International Journal of Multimedia and Ubiquitous Engineering
Vol.9, No.7 (2014), pp.333-346
http://dx.doi.org/10.14257/ijmue.2014.9.7.28

Sentiment Analysis on News Comments Based on Supervised Learning Method

Yan Zhao[1], Suyu Dong[2] and Leixiao Li[3]

[1,2]*College of Management, Inner Mongolia University of Technology*
Huhhot, China
[3]*College of Information Engineering, Inner Mongolia University of Technology*
Huhhot, China
[1]*zhaoyanimut@126.com*, [2]*dongsuyu@126.com*, [3]*llxhappy@126.com*

Abstract

Up to now, sentiment analysis has become one of most active research ares in NLP, many researchers have conducted sentiment analysis for foreign language documents. Compared with the researches of foreign language documents, there are few studies on sentiment classification of Chinese document, and fewer studies on news comments. This paper presents a research of sentiment analysis on news comments. In this paper, we adopt four feature selection methods(DF, IG, CHI, MI), three feature representations(Presence, TF, TF-IDF) and five learning methods(NB, ME, Winnow, C4.5, SVM) for the sentiment analysis of Chinese news comments. The experimental results indicate that, except MI, other three feature selection methods are all suitable for selecting features for news comments, and through comprehensive assessment of feature selection method, CHI is better; TF performs the best calculation of feature weighting; ME outperforms other classifiers for the sentiment classification.

Keywords: *Sentiment analysis, News comments, Machine learning*

1. Introduction

With the advent of Web 2.0, people can check more and more reviews. These reviews are significant for customers, companies and governments. Thus, the sentiment analysis is needed for them. However, vast reviews are commented everyday, the accomplishment of gaining and analyzing these reviews by people is impossible. Sentiment analysis has become one of the key technologies for the solution of this problem. Since the year 2000, sentiment analysis has grown rapidly and become the most active area in NLP [1]. In fact, the sentiment analysis has spread from computer science to management sciences [2]. In recent years, sentiment classification has become the principle research question of sentiment analysis [3]. Sentiment classification aims to classify the polarity of sentiment documents.

So far, most studies of sentiment classification are focused on English documents. In addition, the documents are about movie reviews, product reviews and so on. [4, 5] applied supervised learning method to sentiment classification of movie reviews; [6, [7] had conducted on sentiment classification for product reviews; [8] analyzed the polarity of financial news texts. The researches of sentiment classification for Chinese documents are few, and these studies are about product reviews, restaurant reviews, fewer news comments are used as the corpus for sentiment classification. However, there are more and more reviews and opinions of news, these documents have had an effect on social life of others. Thus, it is necessary for government to conduct sentiment classification of news comments.

In this work, we will conduct experiment on sentiment classification of news comments and analysis following problems:

(1)Which feature selection method performs the best for feature selection (DF, IG, CHI, MI) of news comments?

ISSN: 1975-0080 IJMUE

(2)Which is the best classifier (Winnow, C.5, NB, ME and SVM) for the sentiment classification of news comments?

(3)Which feature representation (Presence, TF, TF-IDF) is the best method regarding news comments classification?

2. Related Works

Supervised learning method and unsupervised learning method are the main technologies of sentiment classification. In this paper, we apply supervised learning method to sentiment classification of news comments. For this method, the key problems are text vectorization and training classifier. Text vectorization contains extracting features and the calculation of feature weighting.

The high feature dimensions are the critical problem of sentiment classification. In the vector space of features, many features are useless for the sentiment classification or even will lower the efficiency and the effectiveness. Hence, dimensionality reduction is important for sentiment classification and a good feature selection method is a good way of dimension reduction. Wrappers and filters are two kinds method of feature selection of machine learning [9]. Wrappers spend plenty of time when it used for feature selection especially for the high-dimensional of space vector. Thus, wrappers are not suitable for feature selection of sentiment classification [10]. Filters are frequently used for extracting features of sentiment classification. They use the evaluation metric to measure the ability of terms for the classification and then to extract features. There are many methods for filters, such as IG, CHI, DF, MI, OR and so on. Up to now, a number of researchers focus on feature selection. [11] evaluated five feature selection methods for text classification and found that IG and CHI were the most effective methods; [12] proved that CHI was the best feature selection method for four classifiers of text categorization; [13] performed the binary classification with SVM and twelve methods of feature selection, the experiment result indicated that new method BNS(Bi-Normal Separation) was the best method; [14] improved Gini index theory and showed that the novel method was better than other feature selection. Compared with the studies of feature selection of text classification, the same researches for sentiment classification are fewer. [15, 16] showed the experiment result of feature selection for sentiment classification, [15] proved that IG outperformed other feature selection methods (DF, CHI, MI), [16] indicated that DF was the most suitable for sentiment classification. A large number of researches proved that various documents employ different methods of feature selection can reach the best accuracy of sentiment classification. This paper will research the feature selection of sentiment classification of news comments.

When we use machine learning method to perform sentiment classification, text feature weighting is necessary after feature selection. Feature weighting methods mainly are Presence, TF, TF-IDF. [17] compared Presence and TF as the feature representation methods of sentiment classification of movie reviews, the result showed that Presence outperformed TF; [18] proved that NB with Presence can achieve the top accuracy for the sentiment classification of Internet restaurant reviews written in Cantonese, SVM with different n-grams need different feature weighting methods to achieving its best accuracy; [16] used Boolean weight and various feature selection to sentiment classification. In this paper, the experiment adopts Presence, TF and TF-IDF to sentiment classification of news comments to gain the best method.

