The Emotion Analysis on the Chinese Comments from News portal and Forums

Jiawei Shen[1, 2], Wenjun Wang[1, 2] and Yueheng Sun[1, 2, a]

[1]School of Computer Science and Technology, Tianjin University, Tianjin 300072, China;

[2]Tianjin Key Laboratory of Advanced Networking (TANK), School of Computer Science and Technology, Tianjin University, Tianjin，300354，China;

[a]yhs@tju.edu.cn

Keywords: Chinese comments, emotion analysis, emotion lexicon.

Abstract. The people's emotion reflect their attitude towards life and society. This paper proposes an approach to measure the positive emotion of Chinese comments from news portal and forums. A large emotion lexicon is first built by the Point mutual information (PMI) and expression symbols, and the method to calculate the emotion values of each comment is also provided based on this lexicon. Experiments on manual annotation dataset show that our method achieves important improvement compared with the baseline. Finally, we investigate the emotional differences among users' comments from different web sources, and find that the more strict the rules of the network platform, the more positive the user comments.

Introduction

Human's emotions are reflected through their spoken and written languages. Internet has become one of the most effective platforms for people to access information and publish their views. When news come out, people express their mood and opinion on news websites or forums in real time. Many comments on different news and subjects have great potential emotion values.

With the deepening of reform and opening up, Chinese pay increasing attention to news which are relevant to the country and their lives. Emotion analysis is an important research topic in public opinion monitoring. Early in 1999, Bradley and Lang proposed Affective Norms for English Words[1]. Dods et al. introduces an technique to measure happiness of large-scale written expression[2] and they tested the robustness and sensivity of their technique[3]. Bol-len et al studied the influence of moods on the stock market in twitter[4]. Lewis et al. investigated geography of America by computing happiness in twitter and showed social media may potentially be used to estimate real-time levels and changes in population-level in population-level measures[5]. Seligman et al proposed the PERMA theory to measure happiness in five dimensions[6]. Chong kuang et al. translated an PERMA lexicon into Chinese and proposed a series method to measure happiness in weibo[7]. Yang Liang et al proposed a method to detect hot events based on emotion analysis[8].

Typical researches work focus on the social media and a lot of researches on happiness via mining large-scale data in Twitter and Weibo get the real time quantitative results. But there are few researches on emotions analysis on the Chinese comments from news portal and forums. Till now, the measurement of emotions is not settled yet. Lin Hongfei proposed the DUTIR which described 7 emotional classification and lexical intensity[9]. DUTIR contains seven types of emotional words: anger, disgust, appreciation, fear, happiness, sadness, and surprise. Each word has an emotion value from 0 to 9. Compare with traditional measurement that takes only happiness emotions into account, DUTIR provides more emotional dimensions to analyze emotions. The DUTIR gradually become one of primary approaches for emotion measurement.

This paper provides a detailed and quantitative approach for the emotion measurement of Chinese texts. We build three different datasets which are from Sina news, Baidu post bar and Tianya forum, respectively. Based on these datasets we expand the emotional words of DUTIR, and then propose a method to compute the positive and negative emotions based on our expanded emotion lexicon.

The rest of this paper is organized as follows. Section 2 describes the methods for emotion lexicon expansion and how to calculate the emotion values of web users' comments. In section 3, we compare our approach with the baseline on manual annotation dataset. Section 4 presents the emotion differences among the comments from different web platforms. Finally, we give concluding remarks in Section 5.

Methodology

Emotion Lexicon Expansion. Our lexicon is based on DUTIR that integrates many Chinese emotional words, idioms and proverbs, and its latest version contains more than 27000 emotional words. Since we are going to analyze the datasets from news portal and forums, we need to add the network emotional languages into the DUTIR according to the dataset's characteristics. The following methods are presented to expand the DUTIR lexicon.

Emotional words expansion based on PMI values and synonym lexicon: Point mutual information (PMI) is one of typical methods for word expansion. However, a certain number of words maybe correlative although they have opposite meanings. So we connect all synonyms and calculate all the PMI values between words in DUTIR and words in vocabulary which we obtained from Sina news' comments, Baidu post bar and Tianya forum, then we get top 4000 candidate words with higher PMI values. Then we set each candidate's emotional category and emotion value to that of its synonym which have maximal PMI value.

