Sentiment Classification for Chinese Microblog

Wen-Hsing Lai, Chang-Hsun Li
Department of Computer and Communication Engineering
National Kaohsiung First University of Science and Technology
Kaohsiung, Taiwan
e-mail: lwh@nkfust.edu.tw

***Abstract*—A sentiment classification method for Chinese microblog is presented. For short sentence microblog, it is very challenging because the information of emotion is very limited. First, an emotion lexicon is built from training corpus. A simple measure – difference ratio is used to choose words from lexicon as features for classification. Support vector machine and voting on counts and accumulated difference ratio are jointly combined as classification method. The experimental results show that our recognition rate is better than the popular method using collocation strength. Our recognition improvement is about 2.06% in testing. Therefore, the difference ratio measure we used and the tactic in constructing the lexicon are proved very effective.**

Keywords- sentiment classification; emotion; support vector machine; microblog; emoticon

I. INTRODUCTION

We can analyze one's emotional state, like happiness, anger, and sadness, by observing a person's body reaction at the time. Imagine applying it in data processing of computer. By means of detecting the characteristics of the computer data, we can foretell the emotional state the information conveyed. Music, video, text are the three computer information data that are often analyzed. Among them, text sentiment analysis has lots of applications. Through the analysis of internet users' comments about news, books, or products, we could know their feelings and do the proper reaction or promotion. The major task, sentiment classification, is to identify users' opinions or emotions as positive or negative. The applications could be very useful. For example, Tao [1] did a sentiment analysis of the news comments, while Wu [2] try to figure out the emotion of the public shareholders by an Internet-forum-based stock market analysis method.

Microblog is a popular broadcast medium. Different from a traditional blog, its content is typically smaller such as short sentences, individual images, or video links. Users can post their texts or comments and use emoticons to express their feelings. The emoticons they used can be considered as a sentiment tag and make microblog a convenient and easy-to-get corpus for short sentence sentiment classification studies. Because, in sentiment classification studies, we generally need sentiment annotated corpus for model training (if we use supervised pattern recognition methods) and for testing the system recognition rate, using the emoticons instead of annotating sentiment corpus by human is very convenient. Therefore, by using emoticons , it can save lots of time of collecting and annotating corpus that are generally very labor intensive. Since the material is short sentence, it is very challenging to recognize its sentiment by utilizing so limited information. However, emoticons are only used in corpus annotating; the objective of our study is to classify emotion based on sentences without emoticons.

Sentiment classification normally needs the help of an emotion lexicon or dictionary that brings together positive and negative words, which can be used as the features in sentiment classification. A good lexicon can help improve the recognition rate a lot. However, one major obstacle to sentiment classification is the lack of a complete sentiment dictionary for many languages. How to overcome the obstacle is very important. Wu [3] combines multi-dictionary and commonsense knowledgebase by gathering nine kinds of sentiment dictionaries as sentiment concept seed, then through sentiment spreading activation from common sense network (ConceptNet) to get more sentiment concepts. On the other hand, Yang [4] built an emotion lexicon automatically from weblog corpora. Considering the difficulty of collecting dictionaries and domain relevance, we will construct an emotional lexicon automatically from corpus.

Further, machine learning or pattern recognition methods are commonly used in sentiment classification. For example, fuzzy association rules is proposed to analyze melancholiac patients' emotion from their daily text messages [5]. Support vector machine (SVM) is also a popular machine learning method and is adopted in emotion classification of weblog [4][6]. Since SVM has accomplished very promising performance in many pattern recognition fields, we also take on this model combined with a voting method.

This paper is organized as follows. The next section will introduce the source of our corpus. Then, the process of establishing the emotion lexicon is presented in the third section. The fourth section describes our sentiment classification method. Experimental results will be shown in the following section. Finally, conclusions and future works are discussed.

II. CORPUS

Collecting and annotating an emotion corpus is very labor-intensive. How to build a source with large amounts of

 ISBN: 978-1-61208-382-7

Figure 1. Plurk.

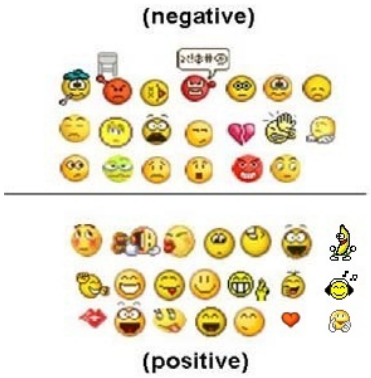

Figure 2. Positive and negative emoticons.

