129	85.128	73.474	RBFNN	41.888	77.106	76.540	77.388	68.231
KNN	41.398	49.723	50.943	65.816	51.970	KNN	41.945	48.106	46.241	66.535	50.707
Average	49.079	67.303	67.370	70.633		Average	54.076	66.399	65.992	70.582	
Data subset 5						Data subset 6					
	DF	CHI	IG	GR	Average		DF	CHI	IG	GR	Average
DT	41.813	46.491	46.491	47.563	45.590	DT	43.262	48.928	48.469	49.311	47.492
NB	45.517	56.433	56.725	56.530	53.801	NB	59.418	63.936	64.931	64.012	63.074
SVM	50.975	66.082	65.205	65.984	62.061	SVM	63.400	76.493	75.804	76.493	73.047
RBFNN	42.983	65.497	66.472	67.641	60.648	RBFNN	45.636	80.934	81.240	80.781	72.148
KNN	42.983	46.979	46.491	47.563	46.004	KNN	45.865	49.770	49.923	61.103	51.665
Average	44.854	56.296	56.277	57.057		Average	51.516	64.012	64.074	66.340	
Data subset 7						Data subset 8					
	DF	CHI	IG	GR	Average		DF	CHI	IG	GR	Average
DT	39.290	47.281	47.503	46.393	45.117	DT	50.141	51.724	51.555	54.438	51.964
NB	51.498	61.154	61.376	61.043	58.768	NB	54.155	62.239	63.030	62.691	60.529
SVM	53.274	69.589	69.811	68.258	65.233	SVM	63.765	71.001	70.831	71.170	69.192
RBFNN	37.736	66.371	66.371	64.373	58.713	RBFNN	43.358	69.248	69.814	68.287	62.677
KNN	37.736	43.729	45.172	44.284	42.730	KNN	42.680	46.128	46.637	51.781	46.806
Average	43.907	57.625	58.047	56.870		Average	50.820	60.068	60.373	61.673	
Data subset 9						Data subset 10					
	DF	CHI	IG	GR	Average		DF	CHI	IG	GR	Average
DT	70.000	71.923	71.923	72.115	71.490	DT	49.904	52.404	52.404	52.308	51.755
NB	68.942	68.462	71.346	68.462	69.303	NB	41.827	50.577	50.192	50.577	48.293
SVM	72.500	76.346	76.346	76.346	75.385	SVM	52.596	56.539	56.539	56.539	55.553
RBFNN	70.192	70.865	70.865	70.865	70.697	RBFNN	47.212	51.442	51.442	51.346	50.361
KNN	70.096	70.577	70.577	70.577	70.457	KNN	47.019	46.731	48.365	46.731	47.212
Average	70.346	71.635	72.212	71.673		Average	47.712	51.538	51.788	51.500	
Data subset 11						Data subset 12					
	DF	CHI	IG	GR	Average		DF	CHI	IG	GR	Average

	DF	CHI	IG	GR	Average		DF	CHI	IG	GR	Average
DT	54.535	55.386	55.474	55.386	55.195	DT	42.002	43.910	43.734	44.115	43.440
NB	50.103	52.216	52.011	51.981	51.578	NB	40.622	43.264	43.264	43.264	42.604
SVM	59.026	64.045	64.015	64.162	62.812	SVM	47.491	50.631	50.455	50.132	49.677
RBFNN	54.007	59.525	59.495	59.965	58.248	RBFNN	32.785	44.761	44.614	44.262	41.606
KNN	51.394	52.245	52.304	53.772	52.429	KNN	36.807	43.000	43.029	43.998	41.709
Average	53.813	56.683	56.660	57.053		Average	39.941	45.113	45.019	45.154	

Table 6

The execution time of each multi-class sentiment classifier corresponding to the best classification accuracy shown in Table 5.