The classification technology is important for sentiment classification. So far, many researches of sentiment classification used machine learning. Naive Bayes, maximum entropy and support vector machine are often used for sentiment classification. [7, 18-20] used SVM and NB to sentiment classification of different documents, [20] showed SVM was better than NB for the sentiment classification of travel reviews; [18,19] proved that compared with NB, SVM was not a universal winner; [7] used more features for sentiment classification and showed that the accuracies were comparable for SVM and NB. [17]

compared SVM, NB with ME for sentiment classification of movie reviews, the experiment result showed that SVM was the best classifier. Fewer researches focused on Winnow and C4.5 for sentiment classification. [6] used Winnow, PA and LM to sentiment classification of product reviews; [21] adopted five classifiers(Centroid classifier, KNN classifier, NB classifier, Winnow classifier, SVM classifier) for sentiment classification of product reviews and found that SVM outperformed other classifiers; [22] proved the effect of sentiment classification of SVM was not better than C4.5 anytime. The above work shows that different documents use different machine learning technologies can reach the best effect of sentiment classification. This paper will compare the utility of five classifiers (SVM, NB, ME, Winnow, C4.5) which is used for sentiment classification of news comments.

3. Theory Model

In this paper, each document is represented as a vector with feature weights. Let {t1,t2,…,tm} be a predefined set of m features that can appear in a document. Let wi is the feature weight in a document. Each document d is represented by the document vector. d={w1,w2,…,wm}

3.1. Data Collection

The data used for our experiment were downloaded from influenced Chinese news website (URL: http : // news.sina.com.cn; http : // news.sohu.com; http : // news.qq.com; http: // news.163.com; http://www.people.com.cn) . We through a crawler acquired 3800 news comments.

To perform the experiment, we trained three students to annotate the polarity of comments. In the whole process of the annotation, non-news comments were firstly excluded, and used for -1, 0, 1 to annotate the polarity of comments. Thereinto, -1 represents negative comments, 1 represents positive comments and 0 represents the comments whose polarities can not be judged. We removed the comments which were annotated with 0 or inconsistent by two students. Finally, there were 1500 positive comments and 1500 negative comments.

To avoid the error from the selection of training set and testing set and guarantee the veracity of our experiment, this paper used 3-fold cross validation, every evaluation index adopts the mean value of three results of experiment.

3.2. Feature Selection

A large number of features are produced by the feature identification, and some features are useless or interferential for sentiment classification. Using these features to represent document, not only the dimension of feature vector space is high but also the effect and efficiency of classifiers will be reduced. Thus, in order to improve the ability of classification, feature selection is needed. This paper will compare the accuracies of sentiment classification which were achieved by four feature selection methods with different feature weighting methods and classifiers. The feature selection methods are DF, IG, CHI and MI.

3.2.1 DF: Document frequency is the number of every feature appearing in all texts (comments). After computing the DF value of every feature, appropriate features are selected through the threshold. If the DF value is too small, the feature is unrepresentative; if the DF value is too large, the feature is not sensitive. By the small threshold and large threshold to wipe out the interferential features.

3.2.2 IG: IG is based on the importance of a certain feature that is measured by the information it provided to category. The amount of information of a feature for

classification is measured by entropy. The IG value of a certain feature t_i is calculated by the following equation:

$$IG(t_i) = Entropy(S) - Expected \quad Entropy(S_{t_i})$$
$$= \left\{ -\sum_{j=1}^{M} P(C_j) \times \log P(C_j) \right\} - \left\{ P(t_i) \times \left[-\sum_{j=1}^{M} P(C_j|t_i) \times \log P(C_j|t_i) \right] \right.$$
$$\left. + P(\overline{t_i}) \times \left[-\sum_{j=1}^{M} P(C_j|\overline{t_i}) \times \log P(C_j|\overline{t_i}) \right] \right\}$$

Thereinto, $P(C_j)$ indicates the probability of a document belonging to C_j. $P(t_i)$ indicates the probability of a document which contains feature t_i. $P(C_j \mid t_i)$ indicates the probability of a document which belongs to C_j if it contains feature t_i. $P(\overline{t_i})$ indicates the probability of a document which does not contain feature t_i. $P(C_j|\overline{t_i})$ indicates the probability of a document which belongs to C_j if it does not contain feature t_i. M indicates the number of classifications.

3.2.3 CHI: CHI measures the relevance between feature t_i and class C_j. It assumes that feature t_i and class C_j match the Gamma distribution with the first-order degree of freedom. The CHI value of feature t_i for class C_j is larger, the relationship between feature t_i and class C_j is more compact, and the ability of feature t_i to distinguish document is stronger. The CHI value of feature are calculated by the following formula (the binary classification):

$$\chi^2(t_i, C_1) = \frac{N \times [N(C_1, t_i) \times N(C_2, \overline{t_i}) - N(C_2, t_i) \times N(C_1, \overline{t_i})]^2}{\left\{ [N(C_1, t_i) + N(C_2, t_i)] \times [N(C_2, \overline{t_i}) + N(C_1, \overline{t_i})] \right.}$$
$$\left. + [N(C_1, t_i) + N(C_1, \overline{t_i})] \times [N(C_2, t_i)] + N(C_2, \overline{t_i})] \right\}$$

Thereinto, N represents the total number of documents, $N(C_j,t_i)$ represents the number of documents belonging to class C_j and containing feature t_i, $N(C_j, \overline{t_i})$ represents the number of documents belonging to class C_j and without feature t_i. For multi-classification, CHI value can use two methods to calculate. One method is that, compute the CHI value of feature t_i for every class and calculate the CHI value with training set, the formulation is $\chi^2_{MAX}(t_i) = \max_{j=1}^{M} \{\chi^2(t_i, C_j)\}$, the M denotes the number of class, select features which is greater than threshold. The other method is calculating the mean value of every class, the formulation is $\chi^2_{AVG}(t_i) = \sum_{j=1}^{M} P(C_j)\chi^2(t_i, C_j)$.