了不起的概念啊。
(Great concept.)
(a)

現在是怎麼了，這種犯人這樣就可以放過了。
(Now what, let such prisoner get away with this.)
(b)

Figure 3. Examples of (a) positive and (b) negative sentences.

diverse text and annotated sentiment is a very tough job. Fortunately, internet users have developed visual cues,generally known as smileys or emoticons, to express their emotion. The text on Internet with emoticons becomes a convenient and useful source for building the emotion corpus [7].

We, therefore, select Plurk [8] as the source. Plurk is a net for microblog, as shown in Figure 1, where users can publish their text, which is limited to 210 Chinese characters or English alphabets. Short sentences from political and news source with emoticons are collected.

Then, we need to determine our emotion model. Plutchik's [9] 2D and 3D emotion model is very popular. Thayer [10] also developed a two-dimensional emotion space with four quadrants. One axis is valence from negative to positive. The other axis is arousal from silent to energetic. In this paper, for simplification and consideration of corpus size, we focus on only positive and negative sentiment.

41 emoticons, including 21 positive emoticons and 20 negative emoticons, as shown in Figure 2, are adopted. Sentences with positive emoticons are annotated as positive and sentences with negative emoticons are annotated as negative. Totally, there are 8520 sentences in our training emotion corpus, comprising 4578 positive sentences and 3942 negative sentences. Testing corpus contains 4229 sentences, involving 2470 positive sentences and 1759 negative sentences. Examples of positive and negative sentences are shown in Figure 3.

III. EMOTION LEXICON

An emotion lexicon is automatically built for sentiment classification from training corpus. Word segmentation is firstly applied. Since two-syllable, three-syllable, and four-syllable words are the most frequently appeared, only the three types of words are collected.

Then, the numbers of words are counted by using a word counting program [11] to get the word count that appeared in the positive and negative sentences, namely N_p and N_n. If one word appears in positive and negative sentences just as frequently, we assume this word does not have discrimination in emotion and discard it. Such method of setting a threshold to discard words without discrimination can reduce wrong classification and increase system performance [12]. However, word application is always context dependent. Thus, the word applications in themselves are highly revealing of the context in which they are being used. So, in our future studies, we will reconsider those words discarded and divide them into further categories - like incidental constructs (not really possessing any significance), and context-revealing constructs (possibly adjectives which reveal the positive-negative weighting purely through application context). The remaining words will be classified according to their word counts in positive and negative sentences, that is, they will be classified into positive emotion lexicon if they appear more frequently in positive sentences, and be classified into negative emotion lexicon if they appear more frequently in negative sentences. That is,

$$If \begin{cases} N_p = N_n & Discard \\ N_p > N_n & Positive \\ N_p < N_n & Negative \end{cases}. \quad (1)$$

The final step of building an emotion lexicon is word sorting according to their significance. There are several measures to determine the word significance. Pointwise mutual information (PMI), which measures the strength of the association of two samples, is widely used in relation extraction, word collocation, and word sense disambiguation in natural language processing. A variation of pointwise mutual information, which measures the collocation strength $co(e,w)$ between an emotion e and a word w, is defined as [4] :

 ISBN: 978-1-61208-382-7

$$co(e,w) = c(e,w) \times \log \frac{P(e,w)}{P(e)P(w)}. \quad (2)$$

c(e,w) is total co-occurrences of *e* and *w*. Normalized collocation strength is also proposed and used in [12].

Instead, in our experiment, we use a very simple measure – difference ratio (DR), which is defined as

$$\frac{(N_p - N_n)^2}{N}, \quad (3)$$

where N is the sum of N_p and N_n. When the total count N is the same and the difference of N_p and N_n is higher, the difference ratio is larger and the word is more significant. If the word appears in positive and negative sentences nearly equally, the difference ratio is low and the word is insignificant. If the difference of N_p and N_n is the same, the word appears less frequently got larger difference ratio and is more significant.

Words are sorted according to difference ratio and we got an emotion lexicon arranged with significance from high to low. The lexicon contains a total of 10476 words, including 5954 positive words and 4522 negative words. The procedure of building an emotion lexicon is presented in Figure 4.

From the emotion lexicon generated, we found out that there are some words unexpected, like people's name, or place's name. They are originally neutral, but classified into positive or negative emotion by learning. From Yu's study [13], we know that emotion lexicon generally includes two kind of words. One is *domain dependent word* and the other is *domain independent word*. *Domain independent word* is word with emotion linguistically and *domain dependent word* is from learning in accordance with the training field. Therefore, those people's names, or place's names are *domain dependent words* from learning. Since our corpus contains political and news contents, it causes politicians' and place's names appear in our emotion lexicon as emotion words.

