

Sentiment polarity Analysis on Microblogging Hot Topic

XU Yabin^{1,2} and ZHANG Guanglei²

¹Beijing Key Laboratory of Internet Culture and Digital Dissemination Research,
Beijing Information Science & Technology University, Beijing 100101, China

²Institute of Computer, Beijing Information Science & Technology University,
Beijing 100101, China

xyb@bistu.edu.cn, guanglei0404@126.com

Abstract

Sentiment polarity analysis on the microblogging hot topic can better understand the Internet user's attitude and tendency toward a specific event, so that government can effectively guide the public opinion. Different from other methods, the sentiment polarity analysis method put forward in this paper gives full consideration to the expression characteristics of microblogging, a particular network media. It particularly considered the network languages and emoticons when calculated the value of sentiment polarity. In this method, we calculate the value of sentiment polarity of each word firstly by combining Pointwise Mutual Information (PMI) and HowNet. Secondly, modify the sentiment polarity value of a word through syntactic dependencies. Finally, accumulate the sentiment polarity value of each word, so the sentiment polarity value of a microblogging can be obtained. The sentiment polarity value of a hot topic can be obtained in the end, through accumulating the sentiment polarity value for all microblogging in a hot topic. Contrast experiment results show that, the method can obtain the sentiment polarity value of a hot topic more accurately and effectively.

Keywords: *microblogging; hot topic; sentiment polarity; syntactic dependencies; HowNet; PMI*

1. Introduction

Microblogging hot topic is a topics page that formed by presenter about current public events and the focus on most attention, which contains a large number of related microblogging. Micro blog can enter the page to send microblogging and discuss each other, while the topic page will automatically include the topic-related microblogging.

The microblogging hot topic has the following features:

(1) Large amount of data. All of these topics in today's society is the most attention and the most influential events, easily resonate. Therefore, each of the hot topics are related from tens of thousands to hundreds of thousands of microblogging. The numbers of microblogging per minute on hot topics are constantly growing.

(2) Strong real-time. Internet users have different concerns at different times, so some outdated hot topics will soon be replaced by new hot topics. Microblogging hot topics are replaced by hours.

(3) Obvious sentiment polarity. Most of the comments are from human heart, the purpose is to express their views, and therefore generally hold a clear tendency.

For all levels of government, sentiment polarity analysis for microblogging hot topic can not only obtain timely overall response on the current hot events of Internet users, extensively understand the feelings of the people and public opinion, but also to track the trend of public opinion, to provide data support for managers to make decisions, to

provide technical assistance to effectively guide public opinion, and thus to maintain social stability and national security.

2. Related Work

In recent years, with the extensive application of Internet, more and more users through various media forms, such as websites, blogs, forums, and so on, express their views and opinions, and thus result in a surge in the amount of information. Therefore, it is very time consuming and difficult to identify the sentiment polarity that the article expressed. Therefore, the automatic analysis technology of text tendency arises at the historic moment. In recent years, some progress has been made in related research^[1]. Common methods are based on emotional word statistical and machine learning. In contrast, the latter is more efficient and better. Therefore, the machine learning method has become the mainstream method.

As for Microblogging, this new social media, its words are generally limited to 140 words early this time. Zhiming Liu^[2] even point out that the actually average word number of Microblogging is only about 40 or so. This shows that Microblogging is a specific short text. In this way, the text recognition technology based on machine learning will appear the problem of sparse feature vectors. It will affect the accuracy and reliability of the calculation of sentiment polarity, and may cause results not available.

Therefore, it is more suitable to use the semantic dictionary for Microblogging's sentiment analysis^[3]. In the implementation process, it is need to collect the positive emotion words and negative emotion words in advance, constructing an emotion dictionary. In the process of calculation, first of all, it is need to calculate the number of positive emotional words and the number of negative emotional words in the text to be detected; Then, according to the amount of the number respectively to determine the sentiment polarity.

Shen Yang *et. al.*,^[4] by defining attitude dictionary, weight dictionary, negative dictionary, degree dictionary and exclamatory dictionary, detects every kinds of words contained in microblogging. Finally, calculating the sentiment index of each microblogging through weighted sum. However, due to the subjectivity of the weight, and thus may produce a certain deviation. He Fengying^[5] propose a method based on the emotional words in HowNet, calculating the similarity between words by HowNet to obtain the sentiment polarity of words. Ding Jianli *et. al.*,^[6] put forward a method to calculate the sentiment polarity of words by combining HowNet and synonyms, and taking into account the influence of adverbs on the sentiment polarity. He Fengying and Ding Jianli only take into account the relationship between the words to be calculated and the benchmark words, and did not take into account the relationship between the words to be calculated and the synonyms of benchmark words.

