Network Public Opinion Analysis Based on Text Mining

WanLe Chi

Department of Information Technology, Wenzhou Vocational &Technical College, Wenzhou, China

358455713@qq.com

Abstract

In order to control and guide the bad emotions of netizens in the process of network group emergencies, it is necessary to analyze the emotional tendencies of users based on network public opinion, so as to provide theoretical support and countermeasures for the government to effectively grasp and monitor network public opinion emergencies. Taking "Luo Yixiao" (girl with leukemia dies after fundraising effort) network hot topic event as an example, the public opinion information is analyzed and tracked. The main research work includes collecting and analyzing micro-blog data related to the incident. Based on the dictionary of How Net and other dictionaries, the emotional words are extended to build a more comprehensive dictionary of emotional classification. The emotion tendency analysis model is constructed to judge the emotional type and the statistical affective word frequency of the network public opinion. In addition, the user's emotion in this event is excavated and visualized. By using the empirical analysis, the evolution stage of the public opinion of the event is divided. The evolution characteristics and rules of the users' emotion in each stage are analyzed, and the reference basis is provided for the proposal of the follow-up network public opinion emotional guidance.

Keywords

Network public opinion, data mining, emotional classification.

1. Introduction

With the popularity of instant messaging tools, the emergence of "two micro and one client" (micro-blog, WeChat, news client) and the wide application of mobile Internet, the Internet platform no longer plays the one-way information transmission function simply. The platform has more advantages and characteristics and become an important carrier for netizens to express their opinions and participate in the interaction. With the help of all kinds of social media and self-media, the netizen can interact more actively with other users through a variety of forms, such as text and video. Their opinions, appeals and emotions show the characteristics of multi-directional communication, interactive communication and diversified content form, forming a complex network public opinion. Network public opinion refers to that netizens express their different views on the social hot events through the Internet platform. It is a collection of netizens' attitude, cognition, behavior and emotional tendencies. It is a collection of netizens' attitude, cognition, behavior and emotional tendencies. However, emotional tendencies are the barometer of public opinion, reflecting people's attitude towards events and emotional trends. It is a true reflection of the opinions and attitudes of all social strata and should be highly valued by the relevant departments and the media. Because emotional information affects the trend of social public opinion. Once losing control, it will lead to the emergence of extreme emotions or attitudes, and then produce the phenomenon of group polarization. In severe cases, it may cause violence in real society.

In recent years, network public opinion events have been frequent and cause huge repercussions, such as " Yitel Hotel attack ", "Female driver beaten up in Chengdu", "Luo Yixiao incident" and " Female tourists beaten up in Lijiang". These events have caused widespread concern and general discussion on the network.

2. Construction of emotional tendency analysis model

2.1 Construction of the emotional tendency judgment word list

Text mining, as one of the research fields of natural language processing, is to deal with semi structured or unstructured natural language text. Then, a certain technology is used to find and extract specific information from it. The network text mining first sets up the target text set by collecting the network text resources. Then, the text sets are processed by the techniques of text preprocessing, feature selection, feature representation and data mining. In addition, the specific information needed by the user is obtained.

Network text has no fixed data structure and model description, which belongs to unstructured or semi structured text. Therefore, it is necessary to convert and store the data of the network text information so as to complete the mining work. The data collected in this article are derived from the Sina micro-blog text. It is an unstructured text that contains unstructured data, such as pictures, hyperlinks and video. First, we need to remove unstructured data such as pictures and Web links to denoise the text. At the same time, we will process and save the structured data such as the release time of micro-blog, user publishing and publishing contents, so as to provide a basis for subsequent text mining and sentiment analysis. The length of micro-blog text is generally short. If each micro-blog is viewed as a text fragment, micro-blog is actually a collection of short text. However, the emotional information needed in this study is included in these micro-blog short texts. Therefore, it is necessary to use text mining techniques and methods to excavate emotional information by dealing with micro-blog texts. One of the research purposes of this paper is to solve the emotion information mining of short content in Weibo by using the text mining methods such as classification, matching, indexing and statistics.