Data subset 1						Data subset 2					
	DF	CHI	IG	GR	Average		DF	CHI	IG	GR	Average
DT	0.62	18.81	2.03	1.12	5.65	DT	0.47	24.71	1.36	1.75	7.07
NB	0.16	0.20	0.09	0.08	0.13	NB	0.21	0.93	0.18	0.21	0.38
SVM	1.39	10.76	1.60	1.34	3.77	SVM	3.20	17.66	1.64	2.93	6.36
RBFNN	1.48	0.59	3.24	1.85	1.79	RBFNN	10.25	1.44	3.71	8.96	6.09
KNN	0.01	0.01	0.01	0.01	0.01	KNN	0.01	0.01	0.01	0.01	0.01
Average	0.73	6.07	1.39	0.88		Average	2.83	8.95	1.38	2.77	
Data subset 3						Data subset 4					
	DF	CHI	IG	GR	Average		DF	CHI	IG	GR	Average
DT	0.64	7.74	2.33	0.62	2.83	DT	7.30	33.13	3.53	5.35	12.33
NB	0.13	0.65	0.05	0.13	0.24	NB	0.20	13.52	0.22	0.22	3.54
SVM	0.78	9.04	1.26	0.78	2.97	SVM	4.25	21.54	8.34	4.25	9.60
RBFNN	1.08	0.06	1.81	0.85	0.95	RBFNN	9.78	1.57	2.04	5.09	4.62
KNN	0.01	0.01	0.01	0.01	0.01	KNN	0.01	0.01	0.01	0.01	0.01
Average	0.53	3.50	1.09	0.48		Average	4.31	13.95	2.83	2.98	
Data subset 5						Data subset 6					
	DF	CHI	IG	GR	Average		DF	CHI	IG	GR	Average
DT	2.20	5.86	0.19	1.28	2.38	DT	1.28	30.78	0.91	0.62	8.40
NB	0.06	4.90	0.04	0.03	1.26	NB	0.12	0.24	0.12	4.20	1.17
SVM	1.17	5.81	1.16	1.26	2.35	SVM	2.95	17.45	0.66	2.50	5.89
RBFNN	0.95	4.33	0.41	0.61	1.58	RBFNN	3.30	2.90	3.38	3.44	3.26
KNN	0.01	0.01	0.01	0.01	0.01	KNN	0.01	0.00	0.00	0.01	0.01
Average	0.88	4.18	0.36	0.64		Average	1.53	10.27	1.01	2.15	
Data subset 7						Data subset 8					
	DF	CHI	IG	GR	Average		DF	CHI	IG	GR	Average
DT	3.15	17.33	2.66	0.30	5.86	DT	28.69	43.49	2.32	18.00	23.13
NB	0.03	0.55	0.02	0.02	0.16	NB	0.12	0.30	0.10	0.21	0.18
SVM	0.60	9.90	0.60	0.61	2.93	SVM	3.15	43.47	3.42	3.42	13.37
RBFNN	0.52	0.48	0.48	0.45	0.48	RBFNN	3.09	0.10	2.48	4.16	2.46
KNN	0.01	0.01	0.01	0.01	0.01	KNN	0.01	0.10	0.00	0.01	0.01
Average	0.86	5.65	0.75	0.28		Average	7.01	17.49	1.66	5.16	
Data subset 9						Data subset 10					

	DF	CHI	IG	GR	Average		DF	CHI	IG	GR	Average
DT	0.31	0.72	0.14	0.32	0.37	DT	0.81	18.59	0.17	0.38	4.99
NB	0.01	0.01	0.01	0.01	0.01	NB	0.01	0.02	0.01	0.01	0.01
SVM	0.53	0.93	0.51	0.51	0.62	SVM	0.67	1.05	0.64	0.65	0.75
RBFNN	0.44	1.35	0.44	0.44	0.67	RBFNN	0.42	0.07	0.42	0.42	0.33
KNN	0.01	0.01	0.01	0.01	0.01	KNN	0.01	0.01	0.01	0.01	0.01
Average	0.26	0.60	0.22	0.26		Average	0.38	3.95	0.25	0.29	
Data subset 11						Data subset 12					
	DF	CHI	IG	GR	Average		DF	CHI	IG	GR	Average
DT	6.34	11.69	7.87	7.61	8.38	DT	2.40	4.87	2.14	2.31	2.93
NB	0.08	0.05	0.08	0.10	0.08	NB	0.13	0.43	0.16	0.13	0.21
SVM	6.17	9.90	6.83	6.32	7.31	SVM	5.76	9.95	5.44	5.65	6.70
RBFNN	3.73	2.97	4.77	4.03	3.88	RBFNN	1.53	3.49	2.00	1.33	2.09
KNN	0.00	0.01	0.01	0.01	0.01	KNN	0.01	0.01	0.01	0.01	0.01
Average	3.26	4.92	3.91	3.61		Average	1.96	3.75	1.95	1.89	