3.2.4 MI: MI is the frequently-used of computational linguistics model analysis, used for measuring the correlation between two objects. The basis idea is that: the larger the MI value is the higher co-occurrence between feature t_i and class C_j is. So, a number of terms with largest MI value will be selected for feature. The MI value calculated by the following formulation (the binary classification):

$$MI(t_i, C_j) = \log \frac{P(t_i, C_j)}{P(t_i) \times P(C_j)} = \log \frac{P(t_i|C_j)}{P(t_i) \times P(C_j)}$$

$$\approx \log \frac{N(C_1, t_i) \times N}{[N(C_1, t_i) + N(C_1, \overline{t_i})] \times [N(C_1, t_i) + N(C_2, t_i)]}$$

Thereinto, the interpretation of N, $N(C_j, t_i)$ and $N(C_j, \overline{t_i})$ is same with CHI. If feature t_i and class C_j is irrelevant, then $P(t_i, C_j) = P(t_i) \times P(C_j)$, and the MI value is zero. Being similar with CHI, for multi-classification, the MI value can be calculated through following

formulation: $MI_{MAX}(t_i) = \max_j^M [P(C_j) \times MI(t_i, C_j)]$ and $MI_{AVG}(t_i) = \sum_{j=1}^{M} P(C_j) MI(t_i, C_j)$, *M* denotes the number of class.

3.3. Feature Weighting

The ability of every feature to distinguish document is different, and this ability can be measured by feature weighting. Feature weighting get from the statistical information of documents. This paper will compare three feature weighting methods: Presence, TF and TF-IDF.

3.3.1. Presence: Presence is based on the feature whether or not appears in the text. If the feature appears in the document, the value is 1, otherwise the value is 0. So, Presence can not represent the effect of features for the document, and Presence is often replaced by other more accurate feature representation methods in practical application. However, in different applications, Boolean is better than other feature weighting methods, many researches of sentiment classification use Presence [16-18].

3.3.2 TF: TF uses the times of feature appearance in the text to represent the documents. In the documents, sometimes many low-frequency features perhaps have the greater ability to distinguish the document; on the contrary, the ability of many high-frequency features is weak. TF maybe ignore some low-frequency features. However, researchers often use TF and it performs well for sentiment classification [19, 23].

3.3.3 *TF-IDF*: *TF-IDF* is the most widely used feature weight calculation method for the text classification. It is based on the idea: if one feature has high-frequency, and rarely appears in other text, then the feature has a good ability to distinguish. Although its ideas and structure of statistics are very simple, but its performance is very good. The *TF-IDF* value of a certain feature is calculated by the following equation: $w_{ij} = tf_{ij} \times \log \frac{N}{n_i}$

Thereinto, w_{ij} indicates the weight of feature t_i in document d_j. tf_{ij} indicates the frequency of feature t_i in document d_j. n_i indicates the number of document which contains feature t_i. N is the number of all documents.

3.4. Classifiers

3.4.1. Naive Bayes: Naive Bayes classifier is widely used in the text classification, it use the Bayes formula to calculate the probability of document d belongs to C_i, the equation is $P(C_i|d) = \frac{P(d|C_i) * P(C_i)}{P(d)}$. $P(C_i)$ denotes the probability of a document belonging to C_i. On the basis of the assumption of independence conditions, NB uses the joint probability between features and categories to estimate the probability of categories given a document, namely that $P_{NB}(C_i|d) = \frac{P(C_i)(\prod_{t_i \in V} P(t_i|C_i)^{W(t_i,d)})}{\sum_j [P(C_j) \prod_{t_i \in V} P(t_i|C_j)^{W(t_i,d)}]}$. Thereinto, feature t_i is independent of document d, $W(t_i,d)$ indicates the weights of feature t_i in document d. $P(t_i \mid C_i)$ indicates the Laplacean probability estimation value of conditional probability of documents belonging to C_i if it contains feature t_i. $P(t_i \mid C_i)$ is calculated by the following equation: $P(t_i|C_i) = \frac{1 + W(t_i, C_i)}{|V| + \sum_j W(t_j, C_i)}$. $W(t_i,C_i)$ indicates the number of documents containing features t_i and belonging to C_i. $|V|$ is the size of $\{t_1,t_2,\ldots,t_m\}$, which are all features coming from all documents.

Although the assumption is harsh, NB performs well and is efficient in the text categorization. For example, [18] showed that using machine leaning to sentiment classification of restaurant reviews written in Cantonese, NB achieved the top accuracy.