It provides a theoretical framework and research foundation for subsequent users' emotional tendency and evolution analysis based on network public opinion data. First of all, it is classified according to the content of the emotional annotation based on target vocabulary. Finally, the emotion tendency judgment word list is formed.

Based on the How Net ontology, all parts of emotion words are labeled with part of speech, word meaning and word class. Among them, emotion annotation is starting from the emotion definition, the part of speech, meaning, intensity, polarity and emotion classification of emotion word are marked. Then, the ambiguous emotion words are removed and the final emotional words are obtained with a total of 34723. At the same time, with reference to the emotion words of Dalian University of Technology, the intensity, polarity and emotion classification of emotion word are marked.

Emotional polarity: It is used to describe the views for appraising characteristics and is divided into positive, neutral and negative；

Emotional intensity: It is divided into five levels of 1,3,5,7,9. 9 indicates the maximum strength and 1 indicates the least strength；

Emotional classification: At present, there are many emotional classification systems at home and abroad. It includes the two-category system, the three-category system and multi category system, and a unified classification standard has not been formed. Based on the 6 kinds of emotion classification system with great influence proposed by Ekman and the traditional classification of Chinese seven emotions, the emotional type in lexical ontology is divided into 7 primary categories, including happy, good, surprise, sadness, anger, fear, evil. Then, according to the emotional classification method proposed by Xu Xiaoying and others, the emotion is subdivided on the basis of 7 primary categories. Finally, a multi class emotional system with 21 emotional subcategories is formed.

The comprehensiveness and accuracy of the emotional vocabulary is of particular importance to the judgment of emotional tendency. Therefore, based on the specific event corpus, the emotional vocabulary is extended. The main processes include word division, merging, filtering and marking.

Finally, an emotional classification word list that can fully reflect the "Luo Yixiao" incident is obtained. The specific information is shown in table 1.

Table 1. Extended emotional classification word list

Subdivision	Code	Number of emotional words	Subdivision	Code	Number of emotional words	Subdivision	Code	Number of emotional words
Happy	PA	1726	Surprise	PC	583	Sadness	NB	1894
Peace	PE	956	Restlessness	NI	732	Anger	NA	736
Respect	PD	1169	Fright	NC	1052	Depression	NE	1367
Praise	PH	8985	Shy	NG	439	Hate	ND	1870
Believe	PG	816	Miss	NF	595	Blame	NN	7915
Affection	PB	1361	Disappointment	NJ	828	Envy	NK	386
Wish	PK	571	Guilt	NH	465	Doubt	NL	480
Total number of emotional words						34926		

2.2 Emotional tendency analysis model

First, the micro-blog text is decomposed into sentences, and the emotional polarity of each sentence is further analyzed. However, the emotion polarity of sentence can be calculated according to polarity of emotion words in the sentence. Then, the emotion value of all sentences is combined to calculate the emotional polarity of the text. The importance of a sentence to the whole text can be expressed by the weight calculated by the comprehensive polarity of the sentence in the text. The problem is described in the form of a formula: For a given text D, the sentence is composed of a sentence set S. First, the emotional value F(Si) of each sentence Si is calculated. As shown in formula (1) and formula (2), the emotional values of the micro-blog text D can be determined by the emotional values of the sentence.

$$\mathrm{F}(S_i) = \sum S_{w_i} \tag{1}$$

$$\mathrm{F}(S) = \sum F(S_i) \tag{2}$$

In the formula, Swi is the emotional value of the emotion word wi in a sentence. If F(S)>0 is, the text expresses positive emotions. If F(S)<0, the text expresses negative emotion. If F(S)=0, then the text expresses neutral emotion.

