# Sentiment Analysis of Microblog Combining Dictionary and Rules

Ding Yuan[1], Yanquan Zhou[1, 2], Ruifan Li[1, 2], Peng Lu[1, 2]
1, School of Computers
Beijing University of Posts and Telecommunications
2, Engineering Research Center of Information Networks
Ministry of Education
Beijing, China
yuanding_bupt@163.com, zhouyanquan@gmail.com, rfli@bupt.edu.cn, lupeng@bupt.edu.cn

***Abstract*—Microblog has become a daily communication tool in recent years. Researches on microblog have drawn more and more attention. Microblogging emotional classification is a major research of user intent analysis based on User-Generated Content (UGC). This paper focuses on the discrimination on two emotional tendencies: positive and negative. Firstly, the system cleared the noisy elements in the microblog, then extracted the features of the microblog and finally classified the microblog using Support Vector Machine (SVM). Furthermore, we improve the algorithms of feature extraction and weight computing combining dictionary approach and rule based approach. The result of experiment shows that the method is effective.**

***Keywords—emotional classification; feature extraction; weight computing; support vector machine***

## I. Introduction

With the rapid development of the Internet and communications industries, the Internet has been increasingly deep into people's lives, more and more users begin to communicate with friends on the Internet, express their views, share their life experiences, or get live information, so a large number of the user generated content (UGC) was born at the right moment. User intent analysis based on UGC as an important research of a lot of contents of UGC has been widespread concern by research-related researchers. Based on semantic analysis of text, user intent analysis mainly study the text created by users analyze user intent content creation by emotional tendencies. Research based on entity extraction and emotional detection technology, combined with the behavior of user, ultimately achieves the purpose of the classification of the UGC's intention. Specific research divisions are: evaluation object and evaluation word extraction, the emotional tendencies detection of evaluation word, sentence emotional detection, and user intent recognition. Microblog sentiment analysis represented as sentences emotional tendency detection which is the mainly and key research content of user intent analysis based on text semantic analysis requires extensive and deep study.

Microblog as increasingly popular form of communication represented as UGC, with a short length of the text, information diversity, real-time characteristics, provides users with a convenient online communication channels as well as rich and fresh content, and has become one of the most popular Internet applications. Since Microblog has such a large user base, including the large scale data and information, and the traditional manual method has been unable to collect and handle the vast amounts of data and information effectively, analyzing and processing by computer technology is absolutely necessary.

Sentiment Analysis infer the emotional state of the speaker in the convey information, and determine its emotional polarity process by analysis and processing to the text with subjective emotional color. Sentiment Analysis of the large scale microblog data will be helpful to improving the Internet public opinion monitoring system and user intent analysis model based on UGC.

Current research work on emotional analysis for Chinese microblog is still in its infancy, while sentiment analysis for English microblog has made important progress. However, due to the huge difference between English and Chinese in syntax, semantics and pragmatics etc., we face more problems in the processing of Chinese microblog. Chinese microblog specified length of text must be less than 140 words, which makes it a microblog text can contain multiple sentences, but topics and emotions expressed by every sentence can be different, even opposite, which put forward higher requirements to the sentiment analysis for Chinese microblog.

In this paper, sentiment analysis for Chinese microblog will be seen as two classification problems, namely positive and negative emotional tendencies. The measure of extraction of feature words from the microblogging text was improved chi-square statistic algorithm, and each feature vector of microblog was calculated based on the statistical results. Then the weights of features were calculated by the improved TF-IDF. Finally the model of SVM learned the training sample set, and gave out the result of emotion classification on the test set.

Section II will describe the related work of sentiment analysis. Section III will elaborate on the principles of methods and the design of algorithms in this paper. Section IV will show the experimental results and related analysis. The

This work was supported by the National Natural Science Foundation of China with grant No. 71231002 (Key Project) and 61100120, Fundamental Research Funds for the Central Universities, Project No. 2013RC0304.

ASONAM 2014, August 17-20, 2014, Beijing, China
978-1-4799-5877-1/14/$31.00 ©2014 IEEE

last section we will have the conclusion of the present stage of the work and the prospect of the next work.

## II. RELATED WORK

Sentiment analysis using of Natural Language Processing, computational linguistics, text analysis study to extract subjective information in the data source to find the polar expressions of the authors' attitude. Affective Computing, published by Professor Picard from MIT in 1995, combines a number of disciplines of computer science, psychology and cognitive science, in order to make the machine can understand people's emotional state, to change their behavior in response to people's emotions. This method has been widely used in the field of artificial intelligence, and has been improved with the deeply research of human-computer interaction. At present, information processing on the language of computers has accounted for about 85% of computer use, so the sentiment analysis research has very important academic significance. Besides, because of its intuitive commercial value, the sentiment analysis has become a heat research field of Information Science in the world.