5.2. Comparisons of different feature selection algorithms

According to the data shown in Table 5 and Table 6, the average best classification accuracies and the average best execution times of different feature selection algorithms concerning different data subsets are calculated, as shown in Table 7. The average classification accuracies and the average execution times of different feature selection algorithms are calculated, which are shown in the last row of Table 7. According to the data shown in Table 7, we can draw the preliminary conclusion that, in terms of classification accuracy, GR performs best among the four feature selection algorithms, CHI and IG are slightly worse than GR. On the other hand, IG is the one with the shortest execution time, the execution times of DF and GR are slightly longer than IG. To verify the significance of the experimental results obtained by this study, Wilcoxon test is used. According to the process of Wilcoxon test shown in Section 4.2, the test results can be obtained, which are shown in Table 8. It can be seen from Table 8 that, in terms of classification accuracy, significant differences are found in cases of CHI vs. GR, CHI vs. DF, GR vs. DF, and IG vs. DF (with $\alpha=0.01$). Significant difference can also be found in the case of GR vs. IG (with $\alpha=0.1$). But, significant difference can not be found in the case of CHI vs. IG. In terms of execution time, significant differences are found in cases of CHI vs. GR, CHI vs. IG and CHI vs. DF (with $\alpha=0.01$), whereas significant difference can not be found in the cases of GR vs. IG, GR vs.

DF and IG vs. DF. According to Table 7 and Table 8, we can draw the conclusions that, in terms of classification accuracy, the ranking of the four feature selection algorithms is $GR \succ IG \succ CHI \succ DF$, where “ $\succ$ ” denotes “superior than”. In terms of execution time, the ranking of the four feature selection algorithms is $IG \succ GR \succ DF \succ CHI$, but significant difference can not be found in the cases of GR vs. IG, GR vs. DF and IG vs. DF.

Table 7

The average best classification accuracies and the average best execution times of different feature selection algorithms concerning different data subsets.

Data subset	Classification accuracy				Execution time			
	DF	CHI	IG	GR	DF	CHI	IG	GR
Data subset 1	50.964	61.986	61.733	65.823	0.73	6.07	1.39	0.88
Data subset 2	60.720	74.273	74.028	75.130	2.83	8.95	1.38	2.77
Data subset 3	49.079	67.303	67.370	70.633	0.53	3.50	1.09	0.48
Data subset 4	54.076	66.399	65.992	70.582	4.31	13.95	2.83	2.98
Data subset 5	44.854	56.296	56.277	57.057	0.88	4.18	0.36	0.64
Data subset 6	51.516	64.012	64.074	66.340	1.53	10.27	1.01	2.15
Data subset 7	43.907	57.625	58.047	56.870	0.86	5.65	0.75	0.28
Data subset 8	50.820	60.068	60.373	61.673	7.01	17.49	1.66	5.16
Data subset 9	70.346	71.635	72.212	71.673	0.26	0.60	0.22	0.26
Data subset 10	47.712	51.538	51.788	51.500	0.38	3.95	0.25	0.29
Data subset 11	53.813	56.683	56.660	57.053	3.26	4.92	3.91	3.61
Data subset 12	39.941	45.113	45.019	45.154	1.96	3.75	1.95	1.89
Average	51.479	61.078	61.131	62.457	2.05	6.94	1.40	1.78

Table 8

Wilcoxon tests of different feature selection algorithms in terms of classification accuracies and execution times.

