Emotion Mining Research on Micro-blog

Yang SHEN[#1],Shuchen LI[*2],Ling ZHENG[**3] ,Xiaodong REN[**4],Xiaolong CHENG[*5]
[#]School of Information Management, Research Center for Chinese Science Evaluation
Wuhan University, WuHan; China
[1]yshen@whu.edu.cn
[*]International School of Software,Wuhan University,WuHan; China
[2]6043923@qq.com
[5]284559295@qq.com
[**]School of Geodesy and Geomatics,Wuhan University,WuHan; China
[3]619284309@qq.com
[4]411649845@qq.com

***Abstract* —In order to identify the emotion expressed in corpus and to estimate the feelings conveyed by micro-blog data, in this paper, we categorize emotional words, build attitudinal words weight dictionary(WD) which consists of 1342 words, and construct self-defined negative words dictionary(NWD), degree words dictionary(DWD) and interjection words dictionary(IWD). We then process classified statistics on 2213 micro-blog items, finding that items whose first sentence express the main idea account for 23.8% of all the complex sentences, and items whose last sentence express the main idea 51.3%, respectively. We first calculate the weights of each clause in a micro-blog item, then take special treatment to the first and last sentence, finally we add up all the weights to get the emotional index(EI) of the item. Test results from the micro-blog emotion weight calculator(MBEWC) developed in C# are cross-checked and reach an accuracy rate of 80.6%.**

Keywords—attitudinal words; weight dictionary; emotional index;weight calculating

Ⅰ. INTRODUCTION

Micro-blog is an emerging multi-media mini-blog. Users can post information on the Internet at anytime and anywhere in various ways such as mobile phone, QQ, Skype and Web API. Micro-blog gains more and more attention and recognition for its real-time characteristic and short format. In China, research on the micro-blog is still at an initial stage, and is mainly confined to discuss the characteristics and use of micro-blog. Extract and analyze the content of micro-blog, and with the help of ideas in Chinese semantic research and perception science, we can find out the emotional trend of the micro-blog publisher. Emotion evaluation plays an important role in the study of authors' emotions, readers' feedback and readers' comments on a particular affair, and it has a considerable promoting effect on the establishment of emotional and psychological evaluation mechanism based on the Web.

Ⅱ. BACKGROUND AND RELATED WORK

In the research domain of human-computer interaction, emotional impact on reaction has been concerned for a long time, especially from the Efficiency of Work &Output perspective [1]. In 2006, Chung-Hsien Wu introduced a new method, including the construction of emotion rules, the representation of semantic tags and attributes, the building of emotional relevance rules and the use of independent mixture model, in order to automaticly identify the emotions in the text, and simplify emotions into 3 categories: happy, unhappy and neutral[2]. In the same year, Jon Oberlander began using personal blog corpus to assort the emotions of blog authors[3]. In 2007, Kazuyuki Matsumoto tried to determine the emotion in dialogue texts through the construction of emotion dictionary. They determine a word's weight value according to its probability of appearing in a particular emotional environment. At last, they got an accuracy of 80%[4]. In 2008, Alastair J. Gill, the fellow researcher of Jon Oberlander, processed emotional evaluation of the blog text, and found that when key words in the text have a strong preference, the recognition gets more accuracy as the length of the text grows. While the accuracy rate turns out to be low when most key words tend to be neutral[5]. In this paper, we combine both the ideas of Kazuyuki Matsumoto and Chung-Hsien Wu, that is, building weight dictionary to identify the emotions expressed in the corpus, or, texts. In the book General Psychology, emotions are classified in four categories: happiness, anger, fear and sorrow[6]. We then incorporate the later three into one category, and define emotions as positive and negative. And in regard to the limited amount of information of micro-blog, we have added the third type--neutral.

Ⅲ. Mining Method

Emotion mining assign weight values to each data item of micro-blog according to the semantic rules in the texts, and then determine the emotion by the values. We define this value as emotional index(EI)---the emotional indicator of each statement. A positive value of EI greater than zero means a positive emotion; A negative value less than zero indicates a negative mood; and zero stands for neutral emotion. The greater an EI is, the more positive an emotion would be.vice versa.