Because users' specific emotions cannot be expressed in declarative sentences, this paper uses rhetorical questions, interrogative sentences, exclamatory sentences and hypothetical sentences to predict emotional tendency. Sentence patterns can be identified by some characteristic words or punctuation marks. As shown in formula (3), the emotional value F'(S) of the micro-blog text is calculated by F'(Si).

$$\mathrm{F}^{'}(S) = F^{'}(S_i) \tag{3}$$

3. Analysis of emotional tendency of network public opinion based on Text Mining

First, the micro-blog text content of "Luo Yixiao" incident is collected and processed. Then, text mining and emotional visualization technology are adopted, and the word frequency of the important emotion words in the incident is counted. Then, the emotional characteristics of users' emotional types and polarity intensity are counted. Therefore, the emotional types and intensity of network public opinion are grasped as a whole.

3.1 Data acquisition and processing

With the popularity of 4G network and the development of Web 2.0 technology, micro-blog has become the fastest growing social network platform. It can not only satisfy users' needs for information acquisition, but also enable people to express their thoughts, emotions, attitudes and opinions through the way of releasing micro-blog. Micro-blog has a good communication effect. Moreover, the speed of information dissemination is very fast. Its forwarding function makes the user's emotion spread quickly in a short time, and has formed a wide and profound influence.

The "Luo Yixiao" incident occurred in November 2016, as one of the ten family events in 2016, has good research value in the tendency and evolution analysis of network public opinion. Taking "Luo Yixiao" as search keywords, the octopus data collector is used to collect the micro-blog text as a data source. The collection time is from 0:00 on 29th, Nov 2016 to 0:00 on 31th, Dec 2016. The collected data from Sina micro-blog, including original micro-blog and forwarding micro-blog, has 78536 network public opinion data related to the event.

First of all, the collected data should be de-weighted and removed. Then, the incomplete data or the empty data are deleted. Finally, the micro-blog data are repeatedly cleaned and checked. In the end, there are 74025 effective micro-blog data related to the "Luo Yixiao" incident. It includes the original micro-blog and the forwarding micro-blog. Among them, the data content of original micro-blog includes user name, user homepage, publishing time, publishing way, publishing content, forwarding number, comment number, publishing website and publishing location. The data field of forwarding micro-blog includes forwarding user's username, user homepage, forwarding time, forwarding text content, original micro-blog release time, original micro-blog content.

Before analyzing the collected public opinion data, we should first preprocess the collected data. That is, the collected data should be de-weighted and removed. First, the same content microblog released by the same user at the same time is deleted to avoid duplication of data affecting the validity of subsequent emotion analysis results. Then, the survey process needs to delete the wrong micro-blog data, that is, to delete incomplete or empty data. Furthermore, it is necessary to delete a large amount of junk information, such as a large number of advertisements, which are produced by adding the heat of "Luo Yixiao" incident. Finally, the micro-blog data are repeatedly cleaned and checked, and a total of 74025 effective micro-blog data related to the "Luo Yixiao" event are finally obtained.

3.2 Emotion word frequency analysis

The first 30 emotion words in the micro-blog text of "Luo Yixiao" incident are counted. These high-frequency affective words can reflect the main emotions of users in the whole event, and show the main concerns of online users, such as "love", "kindness" and "help". From this, it shows that the most expressed emotion is love when users face this social event.

After dividing 74025 micro-blog texts issued by 64516 micro-blog users, a total of 27226 words are obtained. The content of micro-blog text is matched with the emotional word list established in this article for emotional words. There are 6876 emotional words that are finally extracted. The word frequency gap between each emotional word is large, ranging from 1 to 4405 times. After counting the word frequency of emotional words, it is found that some emotion words such as love, kindness, help, donation, people, fraud, assistance and care reach more than 200 times. It shows that these high frequency emotion words are the dominant emotions of the whole event. However, some emotion words such as "ulterior motive”, “no ground for blame", “miserly", "incurable disease" and “recriminate” belong to low frequency words, and these words show the user's own feelings. The low frequency words such as "helpless", “peace” and "indeed" can reflect the user's own attitude and some hidden emotions compared with high frequency words.