Emotional words expansion based on the co-occurrence with emoji (expression symbols): We assume that people use emoji to enhance the tone of the sentence. So we compute the times of emotional words of DUTIR's Co-occurrence with emoji, then we get top 100 emoji. Then we set each emoji's emotional category and emotion value to that of the emotional word which have maximal times of Co-occurrence.

We also do some manual work to filter out the noisy words. Finally we expand the DUTIR to a larger lexicon which we call it DUTIRII. The statistics of DUTIR and DUTIRII is shown in Table 1.

Table 1 Statistics of DUTIR and DUTIRII.

Lexicon	Anger	Disgues	Sadness	Fear	Appreciation	Happiness	Positive	Negative	Sum
DUTIR	387	10240	2305	1170	11067	1956	13023	14102	27125
DUTIRII	426	11563	2589	1286	12383	2208	14591	15864	30455

Methods of calculating the emotion values. Many people post comments to express their opinions and feelings on news and forums. Our basic assumption is that people express their emotions through their comments' text with emotional words intentionally or unintentionally. We only consider the text of comments and assume that the extent of positive emotion in each comment is independent.

We give a formalized definition to each comment. A comment is composed with words. After Chinese Word Segmentation a text of comment become a sequence of words.

$$c_i = \vec{w}_i = < w_{i1}, w_{i2}, w_{i3}, \dots, w_{ij} \dots w_{in} >, c_i \in C \tag{1}$$

Where n is the number of words in comment c_i, w_{ij} is the jth word in comment c_i. The C represents the comments sets, which are assigned with different meanings in different situations. For example, when we compute the value of positive emotion value of comments in one news report, the C represents all comments on the website of the news report. When we compute the positive emotion value of comments on sports news , C represents all comments on sports news.

We use $m(w_{ij})$ to represent the emotional polarity of the word w_{ij}. We define appreciation and happiness as positive emotion, in contrary, anger, disgust, fear and sadness as negative emotion. Since surprise emotion can not be distinctly defined as positive or negative, so we do not define it as positive or negative emotion.

We use $p(w_{ij})$ to represent the positive emotion value of word w_{ij} and $n(w_{ij})$ to represent negative emotion value.

$$p(w_{ij}) = \begin{cases} v(w_{ij}) & \text{if } m(w_{ij}) \text{ is positive} \\ 0 & \text{else} \end{cases} \quad (2)$$

$$n(w_{ij}) = \begin{cases} v(w_{ij}) & \text{if } m(w_{ij}) \text{ is negative} \\ 0 & \text{else} \end{cases} \quad (3)$$

Where $v(w_{ij})$ represents the emotion value of w_{ij}.The values of $m(w_{ij})$ and $v(w_{ij})$ can be obtained from DUTIR or DUTIRII.

Based on each word's emotion value, the positive and negative emotion values of comment c_i and comment set C can be defined as Equation 4~7.

$$p(c_i) = \sum_{w_{ij} \in \vec{w}_i} p(w_{ij}) \quad (4)$$

$$n(c_i) = \sum_{w_{ij} \in \vec{w}_i} n(w_{ij}) \quad (5)$$

$$p(C) = \sum_{c_i \in C} p(c_i) \quad (6)$$

$$n(C) = \sum_{c_i \in C} n(c_i) \quad (7)$$

Because the long sentences have more emotional words, which lead to the higher weight in computing average positive emotion of dataset, so we define $pa(c_i)$ to represent the average positive emotion value and $na(c_i)$ the average negative emotion value of c_i.

$$pa(c_i) = \begin{cases} \frac{p(c_i)}{ne(c_i)} & \text{if } np(c_i) > 0 \\ 0 & \text{else} \end{cases} \quad (8)$$

$$na(c_i) = \begin{cases} \frac{n(c_i)}{ne(c_i)} & \text{if } nn(c_i) > 0 \\ 0 & \text{else} \end{cases} \quad (9)$$

Where $ne(c_i)$, $np(c_i)$ and $nn(c_i)$represents the number of emotional words, number of positive emotional words and number of negative emotional words in c_i, when there is no emotional word in c_i, both $pa(c_i)$ and $na(c_i)$ is 0.

