5-26-2012

An Empirical Study of Sentiment Analysis for Chinese Microblogging

Zhiming Liu
School of Economics and Management, Beihang University, Beijing 100191, China

Lu Liu
School of Economics and Management, Beihang University, Beijing 100191, China

Hong Li
School of Economics and Management, Beihang University, Beijing 100191, China

Follow this and additional works at: http://aisel.aisnet.org/whiceb2011
An Empirical Study of Sentiment Analysis for Chinese Microblogging

Zhiming Liu*, Lu Liu, Hong Li
School of Economics and Management, Beihang University, Beijing 100191, China

Abstract: This paper used three machine learning algorithms, three kinds of feature selection methods and three feature weight methods to study the sentiment classification for Chinese microblogging. The experimental results indicate that the performance of SVM is best in three machine learning algorithms; IG is the better feature selection method compared to the other methods, and TF-IDF is best fit for the sentiment classification in Chinese microblogging. Combining the three factors the conclusion can be drawn that the performance of combination of SVM, IG and TF-IDF is best.

Keywords: microblogging; sentiment analysis; machine learning; feature selection; term weight

1. INTRODUCTION

With the development of the internet, specially the rise of the social media such as blogs and social networks, it is very convenient for users to publish comments to different products or topic events. The reviews for the products from these channels are very valuable for both the sellers and the buyers, while the comments for topic events also is very important for management agencies to learn the public opinions in time. There is rich subjective information in these comments. Automatically mining the opinions in these subjective sentences can generate potential applications. A lot of attentions have been paid to sentiment analysis as an emerging technology in recent years [1-3]. Sentiment analysis refers to the application of natural language processing, computational linguistics, and text analytics to identify and extract subjective information in source materials. There are two main methods for performing sentiment analysis: machine learning methods [1, 3] and semantic-oriented methods [4-5]. The methods based machine learning algorithms look the sentiment analysis as a classification problem, and classify the sentiments through building a sentiment lexicon including positive words and negative words. The prior researches have shown that the performances of the machine learning methods are better than the performances of the semantic-oriented methods [1-2].

Microblogging is an emerging web 2.0 application and developed rapidly in recent years. Microblogging is a broadcast medium in the form of blogging, and it allow users to exchange small elements of content such as short sentences, individual images, or video links. A microblogging differs from a traditional blogging in the following five aspects:

1) Length: Compared with the conditional reviews, the length of the microblogging generally is limited to 140 characters. According to our collected data the average length of the microblogging is 40 characters. For the shorter length of the comments it is easier to understand the opinion in these comments;
2) Easy accessibility: Mostly of the current microblogging provide the API, by which we can easily obtain much information for studying;
3) Unique expressional style: Compared with the conditional blogs, microblogging contains more popular internet words, such as “:)” which presents the happy;
4) Information richness: People in microblogging might come from different fields, therefore, the information in microblogging is very rich. People can publish product reviews, the opinion to the topic events, etc. The microblogging generally provides keywords search function, so we can easily search related information through defining topical keywords;
5) Timeliness: Like send a short message, people can publish microblogging information through...
intelligent phone. Therefore, the information from microblogging is timelier than conditional blogging. This unique characteristic of microblogging makes the microblogging a better timely information source.

Considering the characteristics discussed above, we think that the sentiment analysis for microblogging is very meaningful. Though many researches studied the sentiment analysis for the English microblogging the study for the Chinese microblogging is very deficient [6-7]. Yang shen [8] proposed a semantic-oriented method, which calculates the sentiment index of microblogging through defining five lexicons including attitude lexicon, weight lexicon, negative lexicon, degree lexicon, and conjunction lexicon. However we could not find researches for studying the sentiment analysis with the machine learning methods. For filling in this gap we study the sentiment analysis for the Chinese microblogging with three method learning methods, three feature selection methods, and three feature weighting methods, and compare the commonality of the sentiment classifiers between the microblogging and ordinary reviews.

2. METHODOLOGY

2.1 Machine Learning Method

2.1.1 Support Vector Machine

SVMs (Support Vector Machine) are a novel machine learning algorithm based on structural risk minimization theory [9]. It has been applied in text classification and face recognition. In text classification SVM has been proved better effective compared with conditional methods [10]. In this paper we LIBLINEAR are employed to train and test the classification model. LIBLINEAR is proposed by Rong-En Fan aims to solve large-scale linear text classification and it is especially effective in high-dimension and sparse dataset [11].

2.1.2 Naive Bayes Classification Algorithm

Naive Bayes classification is a frequently used text classification method. It is a simple probabilistic classifier based on applying Bayes theorem with strong independence assumptions. Though the model is very simple it was used widely in text classification field [12]. There are two different Bayes model according to text classification: multinomial model and multivariate bernoulli model.

