Chinese Micro-blog Sentiment Analysis based on Semantic Features and PAD Model

Fei Gao[1], Xiao Sun[1*],Kunxia Wang[2], Fuji Ren[1,3]
[1] Hefei University of Technology, Hefei 230009, China
[2] Department of Electronic Engineering, Anhui University of Architecture, Hefei, China
[3] Faculty of Engineering, University of Tokushima, 2-1 Minami-Josanjima, Tokushima 770-8506, Japan
gaofeisxt@gmail.com, sunx@hfut.edu.cn,kxwang@ahjzu.edu.cn, ren2fuji@gmail.com

* Corresponding author

***Abstract*—With the increasing impact of social networks, microblog becomes important carrier of information and social interaction for human beings, which contains emotional states that have important research significance. We try to analysis the microblog text with the methods of emotional vocabulary, combining domain knowledge of psychology and affective computing, continuous dimension of emotion psychology PAD model which is adopted as basis of sentiment analysis. Emotional state inherent in the text is analyzed to obtain a more accurate result and achieve purposes of emotional analysis. At the same time, to achieve emotional microblog text computability from the aspect of personal characteristics. Experimental results show that the method can improve the microblog text sentiment analysis accuracy and precision. The method is able to get a good application in the different themes and different emotional features.**

Keywords—Emotional word; PAD emotional model; emotional quantification; Chinese microblog; sentiment analysis

I. INTRODUCTION

The social network has risen to become one of the main applications of the Internet. A wide range of social networks are changing the traditional way of life of the people. People are gradually accustomed to get and exchange information and expression emotion on social network. Microblog, compared to traditional ones, is a product of the digital information age and a platform depends on social network information transmission. Take Sina microblog platform as an example: (1) Registered user number is more than 500 million; (2) Daily microblog post number is more than 100 million; (3) A microblog can only accommodate 140 Chinese characters; (4) People usually use only one or two sentences to express all the meanings because of the word limit and the proportions of non-standard language are highly(5) real-time and diversity.

Researches on microblog have become an important aspect of natural language processing and the analysis of emotion in microblog text has turned into vital areas of focus. A large number of experts and scholars were attracted by the huge day sent, quickly spread and fast updated microblog. Meanwhile, there are a lot of microblog with emotion occurring in microblog contents. There are two main missions in text emotion: Information opinion mining and emotional analysis and classify. For example: producer will get feedback when focus on mining user comment information [1]; Text automatic response system [2]; Text automatic generation technique [3]; Recently there are about two types of classification on emotional information: dictionary and rule based method, machine learning and statistics based method.

Studies have demonstrated that in the process of the expression of natural language, people's words and expression not only contains one emotion, a lot of statements are likely to be mixed with expressive emotion. In the microblog text, there are many cases contains a lot of polarity during different emotional words. It is not appropriate to tag the emotional state by a single emotion label. Emotional dimension space description is a method to describe emotion, also known as the passionate dimension theory. Dimensional theory takes the emotions smoothly and gradually in the dimension of space between different emotional distances to measure the similarities and differences of each other [4]. By studying the relationship between atypical emotional words expressing could obtain more accurate text emotion. In this approach, the text emotion can be calculated.

In all kinds of expressive dimension model, proposed by Mehrabian PAD three-dimensional emotional model [5] is among the more mature. The dimensions of emotional description model, it will be divided by P (Pleasure - displeasure), A (Arousal - noncausal) and D (Dominance - Submissiveness) consisting of three dimensional spaces. By calculating psychology, emotion, [6], and utilization of P, A, D, the result shows that this theory can effectively describe and explain human emotion, distinguish the distinct emotional state, and can get all kinds of feelings.

