Asian Journal of Information and Communications 2017, Vol. 9, No. 2, 43-57

KISDI
KOREA INFORMATION SOCIETY DEVELOPMENT INSTITUTE

Asian Journal of Information and Communications

Social evaluation of medical and health treatments based on big data analytics*

Xinshuang Zhao** Fudan University
Yiregedai*** Fudan University
Li Deng**** Fudan University
Huajuan Mao***** Changhai Hospital, Second Military Medical University
Zaiping Jing****** Changhai Hospital, Second Military Medical University

Abstract

With the development of information technology and social media, people are getting used to sharing experiences on social platforms. Social media like Weibo and online forums usually carries a large amount of comments about medical and health treatments, and provides valuable supplementary information for comprehensive understanding of the practical effects as well as the social attitudes of those treatments. In this paper, we aim to explore the above social evaluation method of medical and health treatments based on big data analytics. Due to the characteristics of non-standard expressions, rich variations, and complex contexts of the comments on social media, it's difficult to conduct a precise content analysis for social evaluation. Therefore, we adopt the sentiment analysis and affective computing technique to deal with the comment texts, and present an effective evaluation method which can better reflect the overall social feelings and attitudes to a certain medical and health treatment. The proposed method and its big data analytics are studied and applied to the cases of treatment evaluations of lung cancer, hepatitis B, and diabetes mellitus in this paper.

Keywords: medical and health, social evaluation, big data analytics, sentiment analysis, affective computing

1. Introduction

The surveillance and information analysis of diseases are of paramount importance for the government's medical and health department to keep track of the national health condition and deploy disease control. In general, reliable and analyzable medical and health data are mainly obtained from the report system of medical institutions and statistical investigation of health departments. In China, the large population and uneven distribution of medical resources result in that the capability of collecting and analyzing medical and health data varies largely among medical institutions. The lack of information sharing and combination between medical institutions also leads to problems like data missing and redundancy. Apart from the difficulty along with medical

* This research work was supported by Online Education Research Center, Ministry of Education of China (Qtone Education, No. 2017YB115), Shanghai Pujiang Program (No.16PJC007), and Xiyuan Program of Fudan University(No. 16509).
** 14300690080@fudan.edu.cn
*** 12307100044@fudan.edu.cn
**** 14300690072@fudan.edu.cn
***** Corresponding author: czymm8@163.com
****** jingzp@xueguan.net

data, the process of medical and health evaluation is highly sensitive to the application environment and individual difference. Thus, the professional medical and health evaluation method might not be capable of reflecting the comprehensive curative effect, which also involves patients' feelings and attitudes, or rather, the social evaluation.

Big data analytics can play its part in the social evaluation of medical and health treatments. Owing to the development of information technology and popularization of mobile applications, more and more people are sharing their medical and health experience through social media, such as forums and Weibo, which makes the latter a valuable repository of information concerned with health and medical treatment. For example, Google launched Google Flu Trends or GFT in November, 2008, which can estimate the extent of epidemics in that area based on the number of searches on related items, such as flu, sneeze and fever. The mechanism is that when people or someone they know of suffer from flu-like symptoms, they are likely to look out for prevention and therapeutic methods online. Accordingly, the frequency of related items searched online increases dramatically during the influenza peak. Thus, the regional flu trend can be predicted by computing the searching frequency of related keywords.

The big data analytics not only has played its role in the forecast of epidemics, but is gaining attention in various fields. In the medical and health industry, big data analysis is playing an increasingly important role due to the high utility value of medical and health data. For instance, big data analytics can be applied into the development of clinical pathway and disease prediction. It also has the application potential in social evaluation of medical and health treatments. Some researches already made attempts in using big data analytics in the content analysis of network data to obtain useful information. However, due to the complexity of texts from social media, which include a large amount of non-standard expressions and rich variations, it is difficult to make an accurate content analysis.

In this paper, to present an effective evaluation method which can better reflect the overall social feelings and attitudes to a certain medical and health treatment, we will adopt the sentiment analysis and affective computing technique, along with big data analytics, to deal with the comment texts retrieved from Weibo and medical forums. In order to illustrate this method, we will calculate the affective polarity of certain Chinese words, and then make a sentiment evaluation analysis of three types of diseases respectively: lung cancer, hepatitis B and diabetes mellitus. The next section is a review of the current progress and achievements in related areas. Section 3 and section 4 will introduce the process of collecting data and arranging them into useful information under sets of rules and calculation. Section 5 is a brief summary and discussion.

2. Related works

Big data analytics can be applied in many fields. For enterprises, data-oriented decision making and forecasting have become a critical procedure, where big data analytics is implemented in obtaining useful information. Up till now, big data analysis in the medical and health field mainly focuses on data analysis and resource management. It has provided some research methods and made many achievements.

In terms of digging information from DNA sequences, Zhu and Xiong (2007) introduced different study stages, including data mining based on statistics and special data mining method for DNA sequences analysis. Their work also discussed the essential data mining techniques employed in the analysis of DNA sequences and its importance in biological application. Yue and Jing (2009) applied data mining into the classification of DNA sequences. They developed a new judgement algorithm which is of both concision and high accuracy.

