Computer Science and Information Systems 14(3):805–821 https://doi.org/10.2298/CSIS170413030D

Social evaluation of innovative drugs: A method based on big data analytics

Genghui Dai[1], Xinshuang Fu[2], Weihui Dai[3], and Shengqi Lu[4]

[1] School of Marine Sciences, Sun Yat-Sen University, Guangzhou 200433, China
daigengh@mail2.sysu.edu.cn
[2] School of Management, Shanghai University, Shanghai 200444, China
gracief@126.com
[3] School of Management, Fudan University, Shanghai 200433, China
whdai@fudan.edu.cn
[4] School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China
shengqilu@fudan.edu.cn

Abstract. The evaluation of drugs is a professional and time consuming process which involves a series of clinical trials and evidence-based verifications. However, an innovative drug may still suffer from unpredictable risks after coming into market due to the complex circumstances in practical utilization. Owing to the popularization of information networks and social media, big data analytics exhibits a new perspective of social evaluation as the supplementary means on this issue. This paper designed a Hadoop platform for data collection and processing, and explored the social evaluation of innovative drugs based on big data analytics. Through the analysis of mined data and affective computing on online comments, a new Chinese drug extracted from marine organisms can be evaluated comprehensively by the proposed method. Furthermore, the potential utilization of fullerene materials may be considered for improving its curative effects. Research work of this paper provides a big data analytics method for social evaluation of innovative drugs as well as their promising improvements.

Keywords: Big data analytics, social media, innovative drug, social evaluation, marine biology, fullerene materials.

1. Introduction

The full evaluation of drugs includes pre-marketing evaluation and post-marketing evaluation. In pre-marketing evaluation, an innovative drug has to pass IND(Investigational New Drug) and NDA(New Drug Application) procedures before it can be approved of coming into market. It is a professional and time consuming process, for example, the median time of a standard review is 384 days on IND, and 846 days on NDA by China Food and Drug Administration [30]. Although the above evaluation is based on a series of clinical trials and evidence-based verifications, but there are probably some unpredictable risks which may cause serious adverse reactions for an innovative drug in practical utilization [5][25] Therefore, post-marketing evaluation is the necessary and vital part in a full evaluation of innovative drugs.

As we known, current post-marketing evaluation of innovative drugs is mainly based on the statistical analysis of investigated samples or depends on special reporting channels such as the reporting system of health care institutions [25]. Nevertheless, a lot of disadvantages have been found in the existing method, such as limited samples, poor timeliness, inefficiency and the influence of uncertainty factors [2]. Actually, the practical curative effects and adverse reactions of drugs are mostly related to patients' individual conditions, living habits, and environmental factors. Especially for Chinese drugs, a reliable evaluation usually requires the comprehensive reviews from complex circumstances because there are important differences in different cases. It is difficult to be implemented through the current evaluation system.

Owing to the development of information technology and popularization of mobile applications, more and more people share their experiences of daily life by social media, such as shopping, tourism, medical treatment, and so on. In the meantime, the network of social media has appeared as a new platform and provides the valuable repository for scientific research and social study. As one of the most popular topics on social networks, health care and medical treatment attracts extensive concerns, and thereupon expedites the flourishing of various medical and health forums. The valuable information about innovative drugs in practical utilization can be mined from multifarious online comments and posts on the above forums through big data analysis. Therefore, big data analytics exhibits a new perspective of social evaluation of innovative drugs, which can be applied as the supplementary means to post-marketing evaluation. This paper aims to propose a Hadoop platform for data collection and processing from social media, and explore the social evaluation of innovative drugs based on big data analytics. It is organized as follows: Section 2 introduces the related works; Section 3 designs a Hadoop platform and studies the big data analytics from social media; Section 4 proposes the social evaluation method of innovative drugs; and Section 5 is the discussion and conclusion of this paper.

2. Related works

In recent years, big data analytics has been successfully applied in various fields such as financial markets, social management, production and manufacturing, as well as precision medicine, and shows superiorities over traditional methods in many aspects. In modern medicine and pharmacy, the classification of drugs is becoming more complicated than before, beyond the limitations to diseases or symptoms. As well, the ingredients of drugs are no longer invariants [25]. Those circumstances bring new difficulties and risks on the evaluation of innovative drugs.

Generally speaking, the evaluation of an innovative drug is based on a series of clinical trials and the comprehensive reviews on its effects [2][24]. For example, the test of drug allergy is carried out on extracts of natural drugs [11], and pharmacodynamic test is used for the evaluation of genetic engineering products [10]. However, the innovative drugs may still suffer from unpredictable risks after coming into market, and should be evaluated comprehensively through a professional and time consuming process. In recent years, the outbreak of new epidemic diseases such as influenza A (H1N1) has made the evaluation of innovative drugs faced with great challenges. In order to cope with this problem, many solutions have been proposed, one of which is the big data analytics. Up to now, many achievements have been made with the help of big data analytics [18][27]. It also

provides a new research methodology in medical and health fields, such as the analysis of diabetes cases, the study of regional characteristics of infectious diseases, the mining of disease causing factors, and so on. Zhu et al. summarized the research status and progress on the data mining of DNA sequence, and pointed out its significance in biological application [35]. Yue et al. applied data mining technology to study the classification of DNA sequences, and proposed a new judgment method to explore their classifications [33]. Li designed a health risk model for the assessment of Chinese people from the analysis of big data [16]. Karaolis et al. developed a data mining system to study the pathogenic factors of heart disease using association analysis algorithm [12]. Chang et al. adopted artificial neural networks to predict the outcome in the diagnosis of Parkinson's disease [3]. Dreiseit et al. proposed an improved method which combined artificial neural networks with regression analysis and decision tree to estimate the mortality in diseases [7].

