A LITERATURE REVIEW OF SENTIMENT ANALYSIS ON CHINESE SOCIAL MEDIA

Zhang Yuan¹, Gu Yuwei¹, Zhao Feng¹, Wang Reimei¹*

¹School of Management, Capital normal University, Beijing, China
*5129@cnu.edu.cn

Abstract: Weibo (The mainstream social media of China) has grown to a huge scale since founded in 2014 and has become a platform for most people today to access information and express opinions. By analysing comments and forwarded posts in Weibo, we can dig deeper into the users’ perceptions and opinions about products, which is of a great commercial value. This paper summarizes the related researches on the sentiment analysis of Weibo in China and abroad since 2013, and analyses the analysis of Weibo sentiment analysis based on different methods.

1. Introduction
With the rapid development of the Internet, the number of Weibo’s monthly active users has increased to 411 million, and Weibo has become the mainstream platform for Chinese people to acquire, disseminate, and share information. With the sentiment analysis on Weibo, companies can understand customer evaluations of products and services as well as product preferences; the government can understand the people’s evaluation of new policies, the dissatisfaction with old policies, and the new needs of the society. This paper will sort out the current procedure of Weibo sentiment analysis research, summarize the existing problems, and further explore the research prospects of Weibo.

1.1 Sentiment Analysis
Sentiment analysis, also known as opinion mining. The goal of this research area is to analyse the views, emotions, opinions, attitudes, and emotions expressed by people about entities and their attributes. These entities can be a variety of products, services, institutions, individuals, events, issues or topics [1].

1.2 Weibo Sentiment Analysis Process
Step1. Acquire the data. There are two main ways to get data. One is through the Weibo open platform API, the other is through web crawling. The former is an official development tool of Weibo, which can obtain Weibo data efficiently and conveniently. However, due to API access restrictions, only part of the data can be obtained through the API. The latter combines APIs with web crawlers to get a lot of data. Thus, the latter is the main way to get Weibo data now.

Step2. Perform data preprocessing. Data preprocessing techniques include natural language processing techniques such as word segmentation, part-of-speech tagging, and syntax analysis. In addition to allowing users to post text messages, Weibo can also post images, emoticons, videos and more. Due to the instability of the meaning of the latter, we usually pre-process the data before the sentiment analysis to screen out all the data of the latter. We can see from figure 1 that the format of
the image is different from the text in the crawl of the comment, so it can be directly excluded during the pre-process. Most of the expressions are in the format of “/ + text”, which can be filtered out when we process data. The video is in the form of a hyperlink, which can be screened out while the data is being processed with the expression.

Figure 1 Weibo Page

Step3. Sentiment analysis. At present, there are two different methods of sentiment analysis: sentiment analysis based on sentiment dictionary and machine learning. These research methods are common issues in today’s sentiment analysis research and will be introduced in detail later.

Step4. Result presentation. There are many ways to display the sentiment analysis results. Most of them are now textual, and there are also few research works on visualization.

1.3 Research Scope
Sentiment analysis is involved in this field is very broad, including many related research tasks, such as emotional analysis, opinion mining, opinion analysis, opinion information extraction, emotion mining, subjective analysis, orientation analysis, and comment mining. These tasks are slightly different among each other, but they can all be classified as sentiment analysis tasks.

1.4 Different Methods of Sentiment Analysis
Comprehensive 245 research literature on Weibo sentiment analysis, the main methods of Weibo sentiment analysis are sentiment dictionary and machine learning. Below the authors describe the two methods of sentiment analysis separately.

2. Method Based on Sentiment Dictionary
The analysis method based on sentiment dictionary is based on the sentiment orientation of the words provided by the sentiment dictionary, so as to analyse the text under different granularities.

The sentiment dictionary method analyse the text with different granularities based on the sentiment orientation of the words provided by the sentiment dictionary.

