Expert Systems with Applications 38 (2011) 8696–8702

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A feature selection method based on improved fisher's discriminant ratio for text sentiment classification

Suge Wang [a,b,*], Deyu Li [b,c], Xiaolei Song [a], Yingjie Wei [d], Hongxia Li [a]

[a] School of Computer and Information Technology, Shanxi University, Taiyuan, 030006 Shanxi, China
[b] Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Taiyuan, 030006 Shanxi, China
[c] School of Mathematics Science, Shanxi University, Taiyuan, 030006 Shanxi, China
[d] Science Press, 100717 Beijing, China

ARTICLE INFO

Keywords:
Fisher's discriminant ratio
Feature selection
Text sentiment classification
Support vector machine

ABSTRACT

Owing to its openness, virtualization and sharing criterion, the Internet has been rapidly becoming a platform for people to express their opinion, attitude, feeling and emotion. As the subjectivity texts are often too many for people to go through, how to automatically classify them into different sentiment orientation categories (e.g. positive/negative) has become an important research problem. In this paper, based on Fisher's discriminant ratio, an effective feature selection method is proposed for subjectivity text sentiment classification. In order to validate the proposed method, we compared it with the method based on Information Gain while Support Vector Machine is adopted as the classifier. Two experiments are conducted by combining different feature selection methods with two kinds of candidate feature sets. Under 2739 subjectivity documents of COAE2008s and 1006 car-related subjectivity documents, the experimental results indicate that the Fisher's discriminant ratio based on word frequency estimation has the best performance respectively with accuracy 86.61% and 82.80% under two corpus while the candidate features are the words which appear in both positive and negative texts.

1. Introduction

With rapid development of web technology, the Internet has become a very important source from which more and more people obtain information. At the same time, it is also rapidly becoming a platform for people to express their opinion, attitude, feeling and emotion. Facing with promptly increasing reviews on the Web, it has been a great challenge for information science and technology that how people effectively organize and process document data to obtain the latest information to meet with particular needs and distinguish useful and worthless information. Text sentiment classification is aim to automatically judge what sentiment orientation, the positive ('thumbs up') or negative ('thumbs down'), a subjective text is with by mining and analyzing the subjective information in the text, such as standpoint, view, attitude, mood and so on. An automatically text sentiment classification can be widely applied to many fields. Firstly, it could help users quickly to classify and organize on-line reviews of goods and services, political commentaries, etc. Secondly, as the reviews are often too many for customers to go through, so an automatically text sentiment classifier may be very helpful to a customer to rapidly know the review orientations (e.g. positive/negative) about some product for customers' decision making online or offline. Thirdly, it could also be used to filter email and other messages. Finally, it could be used to public opinion analysis, question–answer system and text summarization. However, unlike structured data, the subjectivity texts on the Web, such as those on BBS, Blogs or forum websites are often non-structured or semi-structured. Consequently, feature selection is a crucial problem to the non-structured or semi-structured data classification. Although there has been a recent surge of interest in text sentiment classification, the state-of-the-art techniques for text sentiment classification are much less mature than those for text topic classification. This is partially attributed to the fact that topics are always represented by keywords objectively and explicitly while the topic sentiments are expressed by a subtle manner. Furthermore, the text sentiments are hidden in a large of subjective information in the text, such as standpoint, view, attitude, mood and so on. Therefore, the text sentiment classification requires deeper analyzing and understanding of textual statement information and thus is more challenging. In recent years, by employing some machine learning techniques (e.g. Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machine (SVM)), many researches have been conducted on English text sentiment classification (Chaovalit & Zhou, 2005; Kennedy & Inkpen, 2005; Michael, 2004; Pang, Lee, & Vaithyanathan, 2002; Tony & Nigel, 2004;

* Corresponding author at: School of Computer and Information Technology, Shanxi University, Taiyuan, 030006 Shanxi, China. Tel.: +86 351 7010555.
E-mail addresses: wsg@sxu.edu.cn (S. Wang), lidy@sxu.edu.cn (D. Li).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.01.077

Turney & Littman, 2002; Turney & Littman, 2003) and on Chinese text sentiment classification (Tan & Zhang, 2008; Wang, Wei, Zhang, Li, & Li, 2007; Ye, Lin, & Li, 2005; Ye, Zhang, & Rob, 2009).

One major particularity or difficulty of the text sentiment classification problem is the high dimensionality of the features used to describe texts, which raises big hurdles in applying many sophisticated learning algorithms to text sentiment classification. The aim of feature selection methods is to obtain a reduction of the original feature set by removing some features that are considered irrelevant for text sentiment classification to yield improved classification accuracy and decrease the running time of learning algorithms (Ahmed, Chen, & Salem, 2008; Tan & Zhang, 2008; Wang et al., 2007; Yi, Nasukawa, & Bunescu, 2003).

In this paper, from the viewpoint of the contribution of a candidate feature to distinguishing text sort, a kind of effective feature selection method based on improved Fisher's discriminant ratio is proposed for text sentiment classification. By considering two kinds of probability estimations, i.e., Boolean value and word frequence, and two kinds of candidate feature manners, four kinds of feature selecting techniques are then proposed. By using SVM to construct the classifier, the experiments are conducted under two corpus, 2739 subjectivity documents of COAE2008s and 1006 car-related subjectivity documents. The experimental results indicate that the proposed method is effective for review text sentiment classification.