Microblog text sentiment analysis can be seen as a two-polarity classification problem, namely positive emotion and negative emotion. Using traditional machine learning methods such as Naive Bayes, Maximum Entropy, K Nearest Neighbor algorithm, Support Vector Machine and so on, we can construct an emotion classifier to recognize the sentiment of micro-blog texts. In 2002, Pang and Lee [1] used supervised learning method to mine the text emotion and get the user's subjective point hidden in the text. They did preprocess of the text, including part of speech tagging, extracting the location information, negative word extraction, uni-gram and bi-gram word extraction, and treated these as features, using methods of machine learning to identify sentiment polarity. The experimental results show that compared to the Naive Bias and Maximum Entropy algorithm, Support Vector Machine has higher classification accuracy rate. In the field of Chinese text processing, Chinese scholars also used traditional machine learning methods to do a lot of experiments on Chinese text. In 2011, Liu Zhiming [2] used SVM, Naive Bias and N-gram model to classify the same Chinese micro-blog text data. Comparing the experimental results, they found SVM getting better results than Naive Bias and N-gram model. In addition, they did text categorization experiments with emotion by making classification methods combine with different weight calculation algorithms, and results show that when using the TF-IDF algorithm, SVM can get better performance.

The sentiment analysis method based on dictionary mainly depends on the judgment of reference words on the experts dictionary. English emotional vocabulary mainly comes from Word Net, a large English vocabulary database, in which nouns, verbs, adjectives and adverbs are organized into a set of cognitive synonymy. Each syn-set represents a basic semantic concept and every syn-set establishes a connection with each other according to various relations. Because each set has a hierarchy of its own, it can constitute a synonym network. In the process of Chinese text, dictionary source is HowNet. The main idea of this method is that according to the collection of reference words with known polarity, the method compares semantic similarity between unknown words and reference words in text, sets threshold, judges sentiment; then, gets scores through the accumulation of positive and negative emotional words in the text and compares the score with the threshold value to get the text sentiment. Zhu Yanlan [3] used manual selection of a few of benchmark words, and then calculated the similarity between words and standard words by semantic similarity and semantic related field on HowNet to distinguish the sentiment of new words. Yao [4] et al in the calculation of Chinese word emotion not only considered the sentiment of words in dictionary, but also analyzed the sentiment of words in context. Hu and Liu [5] got the sentiment of emotion words through the relations between synonym and antonym on WordNet and then according to the sentiment of main emotional words, they got the sentence sentiment orientation. In addition, Peter Turney [6] used mutual information to process text sentiment based on emotional reference words.

In our experiment, we applied dictionary-based and rule-based method into machine learning method. It is better to extract feature words through dictionaries and rules for assistance and expression of rules can make the weight calculation more attention to natural language, which makes the results more reasonable.

## III. MICROBLOG SENTIMENT ANALYSIS

In this paper, the task is treated as a classification problem. We used vector space model to represent microblog. We first abstracted the feature of the microblog, and then mapped the microblog into a N-dimension feature vector, finally used the SVM to complete the classification. There are four main steps: data preprocessing, feature extraction, weight calculation and classification. The main purpose of the data preprocessing is to delete the noisy data in the microblog. Feature extraction is constructing feature vector to represent the microblog. Weight calculation is to calculate the weight of the feature vector. According to the particularity of sentiment classification, we improved the feature extraction algorithm and weight calculation algorithm by integrating dictionaries and rules.

We used the emotional dictionary provided by Intelligence Science and Technology Center, Beijing University of Posts and Telecommunications. We selected the sentiment words having more strong emotional tendency and degree adverbs having more strength to build the motional dictionary by referring to HowNet and other general dictionaries. The emotional dictionary we used is different from other dictionaries because it is only composed of emotional words, so it is suitable to our experiment. The final dictionary contains more than 8500 of the commendatory, more than 7200 of the derogatory term, and more than 200 degree adverbs or dynamic emotion words.

