Research on netizens' emotion evolution in emergency based on machine learning

Zhongjie Ren[1], Peng Zhang[1], Jing Liu[2] and Yuexin Lan[3]

[1]Department of Fire Commanding, People's Police University of China, Langfang, Hebei, China
[2]Department of Literature and Law, North China Institute of Aerospace Engineering, Langfang, Hebei, China
[3]Department of Border Control, People's Police University of China, Langfang, Hebei, China

E-mail: zp_1981@aliyun.com

Abstract. [Purpose/significance] By analyzing the emotional evolution of Weibo users after emergencies, we can find out the law and potential risks of public opinion evolution and provide instruction for the government to control and guide network public opinion. [Method/process] We proposed an analysis model of public opinion evolution based on Emotional Analysis and GBRT. With the help of Python, a web crawler was developed to collect Weibo comments. After that, a Naive Bayesian Classifier was used for emotional analysis. According to public emotion and the number of comments, we divided the evolution process into fever period, persistence period, incubation period and extinction period. Statistical and visualization methods were used to study the evolution characteristics of word cloud, emotional tendency and age groups. Finally, correlation analysis and GBRT were used to predict each individual's emotions. [Results/conclusion] Taking the dangerous chemical explosion accident in Tianjin as an example, we can validate our model. Results shows that the model can reasonably divide the evolutionary stages, find out the law of public opinion evolution in different stages, and accurately predict users' emotional tendency.

1. Background

Nowdays, the social media platform dominated by micro-blog has gradually become the main space for the evolution of public opinion [1-3]. Weibo users' emotions can be infected by other's quickly and lead to a heated debate on the internet [4-5]. Grasping the law of periodic fluctuation of Weibo users' sentiment and further predicting it can prepare the government for potential risks.

In researches about the evolution of Weibo public opinion, Youyonghua divides the evolution process into rapid diffusion stage, stable development stage and declining stage according to two indicators of attention and attention increment [6]. Wu Xiaojuan divides the development of public opinion of Blue Qianjiang arson case into initial stage, the outbreak stage, the repeated stage and the long tail stage, and researches the themes of each stage [7]. Lan Yuexin and others constructed the logistic model of public opinion propagation process, dividing the incident of "Chengdu men beating women drivers" into five stages [8]. At present, the division is mostly according to micro-blog comment volume, number of followers and number of forwards, without considering emotional characteristics. In this paper, emotional inclination index is introduced into the criterion of stage

division of public opinion. There are mainly two kinds of emotional orientation research methods, one is emotional dictionary, the other is machine learning. This paper uses naive Bayesian to conduct emotional analysis.

At present, most of the literature on public opinion emotion prediction is the prediction of group emotion trend. Through time series and gray prediction method, the short-term trend of public sentiment is predicted. Du zhitao et al. predicted online public opinion trend by establishing a grey prediction model [9]. Qian ailing et al. used multi-time series association rules to analyze the BBS trend [10].Zhang heping established an improved Gray Verhulst Markov model to predict the development of public opinion [11].

The current research mostly adopts the time series method, which is used to predict Baidu Index or the average emotional tendency of the public. The time series is simple to use and the system is mature. When the trend of public opinion changes steadily, it can well predict the short-term trend of public average emotion. However, the public opinion is complex, changeable, so the time series is difficult to carry on the long-term forecast; besides, mostly the target of prediction is the average emotion of the public, which cannot reflect the emotion distribution cannot predict the emotion of each member of the group. In addition, users' distinctive features, such as education background, age, gender, number of fans, active status of Weibo, often have a great influence on users' emotional tendency, which cannot be reflected in the methods of time series.

Based on the emotional analysis, statistical analysis, data visualization methods and machine learning, we research the evolution of public sentiment of web users and build a model, which has been validated by Tianjin 8•12 accident cases.

2. Web users' sentiment evolution model

2.1. Research methods

The methods involved in the model are described below.