A. Construction of Weight Dictionary

The selection method of corpus is related to the coverage of the corpus. Coverage refers to the distribution or spread of corpus in different areas and domains, and different areas

978-1-4244-4158-7/09/$25.00 ©2009 IEEE

usually refer to the four-dimensional model consisting of timeline (to reflect the characteristics of the times), the space-axis (reflecting the geographical features), the subject-axis (reflecting the characteristics of knowledge), style-axis (reflecting the stylistic characteristics)[7]. However, due to the length restriction(140 characters) of micro-blog, and that currently the main concern of most users is daily topics such as whether, life, movies, love, emotion, etc. and a large number of isolated users who do not communicate with others[8], micro-blog does not have complete timeline, space-axis, subject-axis or style-axis. Therefore, in this paper, we extract keywords directly from the micro-blog and make the calculation so as to get the emotional index(EI) of the entire micro-blog. Moreover, there are a lot of spam in the micro-blog, so, before the calculation of EI, it is necessary to remove all the items that have nothing to do with emotion, like advertisement or promotions. In 2008, Linhong Xu did some research on the construction and analysis of emotion corpus[9], and had collected a large number of corpus from textbooks and literature books. This study took passages as research object, marking sentences; while micro-blog make sentences as a unit, we marking the words. At the same time, through the construction of corpus, the study has proved that words without emotions are the most common ones. This from another aspect proves the feasibility of the definition of a neutral emotion. The limitations of this study are that it did not give a corresponding solution to the division of negative words and degree adverbs. In fact, a negative word also has a big impact on the emotional coloring of a sentence, especially for the appraise tendency[10]. In view of this, we construct a negative words dictionary. Multi-denial is common in Chinese, when the number of times the negative words appears is odd, it means negative; when the number is even, it conveys positive meaning.

Firstly, we define the weight of words, introduce the concept of attitudinal words, and break the traditional classification of verbs, nouns, adjectives. Attitudinal words refer to the words that represent people's attitudes, such as joy, sadness and so on. Most of the attitudinal words are verbs, adjectives and adverbs, sometimes nouns. But not all the verbs, adjectives and adverbs can represent an attitude. Each attitudinal word has a corresponding weight that represents a degree of emotion. Define the value scope as [-20, 20], for example, ecstasy has a weight of 20, while desperate has a weight of -20. positive value means a positive emotion; a negative value indicates a negative mood; and zero stands for neutral emotion. The greater a value is, the more positive an emotion would be.vice versa. Secondly, we use the micro-blog extraction tool to extract data from a micro-blog named 'Fanfou', randomly select 2000 items from the extracted data and filter out ads, referrals, untyped text items, and there are 1403 data items remained. According to the definition of attitudinal words, we have 3 different individuals marking the words manually, the weighted average is the weight of words. From the 1403 data items we have marked out 524 attitudinal words initially and use the weighted words to build a weight dictionary. And we also composite the initial negative words dictionary(NWD) which contains 13 most common negative words. In further experiments, we have artificially expand the vocabulary of weight dictionary using a synonyms dictionary, then we have 7 individuals cross-marking the words, the weighted average values are used to construct the now-existing weight dictionary which contains 1342 attitudinal words. The building process of weight dictionary is as shown in Figure 1:

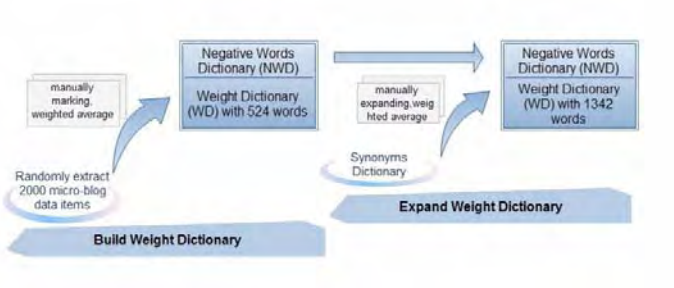

Figure1. The building and expansion process of weight dictionary.