At present, Chinese microblog sentiment analysis research has archived less. Most researchers are based on text sentiment analysis methods. In this paper, we study the Chinese microblog sentiment analysis method, combining with the PAD emotional modeling method and the text sentiment analysis method based on dictionary. This paper proposes a text sentiment analysis method based on distance method named PAD. Through the study of the modeling of emotion, we try to project emotional words in the text to the responsive space, and measure the distance and cluster, so as to realize the quantitative analysis of the text emotion. Experiments

978-1-5090-0806-3/16/$31.00 copyright 2016 IEEE
ICIS 2016, June 26-29, 2016, Okayama, Japan

demonstrate that the set out in the present paper analyzes the text of emotion effectively.

This paper structure is as follows: In the second chapter introduces the related work of text analysis and text emotion modeling; The third chapter detailed elaborated the emotion model based on the text sentiment analysis method; The fourth chapter is the experimental results and related analysis; The fifth chapter of this paper is summarized and the next step is prospected.

II. RELATED WORKS

A. The analysis of Chinese microblog emotion

Chinese microblog sentiment analysis can be divided into three main parts: text preprocessing, feature information extraction and classification of emotional analysis [7]. In emotional information extracting discriminate Zhu Yanlan [8] put forward such as semantic vocabulary tendentiousness similarity calculation method based on HowNet, greatly improving the accuracy of this text. Wang Suge[9] considering the relations between Chinese emotional synonyms, based on synonyms assessment method.

In order to get the microblog emotions expressed by the state microblog emotion classification analysis is to classify the extracted text emotional characteristics[10-11]. Overall, sentiment analysis methods in Chinese and English, there are two general approaches.

The first way is a supervised machine learning methods [12-14]. These kinds of methods mainly adopt the maximum entropy of machine learning, support vector machine (SVM) and the naive Bayesian classifier to analyze the text of the emotion. Zhao et al. [12] based on the CRF model by introducing a "redundant features" to study the classification of emotional problems. Li [13] is presented in the method based on DS-LDA model which is used to review the data of binary classification. The second is built on rules and the method of combination [15-18]. Li Shoushan et al. [15] had studied the specific four kinds of different classification method ' s application in Chinese emotion classification. For microblog sentiment analysis work, mostly microblog emotion classification problem is treated as the problem of text categorization using machine learning model..

B. Text emotion modeling

In the field of emotion modeling, for Chinese words there are less emotional modeling reports of the text that we can find. This problem partly restricts the development of Chinese text sentiment analysis. Literature [6] for the first time take Chinese emotional words in modeling, mapping 88 emotional words to PAD emotional distance metric model, getting 14 emotional categories and their emotional distance relationship.

PAD dimensional model is a relatively frequent use of dimension space identification model of emotional state. PAD emotional model was first proposed by Mehrabian and Russell in 1974. The PAD model is composed of three dimensions; Valence (Pleasure) reflects the emotional nature, showing the polarization degree of emotion. Activation degree (arousal) performance and individual emotional physiological activation level of alertness, this dimension reflects the individual's activity in the environment. Dominance (dominance) under different emotional subjective control level, the emotion is mainly used to distinguish between the state is caused by the external environment, or by their own subjective inspire. Compared with other fuzzy description method of emotion, PAD model has the following features: in PAD model, each kind of emotion has only corresponds to a PAD space coordinates. After PAD parameter s normalization was set, emotion can identify with the only 3 D coordinate, with a high degree of confidence. Through a set of standards, PAD parameters for the independence between the various dimensions of the PAD can make in different emotional dimension of the text. The PAD model can be personalized as emotional factors model.

III. MICROBLOG EMOTION ANALYSIS METHOD OF THE MODEL

A. The construction of microblogging corpora typical emotional vocabulary

Emotion is a subjective physiological phenomenon/ people's emotional expression is through facial expression, semantic, voice and gesture [19]. The result of multimodal function is a kind of subjective impulse from the inside and out of the physiological state. In microblog text, emotion is the main expression of emotional words. For example, the micro blog "录了一个晚上的歌！终于又让我吃到你了！！！除了涮肉我在北京的最爱！吃完就可以安心的睡觉喽！(I recorded the songs for a night! Finally let me eat at you again!! I was in Beijing in addition to rinse meat my favorite! Eat can set his mind at to sleep!), you can see, in the text of emotional words including "最爱(love)", "安心(peace)", according to these words show that this text is positive.