The remainder of this paper is organized as follows: Section 2 introduces the related works. Section 3 introduces the feature selection method based on Information Gain (IG). Section 4 introduces SVM classifier. Section 5 elaborates the proposed feature selection method. Section 6 shows the experimental results. Section 7 concludes the paper.

2. Related work

Sentiment analysis is concerned with analysis of direction-based text, that is, text containing opinions and sentiment (emotions) (Ahmed et al., 2008). We focus on sentiment classification studies which attempt to determine whether a text is contains positive or negative sentiments. Sentiment classification has several important characteristics, including various tasks, features selection, classification techniques, and application domains.

2.1. Features selection

Feature selection is an important part of optimizing the performance of a classifier by reducing the feature vector to a size that does not exceed the number of training cases as a starting point. Further selection of vector size can lead to more improvements if the features are noisy or redundant. Tan and Zhang (2008) presented an empirical study of sentiment categorization on Chinese documents. In their work four feature selection methods, IG, Mutual Information (MI), CHI and Document Frequency (DF) were adopted. The experimental results indicate that IG performs the best for selecting the sentimental terms. Riloff, Patwardhan, and Wiebe (2006) used a subsumption hierarchy to formally define different types of lexical features and their relationship to one another, both in terms of representational coverage and performance. They show that the reduced feature set can improve the performance on three opinion classification tasks, especially when combined with traditional feature selection approachs. Wiebe, Wilson, Bruce, Bell, and Martin (2004) used collocation technique where certain parts of fixed-word n-grams were replaced with general word tags, thereby also creating n-gram phrase patterns. Nasukawa et al. (2003) presented sentiment analyzer (SA) that extracts sentiment (or opinion) about a subject from online text documents. Instead of classifying the sentiment of an entire document about a subject, SA detects all references to the given subject, and determines sentiment in each of the references using natural language processing techniques. Wang et al. (2007) presented a hybrid method for feature selection based on the category distinguishing capability of feature words and IG. Pang et al. (2002) used syntactic (unigrams, bigrams, unigrams + POS, adjectives, and unigrams + position), limited consideration to unigrams appearing at least four times in their 1400-document corpus, and the bigrams occurring most often in the same dataset (the selected bigrams all occurred at least seven times). Hatzivassiloglou and Wiebe (2000) showed that automatically detected gradable adjectives are useful features for sentiment classification, while Wiebe (2000) introduced lexical features in addition to the presence/absence of syntactic categories. Yu and Hatzivassiloglou (2003) used words, bigrams, and trigrams, as well as the parts of speech as features in each sentence. Tong (2001) and Wilson et al. (2005) introduced manual or semiautomatic approaches for generating sentiment lexicons that uses an initial set of automatically generated terms which are manually filtered and coded with polarity and intensity information. The user-defined tags are incorporated to indicate whether certain phrases convey positive or negative sentiment. Riloff, Wiebe, and Wilson (2003) used semiautomatic lexicon generation tools to construct the sets of strong subjectivity, weak subjectivity, and objective nouns. Their approach outperformed the use of other features (e.g. bag-of-words) for objective classification. For the very noisy domain of customer feedback data, Gamon (2004) presented a feature reduction technique based on log likelihood ratio to select the important attributes from a large initial feature vectors.

2.2. Sentiment classification approaches

Text sentiment classification researches have fallen into two categories, i.e., machine learning techniques and score-based approaches. Machine learning techniques train a sentiment classifier based on the training documents which are represented by the selected features. The score-based approaches divide features into two classes, ''positive'' and ''negative'', and then count an overall positive/negative score for a document.

Many studies on sentiment classification have used machine learning algorithms, with SVM and NB being the most commonly used. SVM has been used extensively for movie reviews (Pang et al., 2002; Tan & Zhang, 2008; Wang et al., 2007; Wilson, Wiebe, & Hoffman, 2005), while NB has been applied to reviews and Web discourse (Pang et al., 2002). In comparisons, SVM has outperformed other classifiers such as NB, centroid classifier, K-nearest neighbor, winnow classifier (Pang et al., 2002; Tan & Zhang, 2008). Tan and Zhang (2008) have validated that SVM is the best classifier for Chinese text sentiment classification. So, in this study, we will use SVM classifier. Score-based approaches are typically used in conjunction with semantic features (Turney & Littman, 2003). These approaches generally classify message sentiments based on the total sum of comprised positive or negative sentiment features. All messages with a positive sum are assigned positive sentiment while negative messages are assigned to the negative-sentiment class (Turney & Littman, 2002; Turney, 2002; Turney & Littman, 2003). Score-based methods have also been used for affect analysis, where the affect features present within a message/document are scored based on their degree of intensity for a particular emotion class (Subasic & Huettner, 2001). Sentiment classification has been investigated in different domains such as movie reviews, product reviews, and customer feedback reviews (Gamon, 2004; Pang et al., 2002; Turney & Littman, 2003).

3. Information gain

IG is frequently employed as a term goodness criterion in the field of machine learning (Tan & Zhang, 2008; Yang & Pedersen, 1997). The literatures (Tan & Zhang, 2008; Yang & Pedersen, 1997) validate that IG can achieve very good results and is regarded as one of the most effective feature selection methods. It measures the number of bits of information obtained for category prediction by knowing the presence or absence of a term in a document.

