A Study on Sentiment Computing and Classification of Sina Weibo with Word2vec

Bai Xue
Department of Computer Science and Technology
Beijing Foreign Studies University
Beijing, China

Chen Fu
Department of Computer Science and Technology
Beijing Foreign Studies University
Beijing, China

Zhan Shaobin
Shenzhen institute of information &technology
Shenzhen, China

Abstract

In recent years, Weibo has greatly enriched people's life. More and more people are actively sharing information with others and expressing their opinions and feelings on Weibo. Analyzing emotion hidden in this information can benefit online marketing, branding, customer relationship management and monitoring public opinions. Sentiment analysis is to identify the emotional tendencies of the microblog messages, that is to classify users' emotions into positive, negative and neutral. This paper presents a novel model to build a Sentiment Dictionary using Word2vec tool based on our Semantic Orientation Pointwise Similarity Distance (SO-SD) model. Then we use the Emotional Dictionary to obtain the emotional tendencies of Weibo messages. Through the experiment, we validate the effectiveness of our method, by which we have performed a preliminary exploration of the sentiment analysis of Chinese Weibo in this paper.

Keywords: sentiment analysis, Word2vec, Sentiment Dictionary, semantic distance

I. INTRODUCTION

Social Media, particularly the micro-blogging website Sina weibo, provide a novel way to gather real time data in large quantities directly from users. This data, which is also time-stamped and geo-located, can be analyzed in various ways to examine patterns in a wide range of subjects. Several methods have been already proposed for exploiting this rich information in order to detect events emerging in populations, track possible epidemics, model opinion polls or even infer results of elections.

With the development of Internet social network, more and more people are actively sharing information with others and expressing their opinions and feelings on Weibo. How to classify and analyze people's sentiment and mood from Weibo content they posted online quickly and effectively is getting more important. If we classify Weibo texts using traditional text classification methods, the emotional semantic information will be ignored and the semantic loss leads to the unaccepted classification results. Obviously, the traditional method is not enough for us to classify the Weibo contents according to users' sentiments. In this article, we use sentiment calculation technology to mine and analyze Internet information effectively, identifying different emotional factors and classifying Weibo contents into different sentiment categories. This is an important research topic in present emotional calculation field.

Sentiment classification has three main aspects: Subjectivity Judgment, that is to separate facts and opinions in the content. Polarity Judgment (Orientation Judgment) also known as semantic polarity judgment, that is to decide whether some sentences, words or phrases of texts are positive, negative or neutral.Gradability Judgment is to measure the extent of Subjectivity and Polarity, that is the level of the positive or negative attitude.

In this article, we classify the Weibo contents into three categories: positive, negative and neutral, and also measure the extend of the emotional intensity. First, we use the Paoding word-segmentation tool to divide Weibo contents into separate Chinese words. Second, we use 70% of processed Weibo words to train the Word2vec tool and get an extended Weibo Sentiment Dictionary. We use the distance between words to decide which category they belong to, which is based on the thought of Semantic Orientation Pointwise Mutual Information Algorism (SO-PMI). Then taking account of the effect of negative words, adverbs of degree, exclamatory sentences, rhetorical sentences and emotion icons, the Weibo content's sentiment tendency is calculated by weighing all there factors. Finally, we use the remaining 30% Weibo contents to validate the effectiveness and accuracy of our model.Figure 1. shows the process diagram of our study.

The remainder of this paper is structured as follow. In Section II, we review and discuss the related work. Section III describes the establishment of sentiment dictionary.Section IV discusses sentiment classification. Section V is about experiment results and related analysis and Section VI we talk about conclusions and future works.

Supported by the National Natural Science Foundation of China under Grant No.61170209,61370132; Program for New Century Excellent Talents in University No.NCET-13-0676; Shenzhen strategic emerging industry development funds Grant No.JCYJ20120821162230172; Guangdong Natural Science Foundation Grant No. S2013040012895, Foundation for Distinguished Young Talents in Higher Education of Guangdong, China, Grant No. 2013LYM_0076, the Major Fundamental Research Project in the Science and Technology Plan of Shenzhen Grant No. JCYJ2013032910203205.