Emotional words are relatively important considerations in emotional polarity judging standard. The correct degree of emotional words in the dictionary will have an effect on mood judgment. In this paper, through the analysis of common emotional representation of words in microblog, we select 60 seed words to establish microblog typical emotional seed dictionary corpus. These words and phrases as seed words, through the introduction of Google word2vec model tools [20-22], for these seeds vocabulary extension, training in microblog corpus of 200 million words, obtained with the most relevant 300 words (each word selecting Top5 related words). Through the constant iteration can expand the emotional vocabulary, Part of the emotional seed words as shown in TABLE I.

TABLE I. PART OF THE EMOTIONAL SEED WORDS

Seed words
喜欢(like)，高兴(happy)，愉悦(joviality)，乐意(willi-ng)，关怀(concern)，感激(appreciate)，同情(sympath-y)，兴奋(exc-itement)，悲愤(plaintiveness)，失望(disappointment)，不平(injustice)，心寒(shiver)，自卑(self-abasement)，痛苦(pain)，无聊(bored)，郁闷(depressed)，悲催(tear-inducingmisery)，讨厌(disgusti-ng)，得意(complacent)，欣慰(delighted)， 尊崇(worship)，快乐(cheerful)，舒畅(comfortable)，炫耀(flaunt)

Using above emotional vocabulary in Table I, evaluate PAD with modeling, can get various typical emotional vocabulary. Because traditional method of fixed word cannot extend, has certain limitations, and processing of colloquial words especially new words needs to have a dictionary which can extend continuously. We use the semantic dictionary, based on the seed vocabulary. Some words are shown in Table II below.

TABLE II. NEW EMOTIONAL WORDS

New words	Synonyms	Tendency
善意 (kindness)	笑话(joke)、敬爱 (esteem)、世人(common people)	pos
道听途说 (hearsay)	胡编(cheat)、耳光(a slap on the face)、何故 (wherefore)	neu
听话 (tractable)	家伙(guy)、豆汁儿 (soybean milk)、大人 (adult)	neg
不务正业 (chairwarmer)	菜叶(Leaves)、澳门 (Macao)、浴场 (Kaiserthermen)	neg
盗窃案 (theft)	厂房(plant)、郊区 (outskirts)、站台 (platform)	neg
老龄化 (aging)	估量(appraise)、全球化 (globalization)、产物 (product)	neg
感染力 (infection)	文学(literature)、应和 (responses)、歌唱(sing)	pos
打击报复 (retaliate)	控告人(accuser)、检举 (impeach)、堵塞 (blocking)	neg

B. The modeling emotional words of PAD

In the study of psychology, PAD emotional coordinates of assessment is through a set of complete inventory system. In PAD emotion model, due to the diversity of human language, differentiation, and emotional expression of different typical emotion, the PAD emotion model does not have a unified standard.

Most foreign researchers use Mehrabian compiled complete quantitative table. There are 34 testing programs, including measuring 16 P values, A value and D value all the nine. As the full scale test is more complicated, the researchers have further put forward the simplified scale, to measure the three dimensions using four projects. Chinese PAD model scale of emotional evaluation quantitative table is based on psychological research institute of Chinese Academy of sciences (revised [19], is the semantic difference scale, a total of three dimensions 12, each divided into nine sections. As showed in Table III.

