ISSN 2409-2665
Journal of Logistics, Informatics and Service Science
Vol. 2 (2015) No. 1, pp. 48-58

Classification of Sentiment Sentences Based on Naive Bayesian Classifier

Xiaoheng Ou[1], Yan Cao[1], Xiangwei Mu[1]
[1]Transportation Management College, Dalian Maritime University, China
ouou_19890920@163.com;1035458580@qq.com;dlmussx@126.com

(Received Dec 2014, accepted Apr 2014)

Abstract. This paper is to conduct popular micro-blog for sentiment classification. The Naive Bayesian Classifier is the key in this paper, and study on pretreatment of the text of micro-blog, constructing sentiment dictionary, feature selection, feature weights and expression vector, comes up with some points and conducts the experiment. And the performance of "emoticons + twice sentiment feature extraction +BOOL" is the best pretreatment method. And this experiment gains a relatively satisfactory result.

Keywords: Micro-Blog, Sentiment Classification, Sentiment Dictionary, Feature Selection, Naive Bayesian

1. Introduction

Micro-blog is a platform that sharing, disseminating and obtaining information based on user relationship. Users can build their own community through WEB, WAP and other clients. And micro-blog is the text information that contains about

140 words and realizes real-time update. In 2006, twitter first launched micro-blog service in America, and then micro-blog swept the world. In 2009, with sina, Tencent, NetEasy and other portal sites started to launched micro-blog service, micro-blog is surging in China like a storm.

Micro-blog is a platform, in there you can be a audience and browse information which you are interested in, you can also be a publisher, and other persons can browse information which you release. The information is usually very short about 140 words thus micro-blog gains its name. You can also publish pictures and share videos and so on. The greatest feature of micro-blog is that the information can be published rapidly and spread rapidly. If you have 2 million audiences, your information can be spread to two millions audiences rapidly.

The sentimental analysis is mainly for English micro-blog currently (Barbose et al. 2010, Pak 2011), Sentiment analysis of Chinese micro-blog is still at the beginning (Sun et al. , 2012). Sentiment analysis of English micro-blog mainly carried on the research on the twitter news and it is mainly divided into two types: sentiment analysis that is related to the subject and sentiment analysis that is Irrelevant to the subject. Sentiment analysis of Chinese micro-blog is at the beginning, the current research can be divided into two fields: methods based on sentiment knowledge and methods based on feature classification. Methods based on sentiment knowledge are mainly to establish sentimental dictionary (Li et al. , 2009) or field sentimental thesaurus and judge the sentiment of the text by its sentiment word or the combination of words. The methods based on feature classification is mainly to use machine learning method and treat sentiment analysis as traditional classification, extract feature values and make the judgment.

The method of micro-blog sentiment classification used in this paper is based on Naive Bayesian, Naive Bayesian method mainly has the following three stages. The first stage, Data preparation is mainly to collect data and determine the feature, transfer the feature to feature vector, this stage need human to complete. The second stage, Classifier establishment is to establish Naive Bayesian classifier, Calculate the frequency for each category in the training set, and calculate each feature study of the category division of conditional probability, and record the results. The third stage, application is mainly to classify instances by the classifier that is already built. Input the feature vector of the instance into the Naive Bayesian classifier, and then output the classification category of the instance.

2. The text preprocessing

Sentiment classification for the micro-blog is to determine the category of the text according text content in a given classification system. The text classification is based on machine learning and the classifier that is used is Naive Bayesian classifier. The process of sentiment classification is as picture 1.

