International Journal of Knowledge and Language Processing
Volume 6, Number 1, **2015**

www.ijklp.org
KLP International ©2015 ISSN 2191-2734
pp.40–47

Situation Survey of Chinese Emotion Words using on the Weibo

Pengyuan Liu and Yayun Sun

School of Information Science
Beijing Language and Culture University
No.15, Xueyuan Road, Haidian District
Beijing, China
liupengyuan@pku.edu.cn, sunyayun506@126.com

Received Janurary 2015; revised March 2015

ABSTRACT. *This paper carried though a statistical investigation to the Chinese emotion words using on the Weibo. This is the first attempt as we know to do such Chinese emotion words situation investigation research in Chinese. The data sources contain 543,780,756 items of short text selected from 60 days data of the Weibo. It investigated the frequency distribution of both emotion icons and words and analysed the different type of emotions in the Weibo. It shows that the number of the most frequency used emotion icons is 132 and 43.2% of them are positive. The frequency of the words which express "good" emotion is the most, accounted for 61.64% of the total. It also investigated the different results of the word types with corresponding to the different coverage rate and do some other statistical analysis.*

Keywords: language situation; weibo; emotion icons; emotion words

1. **Introduction.** The Ministry of education of the people's republic of China releases the Green Book: "Language Situation in China"[1,2] every year since 2006. It cover the survey of Chinese words and expressons in newspaper, radio, television and internets (news), the survey of new words of the year, popular words and phrases of the year used by Chinese Media and so on. The survey sums up the relevant policies and work of the language in various fields of the society. It explores the main problems in using language, and can reflect the changes in language using in various fields of the society. It can provide academic support for national and local administrative department, adjust the relevant policies and norms and can provide a reference for the study of language policy. It plays an important position in Chinese language life.

In the language situation in China: 2014[2], Tencent Weibo [3] data analysis report was published for the first time. It investigated a total of 45,000 Tencent weibo users and 20,307,537 items of weibo text. It investigated the basic situation, topic labels, posting behavior, words using on the Tencent weibo. Focus on the 50 high-frequency Chinese characters, the report also made comparative analysis between Tencent weibo and 2012 newspapers, radio and television, network news in the media.

Zeng Xiaobing and Yang Erhong etc. [4] investigated the use of language life in Beijing dialect and colloquial words, new words and expressions, written language vocabulary, named entities and the normalization of Chinese character. It investiaged the real language life in Beijing.

Liang Linlin and Hou Min etc. [5] made statistical analysis on the 44 years Chinese "government work report" from 1954 to 2012. It summed up the laws of words using in the report and confirms the trajectory change of the social life.

Zhangying and Zhaoxue [6] investigated official weibo news and web portal news. It shows that the in the ratio of the written language of official weibo news is higher than that of the web portal news, the topics of official weibo news are more focused than that of web portal news, there are more long sentences in official weibo news and so on.

There are also other language situation related works [7] but our research is the first time as we know to do such Chinese emotion words situation investigation research in Chinese. Emotion words situation is an important part of the whole language life. From the perspective of emotion to investigate the words of the weibo, it can help us to study the characteristics of the language used in the expression of emotion and show the distribution of emotion types expressed by Internet users on the weibo.

The rest of the paper is organized as follows: section 2 is the introduction of the weibo corpus and the emotion dictionary which the paper used to statistic. Section 3 is the investigation and analysis of the weibo corpus. We made the conclusions in Section 4. The last part is the acknowledgment.

2. **The data.**

2.1. **The Sina Weibo [8].** The Sina Weibo is the most popular weibo in China which are registed and used by at least 5 billion people. We download the weibo data from March to August 2013. After that, we random choose 60 days data as our investigation corpus which has 543,780,756 items of weibo. After word segmentation, the basic situation of the investigation corpus is as follows:

TABLE1. THE BASIC SITUATION OF THE INVESTIGATION CORPUS

days	Items of weibo	Total word freqency	Word tokens
60	543,780,756	15,093,998,811	6,630,890

2.2. **The emotion dictionary [9].** This dictionary is developed by Dalian University of Technology. It includes POS categories, emotional categories, emotional intensity and polarity, etc. It divided Chinese emotion into 7 categories and 21 sub-categories. There are total of 27466 emotional words in the dictionary.

3. **The investigation and analysis of the weibo corpus.**

3.1. **Emotion icons.** In the weibo corpus the emotion icons are tagged by "[" and "]" so we can located and statistic them by the marks. After we deleted some noisy data and merged the same emotion icons which divided by one of multi-blanks, we got 486 token of icons and the total frequency of the emotion icons is 158,133,260. We sorted the emotion icons by frequency and got the coverage rate as table 2, got the top 10 emotion icons (泪,哈哈,嘻嘻,心,怒,蛋糕,鼓掌,兔子,汗,偷笑) and their frequencies as figure 1.