Some studies concentrate on improvement of data mining methodology by using techniques such as association rules, artificial neural networks, regression analysis, decision tree, and so on. For example, Chang et al. (2003) used artificial neural networks to simulate the diagnosis process of medical experts. They put dozens of factors into consideration in the simulation procedure and the result proved the artificial neural networks a promising method in the diagnosis of Parkinson's disease. Dreiseitl et al. (2001) adopted regression analysis and decision tree into the artificial neural networks to estimate the mortality in diseases, which improved the algorithm's accuracy and convincingness.

With regard to health management, there are researches on how to build health evaluation system by data mining and improve the health evaluation model. In Li's (2011) work, he devised a health risk model based on big data to explore on the general factors of health risks, which well responds to the Chinese public's demand for health risk evaluation. In a more specific study, Karaolis et al. (2009) applied association analysis algorithm in developing a data mining system to find the pathogenic factors of heart disease.

In general, researches that use network comments to evaluate medical treatment analyze the text contents literally, therefore lack emotional analysis of texts. For instance, Zhou and Srinivasan (2005) employed machine learning and referred to medical terminology dictionary to automatically retrieve medical terms from network text contents. In exploring cases that exist potential misunderstanding or misuse of antibiotics, Scanfeld et al. (2010) mined a thousand twitter status updates for co-occurrence of pairs of neural terms, without further study on words with polarity and intensity. Attard and coulson (2012) made a thematic analysis of patient communication in Parkinson's disease online support forums in their study. The texts they used in analysis were extracted from original data, and then reassembled into non-overlapping themes and subthemes. But the interpretation did not extend beyond the surface meaning of messages.

A commonly used method in analyzing Weibo's network information is to collect Weibo users' basic information and status text contents, then make feature analysis of users through data mining and summarize the characteristics of Weibo's core users. By this means, personalized marketing can be made towards targeted customers precisely (Zhao, 2012). Further studies, like Liu and Liu (2012) adopted machine learning algorithms, feature selection methods and feature weight methods into the sentiment classification of Weibo comments. Pang et al. (2012) put forward an unsupervised sentiment classification method to analyze the emotions in texts from Weibo. These researches have made some explorations on text mining of Weibo, but lack specific analysis on text contents about healthcare.

Human language texts bear emotional inclination and intensity inherently. Therefore, comprehensive emotional analysis of text can better reflect the publisher's attitude tendency. However, the non-standard language used in online social media and its rich variations have greatly increased difficulty in the precise sentiment analysis of texts, which requires that psychology and linguistics knowledge are both involved in order to infer the overall social feelings and attitudes to a certain medical and health treatment. Pan's study (2011) proposed several methods to analyze the emotion tendency of Chinese sentences, including splitting large-scale emotion dictionary, establishing specific rule sets, using multi layer classification algorithm and so on. Dai et al. (2017) in their work explored the social evaluation of innovative drugs through the analysis of mined data and affective computing on online comments. These studies inspire usto adopt the sentiment analysis and affective computing technique to handle the comment texts, and present an effective evaluation method to better reflect the overall social feelings and attitudes to a certain medical and health treatment.

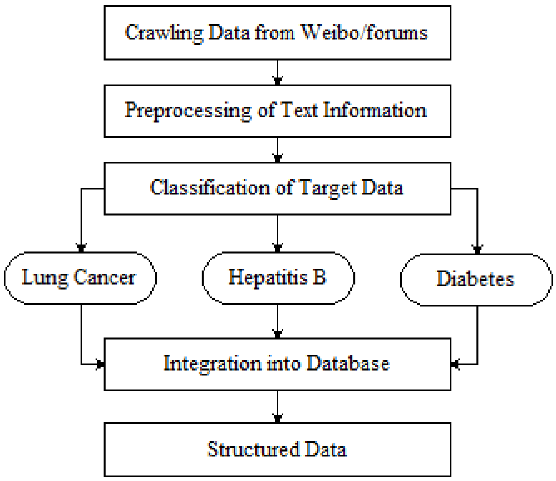
3. Data collecting and preprocessing

The massive medical and health information preserved in Weibo or healthcare forums exists in the form of main posts and comments. For example, patients or their relatives might go to a healthcare website for reference by posting inquiries about a certain disease. The doctors' answers and following replies probably contain recommended treatment and patients' experience. Plus, evaluation of treatment is just one kind of health text information. Other information such as geographical distribution, temporal distribution and causing factors of diseases might also hide in the complex data resource, which needs further research.

3.1 Data collecting method and process

We used GooSeeker to collect text information related to lung cancer, hepatitis B and diabetes mellitus from Weibo and medical forums. First step is to define text crawl rules. Then set the element node so that the crawled data could be imported into the database. The following chart shows the data collection from Kangle Website[3].

Figure 1. Data collecting process



First step is to collect headlines, websites, and numbers of comments of health information by a single node. Eliminate those with less than five replies since they are of little reference value. Then use collected websites as navigation, enter the secondary homepage and gather specific posts, comments and authors' information. Below is the data sample collected from Weibo.