In regard to big data analytics of health and medical information from social media, Zhou et al. used machine learning techniques to realize the automatic retrieval of online text information, and established a social medical terminology dictionary [34]. Ye et al. built a corpus of Chinese medicine, and studied the social evaluation of Chinese medicines in United States from the news reports and social media. Their research showed the social trend of increasing interest and attentions to Chinese medicines by American society and people [31]. Sampathkumar et al. applied Hidden Markov Model to analyze the adverse drug reactions based on the information of online healthcare forums, and provided an effective method for early warning of pharmacovigilance [21]. Existing research findings have indicated that the social evaluation of innovative drugs based on big data analytics can timely reveal the underlying influences and undiscovered effects of the above drugs from patients' feeling and their comments, which are hard to be reflected in the regular post-marketing evaluation.

3. Hadoop platform for data collection and processing

Through an analysis of the related works, we found that affective computing on text information is the useful big data analytics for the study of online comments [1][8][32] .In order to establish the big data environment for social evaluation of innovative drugs, we designed a Hadoop platform [14] to complete the data collection and processing, which can efficiently implement subject extraction and sentiment analysis from online comments. Its framework is designed as in Fig.1.

It includes three layers namely information collecting layer, data storage layer and business analysis layer. Firstly, the related text information are collected by web crawlers and sent to the text server group for preprocessing in information collecting layer. Secondly, the above data will be stored in MapReduce and HDFS in the data storage layer through the interface of HDFS [9]. Finally, text subject extraction, sentiment analysis, and other data analysis will be carried out in the business analysis layer, and all of data changes are executed by calling the data interface system such as HDFS and Hive.

3.1. Run mechanism for big data collection and processing

As big data analytics for social evaluation involves the collection and processing of enormous unstructured data from social media, it is necessary to design an efficient run mechanism carefully for dealing with the data. Hadoop platform has good capacity of distributed

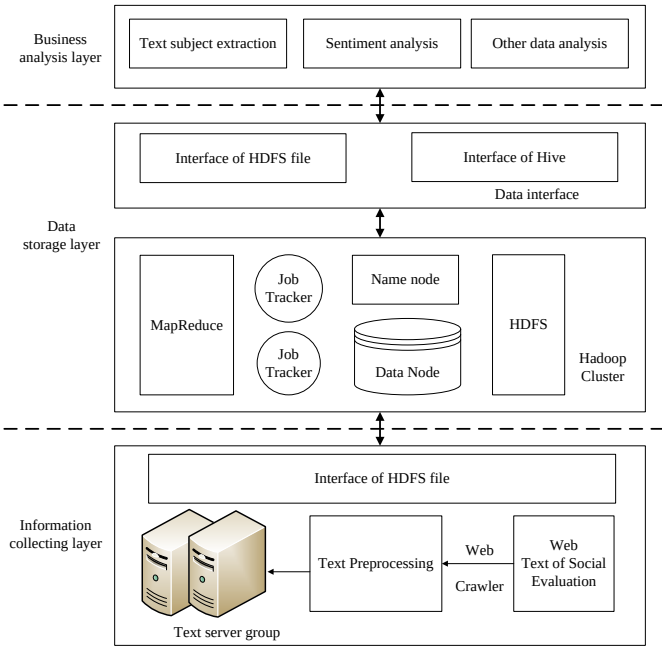

Fig. 1. Hadoop platform for social evaluation of innovative drugs

storage and parallel co-processing. However, its performance depends on the design of an effective run mechanism [22]. The platform includes three main components: master node, client node, and slave cluster, all of which coordinate with each other through the run mechanism to accomplish tasks. In our solution, we designed the run mechanism for big data collection and processing as in Fig. 2.

It can be seen from Fig.2, the Master Node is responsible for job management and resource scheduling, and slave cluster includes a lot of map tasks or reduce tasks for dividing sentiment words, subject extraction and so on. The above run mechanism can be described as follows.

Running mechanism. The running mechanism for data processing includes the following steps.

Step 1. Job submission, Firstly, the client node of the Mapreduce start a JobClient, and send a job with request ID to the JobTracker in Master Node by the JobClient, such as the job of dividing sentiment words.

Step 2. Job initialization, JobTracker puts the job into an internal queue, and hand over the scheduler job for scheduling, and then complete its initialization.

Step 3. Assignment of tasks, JobClient creates the corresponding number of Map tasks and Reduce tasks according to the number of input data, and assigns the Map task and Reduce task to the TaskTracker node in the Slave node.