3.4.2 Maximum Entropy: Maximum entropy classifier (ME) is based on maximum entropy model, [24] was the first application of maximum entropy models in the natural language processing; [25] improved maximum entropy model. [26] found that ME is better classifier than Naive Bayes classifier on text classification. Its basic idea is that it does not make any hypothesis and remain maximum entropy for the unknown information, this is an advantage for maximum entropy compared with Naive Bayes. Maximum entropy model must satisfy the constraint of known information and the principle of maximum entropy. Hence, maximum entropy model is got through solving a optimization problem with constraints. The classical algorithm to solve this problem is Lagrange multiplier method. In this paper, we give the conclusion directly. The result is following: $p^*(C_i|t_i)=\frac{1}{\sum_{C_i}\exp\left(\sum_i \lambda_i f(t_i,C_i)\right)}\exp\left(\sum_i \lambda_i f(t_i,C_i)\right)$

P^* indicates a predictive model for classification; V indicates the feature vectors;C_i indicates the type which the document belongs to. λ_i indicates the feature weight of feature vectors containing many feature t_i. $f(t_i, c_i)$ is a indicator function.

3.4.3 SVM: Support vector machine (SVM) is generally considered as the best classifier for traditional text classification [27], it is usually better than naive Bayes and maximum entropy. Naive Bayes and maximum entropy are based on probability model, support vector machine (SVM) classifier is got by solving the optimal hyperplane represented by vector $\vec{W}$. Hyperplane is used to accomplish classification which can ensure maximum separation between a certain amount of data from the training set and hyperplane. Solving the maximum margin hyperplane eventually is converted into solving a convex quadratic programming problem.

Generally, it translates the above problem into the constrained optimization problem of dual variables through Lagrange Duality. The solution can be written as: $\vec{W}=\sum_{i=1}^{n}\alpha_i C_i \vec{d}_i$. C_i is the correct category for document $\vec{d}_i$. α_i are support vector and greater than zero.

What's more, for linear inseparable problems, kernel function can be used for SVM to convert low dimensional space nonlinear problem to high dimension space linear problem. Mapping of kernel function can be a good control of the computational complexity of nonlinear expansion and can avoid the curse of dimensionality. There are many kernel functions: linear kernel, Gaussian kernel function, radial basis function and so on. In this paper, we used linear kernel function and optimize the parameter of SVM model, which will be used for following experiments.

3.4.4 Winnow: Winnow is a typical liner classifier. Hence, it is simple and easy to realize, it also has the small calculation and storage. Winnow is usually used for classification and show good effect for classification. Using the multiplicative weight updating algorithm, it is suitable for high-dimension, and especially fit many irrelevant attributes. Winnow train weight factor $s=(s_1,s_2,\ldots,s_n)$ for every class, for document $d=(w_1,w_2,\ldots,w_n)$, if $\sum_{i=1}^{n} s_i w_i > \theta$, the document d belong to this class. θ denotes the threshold. Winnow is a mistake-driven algorithm, only when the output and goal is inconsistent, the weight factor will be adjusted, and the adjustment process is following:

(1) If $\sum_{i=1}^{n} s_i w_i > \theta$, but the document do not belong to the predetermined class, s_i should be reduced and it is calculated by $s_i=\alpha s_i(0<\alpha<1)$(the weight s_i not equal to zero), until $\sum_{i=1}^{n} s_i w_i < \theta$;

(2) If $\sum_{i=1}^{n} s_i w_i < \theta$, but the document do not belong to the predetermined class, s_i should be increased and it is calculated by $s_i=\beta s_i(\beta>1)$ (the weight s_i not equal to zero), until $\sum_{i=1}^{n} s_i w_i > \theta$.

3.4.5 C4.5: C4.5 is a well-known classifier, it is based on ID3 algorithm [29] and improved by Quinlan(1993)[28]. C4.5 is a decision tree algorithm which is on the basis of information entropy. To avoid the favoritism from object which is selected by information gain and have more values, C4.5 uses the information gain ratio to select attribute node. The construction of C4.5 contains building tree and pruning tree.

The process of construction of C4.5 is following: (1) Calculating the information expected value I of data classification from training set. Assuming data set S has n training samples, and S is classified for m categories $\{S_1,S_2,\ldots,S_m\}$, the number of samples for every class is n_k. $p_k=\frac{n_k}{n}$ is used for computing the probability of the appearance of S_k. The information entropy or information expected value of m classes is calculated by $I=-\sum_{k=1}^{m} p_k \log_2 p_k$.(2) Computing the information expected value of attribute A when A equal to a_j, $I(A=a_j)$,j=1,2,…m. S is classified for v subsets $\{D_1,D_2,\ldots,D_v\}$, d_j denotes the number of samples from subset D_j which contains the samples that $A=a_j$. d_{kj} denotes the number of samples which belong to D_i and S_k concurrently, and the probability of samples belonging to the k class is $p_{kj}=\frac{d_{kj}}{d_j}$, and the information expected value of subset D_j is $I(A=a_j)=-\sum_{k=1}^{m} p_{kj} \log_2 p_{kj}$. When the probability of samples $A=a_j$ is $p_j=\frac{d_j}{n}$, the entropy of A is $Entropy(A)=\sum_{i=1}^{m} p_j I(A=a_j)$, it means the information expected value of attribute A to divide the current sample set. Gain(A) indicates the information gain of attribute A, namely the information Which A provide for classification, then A obtains the information gain from classifying the current sample is $Gain(A)=I-Entropy(A)$. (3) The information gain ratio of A is calculated by the equation $GainRatio(A)=\frac{Gain(A)}{Entropy(A)}$. On the basis of the different attribute value of pitch point, it construct different branch of decision tree, and divide data to different subsets. For every subset of branch, through recursive fashion to select attribute which has the largest information gain ratio, and the attribute as the decided principle for current pitch point, untile the data of leaf node belong to same class. With the success of construct decision tree, the decision rule can be obtained.