TABLE III. PAD STANDARD PROJECT

NO	Project	NO	Project
S1	愤怒的 (angry)---- 感兴趣的 (interested)	S7	痛苦的 (painful)---- 高兴的 (pleased)
S2	清醒的 (sober)--- 困倦的 (Mondayish)	S8	感兴趣的 (interested)---- 放松的 (relaxed)
S3	受控的 (controlled)---- 主控的 (master)	S9	谦卑的 (humble)---- 高傲的 (arrogant)
S4	友好的 (friendly)---- -轻蔑的 (scornful)	S10	兴奋的 (excited)---- 激怒的 (frenzied)
S5	平静的 (calm)---- 兴奋的 (excited)	S11	拘谨的 (restrained)---- 惊讶的 (surprised)
S6	支配的 (dominant)---- 顺从的 (compliant)	S12	有影响力的 (impactive)-- 被影响的 (effected)

When we get an emotional vocabulary, according to project in Table III, we can evaluate discrimination scores of emotional vocabulary. The Score range is from"- 4 to + 4". Between the final dimension scores, by measuring the dimension of four projects, the parameters calculate the score by the PAD emotional formula which is shown in table IV below.

TABLE IV. PAD PARAMETER NORMALIZATION FORMULA

Dimension	Formula
P	P=(S1-S4+S7-S10)/16
A	A=(-S2+S5-S8+S11)/16
D	D=(S3-S6+S9-S12)/16

In China, the Chinese Academy of Sciences has won 11 typical emotional point locations and PAD reference. By the same text and voice and emotion in the relationship between different expression channels, in this paper, we take 11 categories, followed by neutral, relax, docile, surprise and joy, contempt, disgust, fear, sadness, anxiety and anger. Then we use the K - means to collect the same kind of emotional words. Because word2vec neural network language model is introduced, so all the emotional words are expressed as a semantic vector.

C. Emotional clustering based on the text emotion

After we projecting the extracted emotional vocabulary to PAD space. On the extraction of emotional words and PAD space position of principal component analysis to get the microblog emotions in PAD space position, set as emotional point under test. And then calculate the typical emotional categories to emotional point distance and weight proportion,.

In this paper, we use the calculation method of Euclidean distance, calculation for the emotion of the Euclidean distance between each typical emotional category. Calculation formula as showed in formula (1):

$$s(p1-p2)=\| p1*p1-p2*p2 \| \quad (1)$$

P1, p2 are two test points and p parameter in PAD space observation values. By PAD for the three dimensional space, the final distance between two emotions as follows:

$$s = sqrt(sp*sp+sa*sa+sd*sd) \quad (2)$$

By measuring space distance shows that the smaller the distance between the two kinds of emotion, is both emotional similarity degree is higher. Text analysis of emotion can be carried out on this method. Thus can calculate the emotions under test point to the distance between different emotional categories, get a micrometer posts of this basic emotions.

In this paper, text for each specific emotional categories of Euclidean distance, respectively: S1, S2... Sn. The text under test of emotional weight as follows:

$$M_{\max} = S_{\max} /(s1+s2+...+sn) \quad (3)$$

Mmax is suitable for all kinds of emotional distance recent emotional categories of specific weight.

IV. EXPERIMENTS

In this paper, we used Sina microblog OAuth2.0 interface, use the following method for experimental data set.

"索尼 SONY", "iphone5s" and "人人网 renren", "小时代 small time", "致青春 youth", "毕业季 graduation season", "郭敬明 Guo Jingming", "科比 kobe Bryant", "川菜 sichuan", "必胜客 pizza Hut" . Keywords respectively for fetching the 10 kinds of the theme of the micro blog this data each article 200, Article 2000 tweets on these data. The grabbing text should contain at least 2 or more emotional words (emotional keywords or emoticons).This article takes 2000 microblogs as experimental data.

In addition to the neutral other 10 types of emotional word are set for keywords, fetching 100 articles as experimental data.In order to verify the accuracy of the method proposed in this paper. We adopt the method of contrast test results which will automatically analyze and compare the results of subjective analysis. At the same time parameter Settings in the experiment adopts cross validation to obtain optimal parameters. In this experiment, experiment 1 is in experimental data for 11 kinds of emotional state on the analysis. Experiment 2 only considers microblog gained by the method in this paper.