Table 1. Data collection sample of Kangle Net

Acquisition Rule	Contents	Sample Source
HealthData_Kangle_Level1	Headline, website, and number of comments of articles with keywords	Kangle
HealthData_Kangle_Level2	Comments with keywords	Kangle
HealthData_Kangle_Level3	Related information of comments author, for example, ID	Kangle

3) https://www.kangleweb.com/

Table 2. Text data sample

Blogger Name	Forwarding Date	Contents	Forwarding Amount	Replies
**_sunny	2016/04/23	10 kinds of food diabetics should eat…	79	//Design a diet plan for Dad

3.2 Data preprocessing techniques

The raw data have several defects, including information loss, duplicate items, noise from adverts and data inconsistency. Data preprocessing is to ameliorate the raw data before further analysis process. Liu et al. (2000) summarized several data preprocessing methods, including data reduction based on rough set theory, data concentration based on concept tree, attribute selection method based on statistical analysis, genetic algorithms and so on. Liu et al. (2000) also suggested that data preprocessing should have the functions of data integration, data cleaning, data transformation and data reduction.

The specific preprocessing method varies. In this paper, after data preprocessing we will divide data into three categories based on the disease keywords, namely lung cancer, hepatitis B and diabetes mellitus. Since our aim is to analyze the content of texts rather than to standardize them, most of our work is about data cleaning, which involves dealing with missing data and getting rid of irrelevant data, and eliminate the noise. Below are the rules we used in data cleaning.

- In regard to missing items, fill in them with generic average or most probable information.
- Adverts such as misleading facade or malicious false information from rivals are determined as data noise.
- As to abnormal information, cluster analysis can be used to group similar items and screen out the outliers.
- Manual elimination can be used if the data size is small enough.

4. Affective computing method

4.1 Introduction to the method

Affective computing and content analysis are the two big steps in analyzing the texts or comments posted by netizens. To be brief, firstly identify the tendency and degree of the author's sentiment towards the keywords in the discussed content, and then give a reasonable explanation for their emotional inclinations combined with medical and health knowledge.

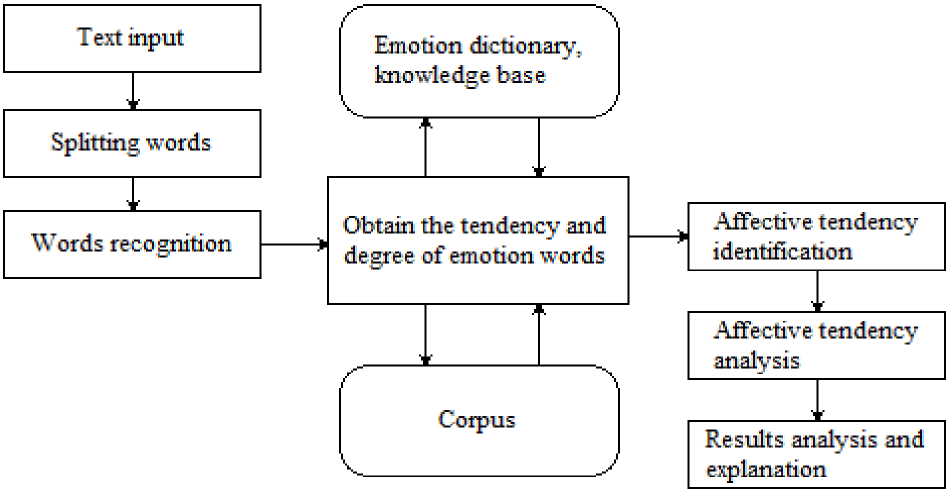
In this paper, sentiment analysis will be achieved by affective computing, including the PMI algorithm and HowNet[4)] algorithm. We will make sentiment analysis through the affective computing algorithm based on the emotion dictionary. Set the emotion dictionary and corpus as a text reference, combine and calculate the affective polarity of the units collected from the collecting software, and analyze the emotion tendencies in the target text.

Under this emotion dictionary calculation algorithm, we firstly do the classification of words of the text, and then split the entire text into words and sentences according to the passages and punctuations. Next, we split the entire sentence text into adjectives, verbs, nouns, adverbs, etc., and put the respective words into the emotion dictionary and corpus to find its emotion tendency and degree. In the end, we add up the emotion degree of these

4) www.keenage.com

words and calculate their mean value (INT), which represents the emotion degree of whole text. The process of calculating emotion dictionary is shown below.

Figure 2. Calculation process of emotion dictionary



Here we use the existing emotion dictionary, “How Net Chinese word library” and “CSC Chinese semantic lexicon” to obtain the affective word tendency degree.

4.2 Comparison of affective computing algorithms

Next we will firstly introduce the commonly used algorithm based on the PMI and the algorithm based on HowNet, and then present the HowNet-PMI algorithm employed in this research to analyze the text and evaluate the results. This article uses HowNet-PMI text emotional computing algorithm, which is put forward by Wang, Wu and Hu (2012) in order to overcome certain limitations of PMI and HowNet. This HowNet-PMI method is the combination of two word polar calculation method. By HowNet’s synonym expansion of benchmark words library, we can replace low-frequency benchmark words with higher-frequency synonyms, which can avoid the problem of low frequency in the corpus of emotional words to some extent. On this basis, the medical evaluation results of content analysis are explained reasonably.