There are two reasons why we use micro-blog but not the existing textbooks, literature works as corpus to construct weight dictionary: (1) Contents expressed in micro-blog is limited. A complete article always consists one or more different ideas, while the 140-character-long micro-blog usually conveys a complete or relatively complete, and even an incomplete idea. Micro-blog belongs to an informal written language, with a clear distinction from formal written language. (2) Web is rich in vocabulary. As micro-blog sees its emergence first in the Internet, there is a wide range of network terms in micro-blog. Those words are rare in formal written corpus but very common in micro-blog. In the follow-up study, we can use the objects of emotion mining to build corpus. Besides, we can use traditional textbooks or literatures to build a general-purpose weight dictionary, so as to process emotion mining on traditional blogs, articles, documents, or even web-comments. This is also of great value.

B. Introduction and improvement of algorithms in emotion calculator

As the weight dictionary been built, using C#, we have developed a micro-blog emotion calculation tool--ROST micro-blog emotion weight calculator(ROST MBEWC). The ROST MBEWC has four versions during improving: (1)beta1--without a negative words dictionary(NWD); (2)beta2--includes a negative words dictionary(NWD); (3) not only beta3--includes negative words dictionary(NWD), but also degree words dictionary(DWD) and interjection words dictionary(IWD), with algorithm further optimized.

In the ROST-MBEWC beta1, we only consider the way to mark the key-words. Negative words are not considered and the weight dictionary is built manually with only 524 words. Subsequently, we manually expand the weight dictionary, it then consists of 1342 words, and a negative words dictionary is also built. ROST MBEWC beta2 has some improvements on the algorithm, it takes negative words into consideration and uses the expanded dictionary. In further study, we have discovered that many attitudinal words contain negative words in themselves, so the original algorithm would calculate the negative words contained in attitudinal words more than once, which may lead to errors.

Improved algorithm is as follows:

(1) Read in a piece of micro-blog text a, according to the punctuation, break it into clauses a1,a2,a3....an;

(2) Refer to the attitudinal words dictionary(AWD), search attitudinal words in a1. Add up weights of all the resulting words, save them as v1, then delete the matched words in a1, result in a1';

(3) Refer to the negative words dictionary(NWD), search negative words in a1', count the number of negative words. When a negative word is matched, delete it from a1', result in a1". Repeat step(3) until there is no negative words contained in a1". If the number of negative words is odd, change v1 into -v1, otherwise, do not change;

(4) When a1 is fully checked, turn to next clause, say, a2. Repeat step(2) and step(3) to calculate v2;

(5) When an is calculated, add up all the numbers from v1 to vn, save as v. So v is the micro-blog's emotional index(EI);

(6) Read in another piece of micro-blog b, repeat step(1) to step (5).

Improved formula to calculate micro-blog's weight is as follows:

$$F = \sum_{i=1}^{n} \sum_{j=1}^{m} A(a_i, w_j) Neg(a_i) \quad \text{(I)}$$

$$Neg(a_i) = \begin{cases} 1 & \text{(when number of negative words in } a_i \text{ is even)} \\ -1 & \text{(when number of negative words in } a_i \text{ is odd)} \end{cases} \quad \text{(II)}$$

In this formula, a is the micro-blog, w is the weight value defined in weight dictionary. Number i stands for the i-th clause in a micro-blog, and j represents the j-th weighted word in a clause.

Improved core pseudocode is as follows:

```
token[n] = microblog[m].spliter;
    //break a micro-blog into clauses
getWeight(Dic, token[n]);
    // calculate the add-up of all the attitudinal words' weights
    in one clause
if countNumber(NegtiveWord)%2 == 1
getWeight = -getWeight;
    //if number of negative words is odd, adopt its opposite
    value
Weight = Weight + getWeight;
    //add up all the emotional values
```

C. Further optimized algorithm and its core code

In further study, degree words and interjection words have prominent impact on the emotions expressed in a sentence, and emotion tool is often error-prone in dealing with these problems. So, based on the previous development, we added a degree words dictionary (DWD) consisting 10 words and a interjection words dictionary (IWD) with 16 words. In addition, according to the Chinese habit of speaking, that usually the first and last sentence is more important, the refined algorithm takes special treatment to the first and last sentence. In order to obtain the correlation between first and last sentence, we have done related research and statistics on 2213 micro-blog items. Extract 2213 micro-blog items from Fanfou, and by statistics, the number of single sentence is 808. In the remaining 1405 complex sentences, sentences whose first clause express the main idea accounts for 23.8% of the total, sentences whose last clause express the main idea counts for 51.3%, and else 24.8%. As shown in the following table:

Version beta4 has further refinement on version beta3, it improves the judgment of degree words and interjection words. At the same time, it has special treatment for first and last sentences.