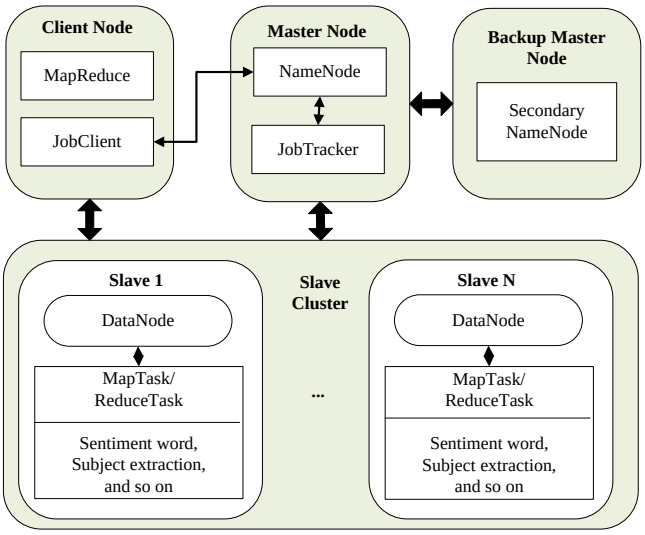

Fig. 2. The run mechanism for big data collection and processing

Step 4. Perform tasks, The TaskTracker node reads the input data stored on the HDFS, at the same time, the TaskRunner task will be created by MapTask and RedcueTask respectively, and the above two tasks will run until the end of task.

In the Hadoop platform, HDFS is responsible for the distributed storage of files in Hadoop cluster, which contains three major parts namely NameNode, DataNode, and Client.

NameNode. It acts as the management role in HDFS and is used to provide a name query service. It is responsible for managing the namespace of the file system, backup and the configuration of the cluster. In addition, the Metadata information stored in NameNode will be loaded into the memory after the NameNode starts.

DataNode. It is the basic unit of file storage, mainly used to save the information block, and will report to NameNode block when the DataNode thread is started, at the same time, it send a heartbeat in every fixed seconds to keep in touch with NameNode. Once NameNode hasnt received heartbeat within a fixed minutes, it means that the DataNode has been lost, and its block should be copied to the other DataNode.

Client. It is a client application to get files in distributed file system, which includes write file, read file and copy file block. The process of read file as follows. Client sends a request to the NameNode to read the file, and the NameNode return the address information of the DataNode that hold the data block, and then the Client calls the read() function to read data from the DataNode. When the Client data read is completed, it will call the close function FSDataInputStream(). In the process of data reading, if Client and DataNode

communication are errors, then Client tries to connect to the next data node. At the same time, the failure of the DataNode will be recorded.

MapReduce. It is responsible for the decomposition of tasks and the summary of the results. The tasks are distributed and completed by each individual node, and all the above nodes belong to a master node, and the final results come from each node through an integration of their intermediate results. The running mechanism of MapReduce is shown as in Fig. 3.

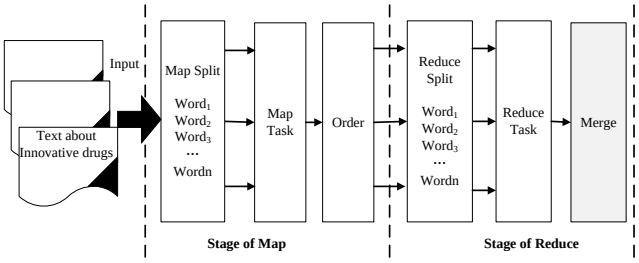

Fig. 3. The processing of the MapReduce

It can be seen from Fig.3, the mechanism of MapReduce include the map and reduce of tasks. In the process of map, data will be split into <key, value> according to the definition of Map function, and will be merged after completing the reduce tasks. It's worthy of mentioning that the Map process and Reduce process can run in parallel.

Task execution. The algorithm of task execution can be described as follows.

```
program Execution of task (Output)
  Init analysis task and hive database connection pool tp;
  begin
      (1) Get the connection from the hive database
          connection tool;
      (2) Connect to Hive, read the task of HQL, and
          send HQL query request to Hive;
      (3) Hive compiles and executes HQL, returns the
          execution result;
      (4) Write the result to local file and upload to
          HDFS path which is assigned by the analysis
          task;
      (5) Read result of the analysis task configuration,
          create new table in the Hive according to the
          configuration;
      (6) Upload the file in step 5 to new table which is
```

```
        create in step 6;
end.
```

3.2. Data analytics for social evaluation

The information about practical utilization of innovative drugs are scattered on microblogs or healthcare forums such as http://www.dxy.com, tieba.baidu.com, 91160.com, and so on in China. The above information are all unstructured texts, for example, the questions and answers, comments on the treatment of a disease or the curative effects of a drug, which are from the patients, family members, and doctors, and usually contain valuable information to be used for social evaluation.

The data analytics for social evaluation of innovative drugs includes text classification and affective computation. The purpose of text classification is to separate and keep the subjective text information for affective computation. It is realized by the subject extraction with a LDA model and the classification based on SVM (Support vector machine) and Bayes classifier. The purpose of affective computation is to calculate the trend and intensity of the above subjective text information for social evaluation. It is realized based on an emotional dictionary, and will be discussed in Section 4 of this paper. The outlined process of data analytics is shown as in Fig. 4.