978-1-4799-5057-7/14 $31.00 © 2014 IEEE
DOI 10.1109/BigData.Congress.2014.59

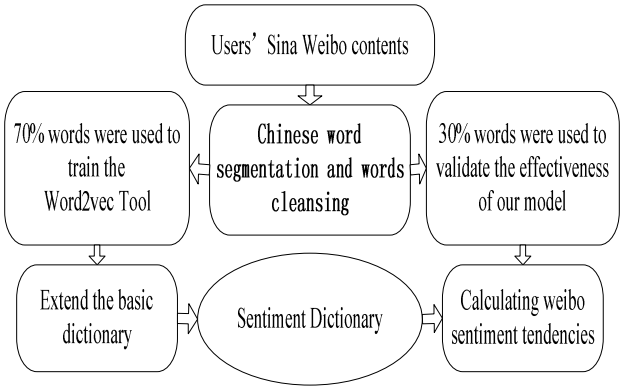

Figure 1. Flow diagram of our method

II. RELATED WORK

This section discusses related work on two key aspects of the article. Firstly, Subsection 2.1 presents sentiment analysis based on emotional dictionary. Then, Subsection 2.2 introduces the word2vec tool.

Sentiment analysis based on emotional dictionary

Foreign text sentiment analysis research began in the 1990s, the early Riloff and Shepherd [1] did some related research to construct the semantic dictionary based on corpus data. Hatzivassiloglou and McKeown [2] considering the restrictive effect of the conjunction on the adjective semantic emotional orientation on a large-scale corpus data, try making emotion tendentiousness judgment of English words. Later, more and more research considers the dependence of emotional words or phrases and feature words. Turney and others[3] use Pointwise Mutual Information (PMI) method to extend the basic positive and negative vocabulary, and then the Semantic Polarity (ISA) algorithm was used to analyze the text of the emotion, when dealing with general corpus data the accuracy rate reached 74%. Tsou and others [4] calculated the semantic orientation of words, at the same time the polarity element distribution, density and semantic intensity of news for statistical treatment in order to measure the public evaluation of politicians.

In Chinese aspect, the domestic Xu Linhong, Lin Hongfei[5] considering vocabulary and structures of sentences, extract nine semantic characteristics influencing emotional statement. Combining manual and automatic access to build an ontology emotional vocabulary, they made a preliminary attempt for sentiment analysis research. Li Dun, Cao Fuyuan adopt "emotions trend definition" priority weights calculation to get words semantic tendency of phrases from the perspective of linguistics, and analyze the characteristics of the word combination, orientation of the central concept to tendency of words calculation, so as to identify the orientation of the phrases and its intensity. The method laid the foundation to a larger size of text sentiment analysis and has its value.

Overall, the use of emotional dictionary and associated information to analyze the text emotion has the advantage of having fine granularity, analysis precision. But limited by natural language processing technology and related extraction technology, the method is easy to lose the important hidden data set pattern, making the future research work have greatly improved space.

Word2vec

The word2vec tool provides an efficient implementation of the continuous bag-of-words and skip-gram architectures for computing vector representations of words. These representations can be subsequently used in many natural language processing applications and for further research. It takes a text corpus as input and produces the word vectors as output. It first constructs a vocabulary from the training text data and then learns vector representation of words. The resulting word vector file can be used as features in many natural language processing and machine learning applications.

There are two main learning algorithms in word2vec: continuous bag-of-words and continuous skip-gram. And it allows the user to pick one of these learning algorithms. Both algorithms learn the representation of a word that is useful for prediction of other words in the sentence.

A simple way to investigate the learned representations is to find the closest words for a user-specified word. The distance tool serves that purpose. For example, if you enter 'france', distance will display the most similar words and their distances to 'France'. The word vectors can be also used for deriving word classes from huge data sets. This is achieved by performing K-means clustering on top of the word vectors. The output is a vocabulary file with words and their corresponding class IDs.

III. ESTABLISHMENT OF SENTIMENT DICTIONARY

Emotional words refer to those having sentiment tendency. They can be verbs, nouns, adjectives, adverbs and idioms[5]. In most cases, the sentiment of text is reflected by the emotional words, so they are fundamental in sentiment analysis. Sentiment Dictionary is the gather of emotional words. In the aspect of emotional tendency, it consists of two parts: positive words and negative words. In the aspect of the origin of the words, it's mainly from three sources: basic emotional dictionary, internet emotional dictionary and Weibo domain emotional dictionary. Figure2 presents the structure of the Weibo Sentiment Dictionary [6].