Affective computing of text polarity based on PMI is to choose both commendatory and derogatory words from the existing corpus as the benchmark of emotion judgment. Calculate the probability when target words and benchmark words appear together in the corpus. If the target word has a high probability to appear with commendatory benchmark words, the target word is regarded as a commendatory word. In the contrary, it would be recognized as a derogatory term. Here the threshold value is calculated through mathematics methods, and the emotion tendency and degree are judged by comparing the probability with the threshold value.

The reference of the commendatory words group is set as:
$WordSet_1$= {$commendatory_1$, $commendatory_2$, … , commendatoryn}
The reference of the derogatory words group is set as:
$WordSet_2$= {$derogatory_1$, $derogatory_2$, …, $derogatory_n$}

Then we can calculate the polarity of target word, $Word_x$, based on PMI:

$$\text{SOPMI(Word x)} = \sum_{i=1}^{n} PMI(Word_x, commendatory_i) - \sum_{i=1}^{n} PMI(Word_x, derogatory_i) \quad (4.1)$$

$$\text{PMI}(Word_1, Word_2) = \log_2\left(\frac{P(Word_1 \& Word_2)}{P(Word_1)P(Word_2)}\right) \quad (4.2)$$

PMI ($Word_1$, $Word_2$) represents the probability that $Word_1$ and $Word_2$ appear together, while P(Wordi) represents the probability that the target word, $Word_x$, appears alone in the selected corpus. Here we set the threshold value $\theta_{PMI}=0.5$. When SO_PMI($Word_1$)>θ_{PMI}, we can say that the target word is commendatory and that the author's emotions are positive. In the contrary, if SO_PMI ($Word_1$) <- θ_{PMI}, the target word is derogatory and the author shows negative emotions. Otherwise, the target word is neutral.

However, the difference between the affective computing of text polarity based on the HowNet similarity and the method used above is that they choose different corpus. If we change the corpus in the above method into the HowNet semantic dictionary to conduct the text emotion calculation:

$$SO_{HowNet(Word_x)} = \sum_{i=1}^{n} Sim(Word_x, commendatory_i) - \sum_{i=1}^{n} Sim(Word_x, derogatory_i) \quad (4.3)$$

Sim($Word_x$, $commendatory_i$) represents the similarity degree between $Word_x$ and $commendatory_i$ based on HowNet similarity calculation, while Sim($Word_x$, $derogatory_i$) represents the similarity degree between $Word_x$ and $derogatory_i$(Wang, Wu and Hu, 2012).

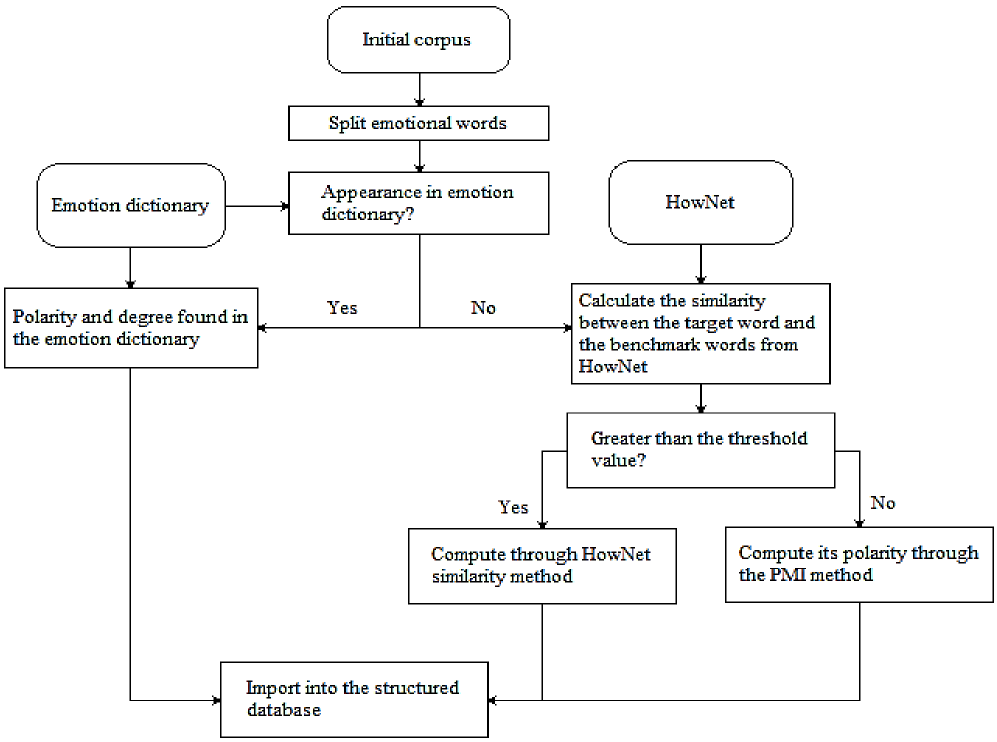
Although both of these methods can calculate affective polarity, the following problems can occur in some cases. The emotion calculation based on PMI relies on the selection of corpus, which is regarded as the benchmark, and the probability of common occurrence with the reference word is regarded as the similarity to the emotional reference word. The higher the probability of co-occurrence, the closer the two words' emotions are in the same context. However, since the process of building up corpus involves selection of words, certain emotional words are less likely to occur in the corpus. In this case, the exact polarity of the target words can not be obtained, and there is high risk of deviation of the emotional analysis.