It can be seen from Fig.4 that the data will be collected from various related websites by crawlers and preprocessed by filter and subject extraction. The specialized subjects will be extracted by LDA algorithm, and then classified by SVM and Bayes classifiers. If it is a subjective text, the affective intensity will be calculated for social evaluation. Otherwise, if it is an objective text, this text will not be processed.

LDA model. In order to extract the related subjects more efficiently, we used LDA model to fulfill this task. LDA model is also called the three layers Bayesian probability model, which includes the layers if words, subjects, and document structures. We hereby divide the above layers into: words, probable subject, and document sets. The matrix model of LDA can be shown in Fig.5 [6].

In Fig. 5, SE refers to the all of social evaluations on innovative drugs, and ϕ refers to the probability distribution of each subject on all terms. Θ expresses the subject distribution of each social evaluation. d_m is the m social evaluation, and w_n is the word of n term, and z_k is the k implicit subject.

In order to obtain the appropriate parameters of LDA model, the preprocessing data are used for training by the following steps:

Step 1. Initialization, randomly assign a subject number z to each of word w from prepared data. Generally, set α is 50/N_{theme}, where, N_{theme} is the number of subject, and β is 0.01.

Step 2. According to the Gibbs Sampling algorithm, collect the subject z from the set of word w, and update this set.

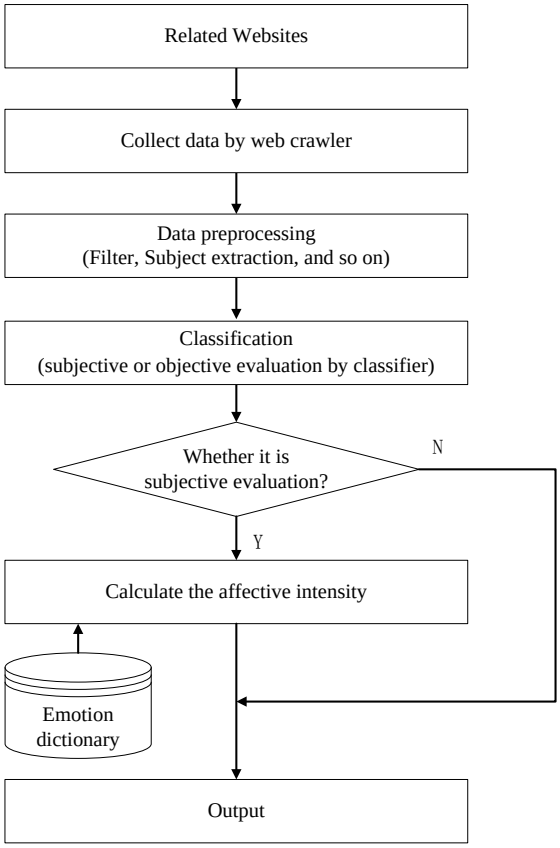

Fig. 4. Process of data analytics for social evaluation

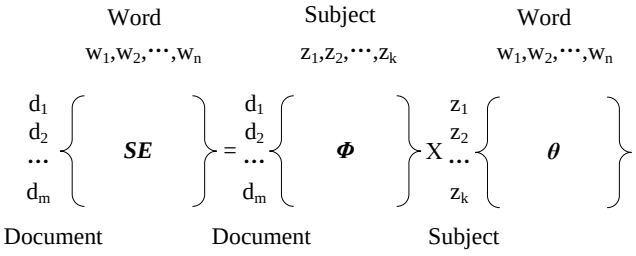

Fig. 5. The matrix model of LDA

Step 3. Repeat step 2 until Gibbs Sampling converges, that is to say, both subject distribution of each comment and word items of each subject are all convergence. After that, the probability distribution function is calculated as follows [19].

$$p(Z_i = k|\overrightarrow{Z}\neg i, \overrightarrow{w}) \infty \frac{n^k_{m,\neg i} + \alpha_k}{\sum_{k=1}^{K}(n^k_{m,\neg i} + \alpha_k)} \bullet \frac{n^t_{k,\neg i} + \beta_t}{\sum_{k=1}^{K}(n^t_{k,\neg i} + \beta_t)} \tag{1}$$

In 2, the probability distribution of subject-topic vector can be described as follows.

$$\theta = \frac{n^k_{m,\neg i} + \alpha_k}{\sum_{k=1}^{K}(n^k_{m,\neg i} + \alpha_k)} \tag{2}$$

As well, the probability distribution of subject-word can be described as follows.

$$\varphi = \frac{n^t_{k,\neg i} + \beta_t}{\sum_{k=1}^{K}(n^t_{k,\neg i} + \beta_t)} \tag{3}$$

Step 4. Calculate the co-occurrence frequency matrix of document-subject-word, and construct the LDA model.

Classified by SVM. Support vector machine (SVM) is a statistical machine learning classification method based on VC dimension theory and structural risk minimization principle. It has been widely used in affective computing on texts and vocal recognition for its superior performance on classification [4][23]. The classification method of subjective or objective comments by SVM is shown as in Fig. 6.