However, the initial decision tree has many branch, it will lead to overfitting. Hence, tree pruning is necessary. In general, there are two methods for tree pruning, the before pruning and the post-pruning. After pruning all the tree will be the candidates for the final decision tree. Using test data to test the result of classification, the decision tree with minimum error rate is reserved.

4. Performance Measures

For this paper, the index to evaluate the experiment result is similar to text classification, they are Accuracy and Precision. Assuming a denotes the number of comments which were correctly assigned to positive; b denotes the number of comments which were incorrectly

assigned to negative; c denotes the number of comments which were incorrectly assigned to positive; d denotes the number of comments which were correctly assigned to negative;

Two methods calculated by the following formulation:

$$\text{Accuracy} = \frac{\text{a}+\text{d}}{\text{a}+\text{b}+\text{c}+\text{d}} \cdot \text{Precision}(pos) = \frac{a}{a+b}, \text{Precision}(neg) = \frac{d}{c+d} \cdot$$

5. Results and Discussion

To complete experiment, we adopt our own implementation for text preprocessing. On the basis of text preprocessing, McCallum's Mallet toolkit [30] implementation of naive Bayes classifier, maximum entropy classifier, Winnow classifier, C4.5 classifier and Chang's LIBSVM [31] implementation of a Support Vector Machine classifier are used for classification. In the process of experiment, using DF to select features firstly and found that there were top 156 features appearing at least five times in our training set. Thus, this paper adopts different feature selection methods to select same 156 features which are convenient to comparing the different situations in the same level of feature numbers.

5.1. Experiment Results of DF Feature Selection

Table 1 shows the experiment result of DF feature selection with different classifiers and feature representation methods to sentiment classification of news comments. Table 1 shows that the accuracies of SVM and ME are all lager than 80% for all feature representation methods. In the all experiments, NB with TF achieves the top accuracy 86.55%, ME with TF-IDF achieves its highest accuracy 85.52%; The minimum accuracy which is achieved by C4.5 with Presence is 73.45%. For all classifiers, with different feature representation methods, the descending order of highest accuracy is NB>ME>SVM>Winnow>C4.5.

Table 1. **Experiment Results of DF Feature Selection**

	Presence(%)			TF(%)			TF-IDF(%)		
	Acc	Pre (pos)	Pre (neg)	Acc	Pre (pos)	Pre (neg)	Acc	Pre (pos)	Pre (neg)
SVM	83.16	84.21	82.01	82.13	81.58	82.73	81.10	78.95	83.45
Winnow	75.86	89.06	64.89	80.00	92.10	66.67	73.79	97.26	50.00
C4.5	73.45	68.97	77.93	77.59	77.40	77.78	78.62	79.61	77.54
NB	82.76	83.02	82.44	86.55	88.41	84.13	79.66	81.05	78.10
ME	84.83	82.76	86.90	80.69	82.14	78.69	85.52	85.26	85.82

5.2. Experiment Results of IG Feature Selection

Table 2 shows the experiment result of IG feature selection with different classifiers and feature representation methods to sentiment classification of news comments. Table 2 shows that for all feature representation methods, the accuracy of SVM, NB and ME are all lager than 80%; On the contrary, the accuracies of Winnow and C4.5 are all less than 80%. In the all experiments, ME with TF-IDF achieves the top accuracy 85.52%, SVM with TF achieves its highest accuracy 84.98%; The minimum accuracy which is achieved by C4.5 with TF-IDF is 71.72%. For all classifiers, with different feature representation methods, the descending order of highest accuracy is ME>SVM>NB>Winnow>C4.5.

Table 2. Experiment Results of IG Feature Selection

	Presence (%)			TF (%)			TF-IDF (%)		
	Acc	Pre (pos)	Pre (neg)	Acc	Pre (pos)	Pre (neg)	Acc	Pre (pos)	Pre (neg)
SVM	84.88	82.89	87.05	84.98	86.84	82.73	83.16	82.89	83.45

Winnow	74.14	88.96	55.12	78.62	84.77	71.94	79.66	88.20	68.99
C4.5	75.86	88.20	60.47	74.83	69.54	80.58	71.72	85.31	58.50
NB	82.76	88.39	76.30	80.69	87.25	73.76	81.03	84.62	76.87
ME	84.14	82.90	85.51	84.48	84.08	84.96	85.52	84.11	87.05

5.3. Experiment Results of CHI Feature Selection

Table 3 shows the experiment result of CHI feature selection with different classifiers and feature representation methods to sentiment classification of news comments. Table 3 shows that for all feature representation methods, except C4.5 classifiers, accuracies achieved by other classifiers are all lager than 80%. In the all experiments, SVM with Presence achieves the top accuracy 86.94%; The minimum accuracy which is achieved by C4.5 with Presence is 74.83%. For all classifiers, with different feature representation methods, the descending order of highest accuracy is SVM>ME>NB>Winnow>C4.5.