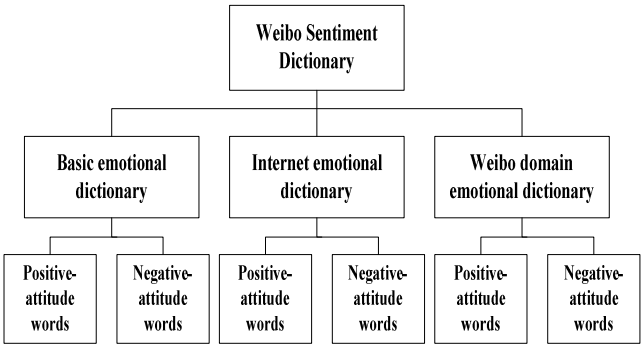

Figure 2. Constitutional diagram of Weibo Sentiment Dictionary

Basic emotional dictionary

We use the ‘ Word Set for Sentiment Analysis (beta version)’ from HowNet as our basic emotional dictionary. After deleting some words that are not commonly used, we obtain 755 positive-attitude words, 1218 negative-attitude words, 3360 positive-judgment words and 3028 negative-judgment words.

Chinese words’ Sentiment Orientation (SO) is defined as the positive or negative attitude toward opinions or some topics. It has two attributes:Polarity (P) and Intensity(I).Polarity presents the positive, negative or neutral attitude. Intensity presents the extend of the attitude. The sentiment orientation of a word is presented as follows:

$$SO(\text{word}) = (\text{P},\text{I}) \quad P \in \{-1,0,1\} \quad I \in \{0,0.5,1\}$$

In P, -1 presents negative, 0 presents neutral and 1 presents positive. In I, the bigger the number, the stronger the attitude. If SO(word).P=0, SO(word) has no value of I.

We have 10 people to label the polarity and intensity of the words in basic dictionary. Each word is labeled by at least 3 people, and the result will be decided by the majority opinions. Through this process we can make sure the accuracy and objectivity of words’ two attributes, which is the foundation of our later work.

Internet emotional dictionary

With the development of Internet, cyberwords emerged quickly. These informal words are different from formal words but also contain strong emotions. And most of cyberwords don’t conform to standard language rules.So taking into account theses cyberwords is also an important part of the construction of the Sentiment Dictionary. There is no ready-made cyberword dictionary, we collect some Internet emotional words from social network, BBS, blogs, comments and popular websites. Collecting cyberwords is a process and should be done in a regular time span.

Weibo domain emotional dictionary

We should admit the fact that basic dictionary is very limited though taking account of cyberwords. Chinese words are changing all the time and different words have different meanings and forms in different situation. So it is necessary and important to construct the Weibo Sentiment Dictionary to identify the Weibo-specific emotional words and improve the accuracy of sentiment classification. Also this dictionary should not be definite but should be extended and improved all the time.

We use the method based on semantic similarity measuring the semantic distance between Weibo words and basic dictionary words to calculate the semantic orientation. The thought of the method is like K-Nearest Neighbour (KNN): Identifying K marked sample points, calculate the similarity between marked points and new points to classify new points. Turney have presented the SO-PMI algorism to construct the sentiment dictionary, analyzing point-wise mutual information with positive words and negative words. But this method may bring data sparsity problem because of the flexible usage of Chinese words and phrases. So we use the similarity distance between words to replace the PMI and construct the new formula (In the article, we use the cosine distance to measure the similarity distance, the bigger the distance value the more similar between two words.):

Pwords= a set of words with positive orientation

Nwords= a set of words with negative orientation

SD(word1,word2)=similarity distance between word1and word2

$$SO\text{-}SD(\text{word}) = \sum_{pword \in Pwords} SD(\text{word},pword) - \sum_{nword \in Nwords} SD(word,nword) \quad (1)$$

In most cases, we take 0 as threshold value:

SO-SD(word) {
- >0 word is a positive-attitude word
- =0 word is a neutral word
- <0 word is a negative-attitude word

We empirically analyze a real social network dataset of about 2 million Weibos (tweets) crawled from Sina Weibo. After word segmentation,we use the word2vec tool to measure the semantic distance between Chinese words and phrases.The training materials are 70% of segmented Weibo contents. After training word2vec, we get distributed representation of Chinese words saved in vectors.bin. The semantic similarity is measured by Cosine value between two vectors. For two n-dimension vectors a(x11,x12,⋯,x1n) and b(x21,x22,⋯,x2n), the Cosine value is calculated as follow, the bigger the value, the nearer the two vectors are in semantic sense.

$$\cos(\theta) = \frac{\sum_{k=1}^{n} x_{1k} x_{2k}}{\sqrt{\sum_{k=1}^{n} x_{1k}^2} \sqrt{\sum_{k=1}^{n} x_{2k}^2}} \quad (2)$$

Through command ‘ ./distance vectors.bin’ , we can get the 40 most similar words from all the words trained in descending order. (Figure3 shows only 10 closest words)