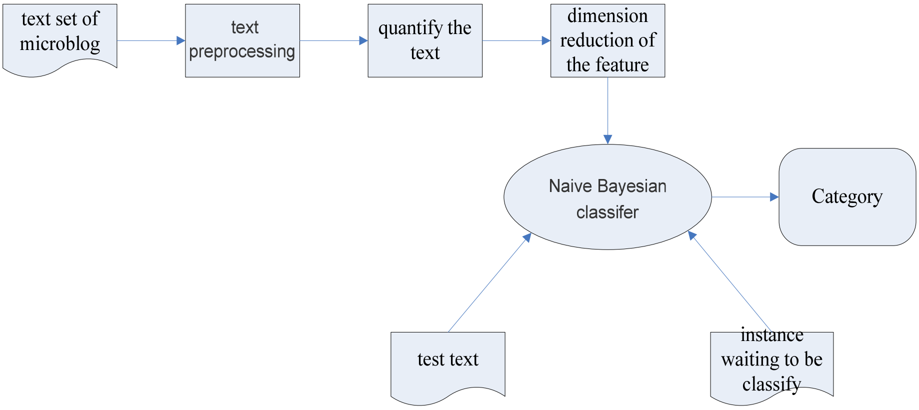

Fig.1. The flow chat of emotional classification

2.1 Micro-blog text segmentation

Automatic word segmentation of Chinese text is that the computer divided the Chinese text into a group of words according to certain rules, which is the first step of preprocessing of Chinese text, micro-blog become a group of words through this process. By certain rules the match phrase and the text content will be roughly consistent. If the word segmentation is not reasonable, the text after segmentation will make some deviation from the original meaning of the text, which will affect the result of the sentiment classification. micro-blog text segmentation uses the maximum matching algorithm, and supported by the “modern Chinese vocabulary table”,set the maximum matching length as 4 and only for Chinese word segmentation processing. The chosen basic emotional lexicon contains 2807 positive words and 2474 negative words, a total of 5281. Considering the specialty of micro-blog text, as there is a large number of network popular words with strong emotional color and emotional images as shown in table 1 in micro-blog text, so we collect network popular words, and forms the network vocabulary dictionary.

Table 1. Emotional images in micro- blog

Positive images	Negative images
😁 smile	😣 grievance
😊 so happy	😭 cry
😃 laugh	😠 curse
🤗 hug	😢 sad
👏 applaud	😡 anger
👍 good	👎 bad

2.2 Emotional feature selection

Emotional feature selection is to select the words which can be the marks of the text, and get rid of the useless words. We can get an extremely sparse matrix after the word segmentation in order to facilitate the calculation and reduce the matrix dimension. Using emotional lexicon for the emotion feature selection is just to choose emotional words as features, this method is simpler, it can be completed in word segmentation. If we use the emotional lexicon while the word segmentation, we can chose the emotional feature conveniently. And Naive Bayesian classifier can modify probability model, it can achieve the goal of classification while training at any time. Feature selection only according to emotional dictionary can neglect the Syntactic relations between the texts, it can cause that the emotional feature and the evaluation subjects are not symmetric. For example, “well, I am speechless to you”, ”well” is positive, but it do not modify the subject “ I ” , if chose “ I ”as the emotional feature, it will affect the result of classification. Syntactic path refers to the syntactic structure which connects between any two nodes in the syntactic tree. The following picture 2 is the syntactic path of s1.

S1:He often reads detective novels.

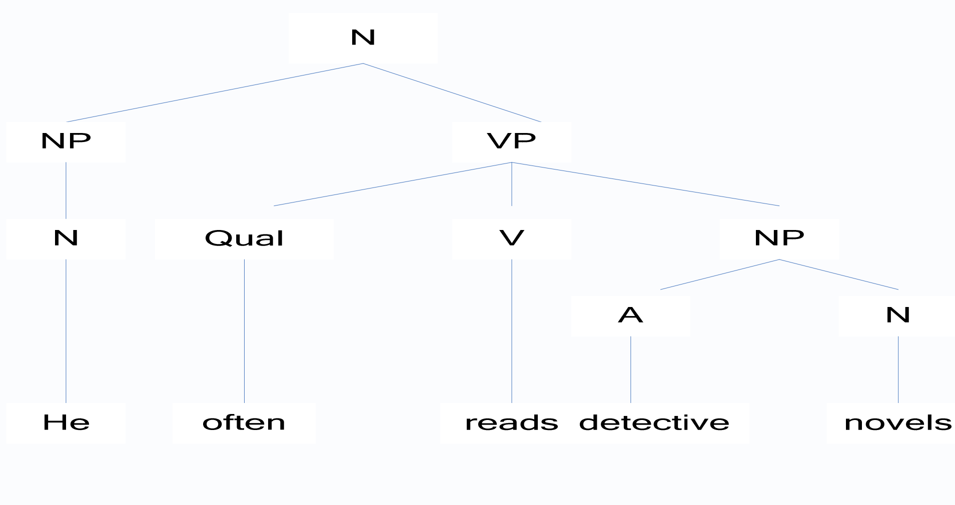

Fig.2. The example of syntactic path

The emotional feature selection based on syntactic path, must first build syntactic path, conduct the syntactic match according to the statistics of syntactic path, and obtain the Syntactic structure of the sentence, on this basis of this, chose the emotional statement which modify the evaluation object as the emotional features.