Through the above analysis, it is found that the high frequency emotion words of users in social networks reflect the dominant emotional tone of the network public opinion events. However, with the decrease of word frequency, emotional words become more subjective and reflect users' emotional experience and subjective evaluation for the whole event.

3.3 Analysis of emotional type

Based on the emotion classification dictionary established in this paper, 6876 emotional words extracted from 74025 micro-blog texts are classified. The statistical results are shown in figure 1. It shows that "praise" is the most important emotional type conveyed in the text of the event. The total frequency of this type of emotion word is close to 50000 times. The second is "blame", the total frequency of this type of emotion word is more than 30000 times. Among the rest of the emotional types with less frequency, "trust" and "sadness" are more prominent. From this, it is concluded that the "Luo Yixiao" incident, as an unpleasant online public opinion event, has received universal condemnation of netizens. But, the real encounter of this girl (Luo Yixiao) also gets the love and help of the vast netizen.

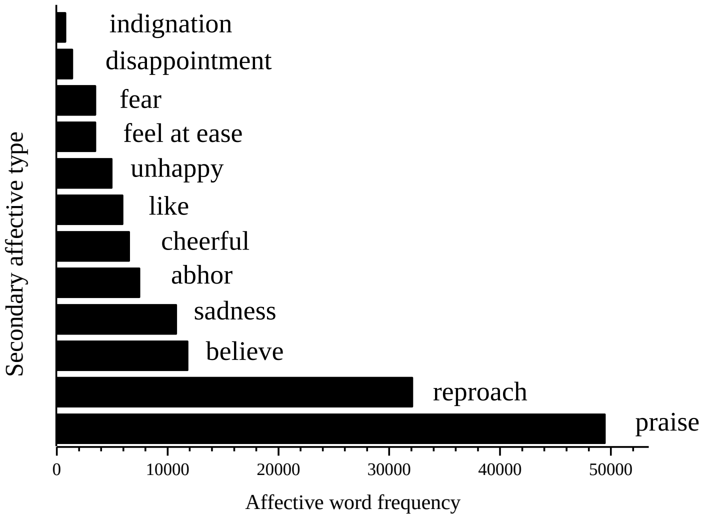

Figure 1. Statistics of the emotional type of “Luo Yixiao” incident (Based on 21 subdivision)

The subdivision emotions in figure 1 are merged into the total emotions, and the emotional type statistics are shown as shown in figure 2. The emotions of the netizens mainly focus on "good" and "evil". After analyzing the content of micro-blog text, most netizens show support, care and help to the misfortune of the unfortunate girl. On the other hand, most netizens dislike and criticize the shameless behavior of her father. Some netizens also express sadness and apathy for the girl.

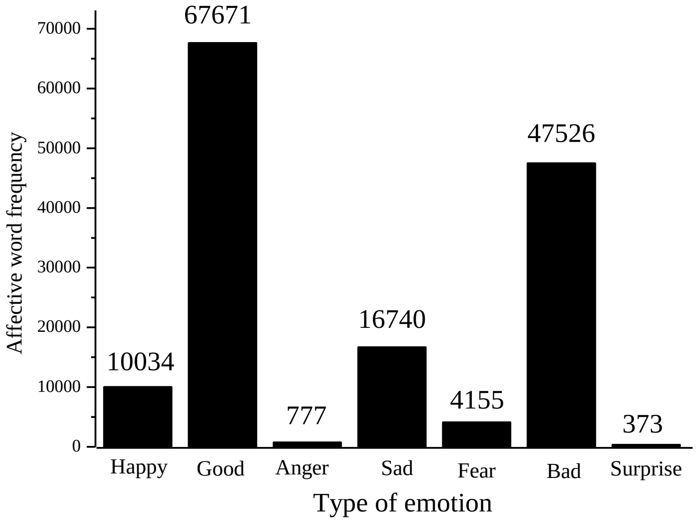

Figure 2. Statistics of the emotional type of the “Luo Yixiao” incident (Based on 21 subdivision)