Enter word or sentence (EXIT to break):

Word: ‘depressed’ in vocabulary: 5037

Word	Cosine distance
bored	0.515712
angry	0.507449
gloomy	0.505774
uncomfortable	0.501681
daze	0.484724
sentimental	0.484203
worried	0.484014
annyoed	0.478599
noisy	0.478171
hurt	0.478157

Figure 3. Result of Word2vec for closest words

Figure4 is the process of determining the semantic orientation of Chinese words and phrases to construct the Weibo Sentiment Dictionary.

```
Train word2vec using 70% of segmented Weibo contents
Obtain vectors.bin    // Distributed representation of Chinese words
For word W(P,I) in vectors.bin:
    IF W is in basic dictionary: Pass
  end IF
    ELSE
        Do command ‘./distance vectors.bin’ to get W’s 100 most
    similar words;
        Find the words in 100 words that are in extended dictionary;
            IF find no words in 100 words: W’s attitude is neutral.
            end IF
            ELSE
                Use the (2) to calculate the SO-SD(W) and to
            identify SO(W).P;
                SO(W).P = SO(W’s most similar word).P;
                Put W into the Weibo Sentiment Dictionary
     end ELSE
  end ELSE
```

Figure 4. Process of constructing the Weibo Sentiment Dictionary

IV. SENTIMENT CLASSIFICATION

In real Weibo social network, a Weibo (tweet) consists of not only emotional words but emotion icons, adverbs of degree, exclamatory sentences and other factors to consider [7]. After pretreatment of Weibo texts, we should build dictionary and decide polarity for every characteristic item in one Weibo sentence: the Weibo Sentiment Dictionary, negation dictionary, degree adverb dictionary. We can get the already-existing dictionary of negation words and degree adverbs on HowNet. Using the weight of every characteristic item in the sentence, we can get the sum of the sentiment tendency, its polarity as well as intensity.

Divide every Weibo text into sentence S1, S2,S3.....Sn. Emotional words of every sentence is W1, W2, W3 ..... Wn. The sentiment orientation of sentence Si is:

$$SO(Si) = \sum_{i=1}^{n} (SO(Wi).P * SO(Wi).I) \quad (3)$$

SO(Wi).P presents the polarity of word Wi, SO(Wi).I presents the intensity of the sentiment also the weight of sentiment value.When there are degree adverbs Di in the sentence, and are also interjections I.

$$SO(Si) = W(I) * (\sum_{i=1}^{n} (SO(Wi).P * SO(Wi).I * W(Di))) \quad (4)$$

W(I) presents the weight of interjections I. W(Di) presents the weight of degree adverb Di.

Furthermore, if the number of negation words before word Wi is odd, the polarity of word shoud be reversed. So the SO (Sentiment Orientation) if a Weibo is following:

$$SO(Weibo) = \sum_{i=1}^{n} SO(Si) \quad (5)$$

V. EXPERIMENT RESULTS AND RELATED ANALYSIS

We empirically analyze a real social network dataset of about 2 million Weibos (tweets) crawled from Sina Weibo. 70% of Weibo content are used to train the Word2vec to decide the sentiment orientation of words and phrases. There are three threshold value we should calculate to measure the result our experiment:

Precision rate(P) = Number of correctly-classified items / Number of all items put in this class

Recall rate(R) = Number of correctly-classified items / Number of all correct items in this class

F1 is the balanced value of P and R to evaluate the overall result of classification.

F1= 2*P*R/(P+R)

Result of classification of sentiment words

TALBE I. Shows the experiment result of sentiment classification of Chinese words and phrases

TABLE I. Result of experiment

	Precision rate	Recall rate	F1
Positive-attitude words	0.94	0.88	0.91
Negative-attitude words	0.96	0.89	0.92
Neutral words	0.85	0.82	0.83

We can see that positive and negative attitude words can get better results because of the characteristics of the words themselves. And also our result is better than experiment result based on traditional SO-PMI algorism.

Result of classification of Weibo contents

We calculate the remaining 30% Sina Weibos' sentiment orientation. Figure 3 and Figure 4 shows the result of our calculation of the sentiment orientation. Comparing the real sentiment orientation of Weibo contents and our calculated results, our method can get the improved and accepted sentiment orientation results. We can see from the result that, our method will get more polarized sentiment value and tends to overvalue the sentiment orientation. This because the Weibo texts in real world can have some mixed effect and can't be just added together in linear process. Words adding together tend to be counter-effective.