Formally, for a term t_k,

$$\begin{aligned} IG(t_k) &= H(C) - H(C|t_k) \\ &= -\sum_{i=1}^{m} p(c_i)\log(p(c_i)) + p(t_k)\sum_{i=1}^{m} p(c_i|t_k)\log(p(c_i|t_k)) \\ &\quad + p(\bar{t}_k)\sum_{i=1}^{m} p(c_i|\bar{t}_k)\log(p(c_i|\bar{t}_k)) \\ &= \sum_{i=1}^{m}\left(p(c_i,t_k)\log\left(\frac{P(c_i,t_k)}{p(c_i)p(t_k)}\right) + p(c_i,\bar{t}_k)\log\left(\frac{P(c_i,\bar{t}_k)}{p(c_i)p(\bar{t}_k)}\right)\right) \end{aligned} \tag{1}$$

where $p(c_i)$ denotes the probability that category c_i occurs, $p(t_k)$ denotes the probability that term t_k occurs, $p(\bar{t}_k)$ denotes the probability that term t_k does not occurs, $p(c_i,t_k)$ denotes the joint probability of t_k and c_i, $p(c_i,\bar{t}_k)$ denotes the joint probability of $\bar{t}_k$ and c_i.

4. Classifier based on SVM

Up to now, it is verified that SVM possesses the best performance for the text sentiment classification problem (Gamon, 2004; Pang et al., 2002; Tan & Zhang, 2008). Therefore, to assess the effectiveness of feature selection methods, we adopt SVM to construct the classifier in the presented paper.

As a relatively new machine learning method, SVM developed by Vapnik (1995) embodies the VC- dimension theory and the structural risk minimization principle. It seeks a decision hyperplane to separate the training data points into two classes and makes decisions based on the support vectors that are selected as the only effective elements in the training set.

Suppose $(x_i,y_i)(i = 1,2,\ldots,n)$ is the set of samples, where $x_i \in R^d$ and $y_i \in \{-1,+1\}$ are the class labels of samples. The general form of a linear discrminant function in a $d-$dimension space can be expressed as $g(x) = w\cdot x + b$. The corresponding separation hyperplane equation can be written as

$$w\cdot x + b = 0 \tag{2}$$

Then, by normalization of the discrminant function, we can make all samples of two classes to satisfy the inequality $|g(x)| \geqslant 1$. In other words, the samples those are the nearest to the separation hyperplane meet $|g(x)| = 1$. This implies that the separation margin equals to $2/\|w\|$. Thus, maximizing the separation margin is equivalent to minimizing $\|w\|$. A separation hyperplane which can correctly separate two-class samples should satisfy

$$y_i[(w\cdot x_i) + b] - 1 \geqslant 0, \quad i = 1,\ldots,n \tag{3}$$

Therefore, within the smallest $\|w\|^2$, the separation hyperplane satisfying Formula (3) is called the optimal separation hyperplane. By using optimum theory, the optimal separation function is obtained as

$$f(x) = sign\{(w^*\cdot x) + b^*\} \tag{4}$$

where w^* and b^* are the solutions of w and b respectively.

5. Feature selection method based on improved fisher's discriminant ratio

Fisher linear discriminant is one of efficient approaches for dimension reduction in statistical pattern recognition (Webb Andrew, 2002). Its main idea can be briefly described as follows. Suppose that there are two kinds of sample points in a d-dimension data space. We hope to find a line in the original space such that the projective points on the line of the sample points can be separated as much as possible by some point on the line. In other words, the bigger the square of the difference between the means of two kinds projected sample points is and at the same time the smaller the within-class scatters are, the better the expected line is. More formally, construct the following function, so-called Fisher's discriminant ratio.

$$J_F(w) = \frac{(\bar{m}_1 - \bar{m}_2)^2}{\bar{S}_1^2 + \bar{S}_2^2} \tag{5}$$

where w is the direction vector of the expected line, $\bar{m}_i$ and $\bar{S}_i^2$ $(i = 1,2)$ are the mean and the within-class scatter of the ith class respectively. So the above idea is to find a w such that $J_F(w)$ achieve its maximum.

5.1. Improved Fisher's discriminant ratio

In fact, Fisher's discriminant ratio can be improved for evaluating the category distinguishing capability of a feature by replacing w with the feature. For this purpose, Fisher's discriminant ratio is reconstructed as follows.

$$F(t_k) = \frac{(E(t_k|P) - E(t_k|N))^2}{D(t_k|P) + D(t_k|N)} \tag{6}$$

where $E(t_k|P)$ and $E(t_k|N)$ are the conditional mean of the feature t_k with respect to the categories P and N respectively, $D(t_k|P)$ and $D(t_k|N)$ are the conditional variances of the feature t_k with respect to the categories P and N respectively.

It is obvious that $(E(t_k|P) - E(t_k|N))^2$ is the between-class scatter degree and $D(t_k|P) + D(t_k|N)$ is the sum of within-class scatter degrees.

In applications, the probabilities involved in Formula (6) can be estimated in different ways. In this paper, two kinds of methods respectively based on Boolean value and frequency are adopted for probability estimation.