3.4 Analysis of the intensity of emotional polarity

According to the calculation rules of the emotional eigenvalue of the sentence, 74025 micro-blog texts related to the "Luo Yixiao" incident are calculated and counted. Furthermore, the polarity of micro-blog is judged according to the value of emotional tendency. The micro-blog of the event is divided into three categories according to polarity: including positive micro-blog (emotional tendency is greater than 0), neutral micro-blog (emotional tendency is equal to 0) and negative micro-blog (emotional tendency is less than 0). As shown in figure 3, in the micro-blog event related to "Luo Yixiao" incident, the number of positive micro-blog is the largest, reaching 48093 and accounting for 65% of the total number. The number of positive micro-blog is even more than twice as much as that the total number of negative micro-blog (17661) and neutral micro-blog (8271). Thus, although the "Luo Yixiao" incident has caused a greater negative impact on the society, on the whole, the netizens have a positive attitude.

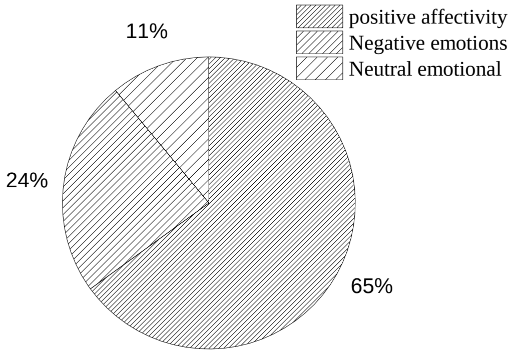

Figure 3. Statistics of micro-blog polarity

The intensity of emotional polarity of micro-blog text is further divided. According to the size of micro-blog's emotional tendency, the intensity of emotional polarity can be divided into general positive (0, 10), moderate positive [10, 20), high positive [20, +∞), general negative (-10, 0), moderate negative (-20, -10] and high negative (-∞, -20]. The number of positive micro-blog and negative micro-blog published in the "Luo Yixiao" incident is shown in figure 4. Therefore, the number of micro-blog at the "general" degree occupies a great proportion. The number of micro-blog in any degree of positive emotion is more than the number of negative emotions.

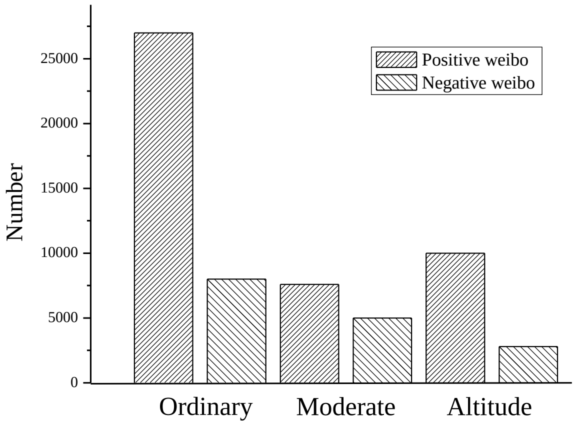

Figure 4. Statistics of micro-blog polarity intensity

3.5 Empirical analysis of the emotional evolution of network public opinion

Based on the online public opinion data of "Luo YiXiao" network emergency, the empirical analysis is used to divide the evolution stage of public opinion according to the theories and methods of network public opinion and emotion analysis at home and abroad. Moreover, user's emotion at each stage is excavated, and user's emotion type and emotion frequency at each stage are further judged and counted to find the evolution characteristics and rules of user's emotion at different stages.

With the "Luo Yixiao" incident network public opinion as the research object, the 74025 micro-blog data related to the event are statistically analyzed. The number of micro-blog, the number of user participation and the daily distribution of micro-blog per person are summarized. The obtained results are shown in table 2. It can be seen from the table that the outbreak and fermentation period have an important position in the evolution of the whole event. The number of micro-blog releases and the number of users participating in the outbreak and fermentation periods are the highest in all evolutionary stages. This is the main data source of sentiment orientation recognition and analysis in this paper. Therefore, it is necessary to analyze the evolution process of these two stages in order to find out the characteristics and rules of emotional evolution, and provide a reference to put forward the emotion guidance strategy of Internet public sentiment in later.

