Classification of Sentiment Sentences Based on Naive Bayesian Classifier

Ou Xiaoheng[1]*, Cao Yan[1], and Mu Xiangwei[2]

Transportation Management College, Dalian Maritime University, China, 116026

e-mail:ouou_19890920@163.com;1035458580@qq.com;dlmussx@126.com

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. 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

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. Considering the specialty of micro-blog text, as there is a large number of network popular words with strong emotional color and emotional images in micro-blog text, so we collect network popular words, and form the network vocabulary dictionary.

2.2 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.

2.3 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. Given the example set T, how to get category c of classification example. The definition is as follows.

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

$$C_{MAP} = \arg\max_{c \in C} P(c \mid T) \tag{3}$$

We can get formula 3 through formula1 and formula 4.

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

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.

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

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.

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

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

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

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.

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

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

4The 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 1: 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

5 Conclusion

We do the emotional classification research on the micro-blog comments based on the Naive Bayesian classifier, and we did the experiment. 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.

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

References

1. BarboseL, Feng J. Robust Sentiment Detection on Twitter from Biased and Noisy Data[C]// Proc of COLING’ 10,2010:36-44.

2. Pak A, Paroubek P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining[C]// Proc of LREC’10,2010:1320-1327.

3. Xie Xieli, zhou Ming,Sun Songmao. Emotional analysis and feature extraction of Chinese micro-blog based on hierarchical structure[J].Chinese information Journal,2012,26(1):73-83.

4. Liu Weiping, Zhou Yanhui,Li Chunliang. The study on the built of Chinese basis emotional dictionary. Computer appliance,2009,29(11):2882-2884.