5.1.1. Fisher's discriminant ratio based on Boolean value

Let $d_{P,i}$ $(i = 1,2,\ldots,m)$ and $d_{N,j}(j = 1,2,\ldots,n)$ denote the ith positive text and the jth negative text respectively. Random variables $d_{P,i}(t_k)$ and $d_{N,j}(t_k)$ are defined as follows.

$$d_{P,i}(t_k) = \begin{cases} 1, & \text{if } t_k \text{ occurs in } d_{P,i,} \\ 0, & \text{otherwise} \end{cases}$$

$$d_{N,j}(t_k) = \begin{cases} 1, & \text{if } t_k \text{ occurs in } d_{N,j,} \\ 0, & \text{otherwise} \end{cases}$$

Let $m_1(n_1)$ and $m_0(n_0)$ be the numbers of positive (negative) texts within feature t_k and without feature t_k respectively. Obviously, random variables $d_{P,i}(t_k)$ and $d_{N,j}(t_k)$ are depicted by the following distributions respectively.

$$P(d_{P,i}(t_k) = l) = m_l/m, l = 1,0, \quad P(d_{N,j}(t_k)) = l = n_l/n, l = 1,0.$$

It should be note that $d_{P,i}(t_k)$ $(i = 1,2,\ldots,m)$ are independent with the same distribution, and so is $d_{N,j}(t_k)$ $(j = 1,2,\ldots,m)$. $d_{P,i}(t_k)$ $(i = 1,2,\ldots,m)$ can be regarded as a sample with size m. Then the

conditional means and the conditional variances of the feature t_k with respect to the categories P and N in Formula (6) can be estimated by using Formulas (7)–(11).

$$E(t_k|P) = E\left(\frac{1}{m}\sum_{i=1}^{m} d_{P,i}(t_k)\right) = \frac{1}{m}\sum_{i=1}^{m} E(d_{P,i}(t_k)) = \frac{m_1}{m} \tag{7}$$

$$E(t_k|N) = E\left(\frac{1}{n}\sum_{j=1}^{n} d_{N,j}(t_k)\right) = \frac{n_1}{n} \tag{8}$$

$$D(t_k|P) = \frac{1}{m}\sum_{i=1}^{m}\left(d_{P,i}(t_k) - \frac{m_1}{m}\right)^2 \tag{9}$$

$$D(t_k|N) = \frac{1}{n}\sum_{j=1}^{n}\left(d_{N,j}(t_k) - \frac{n_1}{n}\right)^2 \tag{10}$$

Hence, we have that

$$F(t_k) = \frac{(E(t_k|P) - E(t_k|N))^2}{D(t_k|P) + D(t_k|N)} = \frac{(m_1 n - m n_1)^2}{mn^2\sum_{i=1}^{m}(d_{P,i}(t_k) - \frac{m_1}{m})^2 + nm^2\sum_{j=1}^{n}(d_{N,j}(t_k) - \frac{n_1}{n})^2} \tag{11}$$

Fisher's discriminant ratio based on Boolean value $F(t_k)$ is subsequently denoted by $F_B(t_k)$.

5.1.2. Fisher's discriminant ratio based on frequency

It is obvious that Fisher's discriminant ratio based on Boolean value dose not consider the appearing frequence of a feature in a certain text, but only consider whether or not it appears. In order to examine the influence of the frequence on the significance of a feature another probability estimation method based on frequence is adopted in Formula (6).

Let $v_{P,i}$ and $v_{N,j}$ be the word tokens of texts $d_{P,i}$ and $d_{N,j}$, $w_{P,i}(t_k)$ and $w_{N,j}(t_k)$ be the frequences of t_k appearing in $d_{P,i}$ and $d_{N,j}$ respectively.

Then the conditional means and the conditional variances of the feature t_k with respect to the categories P and N in Formula (6) can be estimated by using Formulas (12)–(15).

$$E(t_k|P) = \frac{1}{m}\sum_{i=1}^{m}\frac{w_{P,i}(t_k)}{v_{P,i}} \tag{12}$$

$$E(t_k|N) = \frac{1}{n}\sum_{j=1}^{n}\frac{w_{N,j}(t_k)}{v_{N,j}} \tag{13}$$

$$D(t_k|P) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{w_{P,i}(t_k)}{v_{P,i}} - E(t_k|P)\right)^2 \tag{14}$$

$$D(t_k|N) = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{w_{N,j}(t_k)}{v_{N,j}} - E(t_k|N)\right)^2 \tag{15}$$

Hence, we have that

$$F(t_k) = \frac{(E(t_k|P) - E(t_k|N))^2}{D(t_k|P) + D(t_k|N)} = \frac{\left(\frac{1}{m}\sum_{i=1}^{m}\frac{w_{P,i}(t_k)}{v_{P,i}} - \frac{1}{n}\sum_{j=1}^{n}\frac{w_{N,j}(t_k)}{v_{N,j}}\right)^2}{\frac{1}{m}\sum_{i=1}^{m}\left(\frac{w_{P,i}(t_k)}{v_{P,i}} - \frac{1}{m}\sum_{i=1}^{m}\frac{w_{P,i}(t_k)}{v_{P,i}}\right)^2 + \frac{1}{n}\sum_{j=1}^{n}\left(\frac{w_{N,j}(t_k)}{v_{N,j}} - \frac{1}{n}\sum_{j=1}^{n}\frac{w_{N,j}(t_k)}{v_{N,j}}\right)^2} = \frac{mn\left(n\sum_{i=1}^{m}\frac{w_{P,i}(t_k)}{v_{P,i}} - m\sum_{j=1}^{n}\frac{w_{N,j}(t_k)}{v_{N,j}}\right)^2}{n^2\sum_{i=1}^{m}\left(m\frac{w_{P,i}(t_k)}{v_{P,i}} - \sum_{i=1}^{m}\frac{w_{P,i}(t_k)}{v_{P,i}}\right)^2 + m^2\sum_{j=1}^{n}\left(n\frac{w_{N,j}(t_k)}{v_{N,j}} - \sum_{j=1}^{n}\frac{w_{N,j}(t_k)}{v_{N,j}}\right)^2} \tag{16}$$