In addition, there are a number of other scholars' practice. For example, Zhang Shuqing, Zhou^[7] and other present a sentiment analysis method of microblogging based on subject sentences and syntactic dependence. Guibin, Yang Xiaoping *et. al.*,^[8] give a method based on the division of sense group. Huang Ting, Ji Donghong^[9] put forward a method based on graph model and multi classifier. These methods can give us some enlightenment, but they did not take other prominent characteristics into consideration, in addition to the short text feature of microblogging.

Based on the research and analysis of the existing achievements, a new method of calculating the sentiment polarity of the hot topics of Microblogging based on semantic dictionary is presented in this paper. The prominent feature is that, on the basis of the traditional emotion dictionary, the dictionary of the network language and emoticons which has strong sentiment polarity is added. On the implementation method, first of all, analyze the sentiment polarity of each word in a Microblogging about a hot topic according to the sentiment dictionary and emotion dictionary established in advance.

Then, modify the sentiment polarity according to the specific modifications relation in the sentence, through the syntactic dependencies. By accumulating the sentiment polarity value of each word, we can get the sentiment polarity value of a Microblogging. Accumulating all Microblogging's sentiment polarity value about a hot topic, eventually, we can get the sentiment polarity value of a Microblogging about a hot topic.

3. Sentiment Polarity Calculation of Words

3.1. Collection of Emotional Dictionary

In our design, the emotional dictionary contains three parts, positive dictionary, negative dictionary and emoticons dictionary. We use the emotion dictionary of HowNet^[10] and the Chinese emotion dictionary created by National Taiwan University(NTUSD). The network language commonly used and the emoticons of microblogging are collected manually. For example, the emoticons 😂 means "ha ha", with a certain sentiment polarity. Some of the network language such as "Xi Fan" (like) and "idiotic" also have an obviously sentiment polarity.

Table 1 is the part of Emoticons and frequency information used in one million Microblogging, provided by Data Hall .

It is need to mark the sentiment polarity value for other network language and the emoticons. we have carried on the sentiment polarity annotation to the emotion dictionary, the network languages and the emotions, respectively. The range of positive sentiment polarity value is [0,1]. The range of negative sentiment polarity value is [-1,0].

Table 1. Emoticons Example

| No. | emotional text | frequency | tendency |
|-----|----------------|-----------|----------|
| 1 | [haha] | 55579 | positive |
| 2 | [lei] | 42047 | negative |
| 3 | [touxiao] | 28454 | positive |

3.2. Sentiment Polarity Calculation of Words based on HowNet

Mr. Dong Zhendong, the creator of HowNet, made the following definition for HowNet: HowNet is a common sense knowledge base, take the concepts represented by Chinese and English words as describing objects, and the relationship between properties of the concepts as the basic content^[11]. This paper uses the method of Liu Qun and others^[12] to calculate the similarity between words. In the "HowNet", each word has more than one concept, the word used to explain the concept is called the original meaning. The original meaning is the smallest unit used to express the concept. Semantic expression of a notional concept is divided into 4 parts: The similarity of first independent original meaning description is denoted as sim1, the similarity of other independent original meaning description is denoted as sim2, the similarity of the relationship original meaning description is denoted as sim3, the similarity of the symbolic original meaning description is denoted as sim4. The formula is shown as (1):

$$SIM = \sum_{i=1}^4 \beta_i \prod_{j=1}^i sim_j \quad (1)$$

The similarity between the original meanings is calculated as (2):

$$sim(p1,p2) = \frac{\alpha}{\alpha+d} \quad (2)$$

Among them, D is the distance of the two original meaning in the original meaning tree, the sum is a preset parameter.

In this paper, we first select a set of benchmark words, these words have a strong sentiment polarity, and is the HowNet login word. By calculating the similarity between the word meanings of the benchmark words, we can get the sentiment polarity of these words. Here refer to the parameter values used by Liu Qun *et. al.*,^[12]:

$$\alpha = 1.6, \beta_1 = 0.5, \beta_2 = 0.2, \beta_3 = 0.17, \beta_4 = 0.13.$$

Assuming there are n commendatory terms in benchmark word list, denoted as $\{Word_{p1}, Word_{p2}, Word_{p3}, \dots, Word_{pn}\}$. n derogatory terms, denoted as $\{Word_{n1}, Word_{n2}, Word_{n3}, \dots, Word_{nn}\}$. The similarity computation function is denoted as $Sim()$, the sentiment polarity of the word W to be calculated is $Preference(w)$, the corresponding calculation formula is:

Hownet – Preference(W)

$$= \sum_{i=1}^n Sim(Word_{pi}, W) \times Preference(Word_{pi}) + \sum_{i=1}^n Sim(Word_{ni}, W) \times Preference(Word_{ni}) \quad (3)$$

Among them, $Sim(Word_p, W)$ is the similarity of a commendatory term in the benchmark word list and the word to be calculated, the value in $[0,1]$. $Sim(Word_p, W)$ multiplied $Preference(Word_p)$ means the effect of similarity on sentiment polarity. If the similarity of two words is equal to 1, the sentiment polarity of the two words is the same, and can be replaced each other. If the similarity of two words is not 1, the sentiment polarity is weighted by the similarity of words.

The shortcoming of HowNet method is that: the similarity between the word to be calculated and benchmark words mainly rely on HowNet similarity. Because HowNet is sort out manually, it is impossible that all words are included in it, and this may cause this phenomenon: Some words are very similar, but in the meaning of the original meaning tree, the distance between their original meanings is far away, this will result in a low degree of similarity. Taking a lot of tests, we found that if the similarity in two words in HowNet is very high, so they are really similar. But if the similarity of two words in HowNet is not high, they may be considered as synonyms. For example, "fashion" and "beauty" are two very similar words, while their similarity in HowNet is only 0.03 through the HowNet similarity calculation. This problem can lead to the result that the value of the sentiment polarity is not accurate.

Therefore, it is not enough to calculate the sentiment polarity value only by HowNet, it is necessary to consider other methods.

3.3. Sentiment Polarity Calculation of Words based on PMI

Turney presents a pointwise mutual information (PMI) algorithm^[13] which calculate the sentiment polarity. It calculates the co-occurrence probability between words to measure correlation strength, belonging to the calculation method of unsupervised. If the co-occurrence probability is large, in a certain range of two words, shows their strong semantic correlation. On the other hand, if their co-occurrence probability is small, then the semantic correlation between these two words is weak.

The formula for the calculation of PMI of two words is as follows:

$$PMI(Word1, Word2) = \log_2 \frac{p(Word1 \& Word2)}{p(Word1)p(Word2)} \quad (4)$$

Among them, $p(Word1 \& Word2)$ is the co-occurrence probability of two words in the text, $p(Word1)$ 、 $p(Word2)$ is the emergence probability of $Word1$ and $Word2$ respectively in the text.

$PMI(Word1, Word2) =$

$$\begin{cases} >0; \text{ } Word1 \text{ and } Word2 \text{ are related, the higher} \\ & \text{the value, the stronger the correlation;} \\ =0; \text{ } Word1 \text{ and } Word2 \text{ are independent;} \\ <0; \text{ } Word1 \text{ and } Word2 \text{ are mutual exclusion.} \end{cases}$$

This paper defines:

$$p(Word1) = \frac{1}{N} df(Word1) \quad p(Word2) = \frac{1}{N} df(Word2)$$

$$p(Word1 \& Word2) = \frac{1}{N} df(Word1 \& Word2)$$

N is the number of document in the data set, $df(Word1)$ indicates the number of times of which the $Word1$ appears in the document.

$df(Word1 \& Word2)$ is the number of times of which $Word1$ and $Word2$ common occurrence, then formula (4) evolved into:

$$PMI(Word1, Word2) = \log_2 \frac{N \times df(Word1 \& Word2)}{df(Word1)df(Word2)}$$

The method used to calculate the sentiment polarity of words using PMI as mentioned in 2.1 is as follows:

PMI – Preference(W)

$$= \sum_{i=1}^n PMI(Word_{p_i}, W) - \sum_{i=1}^n PMI(Word_{n_i}, W) \quad (5)$$

The method based on PMI has the following deficiencies: the computation of PMI depends on the corpus. If the probability of some affective words in the corpus is low, this method will not calculate the correct sentiment polarity. For example, the word "intelligence" means "smart", and it is a positive evaluation, but the number of times in the corpus appears rarely, resulting that the calculation result of PMI is 0, thus it will be judged as neutral words. So many words, this problem will lead the accuracy of the algorithm to a substantial decline.

It can be seen that, only using the method of PMI to calculate the sentiment polarity is not enough. It is need to further improve the calculation accuracy of the value of sentiment polarity.

