Research on Sentiment Analysis and Abnormal Feature Extraction Technology Based on Comments Data

Hui Yan[1], Xufu Peng[1,2,*] and Xiaowan Zhu[1,3]
[1]College of Computer Science and Technology, Hubei Normal University, Huangshi, 435002, China
[2]College of Arts and Science, Hubei Normal University, Huangshi, 435002, China
[3]College of Educational Science, Hubei Normal University, Huangshi, 435002, China
[*]Corresponding author

***Abstract*—With the promotion of Micro-Blog's influence, the influence of Micro-Blog public opinion has been continuously strengthened. Therefore, based on the low recognition rate of Micro-Blog sentiment words in the traditional basic sentiment dictionary, this study designed and used the Micro-Blog emotional words to expand the basic emotion dictionary. And the fused dictionary is used to analyze the network comment texts. Moreover, according to the results of sentiment analysis, the abnormal criteria is formulated, then, the classical Knapsack model is used to solve the problem of constructing the abnormal comment text collection. Finally, the effectiveness of the sentiment analysis technology based on Micro-Blog dictionary and the method of extracting abnormal text collection using the backpack model are verified by experiments. The emotional tendency of user comments is grasped from the massive data, so as to understand the user's concern, which realized the important practical significance of the management of Micro-Blog public opinion.**

Keywords—basic sentiment dictionary; Micro-Blog sentiment dictionary; sentiment analysis; knapsack model; abnormal text

I. INTRODUCTION

With the rapid development of social media, Micro-Blog-Sina microblog, WeChat, Zhuihu and other platforms have gradually become the source of the spread of hot news. Social media plays an essential role in promoting the development of public opinion with its advantages on fast communication, flexible communication and more convenient interaction [1]. In order to effectively resolve the crisis of network public opinion, the governments at all levels use the network public opinion system to analyze the public sentiment in a timely and effective manner and take corresponding measures, which is the main way in the Internet era.

In the study of network texts, Chinese scholars mainly adopt the following types of analysis, including: Sun Lingfang and Yin Peipei analyzed the shortcomings of traditional technology in dealing with massive data, as well as the similarity characteristics of big data and network public opinion. They used the big data MapReduce model to conduct dig on the opinions and attitudes of the online public, and obtained the public opinion and attitude indicator model, followed analyzed and mined the public opinion and attitude intensity in the hot events [2]. Xing Yunfei, Wang Xiyu and others in the literature 3, research on the emotions of online public opinion users in the new media environment. The thesis constructs the user emotion evolution model from the emotional polarity and emotional intensity theory, and analyzes the user's emotional evolution characteristics and fluctuation laws [3]. In the field of network text analysis research, methods such as statistics or machine learning are the choices of most scholars to classify text categories. In the literature 4, Li Lingyun proposed using a rule-based and statistical method to monitor anomalies using a time series model, which proved to be more effective than ordinary models [4]. Based on sentiment dictionary is a way to approaches statistics, the approach based on sentiment dictionary is to construct a sentiment dictionary with relatively words by finding relations between words [5].

II. SENTIMENT ANALYSIS OF THE REVIEW TEXT

This study proposed a comment text sentiment analysis method based on sentiment dictionary. Our task focuses on constructing a dictionary with wide coverage text matching. And those matching result are marked with an emotion value which is used to reflect emotion degree. The main analysis process is shown in Figure I.

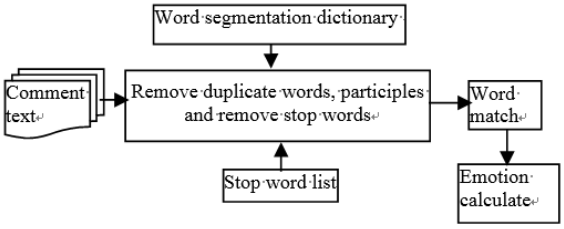

FIGURE I. COMMENT TEXT PROCESSING FLOW

A. Data Preparation

The preliminary work of detecting abnormal emotions in comment text is sentiment analysis of the comment text, and the primary part of the text analysis stage is the preparation of the data. The Micro-Blog-Sina microblog comment text collected by the big data collection platform as an experimental corpus of sentiment analysis. Its original corpus includes news ID, releasing time, comment content, the quantity of comments and forwarding number. The pre-

processing of the comment texts is an important stage of the sentiment analysis. Most of comment texts contain a lot of punctuation and words, some words are not suitable to be treated as a part of emotional text, so these words should be found and removed from the sentence, thus to ensure the sentiment analysis correct.