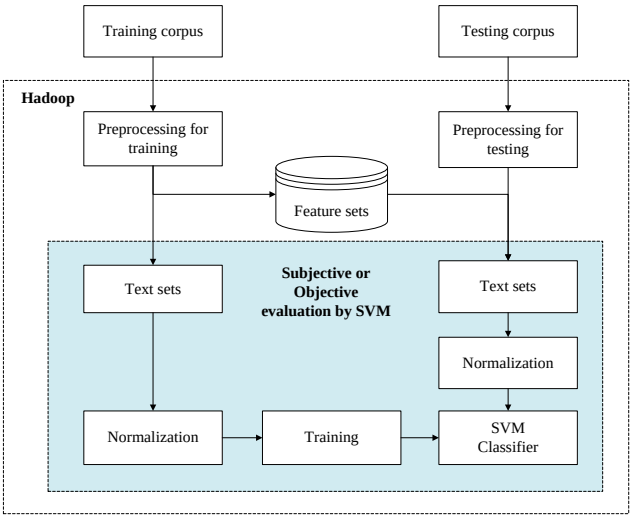

Fig. 6. The classification method by SVM

The classification algorithm can be described as follows.

Set $\{x_i, y_i\}_{i=1}^{n}$ as the set of data sample, where the input data x_i belongs R_d , and the output data $y_i \in$ (-1,1), then the linear discriminant function in d space is $f(x) = \omega \cdot x + b$, and the classification hyperplane equation is $\omega \cdot x + b = 0$. So the method of SVM in a high dimensional space can be described as:

$$y_i[\omega \cdot x + b] = 1 - e_i, i = 1, ..., n \tag{4}$$

Here, ω is the weight, and input x_i is the high dimensional space, b is the error constant, Therefore , the computation of the optimal classification can be converted into dual problem as long as the Lagrange optimization method is used. And the optimal classification function can be expressed as follows

$$f(x) = sgn(\sum_{i=1}^{N} \alpha_i^* y_i K(x_i, X) + b^*) \tag{5}$$

Where, b^* is the threshold of classification, $K(x_i, X)$ is kernel function and it was used the four forms as follows.

RBF kernel function:

$$K(x, y) = exp\frac{-||x - y||^2}{\sigma^2} \tag{6}$$

Linear kernel function:

$$K(x, y) = x^T \cdot y \tag{7}$$

Polynomial kernel function:

$$K(x, y) = [(x \cdot y) + 1]^m \tag{8}$$

Sigmod kernel function:

$$K(x, y) = tanh(u(x \cdot y) + c) \tag{9}$$

Here, the RBF kernel function was used in the SVM in our study.

4. Evaluation method based on big data analytics

As pointed out in this paper, the goal of social evaluation is to provide supplementary information for the comprehensive review on innovative drugs, and makes up the defects of a regular post-marketing evaluation. Therefore, the main role of big data analytics is reflected in the two aspects: new findings of the drug in practical utilization, and feeling and experiences of the drug in practical utilization. It has caused the researchers' attentions that the patients' emotional expressions about a drug possibly indicate its underlying influences and undiscovered effects, as well as the market value. We use affective computing technology to calculate the trend and intensity of emotions from the subjective texts. The above computation is realized based on the emotional dictionary developed by Prof. Lin et al [28], which includes 27,466 emotional words and divided into seven basic categories. At the same time, the collection rules of data should be built in order to get better results.

4.1. Collection rules and word frequency calculation

Collection rules. The valid data can be used for social evaluation should include the complete items: title, content, date of publication, and replying posts. Besides, the data promulgator must be identified, such as patient, family member, or doctor. Table 1 lists the samples of collection data.

Table 1. Sample of data collected

No	Title	Content	Date of publication	Type of promulgator	Count of replying posts
1	cerebral infarction	Butylphthalide is good for the disease · · ·	2017-02-25 13:37:28	Patient	17
2	Haishengsu	Will it affect patient's condition? · · ·	2017-02-24 15:08:42	Family member	22
3	Scopola mine Butylbromide Injection	It is used in the acute gastrointestinal tract · · ·	2017-02-21 15:08:42	Doctor	16
4	Domperidone Tablets	Lead to elevated serum prolactin levels · · ·	2017-02-21 16:08:42	Doctor	15
· · ·	· · ·	· · · · · ·	· · ·	· · ·	

Part-of-Speech. The segmentation methods for Chinese words commonly include forward maximum matching method [29], bidirectional maximum matching method [26] and reverse maximum matching method [20]. We adopted the NLPIR segmentation system [15] for word segmentation and extended it with the POS tagging. Therefore, each word is assigned by a Part-of-Speech as the samples shown in Table 2.

Table 2. Samples assigned by Part-of-Speech

No	Title	Annotation format
1	Nouns	/n
2	Verbs	/v
3	Adjectives	/a
4	Adverb	/d
5	Numerals	/m
6	Punctuation mark	/w
· · ·	· · ·	· · ·

In the processing of word segmentation, if a word is not included in the dictionary, it can't be identified, and should be added to the dictionary by manual. For example, 'Butylphthalide is good for the disease', in which the word of 'Butylphthalide' can't be found in the dictionary, and needs to be added to the dictionary. After processing of segmentation, the online comments still contain a lot of useless words, such as pronouns,

prepositions, determiners, auxiliary, conjunctions, interjections and onomatopoeic words. The above words can't help to extract subjects, but maybe reduce the calculation efficiency of LDA model, and need to be filtered out.