Traditionally, according to the granularity of text, the task of sentiment analysis is categorized into multiple levels such as words, phrases, attributes, sentences, chapters, and topics [1, 2]. Words, phrases, and attributes belong to fine-grained levels. Sentences, chapters and the topic belongs to the coarse-grained hierarchy [3]. In the topic of sentiment analysis of different granularities, The sentiment dictionary plays different roles. At the word and phrase level of sentiment analysis task, the analysis of the emotional tendency of words and phrases is equivalent to constructing a sentiment dictionary; at the sentence and chapter level sentiment analysis tasks, we accumulate the scores of the emotional words or compare them by quantity. The emotional polarity of the sentence is obtained; in the attribute-level sentiment analysis task, the emotional word corresponding to the attribute word is found, and the method of using the word and the phrase to judge the emotion is the emotion of the attribute word.

At the beginning of sentiment analysis, only coarse-grained text analysis is performed. Later, a long article may not only express a point of view, but also may change the emotions of the first half and the second half, and the content of Weibo cannot be limited to 140 words. The chapter-level
sentiment analysis does not apply to short texts, so a fine-grained sentiment analysis was produced. Compared with chapter-level sentiment analysis, words can accurately express positive or negative views, and sentences can express changes in emotions and transitions. The short text of Weibo usually has fewer words, a clear theme, and the evaluation object is obviously suitable for fine-grained sentiment analysis.

Figure 2 2012-2018 research statistics on sentiment analysis based on sentiment dictionary

Figure 2 shows the year analysis of the Weibo sentiment analysis paper. It is concluded from the graph data that the number of Weibo sentiment analysis based on sentiment dictionary is the most in 2016, and it is inferred that 2016 is the heyday of the development of sentiment dictionary. So far, people's research on emotional dictionaries has been quite mature. After 2016, the sentiment analysis based on sentiment dictionary not only corrects the word segmentation precision of the original dictionary, but also adds the current network buzzwords in the original dictionary, and expands the text research granularity on the original basis, and also increases the analysis of sentence relationships and emoticons. Etc. The analysis greatly improves the accuracy of sentiment dictionary analysis.

In the past two years, due to the continuous update of Weibo, many expressions and picture comments have been added. Therefore, there are many papers on how to analyse emotions from punctuation and expressions, and even emotion analysis of pictures. All in all, people are looking for the most appropriate granularity for Weibo text to improve the accuracy of sentiment analysis.

In the Weibo commentary, the emoticon in the comment text has become a trend. The emoticon has a great influence on the emotion of the text, and sometimes even completely changes the emotional polarity in the text. Therefore, the emoticon structure is also constructed. Into the sentiment dictionary. However, the sentiment analysis of emoticons can't add emoticons and text emoticons. Tan et al. proposed a neural network model of emoticon attention mechanism in order to solve the emotional analysis problem caused by emoticons [4].

In Weibo's publication, the form of Weibo is more and more diverse. In addition to text, expression packs, and pictures, it also includes audio and video. Among them, the proportion of pictures is quite large, and almost every tweet is equipped with pictures. In this huge amount of picture tweets, huge commercial value and social value are hidden. For the extraction of the value of pictures, Deng and Tan et al. proposed a method of graphic and textual sentiment analysis based on transfer variables [5]. This method is not only suitable for sentiment analysis of pictures, but also for sentiment analysis of special symbols, including forwarders and emoticons.

In text information, in addition to the emotions contained in the text itself, punctuations also contain certain emotions. Punctuation marks, like emoticons, can change the emotional polarity of text, but unlike emoticons, emoticons themselves are emotional. However, the punctuation itself is not emotional, and the punctuation emotion is dependent on the text information. In view of the shortcomings and shortcomings of the current text sentiment analysis method for the Weibo text sexy analysis, Li et al. use a sentence-level dictionary-based improvement. Syntactic analysis proposes an improved syntax analysis algorithm [6].
It is necessary to construct a good sentiment dictionary. The quality of a sentiment dictionary directly affects the quality of the sentiment analysis. How to construct a sentiment dictionary? Yang et al. proposed that the sentiment dictionary should not only be qualitatively emotionally labelled [7]. It is also necessary to quantitatively identify the emotional intensity; Feng et al. proposed that the construction of the sentiment dictionary is mainly to judge the emotional polarity of lexical emotions[8], that is, the process of dividing vocabulary construction into the derogatory, derogatory and neutral dictionary.