2.1.1. Web crawler

Web crawlers can bulk collect and download data from web pages. At present, many achievements have been made in studies related to it [12-13].To conduct our research ,we made web crawler Specifically applicable to Weibo with Python, which can simulate users' login and search in Weibo, automatically grab microblog comments on relevant emergencies, and simulate the click action to extract more relevant information.

2.1.2. Emotional analysis

Emotion analysis requires word segmentation to be done first. There are some ways: dictionary-based method, statistical method and rule-based method [14]. This paper adopts SnowNLP, a natural language processing module of Python, to conduct word segmentation. Then Naive Bayes classifier was used for emotion analysis. Naive Bayes classifier has high accuracy, reliable classification. Simply and fast, it can make out each sample's classification probability.

Let's say that there are N categories of problems that we're dealing with, $Y = \{c_1, c_2, c_3, c_4 ..., c_N\}$.

According to Bayes theorem, that is

$$P(C|X) = \frac{P(C)\,(X|C)}{P(X)} \tag{1}$$

Where $P(C)$ is the prior probability of the class; $P(X|C)$ is the sample's conditional probability

For each sample, we select the category that minimizes the risk of misclassification on that sample

$$h^* = argminR(C|X) \tag{2}$$

Take the lowest misjudgment rate as the target, and the cost function is as follows:

$$\lambda_{ij} = \begin{cases} 0, & if \quad i = j \\ 1, & otherwise \end{cases} \tag{3}$$

The corresponding conditional risk is

$$R(c|x) = 1 - P(c|x) \quad (4)$$

The classifier whose goal is to minimize error rate is

$$h^* = argmaxP(c|x) \quad (5)$$

Naive Bayes classifier assumes that for a given category, all attributes are independent. Therefore, equation (2) can be written as

$$P(C|X) = \frac{P(C)P(X|C)}{P(X)} = \frac{P(C)}{P(X)}\prod_{i=1}^{d} P\ (x_i|c) \quad (6)$$

So we can assume that:

$$h(x) = arg\ max\ P(c)\prod_{i=1}^{d} P\ (x_i|c) \quad (7)$$

2.1.3. Correlation analysis

Correlation analysis can analyze the relationship between two different variables. The simple correlation coefficient is defined as*:*

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})\left(y_i - \overline{y}\right)}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}} \quad (8)$$

N is the sample size, x_i and y_i are the variable values of the two variables respectively

The test statistic adopts t statistic, which is defined as:

$$t = \frac{r\sqrt{n-1}}{\sqrt{1-r^2}} \quad (9)$$

Simple correlation coefficients, observed values of t-test statistics and corresponding probability values can be calculated.

2.1.4. GBRT model

GBRT is an important algorithm of machine learning. Taking the emotional tendency value obtained in the emotional analysis model as the dependent variable and each attribute of the user as the independent variable ,we can train the GBRT model [15].Gradient promotion tree (GBRT) algorithm, by iterating multiple regression trees to make joint decisions, improves the weak learner to the strong learner and has good generalization characteristics [16].

2.2. Model flow char*t*

The flow chart of the model is shown in figure 1, which includes five stages: data acquisition, data preprocessing, emotion analysis, public sentiment evolution stage division, and GBRT modeling.