Although there are many different domain dictionaries, we didn't find a multiple domains dictionary suitable to multiple topics' sentiment analysis according to the information. Moreover, the difference between one topic's sentiment analysis and multiple topics' sentiment analysis is not strongly relative to the emotional words. In our study, we still used the

TABLE I. THE CLEANING RULES

| ***Rules*** | ***Original microblog*** | ***Post-cleaning microblog*** |
|---|---|---|
| *deleted reply* | 事实，蒙牛坑的人可比切糕坑人多多了//@Angela_不高兴小姐: 蒙牛太差了 | 事实，蒙牛坑的人可比切糕坑人多多了 |
| *delete topic* | #大手机显脸小# 三星 5 年后超苹果！我支持！ | 三星 5 年后超苹果！我支持！ |
| *delete the title* | 【乌克兰冲突又起】乌克兰东南部克里米亚自治共和国首都辛菲罗波尔 26 日发生冲突。 | 乌克兰东南部克里米亚自治共和国首都辛菲罗波尔 26 日发生冲突。 |
| *delete source* | 蒙牛英文广告语变成“没有幸福”遭网友调侃 http://t.cn/zlWICiN （分享自 @头条新闻） | 蒙牛英文广告语变成“没有幸福”遭网友调侃 http://t.cn/zlWICiN |
| *delete website* | 中国滑翔伞第一人坠亡续：未买保险无任何赔偿，原文地址：http://t.cn/zOF5ctU | 中国滑翔伞第一人坠亡续：未买保险无任何赔偿，原文地址： |
| *delete mailbox* | @阮阿绿 女 23 岁 浙江.杭州 征婚 tangcutuzi@163.com | @阮阿绿 女 23 岁 浙江.杭州 征婚 |
| *delete the reference* | 老爸的杰作…@罗建 SVW @lucky-蓉儿 | 老爸的杰作… |

same emotional dictionary to deal with one topic's sentiment analysis and multiple topics' sentiment analysis.

The following describes the main processing steps in order.

## *A. Data preprocessing*

Since the microblog is shorter, more colloquial and noisier than common document [7], so it is necessary to do some pretreatment before classification. Data preprocessing contains considering the respondents emotions. Because the noise data in general have a unified format, so we used regular expressions for data cleaning. Table I lists the main cleaning rules.

Because there are too many oral expressions in microblog, the general segmentation tool cannot do a good job. So, we used the python segmentation tool jieba who fits to cut the micro blog. Removing the stop words is to remove the words that are not useful to this task leads to improvement of the result. As the general stop word dictionary do not specifically consider the sentiment classification, so this paper further screens the dictionary. We removed the adverbs of degree and some special punctuation. These words and punctuation modified its modifier emotional intensity which plays a special role in the feature weight calculation.

## *B. Feature extraction*

Because of microblog text contains a large amount of information, so that the original dimension of the feature space is very high, and there are lots of redundant, hence feature selection is needed. Feature extraction selects several words according to the score getting from an evaluation function. CHI algorithm selects feature words based on the dependence between the words and the classification. It both considers the positive and the negative correlation between the feature and the classification [8]. But it does not consider special nature of sentiment classification. In the sentiment classification, emotional words have more distinguish capability than the non-emotional words. So the paper adopts a new algorithm called MCHI to extract feature words. This new algorithm is a variant of the standard CHI algorithm, which uses more three steps: data cleaning, word segmentation and stop words removing.

The main task of data cleaning is to remove the noise data in the microblog, these noise data mainly includes: reply, topic, and reference. Because these data does not contain emotion, so they are removed as noisy data. For example, the replay of a microblog should be removed, because we should be concerned only published microblog emotion, without emotion information of the words. MCHI algorithm gives each pair of word *w* and class *c* a score to identify the dependence. Calculation formula of MCHI scores is as follow:

$$MCHI\left(w,c\right)=CHI\left(w,c\right)\ +\lambda\times f\left(w\right)\times CHI-E\left(c\right) \qquad (1)$$

*CHI(w,c)* represent the standard CHI score, *f(w)* is indicator function indicating whether the word *c* is an emotion word, *CHI – E(c)* represents the average score between all the non-emotion words and the class *c*, *λ* is a harmonic parameters indicating the importance of the *CHI – E(c)*. In this paper, *λ* is set to 0.3.

According to formula (1), if *w* is non-emotional words, then the MCHI score is equal to the standard CHI score, otherwise it will and a positive score to the standard CHI score. Emotion words more easily are selected as the feature word by the formula (1).

The paper used the following formula (2) to calculate the standard *CHI* score between the word w and the class *c*.

$$\lambda^2(w,c)=\frac{(AD-BC)^2}{(A+B)(C+D)} \qquad (2)$$

*A* defines the number of microblog containing the word *w* and belongs to the class *c*.

*B* defines the number of microblog containing the word *w* but not belongs to the class *c*.

*C* defines the number of microblog not containing the word *w* but belongs to the class *c*.

*D* defines the number of microblog neither containing the word *w* nor belongs to the class *c*.

For each class, this paper gets the *N* highest MCHI score words, then compute the union of the feature words all the classes to form the final feature set.