At present, the sentiment lexicon used in sentiment analysis mainly includes HowNet sentiment dictionary, NTUSD sentiment dictionary, and the emotional vocabulary ontology of Dalian University of Technology[9]. In the library of the emotional and emotional ontology at Dalian Institute of Technology, the emotions are divided into seven categories: music, goodness, anger, sadness, evil, fear, and shock.

Practically, there are some special words in different fields. There are some special social words in Weibo comments, such as "eat chicken" (game players get first in "Jesus Escape: Big Escape"), "Little Milk Dog" (a type of boyfriend, indicating age Small simple and well-behaved), paddling (lazy), etc. If the existing emotional dictionary is used to analyse their emotional polarity, which may lead to a bias in sentiment analysis. Ma et al. generated an emotion dictionary for social networks [10] in order to conduct an emotional analysis of online public opinions in a timely and effective manner. Cui et al. built a special fire-fighting emergency online lyric sentiment dictionary [11]. The existing Chinese and English dictionaries are more mature than the development of minority Chinese word dictionaries. Yan et al. People constructed Tibetan sentiment lexicon for Tibetan [12], and Liu et al. expanded the Uyghur sentiment lexicon [13].

3. Method Based on Machine Learning

![Graph showing research statistics on machine learning related to sentiment analysis]

Figure 3 shows some papers about Weibo sentiment analysis based on machine learning from 2012 to 2019. The number of papers in the graph shows an exponential upward trend. It can be seen that people find that using machine learning to conduct Weibo sentiment analysis based on machine learning can achieve better results.

Unlike extracted from emotion dictionary, machine learning enables computers to simulate or implement human behaviours to acquire new knowledge or skills. Literature [14] points out that machine learning in sentiment analysis can be divided into three types: supervised, semi-supervised and unsupervised. Supervised machine learning can be used to predict the results of new data by learning a function from a given training data set. Supervised machine learning usually requires manual annotation of the actual output of the training set samples. The unsupervised machine learning aims at learning from unlabelled samples to discover the structural knowledge in the set of samples. Supervised and semi-supervised machine learning are currently widely used.
The earliest machine learning method was used only for text classification. In 2013-2014, the research focus of Weibo sentiment analysis based on machine learning is mainly on the methods of feature extraction (including word and word frequency, part of speech, etc.) and the construction of corpus. Some scholars have extended the features of sentiment analysis and tried to add semantic rules to feature extraction. Reference [14] summarizes and introduces three main sentiment analysis models: naive Bayes classifier, maximum entropy algorithm, and support vector machine (SVM) classifier. The maximum entropy algorithm does not assume that the features are statistically independent, so the researchers prefer to use the maximum entropy algorithm, the probability formula of which is as follows:

$$P(C_j|d_i) = \frac{1}{Z(d_i)} \exp\left(\sum_m \alpha_{m,C_j} F_{d_i}(m, C_j)\right)$$

$Z(d_i)$ is a normalized function, $\alpha_{m,C_j}$ is the weight parameter of feature $m$ for category $C_j$, $F_{d_i}(m, C_j)$ is a function of the feature $m$ and the class $C_j$ in the document $d_i$, which takes 1 when $m$ appears in $d_i$, and 0 otherwise.

During this period, SVM and IG (information gain) and TF-IDF (word frequency-inverse document frequency) as feature weights were also explored in 2014 literature [15-16]. The combination of these three has the best effect on Weibo emotional classification. In addition, Sun et al. also proposed emotional features of emoji and semantic association rules [17]. Because of its good generality, high classification accuracy and fast classification speed, SVM has become the focus of later research.