Table 3. Experiment Results of CHI Feature Selection

	Presence (%)			TF (%)			TF-IDF (%)		
	Acc	Pre (pos)	Pre (neg)	Acc	Pre (pos)	Pre (neg)	Acc	Pre (pos)	Pre (neg)
SVM	86.94	86.18	87.77	84.88	86.18	83.45	82.81	81.58	84.17
Winnow	80.69	82.64	78.05	82.41	91.18	70.00	82.41	87.66	76.47
C4.5	74.83	85.97	58.82	77.93	87.82	66.41	76.20	68.10	83.89
NB	83.10	94.52	71.53	83.45	88.05	77.86	85.17	90.51	78.79
ME	86.55	88.30	84.83	83.10	81.43	84.67	82.06	84.66	79.59

5.4. Experiment Results of MI Feature Selection

Table 4 shows the experiment result of MI feature selection with different classifiers and feature representation methods to sentiment classification of news comments. Table 4 shows that for all feature representation methods, SVM with TF and ME with Presence achieve the accuracy 70.45% and 70.34% respectively, the remainder accuracies are all less than 70%. In addition, the accuracy of C4.5 are all less than 60% and the minimum accuracy is 50.69%. For all classifiers, with different feature representation methods, the descending order of highest accuracy is SVM>ME>Winnow>NB>C4.5.

Table 4. Experiment Results of MI Feature Selection

	Presence (%)			TF (%)			TF-IDF (%)		
	Acc	Pre (pos)	Pre (neg)	Acc	Pre (pos)	Pre (neg)	Acc	Pre (pos)	Pre (neg)
SVM	53.26	100	2.16	70.45	80.26	59.71	67.35	75.66	58.27
Winnow	68.62	80.95	55.95	68.28	81.38	55.17	65.86	77.22	52.27
C4.5	59.31	21.85	99	50.69	99.31	2	56.55	17.21	99.3
NB	63.10	41.83	86.86	66.55	47.06	88.32	63.10	44.24	88
ME	70.34	76.77	62.96	69.66	78.42	61.59	67.59	76.14	54.39

5.5. Analysis of Experiment Results

5.5.1. Feature Selection Methods:

(1) As is shown in Table 1-4, using different feature selection methods to sentiment classification of news comments, the accuracy is different for every classifier with the same feature representation. Table 5 shows mean value of accuracies which are achieved by different classifiers with every feature selection method. The statistic result shows that, with different feature selection, SVM with CHI and IG can reach the best accuracy; For

Winnow classifier, CHI is the best feature selection method; The effect of C4.5 with DF and CHI is better than C4.5 with other feature selection methods; For NB classifier, comparing the mean value of accuracies, the highest accuracy was achieved by NB with CHI; The statistic result of ME indicates that ME with IG yields the top accuracy. The statistic data denotes that, the accuracies of SVM, NB, ME with CHI, DF and IG are all larger than 80%. Observing the mean value of accuracies of four feature selection methods, CHI achieves the highest accuracy (82.17%), the accuracy of IG(80.43%) is slightly higher than DF(80.38%). The gap of accuracies between MI and other feature selection methods is lager than 16%.

Table 5. The Mean Value of Accuracies of Different Classifiers with Four Feature Selection Methods

	CHI (%)	DF (%)	IG (%)	MI (%)	Average (%)
SVM	84.88	82.13	84.34	63.69	78.76
Winnow	81.84	76.55	77.47	67.59	75.86
C4.5	76.32	76.55	74.14	55.52	70.63
NB	83.91	82.99	81.49	64.25	78.16
ME	83.91	83.69	84.71	69.20	80.38
Average (%)	82.17	80.38	80.43	64.05	

(2) Table 1-4 show that for different classifiers with four feature selection methods, the positive precision and negative precision are different. Table 6 displays the sum of absolute D-value between negative precision and positive precision. Five classifiers with DF have the minimum sum of absolute D-value between negative precision and positive precision, less than the value of classifiers with CHI and IG. The largest sum of absolute D-value between negative precision and positive precision is achieved by classifiers with MI, and the value (6.5092) is triple of the value from classifiers with other feature selection methods. In addition, using MI feature selection method, there are many extreme results of positive comments and negative comments, such as, for SVM, the positive accuracy is 100% and the negative accuracy is 2.16%. Table 2-5 denote that for most classifiers, the positive precision is larger than negative precision. Through observing the news comments, we find that features for positive comments is more obvious than features for negative comments; moreover, many negative comments use objective expression or sarcasm to express sentiment, hence, the sentiment classification is more difficult.

Table 6. The Sum of Absolute D-value between Negative Precision and Positive Precision

CHI	DF	IG	MI
1.6488	1.2796	1.8005	6.5092

Through the above-mentioned analysis, except MI feature selection method, the results of classifiers with DF, CHI, IG and various feature representation methods are better, and the gap between DF, IG and CHI is small. However, CHI is perhaps the better feature selection method for news comments. Because of the theory of MI, its effect is worst. For MI, when the conditional probability of features is equal, rare features will have higher MI value than the common features, it means that more rare features are selected by MI. However, for most news comments, sentiment is expressed by frequent sentiment words, and the effect of frequent sentiment words is better than low-frequent sentiment words. For DF, IG and CHI, more common features are selected, and the relevance between DF value, IG value and CHI value is strong. Thus, DF, IG and CHI are suitable for news comments and have the similar effect.