Measure	Comparison	R ⁺	R ⁻	Hypothesis	p-Value
Classification accuracy	CHI vs. GR	6.5	71.5	Rejected for GR at 5%	0.011
	CHI vs. IG	32	46	Not rejected	0.583
	CHI vs. DF	78	0	Rejected for CHI at 1%	0.000
	GR vs. IG	65	13	Rejected for GR at 5%	0.041
	GR vs. DF	78	0	Rejected for GR at 1%	0.000
	IG vs. DF	78	0	Rejected for IG at 1%	0.000
Execution time	CHI vs. GR	91	0	Rejected for CHI at 1%	0.001
	CHI vs. IG	91	0	Rejected for CHI at 1%	0.001
	CHI vs. DF	91	0	Rejected for CHI at 1%	0.001
	GR vs. IG	55	36	Not rejected	0.507
	GR vs. DF	23	55	Not rejected	0.209
	IG vs. DF	25.5	65.5	Not rejected	0.162

Besides, corresponding to the best classification accuracy shown in Table 5, the usages of memory and central processing unit (CPU) of each feature selection algorithm are also recorded. To save the space of this paper, the specific data are omitted. The average usages of memory and CPU of each feature selection algorithm are calculated, which are shown in Fig. 7. The space complexity of the four feature selection algorithms can be observed in Fig. 7 since the greater average usages of memory and CPU indicate the higher space complexity of the corresponding feature selection algorithm.

Fig. 7. The average usages of memory and CPU of the four feature selection algorithms.

5.3. Comparisons of different machine learning algorithms

According to the data shown in Table 5 and Table 6, the average best classification accuracies and the average best execution times of different machine learning algorithms concerning different data subsets are calculated, as shown in Table 9. The average classification accuracies and the average execution times of different machine learning algorithms are calculated, which are shown in the last row of Table 9. According to the data shown in Table 9, we can draw the preliminary conclusions that, in terms of classification accuracy, SVM is the best one among the five machine learning algorithms, RBFNN and NB are slightly worse than SVM. On the other hand, KNN is the one with the shortest execution time, the execution times of NB, RBFNN and SVM are slightly longer than KNN. Further, Wilcoxon test is used, and the test results are shown in Table 10. It can be seen from Table 10 that, either in terms of classification accuracy or in terms of execution time, significant difference is found in each pair of the five machine learning algorithms. According to Table 9

and Table 10, we can draw the conclusion that, in terms of classification accuracy, the ranking of the five machine learning algorithms is $SVM \succ RBFNN \succ NB \succ DT \succ KNN$, where “ $\succ$ ” denotes “superior than”. In terms of execution time, the ranking of the five machine learning algorithms is $KNN \succ NB \succ RBFNN \succ SVM \succ DT$.

Table 9

The average best classification accuracies and the average best execution times of different machine learning algorithms concerning different data subsets.

Data subset	Classification accuracy					Execution time				
	DT	NB	SVM	RBFNN	KNN	DT	NB	SVM	RBFNN	KNN
Data subset 1	51.777	63.364	71.227	65.141	49.124	5.65	0.13	3.77	1.79	0.01
Data subset 2	61.179	75.881	82.504	77.948	57.676	7.07	0.38	6.36	6.09	0.01
Data subset 3	49.639	67.758	75.139	73.474	51.97	2.83	0.24	2.97	0.95	0.01
Data subset 4	55.823	70.605	75.947	68.231	50.707	12.33	3.54	9.60	4.62	0.01
Data subset 5	45.59	53.801	62.061	60.648	46.004	2.38	1.26	2.35	1.58	0.01
Data subset 6	47.492	63.074	73.047	72.148	51.665	8.40	1.17	5.89	3.26	0.01
Data subset 7	45.117	58.768	65.233	58.713	42.73	5.86	0.16	2.93	0.48	0.01
Data subset 8	51.964	60.529	69.192	62.677	46.806	23.13	0.18	13.37	2.46	0.01
Data subset 9	71.49	69.303	75.385	70.697	70.457	0.37	0.01	0.62	0.67	0.01
Data subset 10	51.755	48.293	55.553	50.361	47.212	4.99	0.01	0.75	0.33	0.01
Data subset 11	55.195	51.578	62.812	58.248	52.429	8.38	0.08	7.31	3.88	0.01
Data subset 12	43.440	42.604	49.677	41.606	41.709	2.93	0.21	6.70	2.09	0.01
Average	52.538	60.463	68.148	63.324	50.707	7.03	0.61	5.22	2.35	0.01