The number of the features is vital to task. Too little or too much feature both decrease the final accuracy of the classification. Too litter feature words lead to the classification model is too simple, which will occur under fitting phenomenon. Too much features lead to a classification model is too complex, which appear over fitting phenomenon. Experimental results show that the accuracy of classification is highest when the feature number is nearly 4700.Table II lists the accuracy corresponding to different feature counts. The experiment is conducted using 5000 labeled microblog.

TABLE II. CLASSIFICATION ACCURACY RATE UNDER DIFFERENT AMOUNT OF FEATURES

| ***Amount of features*** | ***Classification accuracy*** |
|---|---|
| 298 | 0.512 |
| 627 | 0.56 |
| 1453 | 0.572 |
| 2430 | 0.578 |
| **4692** | **0.692** |
| 5932 | 0.647 |

### *C. Weight calculation*

Weight refers to each component value of a feature vectors, and the value characterize the representation ability to classification. TF-IDF algorithm is a widely used method to compute weight. But this algorithm does not take into consideration the differences of emotional words and non-emotional words in the sentiment classification task [9][10][11]. This paper proposed a new algorithm called MTF-IDF to improve the TF-IDF algorithm by making it more suitable for the sentiment classification task. MTF-IDF algorithm integrates two-language phenomenon into the original TF-IDF algorithm:

- The emotion words have more power than non-emotion words.
- Adverbs of Degree and special punctuation modify emotional word strength.
- When the emotional words are modified by the special punctuation, the emotional words have the different representation ability of polarity classification.

MTF-IDF is calculated as follows:

$$vn_{i,j} = n_{i,j} \times \left(1 - f_1\left(i\right)\right) + n_{i,j} \times f_1\left(i\right) \times f_2\left(i\right) \times f_3\left(i\right) \tag{3}$$

$$TF_{i,j} = \frac{vn_{i,j}}{\sum_k vn_{i,j}} \tag{4}$$

$$IDF_i = log \frac{|D|}{\left|1 + \left\{ j : t_i \in d \right\}\right|} \tag{5}$$

$$TFIDF_{i,j} = TF_{i,j} \times IDF_i \tag{6}$$

$n_{i,j}$ defines the number of word *i* occurring in the microblog *j*; *$f_1(i)$* indicates whether the word *i* is an emotion word; *$f_2(i)$* defines the influence of the degree adverb; *$f_3(i)$* defines the influence of the special punctuation; $vn_{i,j}$ defines the virtual number of word *i* occurring in the microblog *j*; $|D|$ defines the number of all the microblog; *{j: ti ∈ d}* defines the number containing word *i*.

This paper used the libsvm software package to implement the classification. The use of the libsvm is beyond this paper.

## IV. EXPERIMENT

The paper performed three experiments. The first experiment was compared with the baseline method, the second experiment explored the method efficiency one different corpus and the last experiment shows that this method is suitable for three-classification problems.

### *A. Baseline*

This paper designed a classical experiment as a baseline. The classical experiment used the standard CHI and TF-IDF as the method to select the feature words and calculate the weight. The experiment was conducted on a corpus containing about 5000 labeled microblog. The corpus contains multiple topics, including automotive, mobile phone, jewelry and so on. The first 4500 microblog were used for training and the remaining for testing. Each microblog was labeled -1 or +1 indicating the emotional tendency. Table III lists the experimental result.

TABLE III. COMPARATIVE EXPERIMENT BETWEEN BASELINE AND THE METHOD PROPOSED IN THIS PAPER

| | ***Baseline*** | ***Our method*** |
|---|---|---|
| ***POS_P*** | 0.8806 | 0.9198 |
| ***POS_R*** | 0.7824 | 0.8790 |
| ***POS_F1*** | 0.8286 | 0.8989 |
| ***NEG_P*** | 0.8032 | 0.8859 |
| ***NEG_R*** | 0.8742 | 0.9246 |
| ***NEG_F1*** | 0.8372 | 0.9048 |

The results show that the method proposed in this paper is better than the ordinary algorithm. This is because this method integrated emotional information into the feature selection and weighting calculation.

### *B. Experiment on different corpuses*

The new method also ran on two corpuses. The corpuses are different in scale and topic number. Below list these two corpuses in table IV and experiment results in table V.

The experimental result on the corpus B is better than the other corpus. One of the main reasons is that corpus contains only one topic and the other corpus has more topics. The method proposed in this paper has good classification ability on the simple corpus, but should be developed on the complex corpus.