Fisher's discriminant ratio $F(t_k)$ based on frequency is subsequently denoted by $F_F(t_k)$.

By simple deriving, we have the following proposition which depicts the relationship between two kinds of Fisher's discriminant ratios $F_B(t_k)$ and $F_F(t_k)$.

Proposition 1. *For a feature item t_k, if $\frac{w_{P,i}(t_k)}{v_{P,i}} = d_{P,i}(t_k)$ and $\frac{w_{N,j}(t_k)}{v_{N,j}} = d_{N,j}(t_k)$, then $F_F(t_k) = F_B(t_k)$.*

5.2. Process of feature selection

Step 1. Candidate feature set.
In order to compare the classification effects of features from the different regions. We design two kinds of word sets as the candidate feature sets. One of them denoted by U consists of all words in the text set. Another candidate feature set I contains all words which appear in both positive and negative texts.

Step 2. Features used in the classification model.
The idea of Fisher's discriminant ratio implies that it can be used as a significance measure of features for classification problem. The larger the value of Fisher's discriminant ratio of a feature is, the stronger the classification capability of the feature is. So we can compute the value of Fisher's discriminant ratio for every feature and rank them in descending order. And then choose the best features with a certain number.

6. Experiment

6.1. Experiment corpus and evaluation measures

Experiment corpus1: In order to examine the effect of the proposed feature selection method, we used the COAE2008s which is a subset of Chinese opinion analysis evaluation corpus (COAE2008). The COAE2008s corpus is subjectivity texts which has positive or negative sentiments orientation. The total size is 2739 documents that consist of many domains, such as movie, education, finance and economics, house, computer, mobile telephone etc. Each domain category contains positive and negative. There are 1525 positive and 1214 negative documents in this corpus.

Experiment corpus2: We collected 1006 Chinese review texts about 11 kinds of car trademarks published on http://www.xche.com.cn/baocar/ from January 2006 to March 2007. In the corpus there are 578 positive reviews and 428 negative reviews. The total reviews contain 1000 thousands words.

To evaluate the effectiveness of the proposed feature selection methods, three kinds of classical evaluation measures generally used in text classification, Precision, Recall and F_value are adopted in this paper. By PP (PN), RP (RN) and FP (FN) we denote Precision, Recall and F_value of positive (negative) subjectivity texts respectively. These evaluation measures can be calculated according to Table 1 and the following formulas respectively.

$$PP(precision) = \frac{a}{a+b}\times 100\%, RP(recall) = \frac{a}{a+c}\times 100\%, \quad FP(F\_value) = \frac{2\times RP\times PP}{RP+PP};$$

Table 1
Contingency table for performance evaluation.

Predict	Actual	
	Positive texts	Negative texts
positive texts	a	b
negative texts	c	d

$$PN(precision) = \frac{d}{c+d} \times 100\%, RN(recall) = \frac{d}{b+d} \times 100\%, \quad FN(F\_value) = \frac{2 \times RN \times PN}{RN + PN};$$

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} \times 100\%.$$

We performed the experiments in five-fold cross-validation.

6.2. Process of text sentiment classification

The whole experiment process is divided into training and testing parts.

Step 1. Segment preprocess of subjectivity texts.
Step 2. By using the methods introduced in Section 5, obtain the candidate feature sets, and then select the features for the classification model.
Step 3. Express each text in the form of feature weight vector. Here, feature weights are computed by using TFIDF (Yang & Pedersen, 1997).
Step 4. Train the support vector classification machine by using training data, and then obtain a classifier.
Step 5. Test the performance of the classifier by using testing data.

6.3. Experiment result and analysis

For convenience, a kind of simple symbol, Candidate Feature Set + Feature Selection Method, shows which candidate feature set and which feature selection method were adopted in an experiment. For example, U + FB stands for that the candidate feature set U and the feature significance measure $F_B(t_k)$ are adopted in an experiment.

Experiment 1: In order to exam the influence of feature dimension on classification performance 500, 1000 and 2000 features are selected from the candidate feature set I by using the methods in Section 5. The experiment result is shown in Table 2 and Fig. 1 in Experiment corpus2.

Table 2 and Fig. 1 show that almost all evaluation measures are best under 1000 feature dimension. In other words, a larger amount of features need not imply a good classification result. So all the feature dimensions in the succedent experiments are 1000.