Table 10

Wilcoxon tests of different machine learning algorithms in terms of classification accuracies and execution times.

Measure	Comparison	R ⁺	R ⁻	Hypothesis	p-Value
Classification accuracy	NB vs. SVM	0	78	Rejected for SVM at 1%	0.002
	NB vs. RBFNN	11	67	Rejected for RBFNN at 5%	0.028
	NB vs. KNN	73	5	Rejected for NB at 1%	0.008
	NB vs. DT	68	10	Rejected for NB at 5%	0.023
	SVM vs. RBFNN	78	0	Rejected for SVM at 1%	0.002
	SVM vs. KNN	78	0	Rejected for SVM at 1%	0.002
	SVM vs. DT	78	0	Rejected for SVM at 1%	0.002
	RBFNN vs. KNN	77	1	Rejected for RBFNN at 1%	0.003
	RBFNN vs. DT	72	6	Rejected for RBFNN at 1%	0.010
Execution time	KNN vs. DT	14	64	Rejected for DT at 5%	0.050
	NB vs. SVM	0	91	Rejected for SVM at 1%	0.001
	NB vs. RBFNN	0	91	Rejected for RBFNN at 1%	0.001
	NB vs. KNN	66	0	Rejected for NB at 1%	0.003
	NB vs. DT	0	91	Rejected for DT at 1%	0.001
	SVM vs. RBFNN	90	1	Rejected for SVM at 1%	0.002

SVM vs. KNN	91	0	Rejected for SVM at 1%	0.001
SVM vs. DT	16	75	Rejected for DT at 5%	0.039
RBFNN vs. KNN	91	0	Rejected for RBFNN at 1%	0.001
RBFNN vs. DT	1	90	Rejected for DT at 1%	0.002
KNN vs. DT	0	91	Rejected for DT at 1%	0.001

Corresponding to the best classification accuracy shown in Table 5, the usages of memory and CPU of each machine learning algorithm are also recorded. The average usages of memory and CPU of each machine learning algorithm are calculated, which are shown in Fig. 8. The space complexity of the five machine learning algorithms can be observed in Fig. 8 since the greater average usages of memory and CPU indicate the higher space complexity of the corresponding machine learning algorithm.

Fig. 8. The average usages of memory and CPU of the five machine learning algorithms.

5.4. Comparisons between classification accuracies obtained in the existing studies and those obtained in this study

In the existing studies (Pang & Lee, 2005; Goldberg & Zhu, 2006; Wilson et al., 2006; Bickerstaffe & Zukerman, 2010; Cao & Zukerman, 2012; Thelwall et al., 2010; Thelwall et al., 2012), the above data subsets have been used to illustrate the effectiveness of the existing methods for multi-class sentiment classification. Fig. 9 shows the classification accuracies obtained in the existing studies and best classification accuracies obtained in this study, where Thelwall's Unsupervised ssth and Thelwall's Supervised ssth respectively denote the unsupervised version and the supervised version of Thelwall's SentiStrength (Thelwall et al., 2010; Thelwall et al., 2012). It can be seen from Fig. 9 that, with respect to most of the twelve data subsets, the best classification accuracies obtained in this study are higher than the

classification accuracies obtained in the existing studies.