The pro-process includes removing the duplicate content in the comment (spamming, etc.), besides, using the word segmentation tool to separate the words and sentences, then labeling the words after the word segmentation. Finally, deleting the stop word. In the word segmentation, on the one hand, the jieba word segmentation tool is used, word segmentation dictionary uses a custom dictionary(wordbag.txt) that integrates multiple vocabularies, what's more, it collects vocabulary with higher evaluation rate in daily life, which can basically meet the needs of word segmentation. On the other hand, the Stop-List is proposed in this research, stop word list is a word set and collects a lot of neutral words [6]. It should be found and removed. In addition, in feature selection, the sentiment dictionary is used as its pursuant.

B. Building an Emotional Dictionary

In the literature 7, the problem of low coverage of the existing sentiment dictionary in the field of Micro-Blog emotional words is pointed out. This paper proposes a method to improve the SO_PMI algorithm by using the distance mutual information and Laplacian smoothing technique, and to expand the sentiment dictionary in the microblog field. The experiment proves the tendency of the method to judge the words. Compared with the traditional method, there is a big improvement in accuracy [7]. Different from common texts, Micro-Blog comment texts have large amount of emotions and net-words [5], Based on these ideas, the research uses Micro-Blog sentiment dictionary to expand the basic sentiment dictionary, then constructs the fusion dictionaries.

The basic sentiment dictionary includes How Net Chinese emotional dictionary. Li Jun Chinese Derogatory Dictionary of Tsinghua University, Simplified Chinese Emotion Dictionary provided by Taiwan University, and the book of "Emotional Vocabulary Ontology" compiled by Dalian University of Technology. The basic sentiment dictionary consists of negative words, degree adverbs, positive and negative emotional words, and its composition is shown in Figure II.

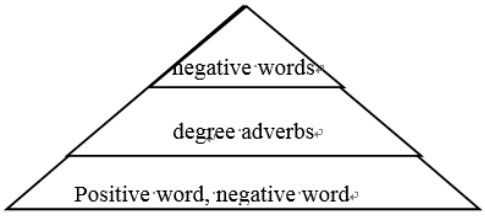

FIGURE II. BASIC EMOTIONAL

The content of Micro-Blog sentiment dictionary is extracted from the collected comment text by SO-PMI algorithm. According to the idea of SO-PMI algorithm, a set of positive and negative words are used as the benchmark words. According to the SO-PMI algorithm, a set of derogatory words and derogatory words are used as the benchmark words. The sentiment tendency of the words to be judged is the mutual information of the word and the derogatory word and the difference between the word and the derogatory phrase. First of all, using the word frequency method to sort the vocabulary according to the order from large to small, and manually select some words with obvious emotional tendency as the emotional seed words. Second, uses the Harbin Institute of Technology to expand the selected seed words, then uses the SO-PMI algorithm to calculate vocabulary sentiment tendency in candidate corpus, and determines the emotional polarity. If the SO-PMI value is less than 0, it is judged as a negative emotional word, if the SO-PMI value is more than 0, judged as a positive emotional word. Finally, the two dictionaries are merged together to construct a sentiment dictionary for project research.

C. Calculate the Emotional Value of Comment Text

Emotional tendency is a subjective evaluation of Micro-Blog content by netizens or a tendency of netizens' inner preferences. Emotional values are measures that describe emotional tendencies and emotional strength. The emotion of Micro-Blog comment texts is mainly reflected by emotional words. In general, most of comment texts can be regarded as a unit sentence. In order to get the emotional score of the comment text, the emotional words and corresponding emotion value should be extracted. For Chinese text, the most appropriate method is to matched with the feature word [6]. Based on the emotional dictionary, the emotion value calculation method is as follows: match the corpus with the sentiment dictionary by words and part of speech after processing the corpus. If the matching is unsuccessful, and the word cannot be reflected, then it can be discarded. If the matching is successful, the position and score of the sentiment word are recorded. Followed, continuing to use each sentiment word as the benchmark, looking forward for degree adverbs, negative words, then to calculate the corresponding scores. Finally, the scores of all the emotional words are summed to obtain the emotional score of the sentence. According to the emotional score of the sentence, the emotional tendency can be analyzed. The weight coefficients of degree adverbs and negative words are shown in Table I.