Further optimized algorithm is as follows:

1) Read in a piece of micro-blog text a, according to the punctuation, break it into clauses a1,a2,a3....an;

2) Refer to the attitudinal words dictionary(AWD), search attitudinal words in a1. Add up weights of all the resulting words, save it as v1, then delete the matched words in a1, result in a1';

3) Refer to the negative words dictionary(NWD), search negative words in a1', count the number of negative words. When a negative word is matched, delete it from a1', result in a1". Repeat step(3) until there is no negative words contained in a1". If the number of negative words is odd, change v1 to -v1, otherwise,do not change;

4) Refer to the degree words dictionary(DWD), search degree words in a1' and count their numbers. When a degree word is matched, delete it from a1', result in a1". Repeat step(4) until there is no degree words contained in a1". When a degree word is found, the clause's attitudinal weight add up one time;

5) Refer to the interjection words dictionary(IWD), search interjection words in a1' and count their numbers. When a interjection word is matched, delete it from a1', result in a1". Repeat step 5) until there is no interjection words contained in a1". Since there is too many interjection words in Chinese, we simply categorize them into two types: positive interjection words and negative interjection words. Interjection words such as Haha, Hehe, Ai, Heng, we can feel the emotion by first sight, so they have dominant impact on the main attitude of a sentence. In this regard, we assign them two special types of property weights, 1 and -1, in order to specify the main emotional attitude of a sentence. If there are negative interjection words contained in a sentence, change v1 into -v1, otherwise, do not change;

6) When a1 is fully checked, turn to next clause, say, a2. Repeat step(2) , step(3),step(4) and step(5) to calculate v2;

7) When an is calculated, according to a specific weight ratio, add up all the numbers from v1 to vn as V;

8) Read in another piece of micro-blog b, repeat step(1) to step(7).

Further improved formula is as follows:

$$F = \begin{cases} \sum_{i=1}^{n} ExcWeight(ai)\left|\sum_{j=1}^{m} A(ai,w_j)Deg(ai)Neg(ai)\right| & \text{(If there are interjection words)} \\ \sum_{i=1}^{n}\sum_{j=1}^{m} A(ai,w_j)Deg(ai)Neg(ai) & \text{(otherwise)} \end{cases} \quad \text{(III)}$$

$$Neg(ai) = \begin{cases} 1 & \text{(when number of negative words in } ai \text{ is even)} \\ -1 & \text{(when number of negative words in } ai \text{ is odd)} \end{cases} \quad \text{(IV)}$$

$$ExcWeight(ai) = \begin{cases} 1 & \text{(positive interjecti on word)} \\ -1 & \text{(negative interjecti on word)} \end{cases} \quad \text{(V)}$$

In this formula, a is the micro-blog, w is the weight value defined in weight dictionary. Number i stands for the i-th clause in a micro-blog, and j represents the j-th weighted word in a clause.

Improved core pseudocode is as follows:

```
if(token[n].hasDegreeWord)
    //check whether a clause has degree words in it
countDegreeNumber(DegreeWord) ;
    //calculate the number of degree word in a clause
getWeight=getWeight*(countDegreeNumber+1) ;
    //one more degree word, emotional value add up one
    time
if(token[n].hasExclamationWord)
    // check whether a clause had interjection words in it
getWeight=exclamationWeight * | getWeight;
    //if interjection words included in the clause, then get
    the abstract value of the clause's emotional weight,
    multiply it with the interjection word's property weight
Total=24%FirWeight+25%∑(MidWeight)+51%LasWeight;
    //Assign weight to clauses according to their positions,
    FirWeight represents for the first clause, MidWeight
    represents for the middle clauses, LasWeight represents
    for the last clause.
```

D. *Basic process of emotion calculation*

The basic process of emotion calculation in micro-blog is shown in figure2. First, break a micro-blog item into clauses according to its punctuation style; Second, search keywords in one clause and add up their weights; Third, search and calculate the number of negative words in a clause in order to determine it is a positive or negative mood. Then, micro-blog in which clause is included in the degree of search words in the dictionary, and the number of statistics, for each additional one, double their weight. Micro-blog and then to find whether the clause contains the interjection, according to the above-mentioned requirements, the right to calculate its value.; And finally, add up weights of all the clauses, then we get the emotional value of the entire micro-blog.