Experiment 2: The aim of this experiment is to compare the effects of 8 kinds of feature selection approaches with 1000 dimension for text sentiment classification in Experiment corpus1. Here IG based on Boolean value is denoted by IGB and that based on frequency is denoted by IGF subsequently. FB and FF stand for $F_B(t_k)$ and $F_F(t_k)$ respectively. The text sentiment classification experiment results are shown in Table 3 using 8 kinds of feature selection methods.

From Table 3, Figs. 2–4, one can see that:

(1) Among two kinds of feature significance measures IG and Fisher's discriminant ratio for feature selection, the accuracy of Fisher's discriminant ratio is better.

Table 2
Classification effects of feature dimension with I + FF in Experiment corpus2.

Dimension	Measure						
	PP	RP	FP	PN	RN	FN	Accuracy
500	82.51	84.47	83.58	79.22	74.97	77.01	80.81
1000	83.04	88.52	85.49	83.58	75.06	78.69	82.80
2000	80.25	88.52	84.10	82.19	70.12	75.21	80.71

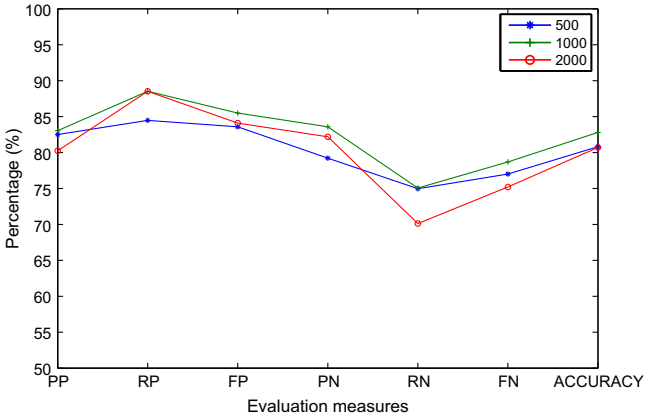

Fig. 1. Classification effects of feature dimension with I + FF in Experiment corpus2.

Table 3
Classification effects of feature selection methods in Experiment corpus1.

Method	PP (%)	RP (%)	FP (%)	PN (%)	RN (%)	FN (%)	Accuracy (%)
U + IGB	80.37	93.90	86.60	90.43	71.19	79.65	83.83
I + IGB	80.14	92.72	85.92	88.89	71.00	78.79	83.10
U + IGF	87.66	86.49	87.06	83.33	84.68	83.98	85.69
I + IGF	88.02	86.62	87.31	83.53	85.17	84.34	85.98
U + FB	79.95	93.69	86.27	90.03	70.47	79.03	83.39
I + FB	79.86	94.18	86.43	90.72	70.16	79.10	83.53
U + FF	87.81	88.07	87.93	85.04	84.61	84.81	86.53
I + FF	87.96	88.00	87.98	84.99	84.86	84.92	86.61

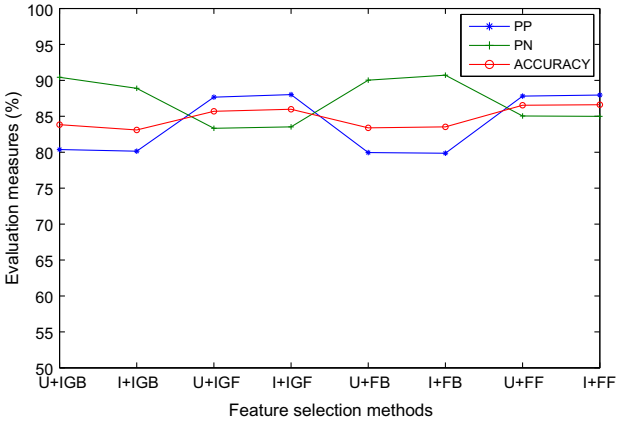

Fig. 2. Precision and accuracy of feature selection methods in Experiment corpus1.

(2) The performances of 8 kinds of variations of IG and Fisher's discriminant ratio rely on the candidate feature set to some extent. For Fisher's discriminant ratio based on frequency and Boolean value the accuracy of they are better when I is adopted as the candidate set.
(3) For 8 kinds of methods, the accuracy of Fisher's discriminant ratio based on frequency is best when the candidate feature set is I.
(4) Among all the 8 kinds of methods, the recall and F_value of positive documents are superior to negative documents, and the precision look quite noisy.

Experiment 3: In order to validate the effectiveness of the proposed methods for sentiment classification, the same experimental

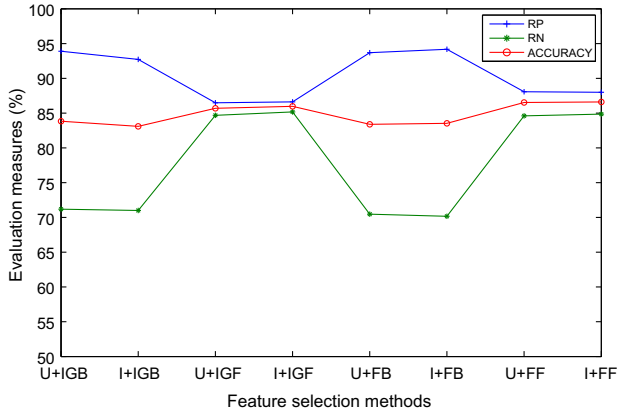

Fig. 3. Recall and accuracy of feature selection methods in Experiment corpus1.