IV. SENTIMENT CLASSIFICATION

There are two stages in our sentiment classification of a testing sentence. The first stage is using SVM to do the classification. Those sentences that SVM cannot classify will pass to the second stage using voting method.

In SVM, the features we used are 100 top words from emotion lexicon with highest difference ratio. If one specific word appears in sentence, the value of feature is set to 1, otherwise, it is 0.

If all the 100 words do not show in the sentence and SVM fail to classify, then we use the whole emotion lexicon to vote. If we cannot find any word of the sentence in lexicon, then the sentence is unrecognizable. Otherwise, we count the numbers of positive and negative words in lexicon appear in the sentence. The larger number determines the sentiment class. That is, if the number of positive words in

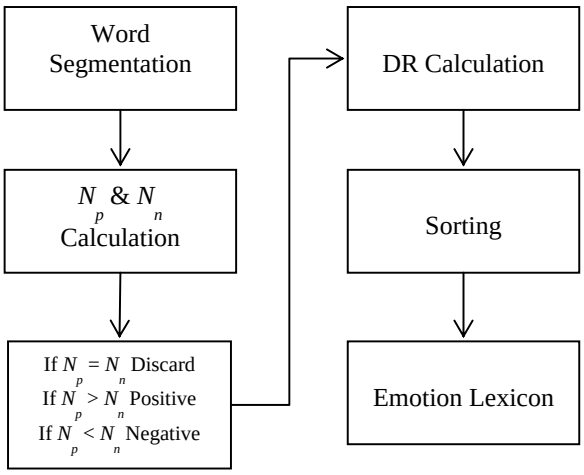

Figure 4. The procedure of constructing an emotion lexicon.

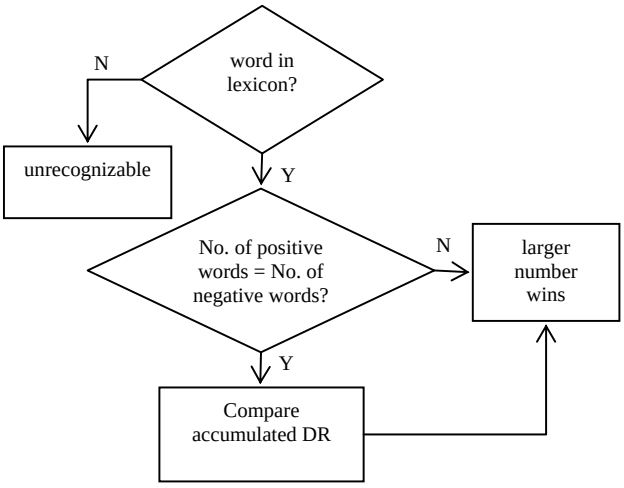

Figure 5. The voting procedure.

太過分了！一想到牠們這麼痛苦的死去，就好難過。
(a)

太|過分|了！ 一|想到|牠們|這麼|痛苦|的|死去，就|好|難過。
(b)

太|過分(N)(too much)|了！ 一|想到|牠們|這麼|痛苦(N)(pain)|的|死去，就|好|難過(N)(sad)。
(c)

Figure 6. An example of (a) a testing sentence "It's too much! As I thought They died such painfully, I am so sad." (b) Word segmentation. "|" is used as word boundaries. (c) Finding positive (P) or negative (N) words.

sentence is greater than the number of negative words, then the sentence is set to positive, and vice versa. If the votes are the same, we accumulate the difference ratio of positive and negative words in the sentence. When the sum of the

difference ratio of positive words is larger, then the sentence is set as positive sentiment, otherwise, it is negative. The voting procedure is shown in Figure 5. An example is shown in Figure 6. First, we have a testing sentence as in Figure 6(a). Then, we applied word segmentation to get the result as in Figure 6(b). Word boundaries are marked by using "|". Assume the SVM fails, then, we check the number of positive and negative words to vote as the procedure in Figure 5. The positive and negative words in lexicon are marked as in Figure 6 (c).