TABLE 2. THE COVERAGE RATE OF THE EMOTION ICONS

Coverage rate	80%	90%	99%
Emotion icon number	39	58	132
Ratio	8.02%	11.93%	27.16%
Freqency	126,150,914	142,533,403	156,566,492

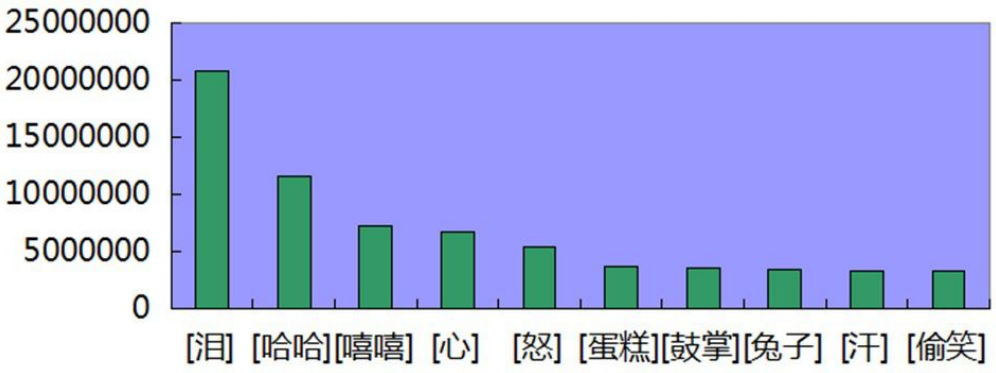

FIGURE 1. THE TOP 10 EMOTION ICONS AND THEIR FREQUENCIES

We selected the first 132 emotion icons for emotion type analysis. Then we examined the distribution of different emotion tendencies. We believe that the users must express their subjectivity more or less when they use emotion icons. The emotion can be divided into positive, negative and neutral. In order to weaken the influence of the individual's subjectivity, we have synthesized the opinion of three students after tagging the emotion of the icon. Finally, the distribution data of the 132 emotion icons are obtained, which is covering 99% of total frequency.

TABLE 3. THE RATIO OF POSITIVE, NEGATIVE AND NEUTRAL

	Positive	Negtive	Neutral

Top 132 emotion icons	43.2%	23.5%	33.3%

3.2. **Words.** We used the 7 categories of the emotion dictionary which are 恶[disgust], 好[good], 惊[surprise], 惧[fear], 哀[sorrow], 乐[happiness] and 怒[anger]. The total frequency of the emotion words is 987,462,074 which is about 6.54% of the corpus frequency shows in table 1. The top ten emotion words are: 喜欢[like], 朋友[friends], 爱情[love], 幸福[happiness], 游戏[game], 快乐[happiness], 帮助[help], 获得[get], 全新[bran-new], 支持[support]. All of these words are positive and come from the "good" or "happy" these two kinds of emotion categories.

TABLE 4. THE TOP TEN EMOTION WORDS.

Words	Frequency	category
喜欢[like]	25,109,684	Good
朋友[friends]	15,510,007	Good
爱情[love]	13,383,128	Good
幸福[happiness]	13,304,572	Happiness
游戏[game]	13,216,028	Happiness
快乐[happiness]	12,698,692	Happiness
帮助[help]	12,060,861	Good
获得[get]	11,518,332	Happiness
全新[bran-new]	10,234,941	Good
支持[support]	9,450,695	Good