In regard to the other method, the calculation problems of the similarity based on HowNet is also affected by artificial selection. Besides, the similarity calculation is measured by the distance from the semantic to the original semantic tree. For example, manual sorting of HowNet Semantics dictionary is not satisfying in terms of the frequency of some benchmark words. Although some words' emotional tendency and degree are very close to each other, the distance between their semantic to the original semantic tree is very far, which causes the inaccuracy in polarity calculation. Because of the possible computational bias of these two methods, this article chooses the HowNet-PMI algorithm to calculate the emotional tendency and degree.

4.3 Affective computing algorithms based on HowNet-PMI

The HowNet-PMI algorithm, based on the principle of the two algorithms, extends the synonym of target words through the semantic dictionary of the HowNet, and replaces the low-frequency words in the PMI algorithm for the polarity calculation. The steps are shown below:

Figure 3. Text emotional computing process based on HowNet-PMI



Firstly, we use NLPIR, a Chinese word segmentation system, to segment original corpus and mark them with properties. For example, from the sentence "Moringa seeds are efficacious in prevention and treatment of diabetes mellitus" we can get "Moringa seeds[n]", "prevention[n]", "treatment[n]", "efficacious[a]", "in[prep]" and other elements. Then select emotional vocabulary from the segments (Word1, Word2,…,Wordn) and look for their polarity and intensity in the emotion dictionary. If these words are not included in the dictionary, calculate the similarity between the target word and the benchmark sentiment words by using formula 4.4 below. Define the threshold value $\theta= 0.8$. Set the maximum similarity as γmax. If $\gamma\max>\theta$, it represents high similarity between target word and benchmark word. Thus, target word can be replaced by a benchmark word of high frequency in the calculation of emotional tendency. In this case, the polarity is calculated as below.

$$\mathrm{SO}_{\mathrm{HowNet}(Word_1)} = Sim\left(Word_1, Benchmark\ Word_j\right) * SO\left(Benchmark\ Word_j\right) \tag{4.4}$$

SO(Benchmark Word j) stands for the intensity of the polarity of benchmark word j(Wang, Wu and Hu, 2012). This method makes up for the insufficiency of some words in PMI by replacement. But if $\gamma\max<\theta$, which means target word and benchmark word are not similar enough to replace each other, the PMI formula shown in 4.1 and 4.2 should be applied.

Having worked out the emotional tendency and intensity of every word in the original text, here is an example of a text analysis about diabetes mellitus. The original text is "I got diabetes around four years ago.

I have been injecting insulin all the time and the blood glucose stays around normal level. Recently I started taking Moringa seeds and I am on the mend. I recommend this medicine to you and hope you get better soon".[5)]

Table 3. Original text segmentation sample

Original Text ID	Keywords	Properties	Keywords ID	Emotion	Emotional Intensity	Emotional Intensity Level
2482	all the time	adv.	3195	0	6.5	7
	normal	adv.	3196	1	4.3	5
	on the mend	adj/adv.	3197	1	4.5	5
	hope	v/n	3198	1	6.7	7

The word's emotional intensity cannot be found in the dictionary, so we used formula 4.5 to find its "synonym" with threshold θ= 0.8.

$$\mathrm{Sim}(Word_1, Word_2) = \frac{\alpha}{Dis(Word_1,Word_2)} \tag{4.5}$$

The letter α describes the distance between words when similarity equals 0.8. The calculation result is 0.83, which means "turn better" can replace "on the mend". The emotional intensity is 5.2, of level 5. Based on formula 3.4 we got the emotional intensity of "on the mend", 4.5, which is also of level 5 and equals to "turn better".

5. Sentiment analysis based on texts

With the emotional tendency of keywords calculated in the text through the HowNet-PMI algorithm, we can compute the weighted average and try to explain the internal logic considering external objective acts. In the case of the last section, the original text of ID 2482's emotion tendency can be computed like this:

$$\mathrm{SOsentence} = \frac{\sum_1^4 Word_n}{4} = (4.3 + 4.5 + 6.7)/4 = 3.8 \tag{4.6}$$

The emotion tendency is positive with the intensity of 3.8, which matches people's emotional cognition of that sentence. In addition, it is noticeable that privatives have essential influence on the polarity and intensity of emotional computing, which is shown in formula 4.6.

$$SO_{sentence} = (-1)\mathrm{n} * Word \tag{4.7}$$

In the formula, n represents the number of words between emotion word and privatives. When the two words are next to each other, n = 1. We do not make further discussion on the effect of conjunctions and degree words in this article.

5) The original text is in Chinese. Here we used "on the mend" for "见好" and "turn better" for "好转". The translation might not express accurately.