5.5.2. Feature Representation Methods: (1)

(1) Table 7 shows the sum of accuracies for classifiers with different feature selection methods and feature feature representation methods. The result indicates that the effect of classifiers with every feature selection method (DF or IG, CHI, MI) and different feature representation methods (Presence, TF, TF-IDF) is similar, and the accuracies of DF, IG and CHI are high. Comparing the mean value of accuracies, the value of TF (3.8699) is slight higher than the value of TF-IDF(3.8222) and Presence(3.8215).

Table 7. The Sum of Accuracies for Different Feature Selection Methods and Feature Representation Methods

	Presence	TF	TF-IDF
CHI	4.1211	4.1177	4.0866
DF	4.0006	4.0696	3.9869
IG	4.0178	4.0360	4.0109
MI	3.1464	3.2562	3.2045
Average	3.8215	3.8699	3.8222

(2) Table 8 shows the sum of absolute D-value between negative precision and positive precision. TF has the minimum sum of absolute D-value between negative precision and positive precision, and slightly less than TF-IDF. Compared with TF and TF-IDF, the value of Presence is higher.

Table 8. The Absolute D-value between Positive Precision and Negative Precision for Different Feature Representation Methods

Presence	TF	TF-IDF
4.349309	3.3797084	3.509046

Summarize the above-mentioned analysis, a descending order of the performance of three feature representation methods is TF>TF-IDF>Presence. However, the gap between TF, TF-IDF and Presence is small. The reason is that texts have a characteristic that the same feature usually appear one time in the same text, a feature appearing more than one time is few. Thus, with same feature selection method, the text represented by three feature representation methods is similar, and the difference of result of three feature representation methods is small.

5.5.3. Classifiers:

(1) Observing table 6, for classifiers with different feature selection methods, the mean value of accuracies is different. ME produces the best mean value of accuracies, which is about two percent larger than SVM and NB, five percent lager than Winnow, and ten percent larger than C4.5, the mean value of accuracies of SVM is slightly higher than NB.

(2) Table 2-5 show that the absolute D-value between positive precision and negative precision is different for classifiers. Table 9 displays the sum of absolute D-value between negative precision and positive precision. We find a descending order of the performances of five classifiers is ME>SVM>NB>Winnow>C4.5. Table 2-5 also show that C4.5 has the biggest gap between negative precision and positive precision.

Table 9. The Absolute D-value between Positive Precision and Negative Precision for Different Classifiers

SVM	Winnow	C4.5	NB	ME
1.5937	2.7170	3.9789	2.1608	0.7877

Combining accuracy, positive precision, negative precision and analyzing the experiment results of sentiment classification of news comments, the descending order of performance of classifiers is ME>SVM>NB>Winnow>C4.5. In a word, ME is the best classifiers for sentiment classification of news comments. Although the effect of SVM and NB is worse than ME, they can be used for sentiment classification of news comments. Winnow and C4.5 are not suitable for sentiment classification of news comments.

In conclusion, DF, IG and CHI can be used for sentiment classification of news comments, and the effect of DF, IG and CHI is similar. Hence, when choosing DF, IG or CHI as the feature selection method, the difference of results comes from classifiers themselves. With the improvement of classifiers, the accuracy of sentiment classification can be improved.

6. Conclusions

This paper compares the feature selection methods, feature representation methods and classifiers of sentiment classification of news comments. The focus is on improvement of sentiment classification of news comments. We find that DF, IG and CHI are effective feature selection methods for every classifiers. Classifiers with DF, IG and CHI achieve different accuracies, however the difference is small. Comprehensive assessment, CHI is the better feature selection method. Compared with DF, IG and CHI, the effect of MI is worst. Thus, MI is not suitable for sentiment classification of news comments. For all classifiers with DF, IG and CHI, TF is slightly better than other two feature representation methods, the gap between three feature weighting methods is small. The reason is that this paper focuses on short document (news comments), the same feature usually appears once in the same document, and the features appearing more than one time in same document are rare. Thus, for three feature representation methods, the text vector of the same text is similar. The experiment result shows that machine learning perform quite well in the domain of sentiment classification of news comments. Comparing five classifiers, SVM, NB and ME are suitable for sentiment classification of news comments, and ME is the best classifier. For the effect of sentiment classification of news comments, combining feature selection methods, feature representation methods, different classifiers and analyzing the experiment result, we find that the influence coming from feature selection methods and feature representation methods is small. Enhancing the classifiers is necessary for improving the accuracy of sentiment classification by a large margin.

This research has some value of practical application and guidance of sentiment classification of short documents. On the basis of this paper, the future researches will focus on analyzing the binary sentiment classification and multi-level sentiment classification of other short documents, and will explore how to enhance classifiers to improve the accuracy and balance of classification.

Acknowledgments

This work was mainly supported by National Natural Science Foundation of China (71363038), Natural Science Foundation of Inner Mongolia, China (2012MS1008, 2013MS1009), Scientific Research Project of Colleges and Universities in Inner Mongolia, China (NJSZ12047).

References

[1] B. Liu, "Sentiment analysis and opinion mining [J]", Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, (**2012**), pp. 1-167.
[2] C. Dellarocas, X. M. Zhang and N. F. Awad, "Exploring the value of online product reviews in forecasting sales: The case of motion pictures [J]", Journal of Interactive marketing, vol. 21, no. 4, (**2007**), pp. 23-45.
[3] B. Pang and L. Lee, "Opinion mining and sentiment analysis [J]", Foundations and trends in information retrieval, vol. 2, nos. 1-2, (**2008**), pp. 1-135.