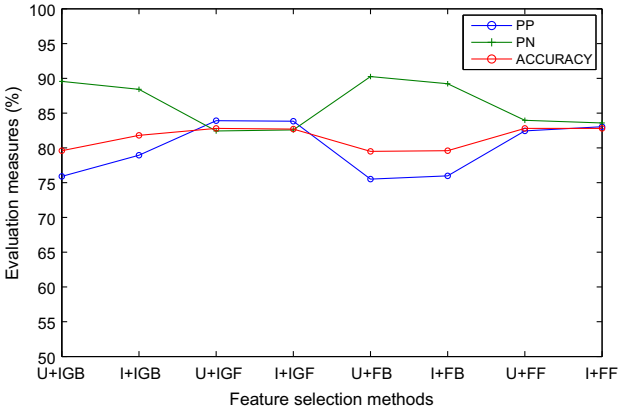

Fig. 5. Precision and accuracy of feature selection methods in Experiment corpus2.

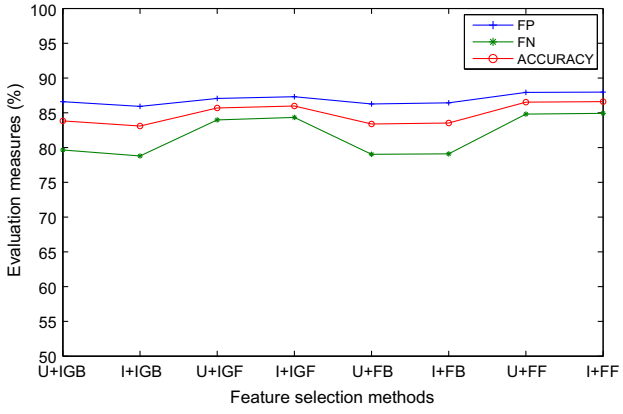

Fig. 4. F-Value and accuracy of feature selection methods in Experiment corpus1.

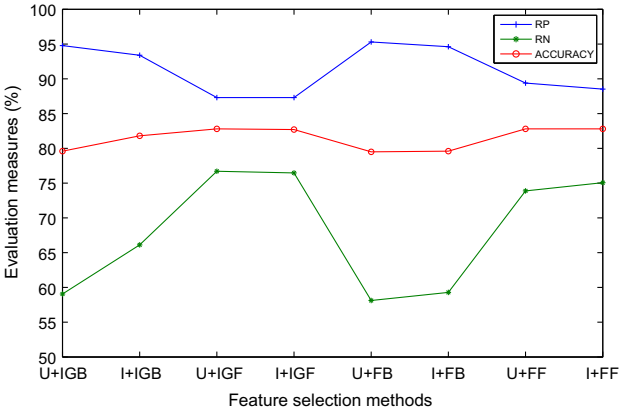

Fig. 6. Recall and accuracy of feature selection methods in Experiment corpus2.

Table 4
Classification effects of feature selection methods in Experiment corpus2.

Method	PP (%)	RP (%)	FP (%)	PN (%)	RN (%)	FN (%)	Accuracy (%)
U + IGB	75.90	94.78	84.26	89.56	59.06	70.93	79.60
I + IGB	78.94	93.39	85.51	88.42	66.12	75.46	81.80
U + IGF	83.91	87.30	85.33	82.43	76.71	79.03	82.80
I + IGF	83.84	87.30	85.22	82.59	76.47	78.88	82.70
U + FB	75.51	95.30	84.25	90.26	58.12	70.63	79.50
I + FB	75.98	94.61	84.23	89.23	59.29	71.03	79.60
U + FF	82.44	89.39	85.68	83.97	73.88	78.39	82.80
I + FF	83.04	88.52	85.49	83.58	75.06	78.69	82.80

settings in Experiment corpus2 will be used. The text sentiment classification experiment results are shown in Table 4.

From Table 4, Figs. 5–7, one can see that each of 8 kinds of feature selection methods has the analogous performance trend both in Experiment corpus1 and Experiment corpus2. The experimental results in Experiment corpus1 seem slightly better than that in Experiment corpus2.

Remarks: (1) The computing time cost of the feature selection process depends upon the size of the candidate feature set, it is advisable to design a smaller candidate feature set such as I in this paper. (2) For text sentiment classification problem many kinds of feature selection methods such as MI, IG, CHI and DF are compared in some literatures (Pang et al., 2002; Tan & Zhang, 2008). Among these methods IG is validated to be best in the past research works. However, the experiments in this paper shown that Fisher's discriminant ratio based on frequency is a better choice than IG for feature selection.

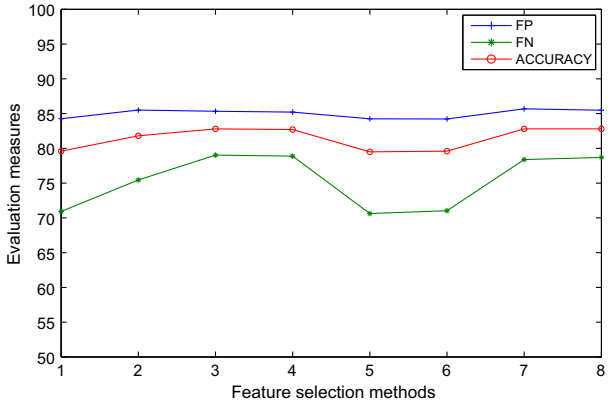

Fig. 7. F_Value and accuracy of feature selection methods in Experiment corpus2.