Table 2. Statistics of microblogrelease about "Luo Yixiao" incident at each evolution phase

Evolution time	Time interval (2016)	The number of micro-blog	Participation in quantity	The average number of micro-blog releases per person per day
Beginning period	11.29	17	17	1.00
Outburst period	11.30	43373	39434	1.10
Fermentation period	12.01-12.02	20890	18793	0.56
Digestion period	12.03-12.23	5057	4758	0.05
Rethinking period	12.24-12.31	4688	4457	0.13

The statistic software Stata12 is used to analyze the basic statistical characteristics of emotional polarity of the 74025 microblogs about "Luo Yixiao” incident. The results are shown in table 3. The average emotional polarity of the incident is 4.606, and the standard deviation is 16.636. The maximum is 388 and the minimum is -122. The difference between the intensities of the Weibo data is significant, showing both extreme positive emotions and negative emotions. In addition, the skewness is 1.862 and the kurtosis is 24.667. Therefore, the distribution of microblog emotional polarity does not meet the normal distribution because the network public opinion information itself has a cumulative effect. Moreover, during the dissemination of public opinion, due to the extremely strong public psychology of netizens, it is easy to have a cluster effect in the network, resulting in extreme fluctuations in the network public opinion information.

Table 3. Description statistics of emotional polarity intensity about "Luo YiXiao" incident

Variable	Obs	Mean	Std.Dev.
Polarity	74025	4.605606	16.63568
Min	Max	Skewness	Kurtosis
-122	388	1.861705	24.66628

The emotional polarity and emotional types of the evolution of network public opinion are analyzed. The results are shown in table 4. It shows that the emotional polarity of "Luo Yixiao" incident in each evolution stage is mainly positive. Moreover, the number of positive micro-blog at each stage is more than the negative micro-blog. Among them, at the beginning stage, because of the small number of micro-blog, the average strength of micro-blog reaches 21.7, becoming the maximum value of all stages. At this stage, the girl, a young patient with a serious illness, is the object of the people's care, support and help. The emergence of "suspicion" emotion has buried the back of the public opinion. From the outbreak to the fermentation period, the average intensity of micro-blog declined obviously because of the exposure of Rolle's personal property. With the further development of the event, after Rolle returned the donation and apologized, the netizen's emotion had been restored to a certain extent. The average polarity of micro-blog increased gradually in the period of digestion and reflection. In the course of the development of the whole event, the beginning period mainly focuses on "praise" and other positive emotions. From the outbreak to the reflection period, both negative emotions, such as "derogatory" and positive emotions such as “praise", are contained. In general, this event is mainly positive. In addition to the "praise" and "blame", "Sadness" is one of the emotions shared by users during the evolution of public opinion.

Table 4. Statistics of emotional polarity intensity and type about "Luo Yixiao" incident

Evolution stage	Emotional polarity				Average polarity strength	Main emotional types		
	Positive	Negative	Neutral	Total				
Beginning period	11	3	3	17	21.7	Praise	Believe	Doubt
Outburst period	28587	9889	4897	43373	4.84	Blame	Praise	Sadness
Fermentation period	12957	5802	2131	20890	4.02	Blame	Praise	Sadness
Digestion period	3560	906	591	5057	4.15	Blame	Praise	Sadness
Rethinking period	2978	1061	649	4688	5.42	Blame	Praise	Sadness
Total	48093	17661	8271	74025	8.03	Praise	Blame	Believe

According to the periodic change process of network public opinion events from occurrence, development to retreat, the various stages of network public opinion evolution are divided. Moreover, the emotional changes of network public opinion in the various stages of evolution are counted and analyzed.