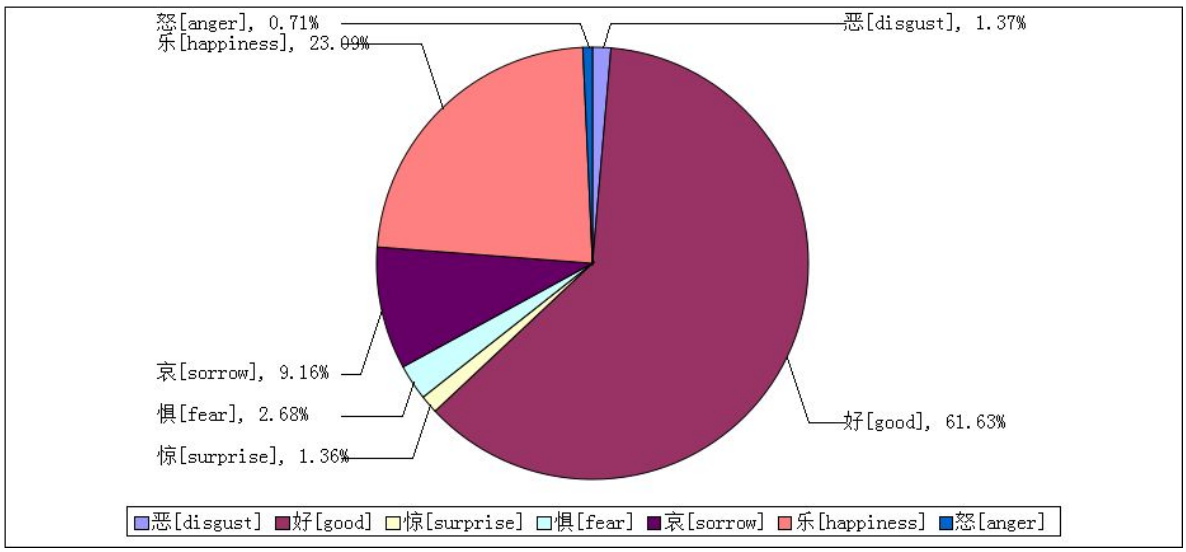

FIGURE 2. THE RATIO OF EMOTION WORD FREQUENCY OF EACH CATEGORY TO THAT OF TOTAL EMOTION WORDS

Emotion words in "Good" and "happiness" accounted for about 85% of the total emotion words using by the weibo people. The reasons may have two aspects: one is in the Chinese vocabulary system, the words which can express positive and joy are advantage; at the same time in weibo, in addition to individual users to express personal emotions, opinions , there also are many business activities and many companies publish advertisements which are always tend to use positive words. Table 5 and Figure 3 shows the distribution of the emotion words in 7 emotion categories.

TABLE 5. THE DISTRIBUTION OF THE EMOTION WORDS IN 7 EMOTION CATEGORIES

Categories	Word types in dictionary	Word types appears in weibo	Coverage ratio （%）	Frequency
恶[disgust]	2369	644	27.18	13,500,051
好[good]	11107	6258	56.34	608,631,702
惊[surprise]	228	124	54.39	13,452,718
惧[fear]	1179	655	55.56	26,441,920
哀[sorrow]	2315	1121	48.42	90,439,518
乐[happiness]	1967	1233	62.68	228,017,118
怒[anger]	388	231	59.54	6,979,047

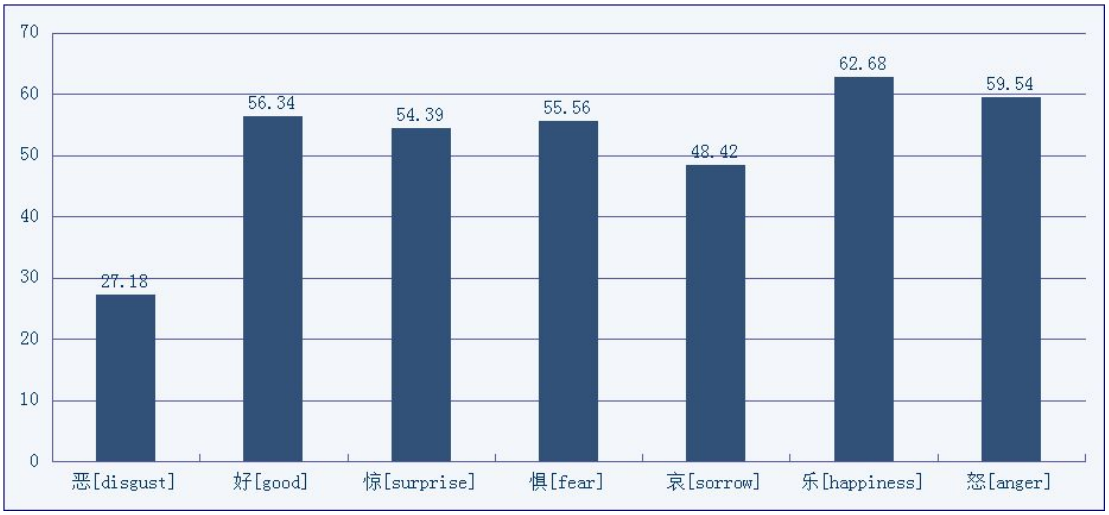

FIGURE 3. COVERAGE RATE OF THE EMOTION WORDS IN 7 EMOTION CATEGORIES

It shows that the emotion words of all the categories being used in weibo corpus and the coverage rate of different category are diferent. The most one is 乐[happiness](62.68%), the least one is 恶[disgust](27.81%) and the average rate is about 52%. We can find that about half of the emotion words in the emotion dictionary is missing in weibo corpus. There are two possible reasons: one is the emotion dictionary we used here is based on writing language whereas the weibo is more like oral language; another is the words of the weibo text are being limited, user always use simplest and most frequently used words to express their emotion. The most frequently used emotion words are listed in table 6 below.