In Experiment 1, extract emotional feature for every microblog first, and then through the PAD emotional space clustering analysis, judge their emotions. At the same time, using subjective evaluation method, keep five reviewers (most principles) directly to grab the microblog, and analyzing the emotional for subjective emotional display of the text, as standard answer. If the subjective evaluation result is similar to the machine sentiment analysis, the judgment is correct. I n Experiment 2, only considers text emotional attributes analysis am correct.

TABLE V. THE RESULT OF EXPERIMENT 1

	P	R	F
放松(relax)	78.10%	67.23%	72.26%
温顺(docile)	75.23%	70.22%	72.64%
惊奇(surprise)	67.97%	71.78%	69.82%
喜悦(joy)	81.54%	71.34%	76.10%
轻蔑(contempt)	64.23%	75.14%	69.25%
厌恶(hate)	81.56%	76.34%	78.86%
恐惧(fear)	69.56%	89.32%	78.21%
悲伤(sad)	79.00%	70.98%	74.78%
焦虑(anxiety)	61.11%	67.34%	64.07%
愤怒(anger)	80.80%	79.42%	80.10%

The result can be seen from table V that based on the PAD model microblog emotion, compared with the subjective judgment. We can confirm that the subjective evaluation of the reviewers, 10 classes generally higher than the current emotional judgments more emotional analysis system. Because of the clustering method of dealing with various emotions, emotional words the accuracy of all kinds of emotions more balanced; the negative emotions can also be seen in table, the accuracy is higher than that of positive emotion. We also conform to the law of subjective judgment. The angry emotion judgment in the table the highest precision, partly because of the angry words in PAD model is more obvious physical distance far than other emotional words, on the basis of also can be concluded that the deep semantic features are introduced to further determine distance between words in PAD model.

TABLE VI. THE RESULT OF EXPERIMENT 2

	P	R	F
放松(relax)	79.56%	71.54%	75.34%
温顺(docile)	77.45%	72.11%	74.68%
惊奇(surprise)	70.12%	72.67%	71.37%
喜悦(joy)	82.21%	73.43%	77.57%
轻蔑(contempt)	70.11%	76.67%	73.24%
厌恶(hate)	81.34%	77.76%	79.51%
恐惧(fear)	70.53%	88.75%	78.60%
悲伤(sad)	78.57%	75.45%	76.99%
焦虑(anxiety)	62.67%	68.65%	65.52%
愤怒(anger)	81.34%	80.54%	80.37%

According to table VI, because only text attribute analysis is correct, comprehensive F value of each emotion classification is improved, because the text emotional attributes

and the emotion word itself exists strong correlation, but also affected by the lexical and syntax, so text emotion judgment F value increase is limited. By introducing the lexical and syntactic information, can further improve the comprehensive index of classification of emotion.

V. Conclusions

With the popularity of microblog in the field of Chinese social interaction, the analysis and research of the micro blog have become the current hot research topic. The emergence of the affective computing technology made the emotional state more computability. In order to improve the level of sentiment analysis precision of the microblog, getting more accurate text emotion, our paper try to use the emotional word to model the continuous emotional space. Moreover we use rules such as emotional dictionary and emotional text feature extraction method to present the PAD emotional model. Through the establishment of PAD emotion model, we analyze the emotional state of different text corpus and description. In details, based on the extraction of microblog, emotional words, emotional characteristics and carries on the analytical judgment, the different position of emotional words in the PAD space allow the computer to obtain microblog emotion more accurate. Experimental results show that the method can be used more accurately in the emotional state compared with previous emotional polarity judgment method.

At present, this method still has many deficiencies. In the future studies, from the several following aspects we try for further research: First of all, in the voice PAD emotion model, PAD emotional state space is not an uniform standard normal distribution. In the text of the emotion has a similar problem. How to acquire the distribution of text emotional relationship will be the future research direction; Second, in the microblog analysis, a lot of emotional conditions such as the relationship of former or the following content, the user's feelings, or even the microblog sending environment at that time. There is a big problem. How to determine the effect of the environment, the people with multi-channel emotions expressed by the state, will help to improve the accuracy of emotional analysis; At the same time, also in text sentiment analysis research, the research achievements of linguistics, psychology and other disciplines will have made a significant impact on the research of this field. As an interdisciplinary field, the next step will be the study of linguistics related to discourse analysis, to find a better way to extract more accurately contained emotional characteristics.