3.4. Sentiment Polarity Calculation of Words Combined HowNet and PMI

Chinese is broad and profound, different people often use different words to express their emotions. If using a strict matching method to statistic the occurrence probability of words will inevitably lead to the problem of data sparsity. The most direct relation between Chinese words is the existence of a synonym. For example, "intelligent" and "smart" two words have synonym relationship in the semantic meaning. But, when calculating the PMI value of the word "intelligent", the value of $P(\text{intelligence} \& \text{benchmark word})$ and the value of $P(\text{intelligence})$ are very small. Thus it will leads to the sentiment polarity value of "intelligence" is zero. Because "intelligence" and "smart" are synonym, they can be replaced in a sentence, and "smart" possess high appearance frequency in the corpus. So, the PMI value of "intelligence" can instead of the PMI value of "smart".

Through the above analysis, it is necessary that each word to be calculated should conduct synonym expansion. Then, how to carry out the synonym expansion? This can be done through knowledge base of HowNet. For example, the description of "intelligence" and "smart" in HowNet knowledge base is shown as Table 2.

Table 2. Description of "Intelligent" and "Smart" in HowNet Knowledge Base

| description of "intelligent": | description of "smart": |
|-------------------------------|-------------------------|
| NO.=035218 | NO.=035228 |
| W_C= intelligent | W_C= smart |
| W_E English word | W_E=English word |
| G_C=adj [cong1 hui4] | G_C=adj [cong1 ming2] |
| S_C=Plus Sentiment | S_C=Plus Sentiment |
| Positive evaluation | Positive evaluation |
| E_C= | E_C= |
| W_E=bright | W_E=bright |
| G_E=adj | G_E=adj |
| S_E=Plus Sentiment | S_E=Plus Sentiment |
| Positive evaluation | Positive evaluation |

The meaning of some fields is as follows: W_C: specific words, G_E: part of speech, W_E: English word, DEF: concept definition. It can be found from Table 3, the content of W_E field of "intelligent" and "smart" is same, and the content of DEF field of "intelligent" and "smart" is same. Through a large number of data analysis, it is known that the "find" is universal. Thus it can be concluded: the words with the same W_E field and the DEF field in the HowNet knowledge base are synonym. In this way, we can find the corresponding synonym.

So, we can use the HowNet knowledge base to expand the synonyms for the emotional word W which is not in the emotion dictionary, and through the PMI calculation method, find a synonym which is most similar to its meaning in the HowNet benchmark word list, and think the sentiment polarity of W is similar to that of the benchmark word, and then replace it.

Thus, the key algorithm steps of the calculation of the sentiment polarity value are as follows:

- (1) For each word W to be calculated, query whether it is in the emotional dictionary.
- (2) If it is in, you can directly obtain the sentiment polarity value, execute step (6); otherwise, expand with the synonyms dictionary, get its synonym set:

$$\text{Synonym}\{W_1, W_2, W_3, \dots, W_j, \dots, W_n\}$$

- (3) add W to the synonym set:

$$\text{Synonym}\{W_1, W_2, W_3, \dots, W_j, \dots, W_n\} \text{ evolve into } \text{Synonym}\{W_1, W_2, W_3, \dots, W_j, \dots, W_n, W\}.$$

- (4) Calculate the similarity of each word in the synonym set and each word of benchmark word list by formula (5), and take the maximum similarity σ_i . At this time, the word is replaced by the corresponding word in the benchmark word list.

- (5) Calculate the PMI value of W and each word in Synonym set by formula (3), the maximum absolute value of which is recorded as $PMI - Preference(W_i)$, thus take it as the sentiment polarity value of the word.

- (6) End.

For a certain emotional word, in the above algorithm, if the sentiment polarity value finally calculated is greater than 0, the emotional word is identified as a commendatory term. If the sentiment polarity value finally calculated is less than 0, the emotional words is identified as a derogatory term. Otherwise, the emotional word is identified as a neutral word.

3.5. Sentiment Polarity Modification of Emotional Words

In Chinese words, if an emotional word is modified by adverbs, then the sentiment polarity of the emotional word may be strengthened. For example, "He is very powerful", the adverb "very" strengthen the sentiment polarity of "powerful". 181 level words chosen from Chinese level words of HowNet, after removing some of unused words, are added to the sentiment polarity word list. This can give a sentiment polarity value for each adverb, which is within [0.5, 2]. Such as "very", "super", and other words of the sentiment polarity is 2, "more", "quite" and other words of the sentiment polarity is 1.5, "reluctantly", "some" and other adverbs of the sentiment polarity is 0.5. In fact, this kind of adverb, whose sentiment polarity is 0.5, weakens the sentiment polarity of the emotional words. Of course, we can also further refine the size of sentiment polarity. However, the analysis and experiments show that, the above values can meet the needs of the calculation of the sentiment polarity.