Now we can compute the percentage of the positive emotion of C, $P(C)$ as Equation 10.

$$P(C) = \frac{\sum_{c_i \in C} pa(c_i)}{\sum_{c_i \in C} pa(c_i) + \sum_{c_i \in C} na(c_i)} \quad (10)$$

Experiment on Manual Annotation Dataset

Annotation dataset. We evaluated the validity of our method by using an annotation dataset. Since there is no existing available dataset, we construct a new one for the evaluation. First 10000 comments in 8 subjects are randomly selected from Sina news, Baidu post bar and Tianya forum. Then 5 Graduate Students are asked to annotate the comments. Each comment is annotated with a positive value ranging from 0 to 10 by two students at least. If the absolute value of difference between two students' annotation is greater than 3, they ask another student to annotate it, and adopt the average of two close values. The positive value represent the degree of the comment's positive emotion, Value of 0 represent none positive emotion in the comment, value of 10 represent very strong positive emotion in the comment. Contrary to positive value, negative value, which equals 10 – positive value, represent the degree of the comment's negative emotion. When the positive value equals the negative value, the comment is an objective comment. The spam comments are first filtered out. The statistics of annotation dataset is shown in Table 2.

Table 2 Statistics of annotation dataset

Positive	Negative	Objective	Sum
2753	4199	1662	8614

Experiment Results. Our approach adopts an unsupervised learning mechanism. Based on DUTIRII, We first compute all comments' positive and negative emotion values, then the average positive emotion value of each comment's emotion in each subject. We use DUTIR lexicon and take no account of intensity of emotional words in each comments as our baseline. Under a simplified situation, we compute positive emotion ratio by Equation 11. We adopt mean squared error(MSE) and mean absolute error(MAE) to evaluate our method.

$$P(C) = \frac{p(C)}{p(C)+n(C)} \quad (11)$$

Table 3 shows the results based on annotation dataset. Our method with DUTIRII get a very close result to annotated dataset and achieve significant improvement in MSE and MAE. Since our method does not take account the emotion of words in combination, it is more suitable for large-scale texts. Please note that our method' MAE value is only 2.2%, so the improvement has significant meaning on analyzing large-scale dataset.

Table 3 Experiment results on annotation

Method	E	F	I	N	M	So	Sp	T	MSE	MAE
noIn+D	0.6222	0.4103	0.3702	0.3825	0.4946	0.3861	0.5730	0.5253	0.0038	0.0463
In+D	0.6961	0.3720	0.3409	0.3968	0.4602	0.3539	0.5312	0.4840	0.0026	0.0445
noIn+D2	0.5936	0.4914	0.4262	0.4852	0.5031	0.4416	0.5818	0.5885	0.0049	0.0645
In+D2	0.5869	0.4094	0.3516	0.4152	0.4432	0.3730	0.5280	0.4965	0.0008	0.0221
dataset	0.5556	0.4188	0.4081	0.4296	0.4470	0.3722	0.4794	0.4842	/	/

In table 3, the 'E', ' F', ' I', ' N', 'M', 'So', 'Sp' and 'T' represent comments in subject of entertainment, finance, internation, nation, military, society and technology. The 'In' and 'noIn' represent the method which take account of intensity of emotional words and the method which do not take account of intensity of emotional words. The 'D' and 'D2' represent the method is based on DUTIR and DUTIRII.

Statistics and Analysis on Large-Scale Dataset

In order to analyze the Chinese emotions in news portal and forums, we perform an in-depth analysis on the large-scale dataset based on our method. Our dataset contains 16,430,730 comments from Sina news, 815,820 comments from Baidu post bar and 7,014,424 comments from Tianya forum.

Emotions in different news categories. We get comments data on 8 categories of news based by web crawling, and compute the P(C) value of comments on each category. The result is shown in Fig.1.

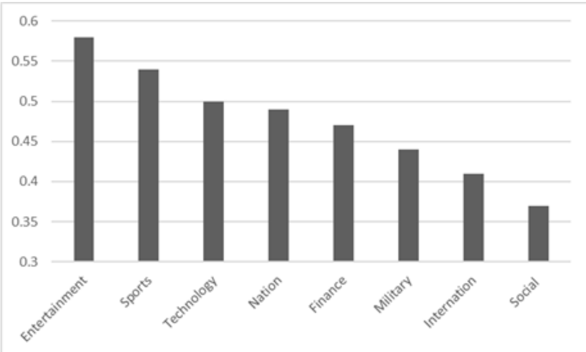

Fig. 1. Comments' P(C) value of. entertainment news, sports news technology news, national news, finance news, military news, international news and social new in Sina news.