It is necessary to point out that several techniques have been proposed in the existing studies, such as metric labeling technique (Pang and Lee, 2005), undirected graph technique (Goldberg and Zhu, 2006), the dependency tree (Wilson et al. 2006) and so on. But, feature selection algorithms are not used in most of the existing studies. Thus, it has the potential to develop more efficient methods for multi-class sentiment classification by combining the framework proposed in this paper with the techniques proposed in the existing studies, such as metric labeling technique (Pang and Lee, 2005), undirected graph technique (Goldberg and Zhu, 2006), the dependency tree (Wilson et al. 2006) and so on.

Fig. 9. The classification accuracies obtained in the existing studies and the best classification accuracies obtained in this study.

6. Conclusions

In this paper, an experimental study is conducted for comparing the performances of four feature selection algorithms (DF, CHI, IG and GR) and five machine learning algorithms (DT, NB, SVM, RBFNN and KNN) in multi-class sentiment classification. A framework for multi-class sentiment classification is proposed, in which the feature selection algorithm is used to select important features of texts, and the machine learning algorithm is used to train the multi-class sentiment classifier. Three public datasets including twelve data subsets are used to measure the classification accuracies of different multi-class sentiment classifiers that combine different feature selection algorithms and machine learning algorithms. The

Wilcoxon test is used to verify the existence of significant difference between each pair of feature selection algorithms and each pair of machine learning algorithms in multi-class sentiment classification. In terms of execution time, the similar comparisons are also conducted. The experimental results indicate the following conclusions:

(1) In terms of classification accuracy, the ranking of the four feature selection algorithms for conducting multi-class sentiment classification is $GR \succ IG \succ CHI \succ DF$, where “ $\succ$ ” denotes “superior than”; while in terms of execution time, the ranking of the four feature selection algorithms is $IG \succ GR \succ DF \succ CHI$, but significant difference can not be found in the cases of GR vs. IG, GR vs. DF and IG vs. DF.

(2) In terms of classification accuracy, the ranking of the five machine learning algorithms for conducting multi-class sentiment classification is $SVM \succ RBFNN \succ NB \succ DT \succ KNN$, where “ $\succ$ ” denotes “superior than”; while in terms of execution time, the ranking of the five machine learning algorithms is $KNN \succ NB \succ RBFNN \succ SVM \succ DT$.

(3) Most of the classification accuracies of the multi-class sentiment classifiers that combine DF with different machine learning algorithms present the increasing tendencies with the increasing numbers of features, whereas the classification accuracies of the multi-class sentiment classifiers that combine CHI, IG or GR with different machine learning algorithms present the tendencies that increase at first and then decrease at some points. In other words, more features do not necessarily lead to better result if CHI, IG or GR is used.

Based on the experimental results, the studies on improving the existing multi-class sentiment classifiers (Pang & Lee, 2005; Goldberg & Zhu, 2006; Wilson et al., 2006; Bickerstaffe & Zukerman, 2010; Cao & Zukerman, 2012; Thelwall et al., 2010; Thelwall et al., 2012) can be attempted, and the new multi-class sentiment classifiers can be developed. Besides, the performances of SVM classifiers could be affected by different kernels and one-vs-one strategies. Thus, experimental comparisons of SVM classifiers with different kernels and one-vs-one strategies can be regarded as future research direction. Further, the Kolmogorov complexity of the multi-class sentiment classifier (algorithm) needs to be studied in the future.

Acknowledgement

This work was partially supported by the National Science Foundation of China (Project Nos. 71371002 and 71571039), the Fundamental Research Funds for the Central Universities, China (Project Nos. N140607001), and the 111 Project (B16009).