If the emotional words are modified by the negative words, the sentiment polarity of the emotional words may change. For example, "I am not happy", the negative word "not" will change the sentiment polarity of "happy". We collected a total of 24 negative words, added to the negative word list. The sentiment polarity value of negative words is -1, need not be refined.

When there is an adverb or a negative word in the vicinity of the emotional word, with their intensity values multiplied the sentiment polarity value of the emotional word, so as to modify the sentiment polarity. If there are both a degree adverb and a negative word, taking the sentiment polarity value of the degree adverb multiplied the sentiment polarity value of the negative word, and then multiplied the sentiment polarity value of the emotional word. The negative result of the adverbs with large sentiment polarity, negative intensity is also big, inevitable. For example, the negative intensity of "very not smart" is obviously stronger than "not smart", their sentiment polarity value is -2 and -1, especially. This shows, the approach is reasonable.

It is similar to the case of a series of adverbs. For example, there are two adverbs in "very very beautiful", which is obviously stronger than that of one adverb. At this time, we use the product of sentiment polarity value of the two degree adverbs as the overall sentiment polarity value. Corresponding to the previous example, the sentiment polarity value is $2 \times 2 = 4$. Visible, this approach is also consistent with common sense.

4. Dependencies Analysis of Words in Microblogging

In front of this part, we introduce the calculation of the sentiment polarity of words. A Microblogging, usually contains a number of sentences, each sentence contains a number of emotional words (0~N) and possible adverbs and negative words. Therefore, it is not only need to find the emotional words in each sentence, but also need to find the adverbs and the negative words. But it's not enough just to find adverbs and negative words, since they can't be sure if they modify the emotional word. Because, it is entirely possible that an adverb or a negative word is used to modify a verb. Therefore, the key is to find a certain emotional word at the same time, but also to find the adverbs and negative words of modifying the emotional word. This requires an analysis of the dependencies between the words in the sentence.

In this paper, the relationship between the emotional words and adverbs and the negative words in the sentence is parsed by the syntactic dependencies. Syntax analysis

tools on the language technology platform (LTP) of Information Retrieval Lab of HIT is used for syntactic dependencies analysis^[10]. There are 4 kinds of syntactic dependencies used in this paper, as shown in Table 3.

Table 3. Dependencies Mark

| Relationship type | sign |
|-----------------------|------|
| verb-object structure | VOB |
| adverbial structure | ADV |
| attributive structure | ATT |

Figures 1 to 4 are the typical samples of syntactic dependencies of adverbial structure (ADV), verb-complement structure (CMP), verb-object structure (VOB), and attributive structure (ATT). It can be clearly seen from the figures, the two nodes connected by an arrow form a dependencies pair. The arc is marked with the type of dependencies, and the arrow points to the child node from the parent node. According to the pointer of dependencies, we can find the corresponding modification degree adverbs.

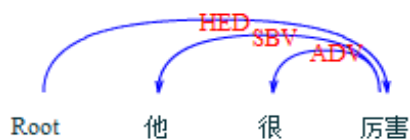


Figure 1. ADV Structure Dependencies Example

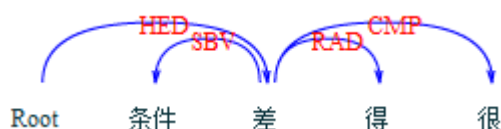


Figure 2. CMP Structure Dependencies Example

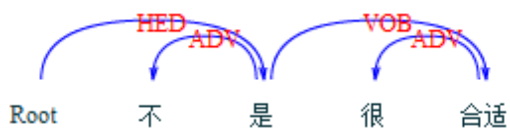


Figure 3. VOB Structure Dependencies Example

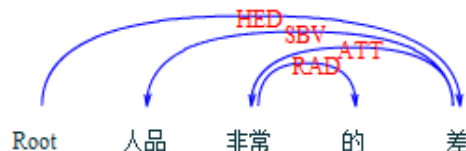


Figure 4. ATT Structure Dependencies Example

The algorithm steps for the analysis of each dependencies and the modification of the corresponding sentiment polarity values are as follows:

- (1) Detection of emotional words based on the results of word segmentation in pre-processing;
- (2) For each detected emotional word, calculate the sentiment polarity value calling the previous algorithm, and calling LTP for the following judgments;
- (3) If there is an ADV structure? If there is, according to the pointer of the ADV structure, find the corresponding degree adverbs and negative words, and find the sentiment polarity word list and negative words list, with multiply their sentiment polarity value multiply that of the emotional word, do sentiment polarity modification;
- (4) If there is a CMP structure? If there is, according to the pointer of the CMP structure, find the corresponding degree adverbs and negative words, and find the sentiment polarity word list and negative words list, with their sentiment polarity value multiply that of the emotional word, do sentiment polarity modification;

(5) If there is a VOB structure? If there is, according to the pointer of the VOB structure to find the corresponding degree adverbs, do sentiment polarity modification;

(6) If there is an ATT structure? If there is, according to the pointer of the ATT structure to find the corresponding degree adverbs and negative words, and find the sentiment polarity word list and negative words list, with their sentiment polarity value multiply that of the emotional word, do sentiment polarity modification;

(7) End.

After the sentiment polarity value of each emotional word in microblogging is calculated through the method proposed in the previous section, use the method presented in this section to modify the sentiment polarity of some words. And accumulate all the sentiment polarity value obtained after the modification, we can get the sentiment polarity value of an entire microblogging. A microblogging hot topic contains a number of microblogging, add all the sentiment polarity value of every microblogging in a hot topic, then we can get the sentiment polarity value of a microblogging hot topic. We can not only determine the sentiment polarity of a hot topic based on whether the sentiment polarity value is greater than 0, less than 0 or equal to 0, but also get the positive or negative emotional strength according to the size of value.

5. Experiment

5.1. Data Set and Data Preprocessing

The experimental data using <Chinese emotion analysis evaluation data > in this paper, provided by NLP&CC2012 (Conference on Natural Language Processing and Chinese Computing) [15]. First, use the LTP tool in the literature [5] to carry on the word segmentation. Add the network languages and emoticons in the user-defined dictionary, using the regular expression remove the advertising messages with website link and the QQ number. Then, mark emotional word by the emotion dictionary, and assign a sentiment polarity value.

5.2. Evaluation Index

Table 4. Evaluation Formula

| | | |
|--|--|--|
| Positive text accuracy rate | Negative text accuracy rate | Positive text recall rate |
| $PP = \frac{TP}{TP + FP} \times 100\%$ | $NP = \frac{TN}{TN + FN} \times 100\%$ | $PR = \frac{TP}{TP + FN} \times 100\%$ |
| Negative text recall rate | Positive text F score | Negative text F-score |
| $NR = \frac{TN}{TN + FP} \times 100\%$ | $PF = \frac{2 \times PP \times PR}{PP + PR}$ | $NF = \frac{2 \times NP \times NR}{NP + NR}$ |

The evaluation index used in this paper is accuracy rate, recall rate and F-Score, which are often used in the sentiment classification study. Among them, the accuracy is the proportion of text number whose experimental results are consistent with the actual known tendency and the total number of text containing tendencies. The recall rate is the proportion of the texts number whose experimental results are consistent with the actual known tendency and the total number of text should. Taking into account the situation that the accuracy rate and the recall rate sometimes appear contradictory, we introduce

the weighted average of the accuracy rate and the recall rate, F-Score, as the comprehensive evaluation value. The accuracy rate, recall rate and F-score are shown in Table 4.

5.3. Experimental Results and Analysis

5.3.1. Experimental Results and Analysis of Words Sentiment Polarity: The experiment result of identification accuracy for commendatory term and derogatory term is shown as Figure 5, using calculation method based on PMI, the calculation method based on HowNet and the combination of the two methods.

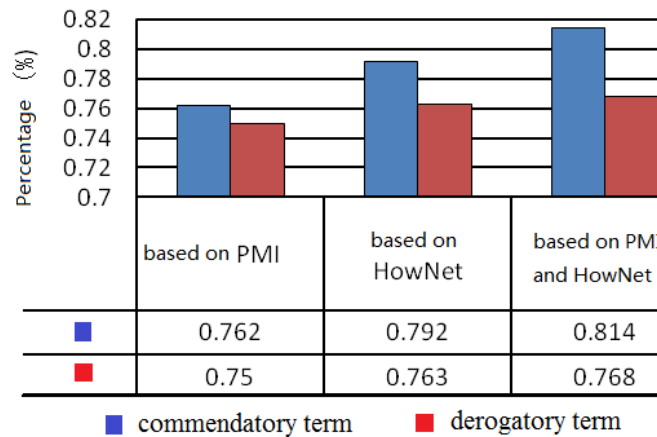


Figure 5. The Comparative Experiment Results of Recognition

From Figure 5 can be seen, the recognition accuracy of commendatory term improve rapidly, using the method of combining PMI and HowNet, but that of derogatory term improve slowly. This is because that a part of words owned derogatory meaning appear with low frequency in data set, and belong to unlogged words in HowNet. Then, these emotional words with sentiment polarity will be calculated as neutral words. But overall it seems, the method of combining HowNet and PMI is feasible for the identification of commendatory terms and derogatory words.