TABLE I. WEIGHT TABLE

Degree w	Weight P(w)	Degree	Weight
Extremely w_e	2	Un+extremenly	0.6
Over w_o	1.7	Un+over	0.8
Very w_v	1.4	Un + very	1.2
More w_m	1.2	Un + more	1.4
Little w_i	0.8	Un+little	1.7
Lack w_l	0.6	Un+lack	2
Negative w_d	-1		

D. Text Classification

The emotional value calculation stage has divided the comment texts into three parts: positive, negative and neutral. This stage aims to use the naive Bayesian classification algorithm to divide the comment texts into the A (event), B (government), C (environment, Facilities), D (netizens reaction), E (others) five categories. In the preparation phase,

some items are manually classified and used as a training set. Then, using the TF-IDF weighting strategy while extracting the characteristic of the data which is to be trained. TF (word frequency) refers to the frequency at which a particular emotional word appears in the corresponding comment text. Its calculation formula is as shown in formula (1):

$$TF_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

The variable n(i, j) represents the amount of occurrences of a particular sentiment word in comment texts, and ∑kn(k, j) represents the sum of the occurrences of all words. IDF (Reverse Document Frequency) is a measure of the importance of a particular word, its value is obtained by dividing the total number of files by the number of files containing the word:

$$IDF_i = \log \frac{D}{\{j : t_i \in d_j\}} \tag{2}$$

After the feature attributes are obtained, calculate the probability of occurrence of each sentiment category in the comment text and estimate the conditional probability of each sentiment word segment for each sentiment category, and output the classifier. Finally, this classifier is used to classify the comment text for the review object.

III. Build an Exceptional Text Collection

A. The Concept of Knapsack

0/1 Knapsack problem: when a traveler is ready to go hiking, he must decide to carry what kind of items. There are N pieces of items to choose from, numbered 1, 2, ..., N. The item of i is w_i kilograms, the value is p_i yuan. The maximum weight he can carry is w kilograms [8]. Which items can achieve the maximum value?

Introduce the variable xi, and set knapsack as P:

$$x_i = \begin{cases} 1, & load\ the\ item\ i\ \text{into}\ P \\ 0, & not\ load\ the\ item\ i\ \text{into}\ P \end{cases} (i = 1, 2...n) \tag{3}$$

Then the mathematical model of this problem is: $\max\ f = \sum_{i=1}^{n} p_i x_i$.

B. Application of Knapsack Model

The knapsack problem is often used in the field of operations research to manage resource allocation, capital budgeting, investment decisions, etc. [8]. Abnormal text collection builds can introduce knapsack issues. In this problem, it can be described as: each pieces of S data, the emotional source of each comment data is V, each comment data contains N words, which comments should be selected as the corresponding emotion of the characteristic text?

The calculation and iterative process of constructing an anomalous text collection by dynamic programming is as follows:

a) *Construct a text matrix R, R = (Rnvw):*. Where N represents the text number (unique identifier of the text), V represents the text sentiment score, and W represents the number of words contained in this text. C is the maximum number of texts that can be accommodated in the word bag, M is the maximum number of words that the word bag can hold, and V is the emotional score of the word bag.

b) *Setting V(i,j) denote the former i (1<=i<=C) pieces of texts, load them into capacity j (1<=j<=M) to obtain the maximum emotional value:* When determining x_i, the (x_1,...,x_{i-1}) has been determined, the problem is in one of the two states:

c) *The word bag capacity is not enough to be loaded into the text i:* Then the maximum value of the first i text is loaded and the maximum value obtained by loading the first i-1 text is the same, 即 x_i=0. The word bag does not increase the emotional value.

d) *The word bag capacity can be loaded into the text i:* If the text i is loaded into the word bag, the sentiment value in the word bag is equal to the emotional value of the first i-1 item in the word bag with the capacity j-w_i adding the emotional score vi of the text i. Then get the recursive formula as in formula (4):

$$V(i, j) = \begin{cases} V(i-1, j) & j < w_i \\ \max\{V(i-1, j), V(i-1, j-w_t) + v_j\} & j \geq w_i \end{cases} \tag{4}$$

IV. Experiments

Micro-blog is one of the largest social platforms in China currently. The experimental data mainly selects 4647 comments about floods event. Unlike the written text, these network comment texts are short and simple, besides the netizen express thoughts directly, and their reviews contain network words and emojis. Data is deduplicated, cleaned, segmented and numbered. Figures IIII and IV are graphs of sentiment values for microblog comment data for flood events.