7. Conclusions

Text sentiment classification can be widely applied to text filtering, online tracking opinions, analysis of public opinion poll, and chat systems. However, compared with traditional subject classification, there are more factors that need to be considered

in text sentiment classification. Especially, the feature selection for text sentiment classification is more difficult. In this paper, under the granularity level of words, a new feature selection method based on Fisher's discriminant ratio is proposed. In order to validate the validity of the proposed method, we compared it with the typical method based on IG while support vector machine is adopted as the classifier. Two experiments are conducted by combining different feature selection methods with 2 kinds of candidate feature sets. The experiment results show that I + FF obtains the best classification effectiveness, its accuracy achieves 86.61% in Experiment corpus1. Our further research works will focus on establishing a sentiment knowledge base based on vocabulary, syntactic, semantic and ontology.

Acknowledgements

This work was supported by the National Natural Science Foundation (Nos.60875040 and 60970014), the Foundation of Doctoral Program Research of Ministry of Education of China (No.200801080006), Natural Science Foundation of Shanxi Province (No.2010011021-1), Shanxi Foundation of Tackling Key Problem in Science and Technology (No.051129) and the Star Program of Science and Technology Office of Taiyuan City (No.09121001).

References

Ahmed, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: feature selection for opinion classification in web forums. *ACM Transactions on Information Systems, 26*(3).

Chaovalit, P., Zhou, L. N. (2005). Movie Review mining: A comparison between supervised and unsupervised classification approaches. In *IEEE proceedings of the 38th Hawaii International Conference on System Sciences* (pp.1–9). Big Island, Hawaii.

Gamon, M. (2004). Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th International Conference on Computational Linguistics* (p. 841).

Hatzivassiloglou, V., Wiebe, J. (2000). Effects of adjective orientation and gradability on sentence subjectivity. In *International Conference on Computational Linguistics (COLING-2000)*.

Kennedy, A., Inkpen, D. (2005). Sentiment classification of movie and product reviews using contextual valence shifters. In *Workshop on the analysis of Informal and Formal Information Exchange during negotiations*.

Michael, G. (2004). Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings the 20th international conference on computational linguistics*.

Yi, J., Nasukawa, T., Bunescu, R., Niblack.W. (2003). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Proceedings of the 3rd IEEE International Conference on Data Mining* (pp. 427–434).

Pang, B., Lee, L., Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification Using Machine Learning Techniques. *EMNLP*'.

Riloff, E., Wiebe, J., Wilson, T. (2003). Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the 7th Conference on Natural Language Learning* (pp. 25–32). Edmonton, Canada

Riloff, E., Patwardhan, S., Wiebe, J. (2006). Feature subsumption for opinion analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 440–448). Sydney, Australia.

Subasic, P., & Huettner, A. (2001). Affect analysis of text using fuzzy semantic typing. *IEEE Transactions Fuzzy System, 9*(4), 483–496.

Tan, S. B., & Zhang, J. (2008). An Empirical study of sentiment analysis for chinese documents. *Expert Systems with Application, 34*(4), 2622–2629.

Tong, R. (2001). An operational system for detecting and tracking opinions in online discussion. In *Proceedings of the ACM SIGIR Workshop on Operational Text Classification* (pp. 1–6).

Tony, M., Nigel, C. (2004). Sentiment analysis using support vector machines with diverse information sources. In Dekang Lin & Dekai Wu (Eds.), *Proceedings of EMNLP-2004* (pp. 412–418). Barcelona, Spain.

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meetings of the Association for Computational Linguistics*, (pp. 417–424). Philadelphia, PA.

Turney, P. D., Littman, M. L. (2002) Unsupervised Learning of Semantic Orientation from A Hundred-billion-word Corpus. Technical Report EGB-1094, National Research Council Canada.

Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Tr Littman Transactions on Information Systems (TOIS), 21*(4), 315–346.

Wang, S. G., Wei, Y. J., Zhang, W., Li, D. Y., & Li, W. (2007). A hybrid method of feature selection for chinese text sentiment classification [C]. In *Proceedings of the 4th International Conference on Fuzzy Systems and Knowledge Discovery* (pp. 435–439). IEEE Computer Society.

Webb Andrew, R. (2002). *Statistical pattern recognition* (2nd ed.). John Wiley and Sons.

Wiebe, J. (2000). Learning Subjective Adjectives from Corpora. In *Proceedings 17th National Conference on Artificial Intelligence (AAAI-2000)*. Austin, Texas.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning subjective language. *Computational Linguistics, 30*(3), 277–308.

Wilson,T., Wiebe, J., Hoffman, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing* (pp. 347–354). British Columbia, Canada.

Yang, Y. M., & Pedersen, J. O. (1997). A Comparative study on feature selection in text categorization. *ICML*, 412–420.

Ye, Q., Lin, B., Li,Y. J. (2005). Sentiment classification for chinese reviews: A comparison between svm and semantic approaches. In *The 4th International Inference on Machine Learning and Cybernetics ICMLC*.

Ye, Q., Zhang, Z. Q., & Rob, L. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert System with Application, 36*(3), 6527–6535.

Yu, H., Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp.129–136).