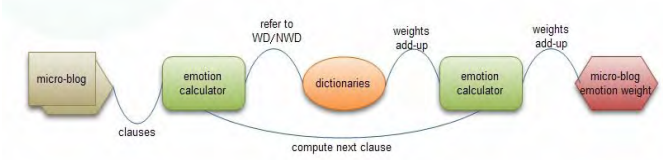

Figure2. basic process of emotion calculation in micro-blog

Ⅳ. Accuracy rate of emotion calculation

In order to verify the accuracy of the ROST micro-blog emotion weight calculator(ROST MBEWC), we designed the following five tests comparing accuracy: 1)beta1(without NWD), WD containing 1342 words and WD containing 524 words; 2)beta2(with NWD), WD containing 1342 words and WD containing 524 words; 3)beta3(with NWD,DWD and IWD, algorithm improved), WD containing 1342 words and WD containing 524 words.

Due to the limitations of the algorithm's accuracy and mistiness to categorize the emotions, in this paper, we consider these calculation results as successful: sentence express positive emotions and the calculated value is greater than 0; sentence express negative emotions and the calculated value is less than 0; sentences express no emotions and the value is 0. Other situations are considered as failure.

A. *Experiment subjects selection and test procedure*

First, select the data created on a particular day from Fanfou's extracted data sets, for example, select the 2582 data items on October 12, 2008. Then, according to the micro-blog spam types mentioned before, use ROST Content Mining (ROST-CM) to split the words, cut the data set into a number of documents in accordance with the time range. Use the traditional TFIDF deformation formula to calculate and figure out the invalid frequency words, and use these words to build a filter words dictionary (FWD). Removed hyper-links, weather forecasting and promoting information with 140 characters, the remained 2213 micro-blog items are the experiment subjects.

Due to the limitation that the software do not support English tests currently, emotion index (EI) of English text is estimated as 0. First, 2125 micro-blog items are cross-checked by 3 different individuals and marked as positive, negative and neutral. We get 605 positive items, 984 negative items and 624 neutral ones. Then, the results are checked by another 5 college students. Remove the corresponding maximum and minimum values, adopt the average number.Results are shown in the following table:

TABLE Ⅰ.CROSS-CHECKING RESULTS OF THE 2215 MICRO-BLOG ITEMS

	positive	**neutral**	**negative**
Total	605	984	624
No.1	576	940	589
No.2	588	884	559
No.3	517	946	561
No.4	575	950	534
No.5	567	924	544
Average	573	945	555

Finally, the 2215 micro-blog items are imported in each version of MBEWC and calculated. The calculating results are compared with manually checked results to determine the correctness of the calculation.

B. *Comparison of experimental results in different versions of calculator*

TABLE Ⅱ.Accuracy rate OF ROST MBEWC

Version	Release Notes		positive	neutral	negative
	WD size	Algorithm feature	Correct ratio	Correct ratio	Correct ratio
BETA 1	1324	Negative words not considered	83.7%	66.6%	51.1%
	524	Negative words not considered	78.5%	66.6%	75.1%
BETA 2	1324	Negative words considered	80.2%	66.3%	55.9%
	524	Negative words not considered	74.8%	66.9%	55.5%
BETA 3	1324	Negative words considered , algorithm improved	80.6%	75.6%	58.0%
	524	Negative words considered, algorithm improved	77.8%	78.6%	58.6%

As shown in table 2, the accuracy rate of ROST MBEWC is 82.5%. In addition, the accuracy rate increases obviously from beta1 to beta3 with the same WD size. In the version beta3, though a smaller size of WD is used, the accuracy is still higher than that of beta1 and beta2, implying that it is the algorithm that plays the key role in the calculation. When special treatment for first and last sentences has been added, the test result is more in line with the actual situation. And the accuracy rate of computing result has been greatly improved.