Of 1,000 texts collected of three diseases respectively, there are 894 valid texts about lung cancer which can be used for emotional computing , the numbers for hepatitis B and diabetes are 797 and 903 respectively. We graded and categorized the keywords used in data crawling. The keywords are the three diseases names. The secondary keywords are about symptoms, treatment, diets and so on. The third-class keywords are a specific version of the secondary keywords. Here we set the column of grade of emotional intensity, namely level 1, level 3, level 5, level 7 and level 9. In the rest of the article, emotion intensity equals to the grade of emotion intensity. Positive emotion tendency is 1; pejorative emotion tendency is -1; neutral emotion or emotions with two tendencies are 0.

5.1 Treatment evaluation of lung cancer

After an emotional analysis of 894 valid texts, the following Table 4 was obtained:

Table 4. Lung cancer keywords' affective tendency and degree

Selected Keywords	Inferior Keywords	3rd Class Keywords	Emotional Intensity	Grade of Emotional Intensity	Emotional Tendency
Lung Cancer	Symptoms	Cough &Hemoptysis	6.9	7	-1
		Chest Distress & Dyspnea	8.5	9	-1
		Fever	3.1	3	-1
		Weight loss	3.0	3	0
	Treatment	Surgery	4.7	5	1
		Drug treatment	5.2	5	1
		Chinese Medicine Diagnosis and Treatment	2.6	3	1
		Chemotherapy	9.1	9	-1
		Radiation therapy	7.4	7	-1
		Biological therapy	2.6	3	1
	Diet	Taboo	7.2	7	1
	Others	Hospital evaluation			
		Drug prices	7.3	7	-1

From the above table we can see that in the symptoms column, the emotional tendencies are all negative, and that the strong emotional intensity symptoms, cough hemoptysis, chest distress and dyspnea show text authors may easily associate both symptoms with lung cancer and are sensitive to the two symptoms. When the two symptoms appear, the author of the text is likely to go to a hospital or clinic for examination.

Emotional tendency in the weight loss column is zero, perhaps it's because the selected text has no obvious emotional tendency towards weight loss. It may also be because the selected text sample related to weight loss is too little, which causes no effective emotional tendency results. In fact, cancer is a common cause of weight loss, but it is difficult to infer a person gets cancer from weight loss.

Fever emotions tend to be negative, but emotional intensity is the same as weight loss. Maybe this is due to fever is common in respiratory diseases and immune system diseases. Similar to weight loss, fever is usually a symptom for auxiliary judgment.

In the treatment column, the affective tendency of surgical treatment is positive, and its emotional intensity is higher than other treatments, indicating that the publisher has higher recognition of surgical treatment. Possible explanation is that choosing surgery therapy is more often in the early stage of cancer, since removed part of the lung will affect breathing and will cause breathing difficulties hereafter, which means relatively high demands on the patient's physical condition at that time. The 5-year survival rate of surgical treatment can reach 30-44%, which is rare for lung cancer that has high mortality rate. However, surgery may not be able to thoroughly remove all cancer cells in the body, and because of cancer cell structure variation, they can transfer fast in the body. Thus, the patient is prone to get cancer recurrence or diseases in other body parts after surgery. The current situation is that most patients with lung cancer are often in its terminal stage when it is found, and with cancer cells metastasis, the surgical treatment is generally not effective at all.

Drug treatment and Chinese medicine treatment are conservative treatments, so their curative effect is limited and are usually associated with other treatments. However, they have some relief and the placebo effect in the short term and there is no huge risk like surgery, so their emotional tendencies are positive.

Radiotherapy and chemotherapy are commonly used in middle-terminal stage of cancer. It is not because the two kinds of therapy are suitable for terminal-stage-cancer treatment, but because that many patients are already in the terminal cancer stage when it's found. So they can only choose radiotherapy and chemotherapy that have obvious effect. The mortality rate in this period is very high, and lung cancer cell metastasis is very fast, so few patients are cured. In the course of treatment, patients often suffer from great physical and mental torture, so it's normal that the authors have negative emotional tendency and high emotional intensity for both treatments.

Biological therapy is a newly arisen therapy, mainly referring to the molecular targeted therapy. Molecular target therapy treatment is a frontier treatment currently. It positions mutated molecules to use medicine. It aims at the cancer cells like tracking missiles. Compared with radiotherapy and chemotherapy, this treatment won't kill "innocent" cells and affect normal cells. This therapy has a higher effectiveness, such as Xalkori (multi-target protein kinase inhibitors) made by Pfizer. Xalkori's disease control rate for ALK positive patients can reach 90%, and side effects are very small, and can well prolong patients' survival, improve their quality of life (Zeng, 2007). However, because of the high cost of molecular targeted therapy and generally not in the health care system, common patient cannot afford it. Also, this therapy is not well understood, so the emotional inclination is positive and the emotional intensity is 3.

In the hospital evaluation column, the selected sample has too many hospitals and the comments are too scattered. Besides, under the condition that sample size is small, it's not accurate and objective to add them up as the publishers' overall evaluation towards our country's hospitals. So we did not deal with this column. In the drug price column, the emotional tendency is negative and the intensity is 7, which indicates that the author generally believes that the drug for lung cancer is overpriced and is very concerned about drug price.