References

- Agarwal, B., & Mittal, N. (2016). Prominent feature extraction for review analysis: an empirical study. *Journal of Experimental & Theoretical Artificial Intelligence*, 28(3), 485-498.
- Balazs, J. A., & Velásquez, J. D. (2016). Opinion mining and information fusion: A survey. *Information Fusion*, 27, 95-110.
- Bertot, J. C., Jaeger, P. T., & Hansen, D. (2012). The impact of polices on government social media usage: Issues, challenges, and recommendations. *Government information quarterly*, 29(1), 30-40.
- Bickerstaffe, A., & Zukerman, I. (2010). A hierarchical classifier applied to multi-way sentiment detection. In *Proceedings of the 23rd international conference on computational linguistics* (pp. 62-70).
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth, CA: Chapman & Hall.
- Cao, M. D., & Zukerman, I. (2012). Experimental evaluation of a lexicon-and corpus-based ensemble for multi-way sentiment analysis. In *Proceedings of the Australasian Language Technology Association Workshop* (pp. 52–60). Dunedin, New Zealand.
- Chang, W. H., & Chang, J. S. (2012). An effective early fraud detection method for online auctions. *Electronic Commerce Research and Applications*, 11(4), 346-360.
- Chen, K., Kou, G., Shang, J., & Chen, Y. (2015). Visualizing market structure through online product reviews: Integrate topic modeling, TOPSIS, and multi-dimensional scaling approaches. *Electronic Commerce Research and Applications*, 14(1), 58-74.
- Chen, S., Cowan, C. F., & Grant, P. M. (1991). Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on neural networks*, 2(2), 302-309.
- Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of*

- Machine learning research*, 7, 1-30.
- Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine learning*, 29(2-3), 103-130.
- Fernández, A., Calderón, M., Barrenechea, E., Bustince, H., & Herrera, F. (2010). Solving multi-class problems with linguistic fuzzy rule based classification systems based on pairwise learning and preference relations. *Fuzzy sets and systems*, 161(23), 3064-3080.
- Fukunaga, K., & Narendra, P. M. (1975). A branch and bound algorithm for computing k-nearest neighbors. *IEEE transactions on computers*, 100(7), 750-753.
- Galavotti, L., Sebastiani, F., & Simi, M. (2000). Feature selection and negative evidence in automated text categorization. In *Proceedings of the ACM SIGKDD-00 workshop on text mining*. Boston, MA, USA.
- Garcia, S., Derrac, J., Cano, J., & Herrera, F. (2012). Prototype selection for nearest neighbor classification: Taxonomy and empirical study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3), 417-435.
- Goldberg, A. B., & Zhu, X. (2006, June). Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing* (pp. 45-52).
- Hastie, T., & Tibshirani, R. (1998). Classification by pairwise coupling. *The annals of statistics*, 26(2), 451-471.
- Hüllermeier, E., & Vanderlooy, S. (2010). Combining predictions in pairwise classification: An optimal adaptive voting strategy and its relation to weighted voting. *Pattern Recognition*, 43(1), 128-142.
- Isa, D., Lee, L. H., Kallimani, V. P., & Rajkumar, R. (2008). Text document preprocessing with the Bayes formula for classification using the support vector machine. *IEEE Transactions on Knowledge and Data engineering*, 20(9), 1264-1272.
- Kang, D., & Park, Y. (2014). Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach. *Expert Systems with Applications*, 41(4), 1041-1050.
- Lam, W., & Han, Y. (2003). Automatic textual document categorization based on generalized instance sets and a meta model. *IEEE Transactions on Pattern Analysis and Machine*