Word frequency calculation. Word frequency calculation is ready for affective computation and evaluation, and fulfilled by the parallel computing on Hadoop platform. Firstly, the type of input and output are built to class Mapper(), and their expressions are as follows: input type is <Object, Text>, and the output type is <Text, IntWritable>. If a task comes up, parallel computing is performed by calling the processes of Map() and Reduce() to complete the word frequency calculation. For example, Fig. 7 shows the word frequency calculation about the comments on fullerene materials.

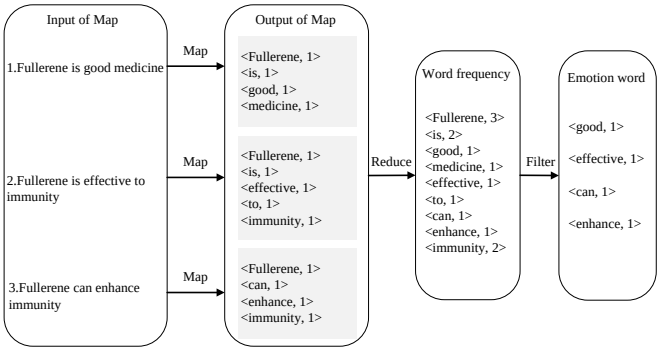

Fig. 7. Word frequency based on parallel computing of MapReduce

It can be seen from Fig.7, there are three tasks about the comments on fullerene materials in the Input of Map, and one task in Reduce. The above three text sections are independently assigned to three map tasks for processing firstly, and then the expression will be transformed into <'word', count value> by the specific function in intermediate process of Output in Map. Among which, count value refers to the total number of a certain word in the text section. Thereafter, the values will be used as input to the Reduce task, and the Reduce task will complete the computation of the total number of occurrences of each word. Finally, the three text sections will be merged into, and output the expression <'word', count value>, such as <immunity, 2> in the case about fullerene materials.

4.2. Subject extraction from online social media

After text preprocessing, the vocabulary dictionary needs to be built as input files for LDA model. In our research, the vocabulary dictionary has more than ten thousand words. Generally speaking, the parameters of the model must be initialized. Here, we set the initial value of the subject number to 5 according to the existing experiences [17]. As

well, we set α equals ten, β is 0.01, and the number of iterations for Gibbs sampling is 2000.

After 2000 iterations of Gibbs sampling, we can get the optimal extraction of the feature words on the five topics. Furthermore, five keywords are extracted respectively on each topic, and the distributions of the above keywords on each topic are shown as in Table 3.

Table 3. Distributions of the keywords on each topic

	Distribution				
1	Comfortable (0.2783)	Good (0.1067)	Body (0.0965)	Anticancer (0.075)	Depression (0.0023)
2	Blood (0.3135)	Complications (0.2149)	Lead to (0.1063)	decline (0.0075)	Form (0.0031)
3	Innovative drugs (0.0843)	Control (0.083)	Blood pressure (0.0645)	Appetite drugs (0.0473)	Dose (0.0163)
4	Eat (0.0873)	Food (0.0584)	Diet (0.1078)	shape (0.0873)	Marine organism (0.0464)
5	Symptom (0.2084)	Study (0.1172)	Technology (0.1070)	Treatment (0.0775)	Development (0.0562)

It can be seen from Table 3 that promulgators on social media pay more attentions to the above five types of subjects. The first subject is 'Effect description', which includes words such as comfortable, anticancer, and so on. At the same time, the other subjects have also been extracted. In addition, the number in the bracket of each word indicates the contribution of the word to this subject.

4.3. Emotional intensity analysis

To facilitate the evaluation, we divide the intensity of emotion into five levels, and assign to the value of 1, 3, 5, 7, and 9 respectively according to previous research [28]. As well, the emotional tendency is also assigned a polarity value. The positive tendency is expressed as 1, and the negative tendency is expressed as -1. The neutral tendency is expressed by 0. After quantized by the above values, the emotions becomes easy to be identified.

Based on big data analytics, we studied the social evaluations of a marine biological medicine and the fullerene materials, which have been reported as with significant curative effects and promising potentials for the treatment of tumors. With the rapid development of ocean resources, marine biological medicine has caused great interest by the developers of innovative drugs due to its natural and special bioactivity. 'Haishengsu', an innovative Chinese drugs extracted from marine organisms, was developed in recent years, and the clinical trials reported its significant anti-tumor effects. This innovative drugs was approved of coming into market in 2013. The emotional tendency and intensity about fullerene materials and 'Haishengsu' are shown as in Table 4.

Table 4. Emotional tendency and intensity about fullerene materials and 'Haishengsu'

First Keywords	Second Keywords	Third Keywords	Emotional intensity	Emotional tendency
Fullerene	Immunity	Lose weight	5	0
		Feel sleepy	3	-1
		Increased resistance	7	1
		Shortness of breath	5	-1
		Vulnerable to the cold	7	-1
Haishengsu	anti-tumor	Significant effect	9	1
		Affect physiological balance	3	-3
		Restrain the disease	5	1
...	...	...	...	...