So immediately in 2015 proposed SVM and CRF multi-feature combination of Weibo sentiment analysis method. Li et al. proved that the effect of affective analysis is optimal by using various text features and selecting different feature combinations to carry out multi-group experiments [18]. The correct rate of SVM model is 88.72% when the feature combination of part of speech, affective word and negative word is selected, and the correct rate of CRF model is 90.44% when the feature combination of affective word, negative word, degree adverb and special symbol is selected.

After 2016, some scholars continue to study Weibo sentiment analysis method which combines machine learning and semantic rules. Jiang et al. proposed that the diversified emotional information obtained by the regular method should be transformed to expand and embed the basic feature template to form a more effective fusion feature template [19]. Another part of the scholars began to use neural networks for sentiment analysis.

In 2018, the new research on Weibo sentiment analysis based on convolutional neural network became an upsurge, which separated the machine learning method from the emotion dictionary method, avoided the artificial intervention in the experiment process as far as possible, and realized the unsupervised machine learning method. Feng et al. in order to accomplish the above purpose, a method based on a convolution neural network and attention model [20], was proposed for sentiment analysis. Experiment results show that the accuracy of this method is obviously higher than that of the traditional machine learning method. Wang et al. improved the possibility of losing too many semantic features in the forward propagation of convolutional neural networks by incorporating the tree-type long-term and short-term memory neural networks [21].

From 2013 to 2018, the accuracy of machine text analysis has been gradually improved and has begun to explore to minimize human participation in sentiment analysis. I believe that at the end of the day, sentiment analysis will become a simple universal thing. When we want to know what Weibo users think about an event, we just have to type in the subject of the event and press the button. The computer can automatically help complete all the calculations.

4. Conclusion
The research methods of machine learning are now rising and gradually replacing the methods of sentiment dictionary. However, through research, we find that whether it is an sentiment dictionary or
a machine learning method, we will learn more or less from foreign sentiment analysis methods. Therefore, the main problems of Chinese Weibo sentiment analysis are

1) The research system is not perfect

   The Chinese Weibo sentiment analysis research is lagging behind in foreign countries for about 9 years. In the early years of Chinese research, many articles directly used the well-known foreign micro-blog sentiment analysis steps and algorithms, but did not find the algorithm problem and then the process of generating new algorithms, resulting in the initial articles mostly in the topic of the topic, between each other There is no association, just a description of the different algorithms. It was not until 2015 that the problem began to improve gradually, but it could be better.

2) Chinese processing problems

   The formation of Chinese and foreign languages, the habits and characters of people using sentences are different. There is a big gap between the sentiment analysis of Chinese Weibo and the sentiment analysis of Twitter. If we simply copy the research method of Twitter, we should not. Develop a true sentiment analysis dictionary and develop our own emotional analysis system.

   The future of Weibo sentiment analysis is evolving along with the continuous development of computer machine learning and artificial intelligence. Sentiment analysis will gradually transformed from manual to fully automatic development, from huge data collection and simulation process to fast and easy acquisition results of sentiment analysis.

   In the past ten years, sentiment analysis has made important progress in theoretical research, technical methods and applications, showing its development potential and application prospects in all aspects of society. China has entered a new heyday in this field. At the same time, there are still a lot of problems remaining unsolved in sentiment analysis. The establishment of Chinese sentiment analysis system and the innovation of technical methods will become important issues in the development of sentiment analysis. We summarize the past research on the sentiment analysis in Chinese Weibo. One of the purposes is to objectively review the past research process, correctly examine the issues of sentiment analysis in Chinese Weibo, and also share these with foreign scholars. Second, We look forward to the rapid breakthrough of the problem of emotional analysis in the future and apply it to every industry to give full play to its value and contribute to the development of our country.

Acknowledgements

This research was funded by the Scientific Research Foundation of Beijing Municipal Education Commission (KM201810028021, KM201610028019).