Acknowledgment

The work is supported by the Natural Science Foundation of Anhui Province(1508085QF119) and State Key Program of National Natural Science of China(61432004). This work is also supported by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR)(201407345). This work was partially supported by the China Postdoctoral Science Foundation funded project(2015M580532). This research has been partially supported by National Natural Science Foundation of China under Grant No.61472117.

References

[1]Pang B, Lee L. Opinion mining and sentiment analysis[J]. Foundations and trends in information retrieval, 2008, 2(1-2): 1-135.

[2]Yu H, Hatzivassiloglou V. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences[C]. Proceedings of the 2003 conference on Empirical methods in natural language processing. Association for Computational Linguistics, 2003: 129-136.

[3]Hu M, Liu B. Opinion extraction and summarization on the web[C]. AAAI. 2006,, 7: 1621-1624.

[4]Wang Zhiliang. Human psychology. Beijing : China Machine Press, 2007.

[5]MEHRABIAN A. Pleasure-Arousal-Dominance: a general framework for describing and measuring individual differences in temperament[J]. Current Psychology, 1996, 14(4) : 261-292.

[6]Mao xia, Jiang lin. A Chinese vocabulary emotion modeling method based on the PAD, the national invention patent. CN102184232A[P], 2011.9.14.

[7]Zhou Shengcheng, et al. Overview on sentiment analysis of Chinese microblogging[J]. Computer applications and software. 2013.3(3):161-164 .

[8]Zhu Yanran, et al. Lexical semantic drift computation based on HowNet [J]. Journal of Chinese information. 2006, 20(1): 14-20.

[9]Wang suge, et al. Based on the methods for identifying synonyms words emotional tendency [J]. Journal of Chinese information. 2009, 23(5): 68-74 .

[10]Xie Lixing. Based on the SVM of Chinese weibo sentiment analysis research [D]. Beijing: Tsinghua university, 2011.

[11]Wang yan . Based on the analysis of co-occurrence chain weibo emotional technology research and implementation [D]. Changsha: national university of defense technology, 2011.

[12]Zhao J, Liu K, Wang G. Adding redundant features for CRFs-based sentence sentiment classification[C] . Proceedings of the conference on empirical methods in natural language processing . Association for Computational Linguistics, 2008: 117-126.

[13]Fangtao Li, Nathan Liu, Hongwei Jin, et al. Incorporating Reviewer and Product Information for Review Rating Prediction[C]. IJCAI 2011.

[14]Dasgupta S, Ng V. Mine the easy, classify the hard: a semi-supervised approach to automatic sentiment classification[C] . Association for Computational Linguistics, 2009: 701-709.

[15]Li Shoushan. Based on the Chinese emotion classification research of Stacking combination classification method [J] . Journal of Chinese information. 2010, 24(005): 56-61

[16]Xie Lixing, et al. More strategy based on hierarchy in Chinese weibo sentiment analysis and feature extraction [J]. Journal of Chinese information . 2012, 26(1): 73-83.

[17]Nasukawa T, Yi J. Sentiment analysis: Capturing favorability using natural language processing[C] . Proceedings of the 2nd international conference on Knowledge capture. ACM, 2003: 70-77.

[18]Ding X, Liu B. The utility of linguistic rules in opinion mining[C]. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007: 811-812 .

[19]Soleymani M, Chanel G, Kierkels J, et al. Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses[A]. Tenth IEEE International Symposium on IEEE, 2008 : 228-235.

[20]Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.

[21]Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS, 2013.

[22]Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. In Proceedings of NAACL HLT, 2013.