It can be seen from Table 4, the value of the emotional tendency include three values (-1, 0, 1). From the value of 1, for example, we can deduce that fullerene has the positive function to increase resistance. The value of 0 means it is not associated with weight loss. As well, 'Haishengsu' has significant anti-tumor effects, and can restrain Heptocellular. However, its effects on physiological balance obtained a weak negative evaluation. The above studies show that social evaluations based on big data analytics may offer supplementary information about the innovative drugs in their practical utilization. It is very helpful for taking a comprehensive review on the innovative drugs, as well as for the improvement of the above drugs. Furthermore, the correlative analysis of evaluations indicates that curative effects of 'Haishengsu' are expected to be promisingly improved if combined with the utilization of fullerene. Therefore, big data analytics exhibits a new perspective of not only the new method for social evaluation of innovative drugs, but also the valuable information for promising development and application of the above drugs.

5. Discussion and Conclusion

Innovative drugs play the important role on promoting the progress of medicine and medical treatments. However, the traditional evaluation method of innovative drugs is a time consuming process, and has a lot of defects such as limited samples, poor timeliness, inefficiency, and the influence of uncertainty factors, especially in the face of sudden outbreak of diseases [13].

This paper designed a Hadoop platform and explored the social evaluation method of innovative drugs based on big data analytics. It aimed to provide the supplementary information for a comprehensive review on innovative drugs, as well as to make up the defects of a regular post-marketing evaluation. The main role of big data analytics is reflected in the following two aspects: new findings of the drug in practical utilization, and feeling and experiences of the drug in practical utilization. Research work of this paper provides a big data analytics method for the evaluation of innovative drugs, and as well, the valuable information for improving their promising development and application.

From the perspective of future research, more data sources such as geography and weather information, historical information about the process of treatments, and the accurate analysis methods such as logical reasoning and meta analysis, may be considered in

big data analytics for improving the precision of evaluations and providing more valuable details. In addition, how to use artificial intelligence to enhance the intelligent analysis ability is of great significance in the future researches.

Acknowledgments. This research was supported in part by Qtone Education of Ministry of Education of China (No. 2017YB115) and Shanghai Pujiang Program (No.16PJC007). Many thanks to Dr. Hongzhi Hu for her assistance to Prof. Weihui Dai and Xinshuang Fu who are the joint corresponding authors of this paper.

References

1. Ahmad, K.: Affective computing and sentiment analysis. emotion, metaphor and terminology. IEEE Intelligent Systems 31(2), 102–107 (2016)
2. Buyck, J.M., Tulkens, P.M., Bambeke, F.V.: Pharmacodynamic evaluation of the intracellular activity of antibiotics towards pseudomonas aeruginosa pao1 in a model of thp-1 human monocytes. Antimicrobial Agents & Chemotherapy 57(5), 2310–2318 (2013)
3. Chang, C.W., Gao, G.D., Chen, H., Li, W.X.: Study on the diagnosis of parkinson's disease with artificial neural network. Chinese Journal of Clinical Rehabilitation 7(28), 3818–3819 (2003)
4. Dai, W., Han, D., Dai, Y., Xu, D.: Emotion recognition and affective computing on vocal social media. Information & Management 52(7), 777–788 (2015)
5. Davis, C., Abraham, J.: The socio-political roots of pharmaceutical uncertainty in the evaluation of 'innovative' diabetes drugs in the european union and the us. Social Science & Medicine 72(9), 1574–1581 (2011)
6. Di, L., Du, Y.P.: Application of lda model in microblog user recommendation. Computer Engineering 40(5), 1–6 (2014)
7. Dreiseitl, S., Ohno-Machado, L., Kittler, H., Vinterbo, S., Billhardt, H., Binder, M.: A comparison of machine learning methods for the diagnosis of pigmented skin lesions. Journal of Biomedical Informatics 34(1), 28–36 (2001)
8. Fleuren, W.W., Alkema, W.: Application of text mining in the biomedical domain. Methods 74, 97–106 (2015)
9. Ghazi, M.R., Gangodkar, D.: Hadoop, mapreduce and hdfs: A developers perspective. Procedia Computer Science 48, 45–50 (2015)
10. Ghobrial, O., Derendorf, H., Hillman, J.D.: Pharmacokinetic and pharmacodynamic evaluation of the lantibiotic mu1140. Journal of Pharmaceutical Sciences 99(5), 2521–2528 (2010)
11. Gómez, E., Torres, M.J., Mayorga, C., Blanca, M.: Immunologic evaluation of drug allergy. Allergy Asthma & Immunology Research 4(5), 251–263 (2012)
12. Karaolis, M., Moutiris, J.A., Papaconstantinou, L., Pattichis, C.S.: Association rule analysis for the assessment of the risk of coronary heart events. In: IEEE International Conference on Engineering in Medicine and Biology Society. pp. 6238–6241 (2009)
13. Koukol, O., Kelnarová, I., Cerný, K.: Recent observations of sooty bark disease of sycamore maple in prague (czech republic) and the phylogenetic placement of cryptostroma corticale. Forest Pathology 45(1), 21–27 (2015)
14. Lee, T., Lee, H., Rhee, K.H., Shin, U.: The efficient implementation of distributed indexing with hadoop for digital investigations on big data. Computer Science & Information Systems 11(3), 1037–1054 (2014)
15. Li, X., Zhang, C.: Research on enhancing the effectiveness of the chinese text automatic categorization based on ictclas segmentation method. In: 2013 IEEE 4th International Conference on Software Engineering and Service Science. pp. 267–270 (2013)
16. Li, Y.M.: Research on Chinese health risks model and risk appraisal. Fourth Military Medical University, Xi'an, China. (2011)