References

[1] Ke, W., Rui, X.: 'A summary of the automatic construction methods of sentiment lexicon'. Acta Automatica Sinica, 2016, 42, (4), pp. 495-511
[2] Guolan, C.: 'Weibo sentiment analysis based on sentiment lexicon and semantic rules'. Intelligence Exploration, 2016, (2), pp.1-6
[3] Xiaobo, C., Guangchao, L.: 'A review of fine-grained sentiment analysis'. Library and Information Service, 2017, 61, (5), pp. 132-140
[4] Hao, T., et al.: 'Weibo sentiment analysis model based on emoticon attention mechanism'. Computer Application Research, 2019, (9), pp.1-7
[5] Pei, D., Changgeng, T.: 'Image fusion based on transfer variables, microblog sentiment analysis'. Computer Application Research, 2018, 35, (7), pp. 2038-2041
[6] Xuehong, L., Hui, G., Yitao, Y.: 'Weibo sentiment analysis based on improved dependency syntax'. Computer and Digital Engineering, 2017, 45, (3), pp. 506-511
[7] Xiaoqing, Y., et al.: 'Automatic Construction and Optimization of Emotional Dictionary Based on Word2Vec'. Computer Science, 2017, 44, (1), pp. 42-47+74
[8] Chao, F., et al.: 'A cross-domain Chinese sentiment dictionary construction method based on word vector'. Data Acquisition and Processing, 2017, 32, (3), pp. 579-587
[9] Songxia, X., Shuyi, Z.: 'A Chinese Emotional Dictionary Expansion Method Based on Mixed Features'. Computer Engineering and Science, 2016, 38, (7), pp. 1502-1509
[10] Bingnan, M., Yongfeng, H., Beixing, D.: 'Emoticon-based social network sentiment dictionary construction'. Computer Engineering and Design, 2016, 37, (5), pp. 1129-1133
[11] Yanbiao, C., et al.: 'Research on the construction of lyric sentiment dictionary for fire emergency'. Intelligence Magazine, 2018, 37, (10), pp. 154-160
[12] Xiaodong, Y., Tao, H.: 'Sentiment Classification of Tibetan Text Sentences Based on Emotional Dictionary'. Chinese Journal of Information, 2018, 32, (2), pp. 75-80
[13] Ruolan, L., Mei, N., Marhaba, E.: 'A Study of Uyghur Emotional Thesaurus Based on Conjunctions'. Chinese Journal of Information, 2018, 32, (3), pp. 49-54
[14] Yan, L.: 'Research on Key Issues in Text Sentiment Analysis'. PhD thesis, Beijing University of Posts and Telecommunications, 2014, pp. 13-15
[15] Shengchen, Z., Wenting, Q., Yingzhi, S., Xunzhi, S., Yunchen, S.: 'A review of Chinese microblog sentiment analysis'. Computer Applications and Software, 2013, 30, (3), pp. 163-164
[16] Zhiming, L., Lu, L.: 'Empirical study on Chinese microblog emotion classification based on machine learning'. Computer Engineering and Applications, 2012, 48, (1), pp. 1-4
[17] Jianwang, S., Xueqiang, L., Leizhen, Z.: 'Chinese Weibo sentiment analysis based on dictionary and machine learning'. Computer Applications and Software, 2014, 31, (7), pp. 177-181
[18] Tingting, L., Donghong, J.: 'Weibo sentiment analysis based on multi-feature combination of SVM and CRF'. Journal of Computer Applications, 2015, 32, (4), pp. 978-981
[19] Jie, J., Ru, X.: 'Weibo Emotion Classification Method Based on Machine Learning and Semantic Rules'. Journal of Peking University (Natural Science Edition), 2017, 53, (2), pp. 247-254
[20] Xingjie, F., Zhiwei, Z., Jinxin, S.: 'Text sentiment analysis based on convolutional neural network and attention model'. Journal of Computer Applications, 2018, 35, (5), pp. 1434-1436
[21] Wenkai, W., Liming, W., Yumei, C.: 'Weibo sentiment analysis based on convolutional neural network and Tree-LSTM'. Journal of Computer Applications, 2019, (5), pp. 1-8