Figures III and IV are graphs of sentiment values.

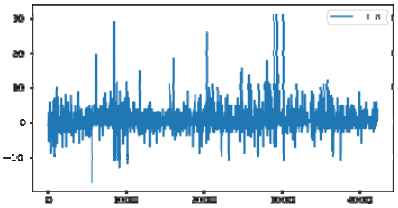

FIGURE III. SCORE

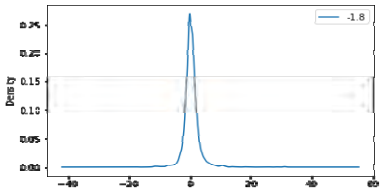

FIGURE IV. SCORE

It's shows in Figure III, the majority of the emotional value of texts is between -10 and +10, and as the absolute value of the score increases, the corresponding amount of text decreases. In addition, the density between 1.8 and +1 is highest.

Emotional scoring for Micro-Blog review data is a preliminary study of emotional data, primarily emotional understanding the distribution of review text, find out some hidden relationships, screening out the targeted data, so as to analyze the reasonable data. Anomaly detection aims at detecting abnormal data samples from data sets. Most of the samples in daily life as normal samples, but the value of abnormal samples is far greater than the normal samples, so using the feature set model of abnormal text backpack.

Abnormalizes, unlike normal, means that a small amount of texts is different from most texts, indicating that the individual is different from the group. Observing the results of the previous research stage, Figure v and Figure vi define the abnormal text of flood class as the comment text whose score is less than -3 and whose score is greater than 3.Using text categorization, 653 flood data can be divided into five categories according to the comment objects of netizens: event itself, government behavior, environmental facilities, netizens' response and others.

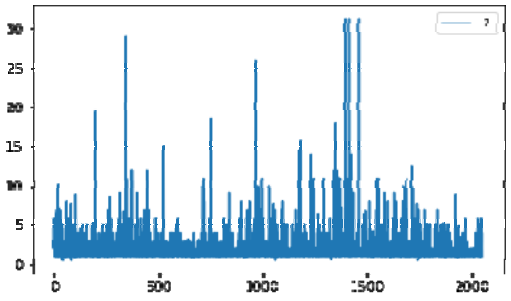

FIGURE V. POSITIVE

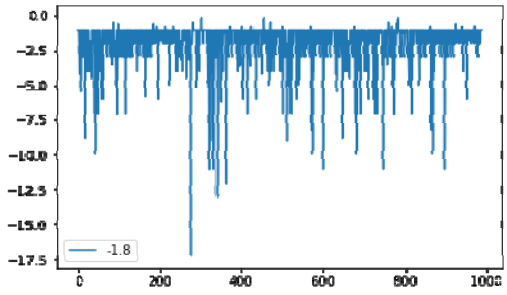

FIGURE VI. NEGATIVE

Anomalous text data collection obtained by dynamic programming, the result is show with Table II and Table III.

TABLE II. ABNORMAL TEXT COLLECTION

	Category	Number of texts	score
negative	Event	147,180,185,188,242,309,310,317,375, 614, 638	63.6
	Government	150,463,1027,1035,2352,2594	66.
	facilities	170,206,219,250,2983	55.1
	response	194,231,580,686,977	54.5
	Other	1990,2193,4509	21.0
positive	Event	151,159,160,204,211,44,300,320,356	53.6
	Government	305,412,417,430,440,696,1117	45.0
	facilities	183,257,1182,1361,1714,2570,2886	49.0
	response	503,846,851,2191,2405,3200,3832,4513	40.
	Other	221,411,487,491,567,631,841,1786,1788	45.5