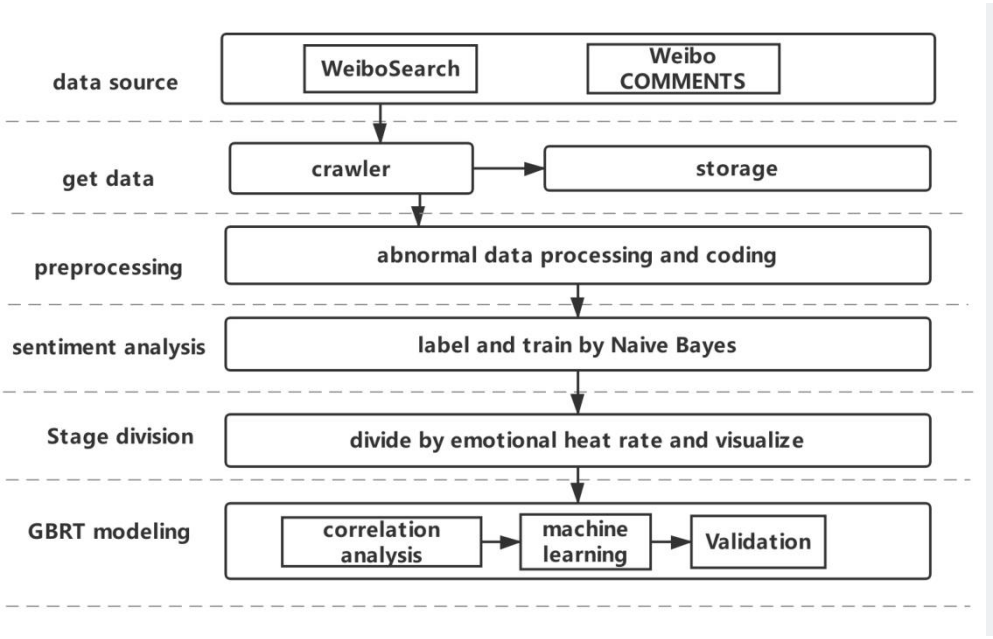

Figure 1. Flow chart of the model.

(1) Data acquisition

Through the analysis of microblog search results, multiple microblog posts with the most comments on the Tianjin incidents were selected, and Python was used to compile crawler for data crawling. The data includes the official microblog of People's Daily, the official microblog of fire department, and other related microblog comments of many official media, as well as the microblog comments of some Internet celebrityies.

(2) Data preprocessing

The data obtained by the crawler may not all conform to the use of modeling, and there may be missing, duplicate and invalid data. Python was adopted to conduct data deduplication, delete missing data, irrelevant data and abnormal data, and store the cleaned data in CSV format.

(3) Emotional analysis

Python language can be used to facilitate the processing of Chinese text data. Firstly, the comment text is processed by word segmentation, and the stop words are removed according to the stop words list to avoid the interference of irrelevant words. We extract some data and manually annotate it , then these samples are regarded as training set, Naive Bayesian classifier was trained to obtains word frequency in each category and further formula out all the comment text emotion inclination degree, the higher the degree of emotional tendency, the more close to the positive emotion, otherwise, the more close to the negative.

(4) Analysis of the evolution stage of public opinion

In various literatures, the number of comments and the number of retweets are often used to measure the popularity of a topic. In general, the more comments, the hotter the topic. However, in order to gain attention, some users use a lot of zombie fan to comment, so the number of comments and retweets cannot effectively and correctly reflect the heat of the event. we introduce the emotion inclination degree here. Considering that the larger the emotional divergence is, the more heated the discussion will be and the hotter the topic will be, the variance of emotion inclination degree can reflect the size of emotional divergence to a certain extent and indirectly indicate the topic heat. In addition, the variance of some dates is relatively large because the number of comments is small, so the variance alone is not enough to fully reflect the fact. After comprehensive consideration, this paper defines emotional heat as follows:

$$\boldsymbol{HOT} = minmax(minmax(num)*q) \quad (10)$$

In formula (3), HOT represents emotional heat, “num” is the number of comments, and q is the variance of emotional tendency. “Minmax” is a standardized treatment. Standardized treatment can remove the influence of dimension and make both factors better reflected in public opinion evaluation.

(5) GBRT model training

Through quantitative data preprocessing and word vectoring process, the user's attributes (including gender, age, location, school/company, the number of fans, the thumb up number, focus on number, the user’s first emotional state, average public emotion inclination degree, nickname) information as input data, emotional tendency degree as the output, we build the GBRT model .