C. Problems in the calculation of emotion weight

In the statistical stage of accuracy, ROST MBEWC has considerable high performance in the calculation of general declarative sentences, while relatively low accuracy in the following situations: interrogative sentences and rhetorical questions; enantiosis, like self-mockery, that the literal meaning are in contrary with the actual meaning; some ancient poetry and proses; dialectical sentence from both sides of the same facts, or expressing same degree of positive and negative attitude to one thing at the same time; and specific vocabulary in some dialects. These error-prone situations need to be further studied.

Ⅴ. CONCLUSION AND FUTURE WORK

In this paper, we define attitudinal words, build weight dictionary(WD) containing 1342 words, negative words dictionary(NWD) containing 13 words, degree words dictionary(DWD) with 10 words and interjection words dictionary(IWD) consisting of 16 words, and give a method to calculate micro-blog's emotional value. This method takes into account the impact of negative words, multiple negatives in Chinese on the expression of emotions. We develop the ROST-MBEWC to calculate the micro-blog emotions in bulk, reaching an accuracy rate of 82.5%. In this paper, we compare the impact of dictionary size, negative words treatment and algorithm on the calculation accuracy. In future, we will carry out further studies in this field to refine the computing tools and improve the accuracy. At the same time, we will adopting the same method to web-comment mining.

ACKNOWLEDGEMENT

This paper is financially supported by National Natural Science Foundation of China (No. 60803080) and Ministry of Education of the P.R.C. Humanities and Social Science Youth Project (08JC870010) and the National Basic Research 973 Program of China (2007CB310806).

REFERENCES

[1] Leysia Palen, Susane Bødker, Don't Get Emotional. *Lecture Notes In Computer Science, Affect and Emotion in Human-Computer Interaction: From Theory to Applications*. 2008. Page(s): 12-22

[2] Chung-Hsien Wu, Ze-Jiang Chuang, Yu-Chung Lin, Emotion recognition from text using semantic labels and separable mixture models, *ACM Transactions on Asian Language Information Processsing(TALIP)*, Volume 5, Issue 2. Jun, 2006. Page(s): 165-183

[3] Jon Oberlander, Scott Nowson, Whose thumb is it anyway? Classifying author personality from weblog text. *Proceedings of the COLING/ACL 2006 Main Conference Poster Session*. 2006. Page(s): 627-634

[4] Kazuyuki Matsumoto, Fuji Ren, Shingo Kuroiwa, Seiji Tsuchiya, Emotion Estimation Algorithm Based on Interpersonal Emotion Included in Emotional Dialogue Sentence, *Lecture Notes in Computer Science*, Volume 4827/2007. 2007. Page(s): 1035-1045

[5] Alastair J. Gill, Darren Gergle, Robert M. French, Jon Oberlander, Emotion rating from short blog texts, Conference on *Human Factors in Computing System, Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computer system*. 2008. Page(s): 1121-1124

[6][EB/OL].[2009-2-20].http://www.pep.com.cn/xgjy/xlyj/xlshuku/shuku13/shuku17/200310/t20031027_60885.htm

[7] Zhang Pu, Some Theoretical Thoughts on Large-scale Real Text Corpus[J].*Applied Linguistics 1999*.Page(s):34-43.

[8] SHEN Yang, TIAN Chen-geng, LI Shu-chen, LIU Shi-chao, The Grand Information Flows in Micor-blog. *SEWM2009*, 2009

[9] XU Lin-hong, LIN Hong-fei, ZHAO Jing, Construction and Analysis of Emotion Corpus, *Transaction of Chinese Information*, Vol. 22, No.1. Jan, 2008. Page(s): 116-122

[10] XU Lin-hong,LIN Hong-fei,Yang Zhi-hao. Recognition Mechanism of Text Tendency Based on Semantic Understanding[J] . *Transaction of Chinese Information*, Vol. 21, No. 1. Jan, 2007. Page(s) :96-100.