2.3 Feature weight calculation

Feature weighting refers to the weight of feature in the text; it is the important basis of classification. In this paper, we conduct the sentiment classification with Boolean and word frequency. And then we conduct feature selection with the basic emotional vocabulary dictionary and the network emotional vocabulary dictionary, because feature selection would be completed during the word segmentation, so the weight is calculated after the feature selection. Boolean and word frequency is the relatively simple method to represent the weight. The Boolean represents the weight as follows .

$$bool(w_j):\begin{cases} 1 \ freq(w_i,d_j) \\ 0 \ freq(w_i,d_j) \end{cases} \tag{1}$$

“freq (wi,dj)” is the frequency of the word wi in the text dj.

2.4 The vector representation of the text

Text d can be represented as the set d = {w1,w2,w3……wn} of some certain words. Feature weights of words are a vectors, thus the text d can be regarded as the matrix of row and vocabulary. In order to save storage space, we use "storage word-index: weight" format in the actual storage, and between each of the two different word vectors we separated with a space. Among them “word-index” is the index, index of test corpus must correspond to the training corpus, “weight” is the word weight in the text, and we separate them with a colon. A text a line, thus form a matrix of text. In general condition, a matrix of text have a matching glossary file, each glossary takes up a line, the line number corresponds to the word index of matrix file, multiple matrixes can also use the same glossary file.

3. Naive Bayesian classifier

The classification principle of Bayesian is to use the Bayesian formula to calculate the subject’s posterior probability through the prior probability of the subject，that is the probability of the category which the subject belonging to, choose the maximum posteriori probability as the object's category. That is to say, Bayesian classifier can optimize in the sense of minimum error rate. There are four kinds of main Bayesian classifier in the present study; they are Naive Bayesian、TAN、BAN and GBN. Bayesian classifier is one of the most valuable methods to address the issue of corresponding machine learning. Especially when conduct the text classification, Naive Bayesian classifier is one of the most effective method. In the paper, we build the emotional classifier with the Naive Bayesian.

Bayesian formula assume that the parameters follow the probability distribution, and make the related reasoning through the existing data, thus to make the optimal decision. Given the example set T, how to get category c of classification example. The definition is as follows.

$$P(c \mid T) = \frac{P(T \mid c)P(c)}{P(T)} \tag{2}$$

P(c) represents the prior probability of c in the training set T; reflect the probability of the test instance belonging to the category c. In case that we do not know the category classification, we can give each category the same prior probability. P (T) represents the prior probability of training set T, P(T|c) represents the probability of training set T on condition that know category c., P (c | T) represent the posteriori probability of c, it represents that the confidence level on condition that we know the training set T. P(T|c) is very easy to get here, but P(c|T) is hard to get., and we are usually more concerned about P(c|T). We can get P(c|T) through P(T|c) with Bayesian. In many cases, the classifier need to find the maximum posteriori assume. c $\in$ C(C is the complete set of the classification), it is the most probable category cMAP on condition we know the training set. We described it by the following formula 3.

$$c_{MAP} = \underset{c \in C}{\arg\max}\, P(c \mid T) \tag{3}$$

We can get formula 3 through formula1 and formula 4.

$$c_{MAP} = \underset{c \in C}{\arg\max} \frac{P(T \mid c)P(c)}{P(T)} \tag{4}$$

We can do not depend any other category to get constant P(T), it can be ignored, formula 3 can be changed like formula 5.

$$c_{MAP} = \underset{c \in C}{\arg\max}\, P(T \mid c)P(c) \tag{5}$$

We can get training characteristics and its analogy through Bayesian, and we can get most probable category when we put in a new instance. Its main idea is to calculate probability of the other category when given a new instance; we think the instance belongs to the category which has the maximum probability.