5.3.2. Experimental Result and Analysis of Microblogging Sentiment Polarity: In the method of PMI and HowNet combination, the value of σ_{HowNet} is very important. Its effect on the results is great. To this end, we need to find one of the most reasonable σ_{HowNet} . We try repeatedly in the range of possible values. At last, we find, when the value of σ_{HowNet} is 0.75, the accuracy rate, recall rate and F-score of positive text are the highest. The accuracy rate, recall rate and F score of negative text are also ideal. The experimental process is shown in Table 5.

Table 5. Results of σ_{HowNet} in Different Values

| σ_{HowNet} | PP | PR | PF | NP | NR | NF |
|-------------------|--------------|--------------|--------------|--------------|--------------|--------------|
| 0.6 | 0.681 | 0.69 | 0.685 | 0.721 | 0.715 | 0.718 |
| 0.7 | 0.68 | 0.695 | 0.687 | 0.738 | 0.74 | 0.739 |
| 0.71 | 0.693 | 0.664 | 0.678 | 0.744 | 0.713 | 0.728 |
| 0.72 | 0.704 | 0.696 | 0.700 | 0.723 | 0.711 | 0.717 |
| 0.75 | 0.715 | 0.742 | 0.723 | 0.738 | 0.715 | 0.726 |
| 0.8 | 0.713 | 0.733 | 0.723 | 0.711 | 0.719 | 0.715 |

It can be seen from Table 5, when computing the sentiment polarity value of a microblogging, the accuracy rate is decreased than the results in Figure 5. The reasons are as follows:

(1) Microblogging has the situation of "commendatory derogatory use" or "derogatory praise use". For example, #award implant advertisement #learning well, not as good as advertising.

(2) The polarity of emotion words is not obvious. For example, #crusade against education system under the national flag # have qualified successors.

(3) The sentiment polarity is obvious, but it does not contain emotional words, using metaphors. For example, #officer property publicity #has anyone considered to be a revolution on Jinggangshan?

In order to judge the validity of the proposed method, we compare this method with the method based on the combination of HowNet and SVM in literature [15].

Comparative experimental results are shown in Table 6. Experimental data is 50000 microblogging of the data set, covering all the hot topics in our data set.

Table 6. The Accuracy Comparison of Microblogging Sentiment Polarity

| method | correctly identified number | accuracy rate |
|--------------|-----------------------------|---------------|
| Paper method | 35369 | 70.7% |

As can be seen from table 6, the accuracy of this method is slightly higher than the method in literature [15]. The analysis shows that, the method in literature [15] is based on the classification method of SVM, and the data sparsity problem exists in the feature extraction stage for microblogging, which is not conducive to feature extraction.

Table 7 is the statistical results of the sentiment polarity of some microblogging hot topics in the data set. If the sentiment polarity value is greater than 0, It is a positive emotion. If the sentiment polarity value is less than 0, it is a negative emotion. If the sentiment polarity value is equal to 0, it is a neutral emotion. Can be seen from table 7, in the incident "Philippine warship malicious impact", the negative emotion is particularly prominent, as there are more emotional words in this event, and the tendency is obvious. In the hot topic "crazy onion", a number of microblogging are determined to be a neutral emotion, as there are more daily languages, so it is not accurate to calculate their sentiment polarity. In the "officer property publicity" hot topic, the negative emotion is more prominent, as the microblogging hot topic has more questioning and abusive words. It can be seen, the above statistical results are basically consistent with reality.

Table 7. Microblogging Hot Topics Sentiment Statistics

| hot topic | Positive emotion | Negative emotion | Neutral emotion | total |
|-------------------------------------|------------------|------------------|-----------------|-------|
| Philippine warship malicious impact | 28.3% | 42.6% | 29.1% | 702 |
| wild onion | 32.9% | 34.4% | 32.7% | 801 |

6. Conclusion

Microblogging hot topic reflects the focus of today's public opinion, analysis of the sentiment polarity not only can get the reaction on hot events of most Internet users, but also can track the trend of public opinion, provide data support for managers to make

decisions, provide technical assistance for effective guidance of public opinion, thereby maintain social stability and national security. So the research is of great significance.