V. EXPERIMENTAL RESULTS

In the SVM stage, the software LIBSVM [14] is used. In training, 5105 sentences out of 8520 sentences are recognizable. That is, there are 3415 unrecognizable sentences that do not contain any of the 100 feature words. In the 5105 sentences, the recognition rate is 59.92%. In testing, 1476 sentences out of 4229 sentences are recognizable, and 2753 sentences are unrecognizable. In the 1476 sentences, the recognition rate is 71.68%.

Those unrecognizable sentences are sent to the second voting stage. In the second stage, 2212 sentences out of 2753 sentences are recognizable. That is, the rest 541 sentences do not contain any word in emotion lexicon. In the 2212 sentences, the recognition rate is 62.79%. Among them, 1871 sentences are decided by vote counts directly, 341 sentences are decided by further comparing the accumulated difference ratio. In the vote counts part, the recognition rate is 64.03%, and in the accumulated difference ratio part, the recognition rate is 56.01%.

Totally, in testing, 3688 sentences out of 4229 sentences are recognizable. In the 3688 sentences, 2447 sentences can get correct sentiment answer and the recognition rate is 66.35%.

By observing the recognition results, most of the expression of the sentences is appropriate. But some are not. The reasons are, firstly, the emoticons annotated are not always very accurate, and, secondly, some blogger would say things with irony (use positive words to express negative feeling), which makes the system hard to recognize.

Because there are few studies focus on short sentence sentiment classification, we will compare our method with Yang's method [4], which is a very effective emotion classification model applied in weblog articles. We will simulate Yang's method - cooperating SVM and Yang's method 3 on our short sentence corpus from microblog and classify the sentiment into positive and negative, because his method 3 outperforms his method 1 and 2 in most cases. Yang's method is applied in another context, so, to be fair, we use Yang's lexicon creation method to create lexicon from our corpus in simulating Yang's method.

First, Yang uses collocation strength to build an emotion lexicon. All collocations (word-emotion pairs) are listed in a descending order of collocation strength. For a specific word, if the collocation strength of positive emotion is larger, then the word is classified as positive emotion and vice versa. The lexicon by using Yang's method from our training corpus is 11316 words including 5892 positive words and 5424 negative words. Since our method discard the words which appear in positive and negative sentences just as frequently, our lexicon is smaller.

TABLE I. COMPARISON OF RECOGNITION RATE AND THE NUMBER OF RECOGNIZABLE SENTENCES IN BRACKETS OF OUR AND YANG'S METHOD.

Method	SVM		Voting		Total
	training	testing	counts	DR/co	
Ours	59.92% (5105)	71.68% (1476)	64.03% (1871)	56.01% (341)	66.35% (3688)
Yang's	62.63% (5336)	64.26% (1950)	62.92% (1502)	54.36% (298)	62.93% (3750)

To be compared with our method, top 100 words in the lexicon by Yang's method are used as features in SVM. The major difference of Yang's sentiment classification procedure and ours is they use accumulated collocation strength to decide the sentiment when the positive and negative votes are the same.

In the simulation of Yang's method, 5336 sentences out of 8520 sentences are recognizable, and the recognition rate is 62.63% in SVM training. In SVM testing, 1950 sentences out of 4229 sentences are recognizable. In the 1950 sentences, the recognition rate is 64.26%.

In the second voting stage, 1800 sentences are recognizable. In the 1800 sentences, the recognition rate is 61.5%. Among them, 1502 sentences are decided by vote counts directly, 298 sentences are decided by further comparing the accumulated collocation strength. In the vote counts part, the recognition rate is 62.92%, and in the accumulated collocation strength part, the recognition rate is 54.36%.

Totally, in testing by Yang's method, 3750 sentences out of 4229 sentences are recognizable. In the 3750 sentences, 2360 sentences can get correct sentiment answer and the recognition rate is 62.93%.

To make it clearer, all the experimental results are summarized in Table 1. Comparing our method with Yang's method, since we discard some words that we think do not have discrimination in emotion, which led to smaller lexicon, and causes the number of sentences that we can process is fewer. But our recognition rate is better than Yang's method using collocation strength. In other words, we both use 8529 sentences for training, and in 4229 testing sentences, our method can hit correct sentiment in 2447 sentences, 66.35% of 3688 sentences, comparing with 2360 sentences, 62.93% of 3750 sentences, by using Yang's method. The difference is 87 sentences. Our recognition improvement is about 2.06% of 4229 testing sentences.

From the results compared to Yang's method, we can conclude that discarding words which appear in positive and negative sentences just as frequently and adopting a measure considering the difference between N_p and N_n and the total word frequency N can help improve the recognition rate.