- Intelligence*, 25(5), 628-633.
- Lee, A. J., Yang, F. C., Chen, C. H., Wang, C. S., & Sun, C. Y. (2016). Mining perceptual maps from consumer reviews. *Decision Support Systems*, 82, 12-25.
- Li, Y. M., Chen, H. M., Liou, J. H., & Lin, L. F. (2014). Creating social intelligence for product portfolio design. *Decision Support Systems*, 66, 123-134.
- Liu, Y., Bi, J.W., & Fan, Z.P. (2017). Ranking products through online reviews: A method based on sentiment analysis technique and intuitionistic fuzzy set theory. *Information Fusion*, 36, 149-161.
- McCallum, A., & Nigam, K. (1998). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization* (pp. 41-48). Menlo Park, CA: AAAI Press.
- Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093-1113.
- Musavi, M. T., Ahmed, W., Chan, K. H., Faris, K. B., & Hummels, D. M. (1992). On the training of radial basis function classifiers. *Neural networks*, 5(4), 595-603.
- Omar, N., Albared, M., Al-Moslmi, T., & Al-Shabi, A. (2014). A comparative study of feature selection and machine learning algorithms for Arabic sentiment classification. In *Asia Information Retrieval Symposium* (pp. 429-443).
- Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics* (pp. 115-124).
- Park, K. J., & Kanehisa, M. (2003). Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics*, 19(13), 1656-1663.
- Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1), 81-106.
- Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann San Mateo, CA, USA.
- Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, 89, 14-46.
- Ravi, K., & Ravi, V. (2016). Sentiment classification of Hinglish text. In *2016 3rd*

- International Conference on Recent Advances in Information Technology* (pp. 641-645).
- Schwenker, F., Kestler, H. A., & Palm, G. (2001). Three learning phases for radial-basis-function networks. *Neural networks*, 14(4), 439-458.
- Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E. (2015). Sentiment analysis: a review and comparative analysis of web services. *Information Sciences*, 311, 18-38.
- Sharma, A., & Dey, S. (2012a). A comparative study of feature selection and machine learning techniques for sentiment analysis. In *Proceedings of the 2012 ACM Research in Applied Computation Symposium* (pp. 1-7).
- Sharma, A., & Dey, S. (2012b). Performance investigation of feature selection methods and sentiment lexicons for sentiment analysis. *IJCA Special Issue on Advanced Computing and Communication Technologies for HPC Applications*, 3, 15-20.
- Šuch, O., & Barreda, S. (2016). Bayes covariant multi-class classification. *Pattern Recognition Letters*, 84, 99-106.
- Tan, S., & Zhang, J. (2008). An empirical study of sentiment analysis for Chinese documents. *Expert Systems with Applications*, 34(4), 2622-2629.
- Tang, H., Tan, S., & Cheng, X. (2009). A survey on sentiment detection of reviews. *Expert Systems with Applications*, 36(7), 10760-10773.
- Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1), 163-173.
- Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12), 2544-2558.
- Tsytsarau, M., & Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24(3), 478-514.
- Vapnik, V. (2000). *The Nature of Statistical Learning Theory*. Springer, New York.
- Wang, G., Sun, J., Ma, J., Xu, K., & Gu, J. (2014). Sentiment classification: The contribution of ensemble learning. *Decision support systems*, 57, 77-93.
- Watanabe, O. (1992). *Kolmogorov complexity and computational complexity*. Springer,

- Berlin.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6), 80-83.
- Wilson, T., Wiebe, J., & Hwa, R. (2006). Recognizing strong and weak opinion clauses. *Computational Intelligence*, 22(2), 73-99.
- Witten, I. H., & Frank, E. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). San Francisco: Morgan Kaufmann.
- Wu, T. F., Lin, C. J., & Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5(Aug), 975-1005.
- Yang, X., Yang, G., & Wu, J. (2016). Integrating rich and heterogeneous information to design a ranking system for multiple products. *Decision Support Systems*, 84, 117-133.
- Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 42-49.
- Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the 14th international conference on machine learning* (pp. 412–420). Nashville, TN, USA.
- Zahorian, S. A., & Nossair, Z. B. (1999). A partitioned neural network approach for vowel classification using smoothed time/frequency features. *IEEE Transactions on Speech and Audio Processing*, 7(4), 414-425.
- Zhang, D., Li, S., Zhu, C., Niu, X., & Song, L. (2010). A comparison study of multi-class sentiment classification for Chinese reviews. In *Proceedings of the 7th international conference on fuzzy systems and knowledge discovery* (pp. 2433-2436).
- Zhang, W., Xu, H., & Wan, W. (2012). Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis. *Expert Systems with Applications*, 39(11), 10283-10291.