In present many researchers studied the text classification with the multinomial model [2, 13, 14], therefore we choose the multinomial model to conduct the experiments in this paper.

Multinomial Bayes classification calculates the probability of the word $W_t$ appears in given class $c_j$ with the following formula:

$$p(w_t | c_j) = \frac{\sum_{i=1}^{n_j} n_{it}}{\sum_{i=1}^{W} \sum_{j=1}^{N_j} n_{it}}$$

(1)

$n_{it}$ Represents the number of word $t$ occurs in document $i$, $N_j$ represents the size of the classification $c_j$ in train dataset, and $W$ represents the size of the word dictionary.

The posterior probability is shown as formula 2:

$$p(c_j | d) = \frac{p(c_j) p(d | c_j)}{p(d)}$$

(2)

2.1.3 N-Gram model

Text classification based on n-gram model is a novel model in natural language processing [15]. Differ from the traditional vector space model, n-gram model looks the document as a sequence of words. An n-gram model is a type of probabilistic model for predicting the next item in such a sequence.

For a string $S = C_1C_2...C_{n-1}C_n$, n-gram model presume that the probability of the n character only depend
on the previous n-1 characters:

\[ p(c_n \mid s_{c1\ldots cn-1}) = p(c_n \mid c1\ldots cn-1) \]  

(3)

2.2 Feature Selection Methods

2.2.1 IG (Information Gain)

IG is frequently employed as a term goodness criterion in the field of machine learning \cite{16}. It measures the number of bits of information obtained for category prediction by knowing the presence or absence of a term in a document. The formula for IG is shown as following:

\[
IG(t) = -\sum_{i=1}^{|C|} P(c_i) \log P(c_i) + P(t) \sum_{i=1}^{|C|} P(c_i \mid t) P(c_i \mid t) \\
+ P(\overline{t}) \sum_{i=1}^{|C|} P(c_i \mid \overline{t}) \log P(c_i \mid \overline{t})
\]

(4)

The \( P(c_i) \) represents the probability of class \( c_i \); \( P(t) \) represents the probability of the term \( t \) occurs; \( P(\overline{t}) \) represents the probability of the term \( t \) not occurs.

2.2.2 CHI

The CHI statistic chooses the term through measuring the dependency between the term and the class. The bigger value means the stronger dependency between the term and the class, on the contrary, the smaller value means the relatively dependency between the term and the class. The formula for CHI is shown as following:

\[
CHI(t, c_i) = \frac{N(N_{1i}N_{00} - N_{10}N_{0i})^2}{(N_{1i} + N_{0i})(N_{1i} + N_{00})(N_{10} + N_{0i})(N_{10} + N_{00})}
\]

(5)

\[
CHI(t) = \max_{c_i} (CHI(t, c_i))
\]

(6)

Where \( N \) is the total number of the documents in train dataset; \( N_{1i} \) is the number of times \( t \) and \( c_i \); \( N_{10} \) is the number of the times \( t \) occurs without \( c_i \); \( N_{0i} \) is the number of times \( c_i \) occurs without \( t \); \( N_{00} \) is the number of the times neither \( c_i \) nor \( t \) occurs.

2.2.3 DF (Document Frequency)

DF is a simplest feature selection method, and it selects a term through set the threshold of the DF. DF is the number of documents in which a term occurs in a dataset. DF methods assume that both rare and common words are either non-informative for category prediction, or not influential in global performance. Therefore, DF removes the words whose document frequency is less than some predefined small threshold or bigger than some predefined large threshold. Though the idea is very simple, the performance of DF in text classification is very good \cite{17-18}.

3. Data Collection

There are not public available Chinese microblogging dataset, therefore, we write a crawler program to collect some microblogging data from Sina microblogging (http://weibo.com). In Sina microblogging we can collect microblogging data from specific subject through defining keywords. For avoiding the experimental results dependant on the specific fields, we collect the data from four subjects: H1N1, movie reviews, earthquake, and sport events. The three members from our team labeled the train dataset; finally we get 2134 piece of reviews including 1002 piece of positive reviews and 1132 piece of negative reviews.

4. EXPERIMENTS
4.1 Experiment Design

In this paper ICTCLAS \cite{19} is employed to perform word segmenting for each review. Firstly, we select specific feature weighting method to build vector space model, secondly feature selection method is employed to select terms, finally, we use three machine learning algorithms to train classification model.

In experiments we use WEKA toolkit (http://www.cs.waikato.ac.nz/ml/weka/) to perform the SVM, Naive Bayes algorithms, and use Lingpipe (http://alias-i.com/lingpipe/index.html) to perform the n-gram algorithm. 10-cross validation is used in experiment and F-SCORE is used to evaluate the classifier performance, the formula of F-SCORE is following:

$$F = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

(7)

Where Recall is the percentage of recall of the algorithm, and Precision is the percentage of accuracy of the algorithm.