5.2 Treatment evaluation of hepatitis B

After an emotional analysis of 797 valid texts, the following Table 5 was obtained:

Table 5. Hepatitis B keywords' affective tendency and degree

Selected Keywords	Inferior Keywords	3rd Class Keywords	Emotional Intensity	Grade of Emotional Intensity	Emotional Tendency
Hepatitis B	Symptoms	Weak & Nausea	2.4	3	0
		Abdominal distension	0.8	1	0
		Liver pain	2.6	3	-1
	Treatment	Drug treatment	4.3	5	-1
		Conditioning	6.7	7	1
	Diet	Taboo	7.5	7	-1
	Others	Infection	7.8	7	-1
		Hospital evaluation			
		Drug prices	7.7	7	-1

As shown in Table 5, symptoms analysis is similar to that of lung cancer. Symptoms like weak, nausea and abdominal distension have low correlation with hepatitis B. while liver pain is relatively a unique symptom, given that it is difficult to distinguish liver pain from other areas' pain for normal patients, its emotional tendency is negative, and the emotional intensity is 3. In hepatitis B treatment, patients should pay more attention to the diet daily life, personal hygiene and so on. They need to exercise appropriately to prevent colds and infections. These activities are easier to accomplish and the effect is obvious, so the emotional inclination is positive and the intensity is 7.

In the diet taboo column, because hepatitis B causes liver function to drop, the patient has to eat healthy food in general, but these foods are relatively less delicious. The reason why emotions tend to be negative and intensity of 7 is that almost all hepatitis B patients have different degree of obesity, and it's also because their poor eating habits result in obesity and then they get hepatitis B. Patients with unhealthy eating habits are more likely to have negative emotions in diet control and it is difficult for them to stick to healthy diet.

In the infection column, emotions are negative and emotional intensity is high. Possible explanation is that hepatitis B has many transmission ways, so patients need to pay attention in their daily life. Also, those who don't have hepatitis B are concerned about the infection of hepatitis B from patients. They have high vigilance, which causes hepatitis B patients feel self-abased. As a result, the emotional intensity is high. The hospital evaluation is the same as lung cancer.

In the drug price column, the emotional tendency is negative and the intensity is very high. It's mainly because the cost of treatment is very high. Currently, the average cost of hepatitis B treatment is as high as 22,464 yuan while China's urban residents' per capita income is only 28,844 yuan. The hepatitis B treatment costs more than 40% of the average income of urban residents. Even for city dwellers, it is a huge expense. Let alone rural patients with an average annual income of just 9,800 yuan, it is a disaster. Rural patients account for more than one half of hepatitis B patients, so the emotional tendency and intensity of drug prices tally with the actual situation. If the price of drugs can be reduced from the current 22,464 yuan to 3,378 yuan, rural residents can afford hepatitis B treatment.

5.3 Treatment evaluation of diabetes mellitus (DM)

After an emotional analysis of 903 valid texts, the following Table 6 was obtained:

Table 6. Diabetes mellitus keywords' affective tendency and degree

Selected Keywords	Inferior Keywords	3^{rd} Class Keywords	Emotional Intensity	Grade of Emotional Intensity	Emotional Tendency
DM	Symptoms	3 more 1 less (Eat more, drink more, urinate more and weight loss)	5.5	5	0
		Obesity & weak	2.3	3	0
	Type	Type I		0	0
		Type II		0	0
	Treatment	Drug treatment	3.6	3	-1
		Insulin therapy	2.9	3	-1
		Exercise therapy	3.3	3	1
		Diet therapy	2.3	3	1
	Diet	Taboo	7.7	7	-1
	Others	Complications	9	9	-1
		Heredity	7.3	7	-1
		Hospital evaluation			
		Drug prices	7.5	7	-1

The medical evaluation of DM is shown in Table 6. DM is a metabolic disease and its symptoms are not obvious. Generally speaking it is diagnosed through regular medical examination. Its effects on people are not too obvious and it is not easy to be found by symptoms, so emotional tendency value is zero. The data of Type I and Type II diabetes are inadequate, so we didn't get reliable emotional tendency and intensity.

The treatment methods are mainly the medicine treatment and insulin treatment. Insulin treatment has no side effects, but it needs to be used for the whole life and the price is not friendly, so the emotional tendency is negative and the emotional intensity is 3, which is not that strong. The effects of the medication on the liver are relatively large, but the side effects are chronic, so the emotional tendency is negative and the emotional intensity is 3. As for exercise therapy and diet treatment as concomitant treatment, the affective tendency is positive, but the intensity is relatively low.

DM has more taboo in diet compared with other two diseases. Patients cannot take in foods with high level of sugar and they have to choose sugar-free, high-fiber foods. Moreover, they have to pay attention throughout their life. Eating food with poor taste throughout the life is a huge test for the patients, so the emotional intensity is high. Diabetes is often associated with multiple high risk complications and has hereditary, so it is highly negative in terms of emotional intensity and tendency. Diabetes treatment requires lifelong use of drugs, so patients are sensitive to drug prices.