 ISBN: 978-1-61208-382-7

VI. CONCLUSIONS AND FUTURE WORKS

This paper presents a sentiment classification method for Chinese microblog Plurk. It is a very challenging job because the material is short sentence and the information of emotion is very limited.

A very simple measure, i.e., difference ratio, is chosen for lexicon building and feature selection. SVM and voting are combined as classification method. The experimental results show that our recognition rate is better than Yang's method using collocation strength. In using the same training and testing sentences, our method can get correct sentiment in 2447 sentences, while Yang's method can only get correct sentiment in 2360 sentences. The difference is 87 sentences. Our recognition improvement is about 2.06% of 4229 testing sentences. Therefore, the difference ratio measure we used and the tactic in constructing the lexicon are proved very effective.

Our method can help intelligent HCI (Human-Computer Interaction) systems sense, i.e., detect and interpret the user's emotional states automatically, and then respond appropriately, which will make the HCI systems more natural, efficacious, persuasive and trustworthy. Furthermore, the emotional data may be used in many applications, including music or consumer products recommendation, training plan or education improvement, and even health care.

However, the size of our corpus and emotion lexicon is still not enough and needs to be extended in the future. Besides, words collocations with strong sentiment orientation are important for the text sentiment analysis [15]. Semantic models can help to identify the sentiment orientations of collocations and improve our sentiment classification in the future.

REFERENCES

[1] F. M. Tao, J. Gao, T. J. Wang, and K. Zhou, "Topic Oriented Sentimental Feature Selection Method for News Comments," Journal of Chinese Information Processing, Vol. 24, No. 3, May, 2010, pp. 37-43.

[2] J. Wu, Y. X. Chen, and D. M. Liu, "Internet-forum-based Stock Market Analysis Method," Computer Engineering, Vol. 38, No. 13, July 2012, pp. 254-256.

[3] H. H. Wu, "Sentiment Analysis Using Multi-dictionary and Commonsense Knowledgebase," Master Thesis, Department of Computer Science and Information Engineering, National Taiwan University, Taiwan, July 2011.

[4] C. H. Yang, K. H. Y. Lin, and H. H. Chen, "Building Emotion Lexicon from Weblog Corpora" Proceedings of 45th Annual Meeting of Association for Computational Linguistics, poster, June 23rd-30th, 2007, Prague, Czech Republic, pp. 133-136.

[5] H. W. Chiu, "Using Fuzzy FP-tree Model to Discover Associations Among Melancholiac Patient's Daily Text Messages," Master Thesis, Department of Electrical Engineering, National Taipei University of Technology, Taiwan, July 2010.

[6] C. H. Yang and H. H. Chen, "A Study of Emotion Classification Using Blog Articles," Proceedings of the 18nd Conference on Computational Linguistics and Speech Processing (ROCLING 2006), 2006, pp. 253-269.

[7] J. Read, "Using Emotions to Reduce Dependency in Machine Learning Techniques for Sentiment Classification," Proceedings of the ACL Student Research Workshop, 2005, pp. 43-48.

[8] Plurk, http://www.plurk.com/ [retrieved: January, 2015].

[9] R. Plutchik, The Emotions, University Press of America, 1991.

[10] R. E. Thayer, The Biopsychology of Mood and Arousal, Oxford University Press, 1989.

[11] Chinese Word Segmentation System, Academia Sinica, Taiwan, http://ckipsvr.iis.sinica.edu.tw/ [retrieved: January, 2015].

[12] Y. T. Sun, C. L. Chen, C. C. Liu, C. L. Liu, and V. W. Soo, "Sentiment Classification of Short Chinese Sentences," Proceedings of the 22nd Conference on Computational Linguistics and Speech Processing (ROCLING 2010), Puli, Nantou, Taiwan, September 2010, pp. 184-198.

[13] H. C. Yu, K. T. H. Huang, and H. H. Chen, "Domain Dependent Word Polarity Analysis for Sentiment Classification," Computational Linguistics and Chinese Language Processing, Vol. 17, No. 4, December 2012, pp. 33-48.

[14] C. C. Chang and C. J. Lin, "LIBSVM -- A Library for Support Vector Machines," http://www.csie.ntu.edu.tw/~cjlin/libsvm/ [retrieved: January, 2015].

[15] S. Wang and A. Yang, "A Method of Collocation Orientation Identification Based on Hybrid Language Information," Journal of Chinese Information Processing, Vol. 24, No. 3, May, 2010, pp. 69-74.

 ISBN: 978-1-61208-382-7