4.2 Result and Analysis

4.2.1 Comparison of the Different Feature Weighting Methods

This experiment is designed to compare the following feature weighting methods:

1) Presence weighting: if the term occurs in the document, the weight is 1, otherwise the weight is 0;

2) TF (Term Frequency) weighting: the term weighting is defined as the number of term occurs in the document;

3) TF-IDF (Term Frequency-Inverse Document Frequency) weighting: this method revised the TF weighting, and the number of the documents containing the terms is considered as a impact factor. The inverse document frequency factor is incorporated which diminishes the weight of terms that occur very frequently in the collection and increase the weight of terms that occur rarely. The formula for calculating the TF-IDF can be written as follows:

$$W(t,d) = tf(t,d) \times \log(N / n_t)$$

(8)

Where $N$ is the total number of training documents, and $n_t$ is the number of documents containing the word $t$.

The prior studies often adopted a specific feature weighting method \cite{2,14}. For example, (1) compared the performances of the Presence and TF in English sentiment analysis, and the results showed that the Presence outperformed better. Xu jun \cite{20} conducted the same experiments in Chinese news sentiment analysis, and the results also showed that the Presence outperformed better than TF. However, to our best knowledge, no similar experiments are conducted in microblogging field. Therefore, in this paper we design an experiment to compare the performances of the three different feature weighting methods.

In this experiment we employ the IG as the feature selection method, and SVM and Naïve bayes are chosen to conduct classification. The Figure 1 and the Figure 2 display the performance curves of three feature weighting methods using SVM and Naïve bayes.
As shown in Fig.1, according to the different machine learning methods, three weighting methods have different advantages. In SVM classification method, the TF-IDF outperforms best across all weighting methods, while the performance is close between the Presence and the TF. In Naïve Bayes classification method, the Presence outperforms best across all weighting methods, while the performance of TF-IDF is worst especially when the number of the features reaches to the 4000.

Taking the classification algorithm and weighting methods into account, we can find that the performance of SVM classification reaches to the peak when adopting the TF-IDF weighting method and the number of features is 2000: the F-SCORE is 87.07. The performance of Naïve Bayes classification reaches to the optimal when adopting the Presence weighting method and the number of feature is 3000: the F-SCORE is also 86.07. Therefore, according to the IG feature selection method, the SVM and the TF-IDF is the best combination.

### 4.2.2 Comparison of the Different Feature Selection Methods

This experiment compares the performance of the different feature selection methods. In experiment SVM is employed as the classification algorithm and the TF-IDF is used to weight feature. The results are shown in Figure 3.

As shown in Fig.3, the IG outperforms best across the three feature selection methods. When the number of feature is close to 2000 the performance of the IG reaches to the optimal. The performance of the CHI and DF is very close. When the number of feature exceeds the 2500 the performance of three methods is stable.

### 4.2.3 Comparison of the Different Machine Learning Algorithms

<table>
<thead>
<tr>
<th>Classification algorithm/weighting method</th>
<th>Presence</th>
<th>TF</th>
<th>TF-IDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>85.10</td>
<td>84.54</td>
<td>87.07</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>86.07</td>
<td>86.41</td>
<td>84.91</td>
</tr>
<tr>
<td>N-GRAM</td>
<td></td>
<td></td>
<td>82.32</td>
</tr>
</tbody>
</table>

The aim of this experiment is to compare the performance of three different machine learning methods. In experiment 1 we find that the performance of classification algorithms depends on the feature weighting methods, thus, we choose three different weighting methods to conduct the experiment. The experimental results are shown in Table 1.

From the table 1 we can learn that the performance of N-Gram model is worst in three classification algorithms, while the performance of the SVM and Naïve Bayes depends on the weighting methods. The performance of the SVM reaches to the optimal when adopting the TF-IDF and the performance of the Naïve Bayes reaches to the optimal when adopting the Presence.

From the prior experiments we can draw a conclusion: the performance of sentiment classification is best when the TF-IDF, SVM, and IG are combined.

## 5. CONCLUSIONS

This paper studies the sentiment analysis in microblogging, and the experiments show that the three machine learning algorithms are effective; especially the performance of classification is optimal when the TF-IDF, SVM, and IG are combined. In future, we will conduct more in-depth analysis for the sentiment mining in microblogging; for example, we can compare the performance of the machine learning methods and the semantic-oriented methods. In addition, we will study whether the conclusions is extended to other fields, such as emergency events.
ACKNOWLEDGEMENT

The research is supported by the National Natural Science Foundation of China under Grant No. 90924020 and the PhD Program Foundation of Education Ministry of China under Contract No. 200800060005.

REFERENCES