The study found that: In the beginning stage, the netizen's attitude to the network public opinion event is complicated. However, the extraction of emotions and views is conducive to further tracking the future trend of events. For example, in the "Luo Yixiao" incident, the rise of "suspicion" of emotions at the beginning stage laid a foreshadowing for the netizens turning to Rolle's exposure and accusation. In the outbreak period, the number of micro-blog is the most, and the user's participation is also the highest. The information of netizens' attitudes, views and emotions can provide a full data basis for the analysis and monitoring of network public opinion. At the same time, the emotional tendency in the outbreak period defines the overall emotional evolution trend of the network public opinion events to a great extent. The relevant departments should pay more attention and guidance to the emotional evolution of network public opinion in the outbreak period. In the fermentation period, the netizens are more sensitive to the new information and new trends of network public opinion events. Positive information disclosure and public opinion disclosure can play a good effect in this stage. In the period of digestion and reflection, the participation degree of users is low. Therefore, the media department still needs to track and report the network public opinion events, avoid rumors, clean up the network environment, and avoid the two fermentations of network public opinion events.

4. Conclusion

Taking the network public opinion of "Luo YiXiao" network emergency as the research object, a more comprehensive emotion classification dictionary is constructed on the basis of the research of theories and methods of network public opinion and emotion analysis by domestic and foreign scholars. At the same time, an emotional tendency analysis model is established. The user's emotion is excavated and visualized in this event, and the user's emotional type and emotional word frequency are judged and counted. This model can provide theoretical support and decision-making basis for the relevant departments to effectively guide the emotional evolution of network public opinion. The emotional expression and dissemination of netizens have a direct impact on the emergence and development of events. If we don't monitor and control negative emotions or negative emotions in a timely manner, netizens' emotion will easily become extreme, which will lead to the emotional polarization. Eventually, there will be a lot of disruptive mass events that will impact or threaten social security and stability.

Acknowledgements

Avoid the stilted expression, “One of us (R. B. G.) thanks...” Instead, try “R. B. G. thanks”. Do NOT put sponsor acknowledgements in the unnumbered footnote on the first page, but at here.

References

[1] Kumakawa, T. (2017). Text mining infrastructure in r. Open Access Library Journal, 04(6), 1-6. https://doi.org/10.18637/jss.v025.i05

[2] Park, K., & Kremer, G. E. O. (2017). Text mining-based categorization and user perspective analysis of environmental sustainability indicators for manufacturing and service systems. Ecological Indicators, 72, 803-820.https://doi.org/10.1016/j.ecolind.2016.08.027

[3] Sheng, X., Wu, X., & Luo, Y. (2017). A novel text mining algorithm based on deep neural network. International Conference on Inventive Computation Technologies (pp.1-6). IEEE.https://doi.org/10.1109/inventive.2016.7824810

[4] Ruch, P. (2017). Text mining to support gene ontology curation and vice versa. Methods in Molecular Biology, 1446, 69-84.https://doi.org/10.1007/978-1-4939-3743-1_6

[5] Kim, D., & Kim, S. (2017). Sustainable supply chain based on news articles and sustainability reports: text mining with leximancer and diction. Sustainability, 9(6), 1008. https: //doi. org/10.3390/su9061008

[6] Olorisade, B. K., Brereton, P., & Andras, P. (2017). Reproducibility of studies on text mining for citation screening in systematic reviews: evaluation and checklist. Journal of Biomedical Informatics, 73, 1.https://doi.org/10.1016/j.jbi.2017.07.010

[7] Wang, F., Xu, T., Tang, T., Zhou, M. C., & Wang, H. (2017). Bilevel feature extraction-based text mining for fault diagnosis of railway systems. IEEE Transactions on Intelligent Transportation Systems, 18(1), 49-58.https://doi.org/10.1109/tits.2016.2521866

[8] Caã±Ada, A., Capella-Gutierrez, S., Rabal, O., Oyarzabal, J., Valencia, A., & Krallinger, M. (2017). Limtox: a web tool for applied text mining of adverse event and toxicity associations of compounds, drugs and genes. Nucleic Acids Research, 45(Web Server issue), W484-W489. https://doi.org/10.1093/nar/gkx462