To predict a new instance of X is to get the most probable category on the condition that we have determined attribute value <a1,a2,a3……an> of training set. It is described as formula 6.

$$C_{MAP} = \arg\max_{c \in C} \max P(a_1, a_2, a_3, \ldots a_n \mid c) p(c) \tag{6}$$

The premise of the Naive Bayesian is that the attribute is independent given the classification instance. It Is described as formula 7.

$$P(a_1, a_2, a_3, \ldots, a_n \mid c) = \prod_{j=1}^{n} P(a_j \mid c) \tag{7}$$

We put formula 7 into formula 6, thus we can get Naive Bayesian formula, it is described as the following formula 8.

$$C_{MAP} = \arg\max_{c \in C} P(c) \prod_{j=1}^{n} P(a_j \mid c) \tag{8}$$

P(c) and P(aj|c) are described as formula 9 and 10.

$$P(c) = \frac{\sum_{i=1}^{n} \sigma(c_i, c)}{n} \tag{9}$$

$$P(a_j \mid c) = \frac{\sum_{i=1}^{n} \sigma(a_{ij}, a_i) \sigma(c_i, c)}{\sum_{i=1}^{n} \sigma(c_i, c)} \tag{10}$$

"N" represents the total number of the training sets, "aij" represents the number j value in the training set n, when ci=c, then $\sigma(c_i, c)$=1, thus $\sigma(c_i, c)$=0.

In most instances, the estimates of the probability are good using the method above, however, in some cases, such as when a feature does not appear in a category, it will generate the phenomenon that the classification result is zero which will reduce the quality of the classifier. Here we introduce Laplace proofreading. When the training sample is large enough, it will not occur the situation that the frequency is zero and it will not affect the classification result. As shown in Equation 11 and Equation 12.

$$P(c)=\frac{\sum_{i=1}^{n}\sigma(c_i,c)+1}{n+n_c} \tag{11}$$

$$P(a_j\mid c)=\frac{\sum_{i=1}^{n}\sigma(a_{ij},a_j)\sigma(c_i,c)+1}{\sum_{i=1}^{n}\sigma(c_i,c)+n_j} \tag{12}$$

4. Evaluation index

The main purpose of this paper is the text sentiment classification, and the goal of is to make the machine automatic classification results the same with the manual classification by evaluating the classification results from the classification accuracy and speed based on the similarity of the machine classification and manual classification. The classification speed depends on the complexity of the classifier, and the classification accuracy depends on the classification results which experts think, make judgments and recognize.

We often use Precision, Recall, Fl as the classification precision index currently. They are described as the following formulas.

$$presion=\frac{the\ nuber\ of\ the\ correct\ text\ belongs\ to\ this\ category}{the\ number\ of\ the\ text\ belongs\ to\ this\ category}*100\% \tag{13}$$

$$recall=\frac{the\ nuber\ of\ the\ correct\ text\ belongs\ to\ this\ category}{the\ number\ of\ the\ text\ belongs\ to\ this\ category\ which\ is\ classified\ by\ human}*100\% \tag{14}$$

$$\text{F1}=\frac{2*precision*recall}{precision+recall}*100\% \tag{15}$$

5. The result of the experiment

We do the experiment with the Test corpus based on method of emotional dictionary and secondary emotional feature extraction. The results of the experiment are described as follows.

Table 2. The result of the classification (F1:100%)

	word frequency		BOOL	
	emotional dictionary	second extraction	emotional dictionary	second extraction
ignore punctuation	70.05	72.88	70.06	72.32
consider punctuation	70.68	73.16	70.31	74.57

6. Summary

We do the emotional classification research on the micro-blog comments based on the Naive Bayesian classifier, and we did the experiment. We do the emotional classification on the top topic on micro-blog, and we achieve the relatively satisfactory result. In the future study on the emotional classification of micro-blog, we should think deeply on the diversity of the object processing, this can be one of the most effective way to improve the performance of the classification. Another micro-blog sentiment classification research in the specific fields can realize the public opinion analysis and find public opinion. We can do the emotional classification of the text with the technology of the emotional classification. We can get the emotional state of the Internet users, a social phenomenon, the preferences of a product and other information, which not only have a certain commercial value but also a help on the social stability. Micro-blog sentiment classification is a promising research work, this paper studies a very small field, there are still many problems to be solved, and I wish to solve these problems in the future study work.

Acknowledgement. This work is supported by Program for Liaoning Innovative Research Team in University, so we express our deep gratitude for this.

References

Barbose L., Feng J. (2010). Robust Sentiment Detection on Twitter from Biased and Noisy Data. Proc of COLING' 10, 2010:36-44.

Pak A, Paroubek P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Proc of LREC'10:1320-1327.

Sun S., Xie X., Zhou M. (2012). Emotional analysis and feature extraction of Chinese micro-blog based on hierarchical structure. *Chinese Information Journal*, 26(1):73-83.

Li C., Liu W., Zhou Y. (2009). The study on the built of Chinese basis emotional dictionary. *Computer Appliance*, 29(11):2882-2884.