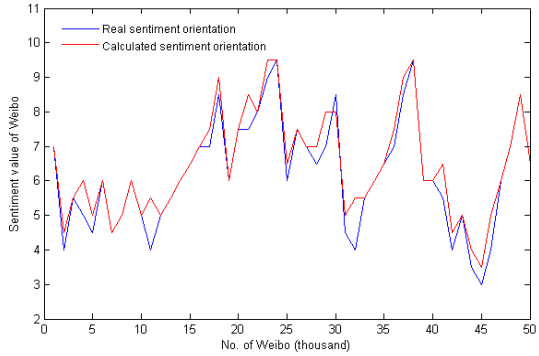

Figure 5. Positive sentiment orientation

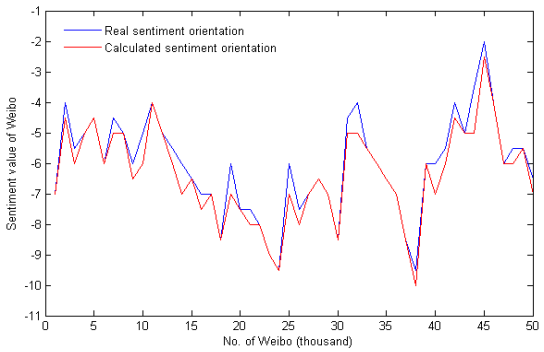

Figure 6. Negative sentiment orientation

The method is more accurate when calculating Weibo sentiment with high sentiment value.The accuracy of 83% which is higher than most of other methods is based on the fact that the algorithm cannot explicitly detect sarcasm and has also difficulties in dealing with multiple topics. For instance,'I like Mike but don' t like Thomas'would be categorized into neutral rather than positive for Mike and negative for Thomas. In order to ad-dress these limitations a more detailed language analysis would be required which is beyond the scope of this work. Also, such analysis would significantly increase the runtime of the algorithm as well as its computational requirements.

VI. CONCLUSION AND FUTURE WORK

This paper presents a new model to build a Sentiment Dictionary using Word2vec tool based on our Semantic Orientation Pointwise Similarity Distance (SO-SD) model. Then we use the Emotional Dictionary to obtain the emotional tendencies of Weibo messages. Through the experiment, we validate the effectiveness and accuracy of our method. Using the Weibo Sentiment Dictionary, we classify the Weibo contents with a relative high accuracy rate of 83%, and get even more accuracy rate when calculating the Weibo contents with strong sentiment.

Future work will include a number of as-pects. Firstly, the existing algorithm is currently being extended to include additional sentiment dimensions. This will also require the develop-ment of context specific lexicons to allow for the fact that for different topics or users certain terms may have different contextual influence on the sentiment. Moreover, such lexicons need to be further evaluated to assess the impact of social and demographic differences. Secondly, the inclusion of user specific aspects needs to be further evaluated to determine the nature of a user's opinion based on an analysis of their previous postings or their social network within the context of social networking mechanism. Finally, a semantic interpretation of the overall sentiment profile, its individual features and an explanation for changes in the profile that are based on time series analysis mechanisms to be able to identify and to track changes in sentiment.

REFERENCES

[1] RiloffE, Shepherd J. A corpus-based approach for building semantic lexicons. in: Proceedings of the secong conference one mpirical methods innatural language processing. 1997. 117~124

[2] Hatzivassiglouv, McKeownK R. Predicting the semantic orientation of adjectives. in: Proceedings of the 35th annual meeting of the European Chapter of the ACL.Morristown, NJ,USA: ACL, 1997. 174~181

[3] Turney PD,Littman M l. Measuring Praise and Critism Inference of Semantic Orientation from As sociation. ACM Trans on Information Systems,2003, 21(4): 315~346

[4] BKY Tsou, RWM Yuen, OY Kwong, et al. Polarity classification of celebrity coverage in the Chinese press . in: Proceeding of the 2005 International Conference on Intelligence Analysis. Virginia, USA,2005.

[5] Zhu Y L, Min J, Zhou Y, et al. Semantic orientation computing based on HowNet[J]. Journal of Chinese Information Processing, 2006, 20(1): 14-20.

[6] Baumgarten M, Mulvenna M D, Rooney N, et al. Keyword-Based Sentiment Mining using Twitter[J]. International Journal of Ambient Computing and Intelligence (IJACI), 2013, 5(2): 56-69.

[7] 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.

[8] Scott A. Golder and Michael W. Macy. Diurnal and seasonal mood vary with work, sleep, and day length across diverse cultures. Science, 333(6051):1878− 1881,September 2011.