6. Summary and discussion

In this paper, we summarized the application status quo in the field of medical and health data and realized that massive information on the medical forums or microblogs remain untapped. However, with the improvement of people's awareness about healthcare and the increase in the need of communication about healthcare information, we believe that the big data resource will be well used in the near future.

This article has innovation in terms of making sentiment analysis of texts related to medical and health treatments and the discussion of social evaluation on three types of diseases, which is supplementary to the professional evaluation of medical and health treatments. For example, we used data mining tools to collect data from medical forums and Weibo and built a structured database. Then we used affective computing of the collected texts based on HowNet-PMI method and made sentiment analysis of three diseases. To sum up, the results of text sentiment analyses are roughly in line with the actual situation, especially when a combination of outstanding emotional tendency and intensity shows up. Take the tendency and intensity of the price column for example, then the figures are all negative (-7) for three diseases, which shows the price of treatment is a big trouble for the patients. Thus, the research results and analyses can reveal some common views about certain treatments of the disease and encourage improvement on some factors in the treatment process such as price.

However, the research also has some limitations. The amount of data we collected and used in the research is limited, which brings difficulty in using precise quantitative model in the statistical analysis. In regard to sentiment analysis, our study focused on the sentiment of words, using the average as the intensity of the whole text without consideration of the contexts. Consequently, the analytic results might be inaccurate at times. Additionally, the characteristics of web comments, such as non-standard expressions, rich variations, and complex contexts of the comments are usually intractable in making text analysis. In this paper we did not conduct further research in this direction. However, since the texts we used in this paper are in Chinese, which is quite different from other languages such as English or Spanish, the effects such as variations of words and changed form of characters are negligible. In this light, we suggest further analysis to proceed with one of those languages.

This article just made a try in the big data analytics' application in the medical and health field. We suggest that future studies can improve from the following aspects:

- Take more data resources into account, such as the combination of clinical data and network data.
- Identify the validity of medical and health information from the internet.
- Improve the credibility of affective computing models.
- Apply the proposed method to other languages such as English and Spanish.

References

Attard, A. and Coulson, N. S. (2012). A thematic analysis of patient communication in Parkinson's disease online support group discussion forums. *Computers in Human Behavior*, *28*(2), 500-506.
Chang, C., Gao, G. D., Chen, H. and Li, W. X. (2003). Study on the diagnosis of Parkinson's disease with artificial neural network. *Chinese Journal of Clinical Rehabilitation*, *7*(28).
Dai, G. H., Fu, X. S., Dai, W. H. and Lu, S. Q. (2017). Social evaluation of innovative drugs: a method based on big data analytics. *Computer Science and Information Systems*, *14*(3), 805-821.
Dreiseitl, S., Ohno-Machado, L., Kittler, H., Vinterbo, S., Billhardt, H. and Binder, M.(2001). A comparison of machine learning methods for the diagnosis of pigmented skim lesions. *Journal of Biomed Inform*, *34*(1), 28-36.
Karaolis, M., Moutiris, J. A., Papaconstantinou, L. and Pattichis, C. S (2009). Association rule analysis for the assessment of the risk of coronary heart events. In *Proceedings of IEEE International Conference on Engineering in Medicine and Biology Society*, 6238-6241.
Li, Y. (2011). *National Health Risk Model and Risk Assessment*. The Fourth Military Medical University.
Liu, M. J., Wang, X. F. and Huang Y. L. (2000). Data preprocessing in data mining. *Computer Science*, *27*(4), 54-57.
Liu, Z. M. and Liu, L. (2012). Empirical study of sentiments classification for Chinese microblog based on machine learning. *Computer Engineering and Applications*, *48*(1), 1-4.
Pan, W. B. (2011). *Sentiment Tendency Analysis of Chinese Sentences Based on the Emotion Dictionary*. Beijing University of Posts and Telecommunications.
Pang, L., Shou-Shan, L. I. and Zhou, G. D. (2012). Sentiment classification method of Chinese micro-blog based on emotional knowledge. *Computer Engineering*, *38*(13), 156-158.
Scanfeld, D., Scanfeld, V. and Larson, E. L. (2010). Dissemination of health information through social networks: twitter and antibiotics. *American Journal of Infect Control, 38*(3), 182-188.
Wang, Z. Y., Wu, Z. H. and Hu, F. T. (2012). Words sentiment polarity calculation based on HowNet and PMI. *Computer Engineering*, *38*(15), 187-186.
Yue, X. and Jing, Y. (2009). Data mining algorithm of DNA sequences. *Journal of Biomathematics*, *24*(2), 363-368.
Zeng, Y. X. (2007). Biotherapy of tumor and molecule-targeted therapy. *Journal of China Prescription Drug*, *46*(1), 51-52.
Zhao, D. (2012). *Research on Core Technologies of Weibo's Precision Marketing Based on Data Mining.* Harbin Institute of Technology.
Zhou, L. and Srinivasan, P. (2005). Concept space comparisons: explorations with five health domains. In *Proceedings of AMIA Annual Symposium*, 874-878.
Zhu, Y. and Xiong, Y. (2007). DNA sequence data mining technique. *Journal of Software*, *18*(11), 2766-2781.