17. Magnusson, M., Jonsson, L., Villani, M., Broman, D.: Parallelizing lda using partially collapsed gibbs sampling. Statistics 24(2), 301–327 (2015)
18. Mochón, M.C.: Social network analysis and big data tools applied to the systemic risk supervision. Ijimai 3(6), 34–37 (2016)
19. Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., Welling, M.: Fast collapsed gibbs sampling for latent dirichlet allocation. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, Usa, August. pp. 569–577 (2008)
20. Qu, H.Y., Zhao, W.: A revised bmm and rmm algorithm of chinese automatic words segmentation. Advanced Materials Research 267, 199–204 (2011)
21. Sampathkumar, H., Chen, X.W., Luo, B.: Mining adverse drug reactions from online healthcare forums using hidden markov model. BMC Medical Informatics and Decision Making 14(1), 1–18 (2014)
22. Seo, Y.D., Ahn, J.H.: Hadoop-based integrated monitoring platform for risk prediction using big data. Applied Mechanics & Materials 826, 113–117 (2016)
23. Silva, C., Ribeiro, B.: On text-based mining with active learning and background knowledge using svm. Soft Computing 11(6), 519–530 (2007)
24. Stegenga, H., Chambers, M., Jonsson, P., Thwaites, R., Garner, S.: A framework to guide the use of real-world evidence to support evaluation of relative effectiveness of new medicines. Value in Health 19(7), A488–A488 (2016)
25. Wang, M.: Progress and attention of new drug evaluation research. Journal of Chifeng University 24(4), 19–21 (2008)
26. Wei, X.U., Zhang, M.J., Xiong, Z.H.: Feature point matching based on bidirectional maximal correlation and parallactic restriction. Computer Engineering & Applications 44(28), 155–157 (2008)
27. Woodard, J.: Big data and ag-analytics: An open source, open data platform for agricultural & environmental finance, insurance, and risk. Agricultural Finance Review 76(1), 15–26 (2016)
28. Xu, L.H., Lin, H.F., Pan, Y., Ren, H., Chen, J.M.: Constructing the affective lexicon ontology. Journal of the China Society for Scientific & Technical Information 27(2), 5–7 (2008)
29. Yang, C.C., Luk, J.W.K., Yung, S.K., Yen, J.: Combination and boundary detection approaches on chinese indexing. Journal of the Association for Information Science & Technology 51(4), 340–351 (2000)
30. Yao, X., Ding, J., Liu, Y., Li, P.: The new drug conditional approval process in china: Challenges and opportunities. Clinical Therapeutics 39(5), 1040–1051 (2017)
31. Ye, Q., Wu, Q.: Study on report of american media of traditional chinese medicine situation and key words. AMIA Symposium proceedings 7(8), 626–629 (2014)
32. Yu, H.: Analysis on the subject and emotion of the medical forum of cerebrovascular disease. Beijing Jiaotong University, Beijing, China. (2016)
33. Yue, X., Jing, Y.: Research based on the algorithm of dna sequences data mining. Journal of Biomathematics 24(2), 363–368 (2009)
34. Zhou, L., Srinivasan, P.: Concept space comparisons: Explorations with five health domains. AMIA Symposium proceedings 2005, 874–878 (2005)
35. Zhu, Y.Y., Yun, X.: (dna) sequence data mining technique. Journal of Software 18(11), 2766–2781 (2007)

Genghui Dai is currently a Ph.D. candidate at the School of Marine Sciences, Sun Yat-Sen University, China. He received his master degree in Zoology from East China Normal University, China in 2005. His research interests include marine biology, microbial mechanism, and innovative drugs. Contact him at daigengh@mail2.sysu.edu.cn.

Xinshuang Fu, is currently a lecturer at the School of Management, Shanghai University, China. She received her Ph.D. in Management Science and Engineering from Shanghai University, China in 2013. Her research interests include knowledge management and information science. Contact her at gracief@126.com.

Weihui Dai is currently a professor at Department of Information Management and Information Systems, School of Management, Fudan University, China. He received his Ph.D. in Biomedical Engineering from Zhejiang University, China in 1996. He serves as a committee member of Shanghai Chapter, China Computer Society, and deputy director of Research and Translational Expert Council, Professional Committee of Endovascology, Chinese Medical Doctor Association. His recent research interests include complex system modeling and simulation, social media and intelligent information processing, social neuroscience and emotional intelligence, etc. Dr. Dai became a member of IEEE in 2003, a senior member of China Computer Society in 2004, and a senior member of Chinese Society of Technology Economics in 2004. His works have appeared in international journals with more than 130 papers. Contact him at whdai@fudan.edu.cn.

Shengqi Lu is currently a Ph.D. candidate at the School of Information Management and Engineering, Shanghai University of Finance and Economics, China. He received his master degree in Software Engineering from Fudan University, China in 2006. His current research interests include Artificial Intelligence and Business Intelligence. Contact him at shengqilu@fudan.edu.cn.

Received: April 13, 2017; Accepted: August 25, 2017.