TABLE III. TEXT FEATURE COLLECTION

	Category	Text Collection
negative	Event	Lying, sad, rainstorm, expert, vocal, facts, drowning, scary, flood, natural disaster, dead, reports, missing, mentally retarded, unfinished, thrust, spray, critical, helpless irresponsible, surprised, disaster, small, powerless, comments,
	Government	Death, disappointment, truthfulness, ordinary people, useful, facts, corrupt officials, rushing corruption, problem solving, disaster, quiet, powerlessness, eyeballs, decision making, sniping, overthrowing, chilling, parents, kung Fu, reconstruction, stubbornness,
	facilities	Shit, build, earth, build, bridge, wound, ignorance, free, keyboard, cross, maintenance, detection, pusher, disaster, question, bridge, care, excited
	response	Police, murder, liberation army, black, homeless, chilly, sad, indifferent, sinned, deserved, low-lying, communication.
positive	Event	Situation, cute, warrior, peace, control, advance, forecast, notice, withdrawal, ones, officer, flood control, news, newspaper, pray, hope, protect the environment, loved
	Government	Country, condolences, rewards, substance, respect, key love, transmission, attention, hand in hand, recovery, overcoming difficulties, water retreat, news, real, face,
	facilities	Drainage system, suitable, thrust, attack, Tofu-dreg construction, bridge, level, high, test, flood relief, dial, project construction, reconstruction, bridge, best effort,
	response	The first time, rescue, concern, hope, worth fundraising, gratitude, danger, truth, report,

The "other" category is a comment unrelated to this event, so it is not included in the analysis results. The "event" refers to the fact that the content of Micro-Blog is actually mentioned in the comments, while the "netizen response" is abstract, and there is no direct discussion about Micro-Blog content, but the emotion expressed is caused by the event, so the "event itself" and "netizen response" can be classified into one category in some degree. According to the feature collection, we can extract valuable information, then predict the sentiment tendency and netizens' concerns of current events from netizens' comment text.

V. Conclusion

Emotion is one of the main factors affecting action. Therefore, mastering the sensational trend can effectively follow the reality and supervise the work of network public opinion. "Research on Sentiment Analysis and Abnormal Feature Extraction Technology Based on Comments Data " uses sentiment analysis technology based on sentiment dictionary to analyze emotions of online comment texts. Implement the emotional value calculation of the comment text. According to the results of sentiment analysis, the abnormal standard is established. Finally, the classic backpack text model is used to construct the abnormal comment text matrix vector. By mining the floodplain netizen comment text, the abnormal text feature set is constructed from positive and negative aspects. The research results verify the effectiveness of the technical method to some extent. But this is only a tentative study. The construction of emotional indicators and abnormal standards has many shortcomings and areas for improvement due to realistic factors. The relationship between netizens' emotions and social events is complicated and involves various factors. The results of this study cannot represent the emotional situation of netizens in all events.

Acknowledgements

This research was supported by an innovative team of excellent young-middle-aged, universities in Hubei province (T201430).

References

[1] Wen XIN. The Influence of Public Emotions on the Dissemination of Public Opinions on Sina Weibo——Taking the Jiangge Event as an Example[J]. News Research Guide,2018,9(09):160+189.

[2] Lingfang SUN, Peipei YIN. Research on Internet Lyric Emotional Strength Based on Big Data Technology[J]. Computer and Digital Engineering, 2018, 46(01): 160-166+181.

[3] Yunfei XING, Xiwei WANG, Yanan WEI, Wei WANG. Research on Emotion Evolution Model of Internet Public Opinion Users in New Media Environment——Based on Emotional Polarity and Emotional Intensity Theory[J]. Information Science,2018,36(08):142-148.

[4] Lingyun LI. Research about Event Real-Time Monitoring Framework and System Based on Micro-Blob [D]. Beijing University of Posts and Telecommunications, 2015.

[5] Xiaohong HAO.. An Automatic Construction Approach for Sentiment Dictionary Based on Weibo Emoticons [A]. Wuhan Zhicheng Times Cultural Development Co., Ltd. Proceedings of 2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018) [C]. Wuhan Zhicheng Times Cultural Development Co., Ltd:2018:9.

[6] Xing WU, Haitao LU, Shaojian ZHUO. Sentiment Analysis for Chinese Text Based on Emotion Degree Lexi-con and Cognitive Theories[J]. Journal of Shanghai Jiaotong University (Science),2015,20(01):1-6.

[7] Xinkai YANG. Research and application of Chinese microblog emotion dictionary [D]. Shanghai Normal University, 2017.

[8] Zhiping SONG. Mathematical model of knapsack problem and its application [J]. Western Development, 42-144.