[4] T. Mullen and N. Collier, "Sentiment Analysis using Support Vector Machines with Diverse Information Sources [C]", EMNLP, vol. 4, (**2004**), pp. 412-418.
[5] T. Wilson, J. Wiebe and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis [C]", Proceedings of the conference on human language technology and empirical methods in natural language processing, Association for Computational Linguistics, (**2005**), pp. 347-354.
[6] H. Cui, V. Mittal and M. Datar, "Comparative experiments on sentiment classification for online product reviews [C] ", AAAI, vol. 6, (**2006**), pp. 1265-1270.
[7] K. Dave, S. Lawrence and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews [C]", Proceedings of the 12th international conference on World Wide Web. ACM, (**2003**), pp. 519-528.
[8] A. Devitt and K. Ahmad, "Sentiment polarity identification in financial news: A cohesion-based approach [C]", ACL, (**2007**).
[9] G. H. John, R. Kohavi and K. Pfleger, "Irrelevant Features and the Subset Selection Problem [C]", ICML, vol. 94, (1994), pp. 121-129.
[10] J. Chen, H. Huang, S. Tian and Y. Qu, "Feature selection for text classification with Naïve Bayes [J]", Expert Systems with Applications, vol. 36, no. 3, (**2009**), pp. 5432-5435.
[11] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization [C]", ICML, vol. 97, (**1997**), pp. 412-420.
[12] M. Rogati and Y. Yang, "High-performing feature selection for text classification [C]", Proceedings of the eleventh international conference on Information and knowledge management, ACM, (**2002**), pp. 659-661.
[13] G. Forman, "An extensive empirical study of feature selection metrics for text classification [J]", The Journal of machine learning research, vol. 3, (**2003**), pp. 1289-1305.
[14] W. Shang, H. Huang, H. Zhu, Y. Lin, Y. Qu and Z. Wang, "A novel feature selection algorithm for text categorization", Expert Systems with Applications, vol. 33, no. 1, (**2007**) July, pp. 1-5, ISSN 0957-4174.
[15] J. Yao, H. Wang and P. Yin, "Sentiment feature identification from Chinese online reviews [M]", Advances in Information Technology and Education, Springer Berlin Heidelberg, (**2011**), pp. 315-322.
[16] H. Wang, P. Yin, J. Yao, and J. N. Liu, "Text feature selection for sentiment classification of Chinese online reviews [J]", Journal of Experimental and Theoretical Artificial Intelligence, vol. 25, no. 4, (**2013**), pp. 425-439.
[17] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques [C]", Proceedings of the ACL-02 conference on Empirical methods in natural language processing, Association for Computational Linguistics, vol. 10, (**2002**), pp. 79-86.
[18] Z. Zhang, Q. Ye, Z. Zhang and Y. Li Y, "Sentiment classification of Internet restaurant reviews written in Cantonese [J]", Expert Systems with Applications, vol. 38, no. 6, (**2011**), pp. 7674-7682.
[19] B. Yu, "An evaluation of text classification methods for literary study [J]", Literary and Linguistic Computing, vol. 23, no. 3, (**2008**), pp. 327-343.
[20] Q. Ye, Z. Zhang and R. Law, "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches [J]", Expert Systems with Applications, vol. 36, no. 3, (**2009**), pp. 6527-6535.
[21] S. Tan and J. Zhang, "An empirical study of sentiment analysis for Chinese documents [J]", Expert Systems with Applications, vol. 34, no. 4, (**2008**), pp. 2622-2629.
[22] E. Gabrilovich and S. Markovitch, "Text categorization with many redundant features: Using aggressive feature selection to make SVMs competitive with C4. 5 [C]", Proceedings of the twenty-first international conference on Machine learning, ACM, (**2004**), vol. 41.
[23] A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification [C]", AAAI-98 workshop on learning for text categorization, vol. 752, (**1998**), pp. 41-48.
[24] A. L. Berger, V. J. D. Pietra and S. A. D. Pietra, "A maximum entropy approach to natural language processing [J]", Computational linguistics, vol. 22, no. 1, (**1996**), pp. 39-71.
[25] S. F. Chen and R. Rosenfeld, "A survey of smoothing techniques for ME models [J]", Speech and Audio Processing, IEEE Transactions on, vol. 8, no. 1, (**2000**), pp. 37-50.
[26] K. Nigam, J. Lafferty and A. McCallum, "Using maximum entropy for text classification [C]", IJCAI-99 workshop on machine learning for information filtering, vol. 1, (**1999**), pp. 61-67.
[27] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features [M]", Springer Berlin Heidelberg, (**1998**).
[28] J. R. Quinlan, "C4. 5: programs for machine learning [M]", Morgan kaufmann, (**1993**).
[29] J. R. Quinlan, "Induction of decision trees [J]", Machine learning, vol. 1, no. 1, (**1986**), pp. 81-106.
[30] McCallum, Andrew Kachites, "MALLET: A Machine Learning for Language Toolkit", http://mallet.cs.umass.edu, (**2002**).
[31] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines", ACM Transactions on Intelligent Systems and Technology, vol. 2, no. 27, pp. 1--27:27, (2011), Software available at http://www.csie.ntu.edu.tw/ ~cjlin/libsvm