TABLE IV. DIFFERENT CURPUSES

| | *Corpus A* | *Corpus B* |
|---|---|---|
| ***Training set size*** | 4500 | 1800 |
| ***Test set size*** | 500 | 200 |
| ***Topic number*** | More than six | Only one |

TABLE V. COMPARATIVE EXPERIMENT BETWEEN ONE TOPIC CORPUS AND MULTIPLE TOPICS CORPUS

| | *One topic* | *Multiple topics* |
|---|---|---|
| ***POS_P*** | 0.9091 | 0.9198 |
| ***POS_R*** | 0.9402 | 0.8790 |
| ***POS_F1*** | 0.9243 | 0.8989 |
| ***NEG_P*** | 0.9278 | 0.8859 |
| ***NEG_R*** | 0.8911 | 0.9246 |
| ***NEG_F1*** | 0.9091 | 0.9048 |

### C. Experiment of three-classification

We also conduct an experiment on a corpus including three labels where +1 indicate positive, -1 represent negative, 0 means this microblog do not contain any emotion. The corpus which also has multiple topics contains about 5000 microblogs, including 1610 positive microblogs, 1763 negative microblogs and 1627 neutral microblogs. The first 4500 microblogs were used for training and the remaining for testing. The experiment result is listed in table VI.

TABLE VI. EXPERIMENT RESULT OF THREE-CLASSIFICATION

| | |
|---|---|
| ***POS_P*** | 0.894 |
| ***POS_R*** | 0.501 |
| ***POS_F1*** | 0.641 |
| ***NEG_P*** | 0.931 |
| ***NEG_R*** | 0.534 |
| ***NEG_F1*** | 0.679 |

The new method can be used to deal with the three-classification problem directly, but the result is not as good as two-classification. The main reason is three-classification is more difficulty than two-classification. The result also suggests that there is still much room for improvement in the three-classification problems.

## V. CONCLUSION

This paper thought the analysis of the emotional polarity of microblog as two-classification problems. We used the VSM model to represents a microblog, and then used SVM to give out the result of classification. Our main operation to the data set was cleaning, Word segmentation, removing stop words, feature selection and classification. The experiment results show that the proposed method of emotional polarity analysis of Chinese microblog is more effective. Of course, there is quite a lot of development of this article practices, We believe that future research could focus on three aspects: 1) we could improve the performance of this method in a multi-topics model by learning from LDA to process the sentiment classification in the shallow semantic layer. 2) Adding part of the semantic features based on the features of the original word will improve the performance of classification. 3) According to the special nature of the problem, we could add the appropriate rules.

## *References*

[1] Pang B, Lee L, Vaithyanathan S. Thumbs up ? Sentiment classification using machine learning techniques. EMNLP '02. Proceedings of the ACL-02 conference on Empirical methods in natural language processing, 2002: 79-86.

[2] Liu Zhiming, Liu Lu. Empirical study of sentiment classification for Chinese microblog based on machine learning. Computer Engineering and Applications, 2012, 48(1): 1-4.

[3] Zhu Yanlan, Min Jin, Zhou Yaqian, Huang Xuanjin , Wu Lide . Semantic Orientation Computing Based on HowNet.Journal of Chinese Information Processing. 2006, 20(1): 14-20．

[4] Yao Tianfang, Lou Decheng. Research on Semantic Orientation Distinction for Chinese Sentiment Words. Chinese study on computing technology and language problems——The Seventh International Conference on Chinese Computing. Wuhang: International Conference On Chinese Computing (ICCC 2007), 2007.

[5] Hu M Q, Liu B. Mining and summarizing customer reviews. KDD’04. Proceedings of the 10th international conference on Knowledge discovery and data mining (KDD). 2004:168-177.

[6] Turney P D. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. ACL '02. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistic. Philadelphia : Association for Computational Linguistics, 2002:417-424.

[7] Liu Xiaojuan, You Bin, Zhang Aiyun. Review on the Data Used in Researches of Microblogs. Journal of Intelligence, 2013, 32(9), 39-45.

[8] Zhu Lina. Research of Feature Selection for Chinese Web Page Categorization. Beijing: China University of Petroleum, 2009.

[9] Ma Wenwen, Deng Yigui. New feature weight calculation method for short text. Journal of Computer Applications, 2013, 33(8): 2280-2282,2292.

[10] Ma Ting, Geng Guohua, Zhou Mingquan. An effective approach to calculate the feature weights. J. of Zhengzhou Univ. (Nat. Sci. Ed.), 2008, 40(4): 48-51.

[11] Qin Shian, Li Fayun. Improved TF-IDF Method in Text Classification. New Technology of Library and Information Service, 2013, 238(10): 27-30.