On the basis of the sentiment polarity analysis of microblogging hot topics, we have fully considered the characteristics of microblogging, the special network media. The sentiment polarity computing of the network languages and emoticons are added, and combine HowNet and PMI to calculate the sentiment polarity of words. On this basis, by the fixed syntactic dependencies template, we modify the sentiment polarity of the combination words. Taking the accumulation of sentiment polarity value, we can obtain the emotional value of whole microblogging. Finally, accumulating the sentiment polarity of all microblogging of a hot topic, we can get the sentiment polarity of the hot topic.

Comparing experimental results verified the effectiveness of this method.

Acknowledgements

The National Natural Science Foundation of China under Grant No. 61370139;

The Beijing Key Laboratory of Internet Culture and Digital Dissemination Research under Grant No. ICDD201506;

The Project of Construction of Innovative Teams and Teacher Career Development for Universities and Colleges Under Beijing Municipality No. IDHT20130519;

The Project of Beijing "3-Communication & 2-Platform" standard demonstration project No. BJW201601.

References

- [1] L. Xiao-jun and D. Lin, "Survey on sentiment orientation analysis of texts [J]", *Journal of Zhejiang University (Engineering Science)*, vol. 7, no. 45, (2011), pp. 1167-1174.
- [2] L. Zhiming and L. Lu, "Empirical study of sentiment classification for Chinese microblog based on machine learning [J]", *CEA*, vol. 1, no. 48, (2012), pp. 1-4.
- [3] Z. Shengchen and Q. Wenting, "Overview on sentiment analysis of Chinese microblogging [J]", *Computer Applications and Software*, vol. 3, no. 30, (2013), pp. 161-164.
- [4] S. Yang, L. Shuchen and Z. Ling, "Emotion Mining Research on Micro blog[C]", //WebSociety, 2009. SWS 09.1st IEEE Symposium, (2009).
- [5] H. Feng-ying, "Orientation analysis for Chinese blog text based on semantic comprehension [J]", *Journal of Computer Applications*, vol. 8, no. 31, (2011), pp. 2130-2133.
- [6] D. Jian-li and C. Xiang, "Orientation analysis of Web reviews [J]", *Journal of Computer Applications*, vol. 11, no. 30, (2010), pp. 2937-2940.
- [7] Z. Shu-qing and Z. Wen, "Sentiment Analysis of Micro Blog Based on the Main Sentence and Syntactic Dependencies [J]", *Journal of University of South China (Science and Technology)*, vol. 1, no. 29, (2015), pp. 109-114.
- [8] G. Bin and Y. Xiao-ping, "Chinese Micro blog Sentiment Orientation identification Based on Sense Group Partition [J]", *Journal of Chinese information processing*, vol. 3, no. 29, (2015), pp. 100-105.
- [9] H. Ting and J. Dong-hong, "Emotional Orientation Analysis of Micro blog Based on Graph Model and Multiple Classifiers [J]", *Computer Engineering*, vol. 4, no. 41, (2015), pp. 171-175.
- [10] HowNet [EB/OL]. HowNet's Home Page. <http://www.keenage.com>. 10/12, (2011).
- [11] D. Fei-long, "HowNet", provision of paths, shared new world-Mr. Dong Zhendong on HowNet and knowledge sharing [J], *PC World China*, vol. 29, (1999), pp. 25-27.
- [12] L. Qun and L. Su-jian, "Word's semantic similarity computation based on Hownet [EB/OL]", <http://ww.keenage.com>, (2002).
- [13] Turney P. and Littman M. L., "Measuring praise and criticism: Inference of semantic orientation from association [J]", *ACM Trans. On Information Systems*, vol. 4, no. 21, (2003), pp. 315-346.
- [14] Chinese microblogging sentiment analysis evaluation [DB/OL]. <http://tcci.ccf.org.cn/conference/2012/pages/page04eva.html>, 12/09, (2012).
- [15] X. Lin-hong and L. Hong-fei, "Text Orientation Identification Based on Semantic Comprehension [J]", *Journal of Chinese information processing*, vol. 1, no. 21, (2007), pp. 96-100.

Authors



XU Yabin was born in 1962. He received the M.S. degree in Beijing from Beijing Jiaotong University in 1989. He is a professor at Beijing Information Science & Technology University. His research interests include Social network, Cloud computing, Future network, *etc.*



ZHANG Guanglei was born in 1990. He received the M.S. degree in Beijing from Beijing Information Science & Technology university in 2015. His research interests include social network, *etc.*