Top 2 $P(C)$ values belong to comments on entertainment and sports, and it is easy to understand that fans always praise their idols. Comments on technology news, national news and finance news are close to neutral. Note that $P(C)$ values of comments on international news and social news are ranked in the last two. We find that international news often report regional and international disputes, while social news often report the unfortunate, unfair or emergent events. Naturally, people' comments are negative.

Emotion difference in news portal and forums. In order to find out whether different network platforms influence emotions of comments, we crawl comments of similar topics from Sina news, Baidu post bar and Tianya forum. The $P(C)$ values of different platforms are shown in Fg.2.

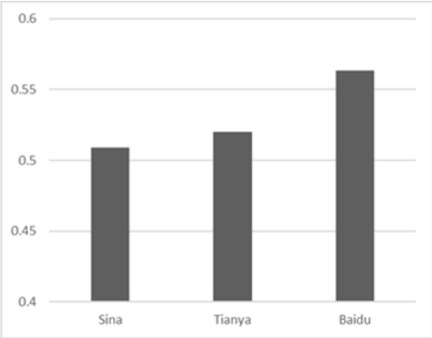

Fig. 2. The comparison of $P(C)$ values of. Sina news, Tianya forum and Baidu post bar

The Fig.2 shows that $P(C)$ values of comments on forums are high than that on Sina news. Customarily, people visit forums to communicate with others who has same hobby and forums are more normative. Besides, there are more rules in forums, basically, swearing is not allowed in most of forums. Also we find that Baidu post bar have more strict rules than Tianya forum. So it's not difficult to understand that comments on forums are more positive.

Summary

This paper proposes a method to compute the emotion values of large-scale dataset based on DUTIR and DUTIRII. We expand the DUTIR based on the large dataset obtained from Sina news, Baidu post bar and Tianya forum. Experiments on manual annotation dataset show that compared with the baseline, our method is more suitable to measure the positive and negative emotions of the comments on network communities. Also we find that the comments' emotion on forums is more positive than that on news portal, which is consistent with our intuitive knowledge.

Acknowledgment

This work was supported by the General Project of National Social Science Fund(15BTQ056) , Major Project of National Social Science Fund(14ZDB153), the National Science and Technology Pillar Program (2013BAK02B06 and 2015BAL05B02), Tianjin Science and Technology Pillar Program (13ZCZDGX01099,13ZCDZSF02700), National Science and Technology Program for Public Well-being(2012GS120302).

References

[1] Bradley M M, Lang P J. Affective norms for English words (ANEW): Instruction manual and affective ratings[R]. Technical report C-1, the center for research in psychophysiology, University of Florida, 1999

[2] Dodds P S, Danforth C M. Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents[J]. Journal of Happiness Studies, 2010, 11(4):441-456.

[3] Dodds P S, Harris K D, Kloumann I M, et al. Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter.[J]. Plos One, 2011, 6(12):: e26752.

[4] Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market[J]. Computer Science, 2010, 2(1):1–8.

[5] Mitchell L, Frank M R, Harris K D, et al. The geography of happiness: connecting twitter sentiment and expression, demographics, and objective characteristics of place.[J]. Plos One, 2013, 8(5):e64417-e64417.

[6] Seligman M E P. Flourish : a visionary new understanding of happiness and well-being[M]. Free Press, 2012.

[7] Kuang C, Liu Z, Sun M, et al. Quantifying Chinese Happiness via Large-Scale Microblogging Data[C]// Web Information System and Application Conference. IEEE, 2015:227-230.

[9] Yang Liang, Lin Yuan, Lin Hongfei. Micro-Blog Hot Events Detection based on Emotion Distribution[J]. Journal of the China society for scientific and technical information, 2012, 26(1):84-90.
[9] Xu Linhong, Lin Hongfei, Pan yu, Ren Hui and Chen Jianmei. Constructing the Affective Lexicon Ontology[J]. Journal of the China society for scientific and technical information, 2008, 27(2):180-185.