TABLE 6 THE MOST FREQUENTLY USED EMOTION WORDS (TOP TEN)

Category	Top ten words	Coverage rate (%)
恶[disgust]	诱惑、不够、怀疑、嫉妒、借口、未必、无情、批评、犹豫、毛病	44.2
好[good]	喜欢、朋友、爱情、帮助、全新、支持、其实、推荐、希望、感谢	45.1
惊[surprise]	原来、神奇、神秘、奇迹、奇怪、好奇、惊讶、奇妙、惊人、不可思议	78.9
惧[fear]	害怕、小心、可怕、困难、紧张、厉害、重点、恐怖、恐惧、悄悄	40.5
哀[sorrow]	失去、痛苦、回忆、伤害、寂寞、消息、眼泪、伤心、孤独、悲伤	28.7
乐[happiness]	幸福、游戏、快乐、获得、开心、起来、升级、青春、成功、轻松	41.4
怒[anger]	脾气、爆发、失落、愤怒、变色、惩罚、投诉、宣泄、上火、发怒	70.0

Table 6 shows that all the top 10 emotion words are commonly used words. All these words except one word “不可思议”[inconceivable] are two-character word which is the most common used word type in Chinese. In addition, surprisely, the coverage rato of 惊[surprise] and 怒[anger] is above 70%. It seems that only 10 emotion words can express above 70% surprise or anger emotion. We did not know if it is a common law for many language or just in Chinese language yet.

We also statistic the number of word types according to the coverage rate in table 7.

TABLE 7. THE NUMBER OF WORD TYPES AND THEIR COVERAGE RATE

Category	Coverage rate above 80%	Coverage rate above 90%	Coverage rate above 99%	All
恶[disgust]	51	90	233	644
好[good]	409	815	2531	6258
惊[surprise]	11	22	56	124
惧[fear]	68	118	297	655
哀[sorrow]	88	164	470	1121
乐[happiness]	68	137	492	1233
怒[anger]	18	37	104	231
Total	713	1383	4183	10266

It shows that different emotional types have very different amount words to express emotion by weibo users. We can find that if we can master 1383 words, we can express 90% emotions though all the emotion words have exceeded 10000.

4. **Conclusion.** This paper carried though a statistical investigation to the Chinese emotion words using on the Weibo. This is the first time as we know to do such Chinese emotion words situation investigation research in Chinese. It investigated the frequency distribution of emotion icons and words and the different type of emotions in the Weibo. It shows that the number of the most frequency used emotion icons is 132 and 43.2% of them are positive. The frequency of the words which express “good” emotion is the most, accounted for 61.64% of the total. Emotion words in "Good" and "happiness" accounted for about 85% of the total emotion words using by the weibo people.

Acknowledgment. This work is supported by National Language Committee Research Project (Grant No. WT125-45).

REFERENCES

[1] Language Situation in China: 2006. The Commercial Press 2006.

[2] Language Situation in China: 2014. The Commercial Press 2014.

[3] http://t.qq.com.

[4] Zeng Xiaobing, Yang Erhong and Qiu Lina. The Word and Character Situation and characteristic of Beijing City language life. Journal of Jianghan University (Humanities Science Edition) [J], 2011(4). In Chinese: 曾小兵，杨尔弘，邱丽娜.北京城市语言生活的字词使用情况及其特征[J] . 江汉大学学报（人文科学版），2011(4).

[5] Liang Linlin, Hou Min and He Yuyin. A quantitative study of word and social transition through Chinese government work report over the years. Guangxi Social Sciences [J]. 2014(4). In Chinese: 梁琳琳，侯敏，何宇茵. 中国历年《政府工作报告》词汇与社会变迁的计量研究[J]. 广西社会科学，2014(4).

[6] Zhangying and Zhaoxue. A Computational Stylistic Analysis of News on Official Microblogs and Portal Webs. Theory and Modernization [J]. 2014(4). In Chinese: 张瑛，赵雪.官方微博与门户网站新闻语体的计量对比与分析[J].理论与现代化，2014(4).

[7] Yu Dongqing. The survey and analysis on space-time distribution of annual neologism [D]. Master degree thesis of Beijing Language and Culture University. 2013. In Chinese: 于东卿.年度新词语使用的时空分布调查与分析[D].北京语言大学硕士学位论文, 2013.

[8] http://weibo.com.

[9] http://ir.dlut.edu.cn.