3. Tianjin 812 accident emotive analysis

On August 12, 2015, a chemical explosion in Tianjin binhai new area attracted widespread attention on Weibo and became one of the hottest topics that year. People have different opinions and sentiments; the discussions admixture of good and evil, and official information is released improperly, resulting in public opinion crisis [17].

In order to test and verify the model, our model was used to analyze this chemical accident.

3.1. Data acquisition

Search keywords such as "Tianjin 8•12" and "binhai new area" on Weibo, select comments of People's Daily, China fire station and other microblogs, compile crawlers and obtain 15,253 pieces of data. To get each user’s information of comments, the crawler is used again to obtain the user's microblog homepage information. The crawling data includes comment content, comment time, commenter ID, nickname, number of thumb up users, number of fans, number of followings, gender, information source, age, gender and province of some users.

3.2. Data preprocessing

Delete duplicate, missing, invalid data. For example, in some data, there are “0 years old”, “105 years old”, etc., which are inconsistent with cognition and belong to abnormal data. The value of abnormal attributes corresponding to the data is deleted, and the remaining attributes remain unchanged. After the process, there are 15164 pieces data remaining.

3.3. Naive Bayesian classifier training and prediction

In this paper, Naive Bayesian model is used for sentiment analysis. 300 comments of positive emotion and negative emotion were randomly selected, and the positive and negative emotional tendency degree was manually calibrated as training data. After segmentation, we remove the stop words. Naive Bayes classifier was trained to calculate the emotional propensity of all other comments. Taking the comment "we pray for you, and we firmly support you" as an example, the positive emotion is denoted as S, and the negative emotion is denoted as N, then:

$$P(S|”祈福”,”天津”,”坚定”,”支持”)=\frac{P(”祈福”,”天津”,”坚定”,”支持”|S)}{P(”祈福”,”天津”,”坚定”,”支持”)} \quad (11)$$

$$P(N|”祈福”,”天津”,”坚定”,”支持”)=\frac{P(”祈福”,”天津”,”坚定”,”支持”|N)}{P(”祈福”,”天津”,”坚定”,”支持”)} \quad (12)$$

Then, according to the conditional independence hypothesis of Naive Bayes classifier, we assume that each word is independent, then

$$P(”祈福”,”天津”,”坚定”,”支持”|S)=P(”祈福|S)*P(”天津”|S)*P(”坚定”|S)*P(”支持”|S)$$

$$S=\frac{P(S|”祈福”,”天津”,”坚定”,”支持”)}{P(N|”祈福”,”天津”,”坚定”,”支持”)} \quad (13)$$

S is the probability of positive emotion. As S is calculated through the positive and negative word frequency, it can fully reflect the degree to which emotions tend to be positive, and part of the results are shown in figure 2.

['祈祷']	0.829446
['看第一幅图就好心疼']	0.867565
['回复:对呀，只能祈福了']	0.043149
['居民楼都炸成这样了？才死这么点人？楼里都没人吗？我只想说，骗子死全家。	0.004117
['这是化学物品爆炸，出动防化部队没错，真不知道这有什么好喷的，怎么救如何	0.838826
['回复:你根本不知道现在消防员的工资有多少，他们连最起码的高温补助都没有。	0.002227
['回复:我不满意别人对消防员的不尊重难道有错？你真是狭隘，没有经过了解就	0.637814
['回复:你不满意，所有的人就不满意了？']	0.190414
['为什么每次火灾都会牺牲很多的消防员 他们都那么年轻 到底是一些领头人为	0.10705
['这就是职能部门的职责问题。这样危险极高的场所负责安全的部门尽责了吗？领	0.000949
['看一次眼泪掉一次，真的乞求不要再多的伤亡了，愿一切安好?????????????']	0.807676

Figure 2. Results of Naive Bayes prediction model.

According to the graph, the result of emotion analysis accords with cognition.

3.4. Evolution stage division

3.4.1. Distribution rule and analysis of emotion inclination degree

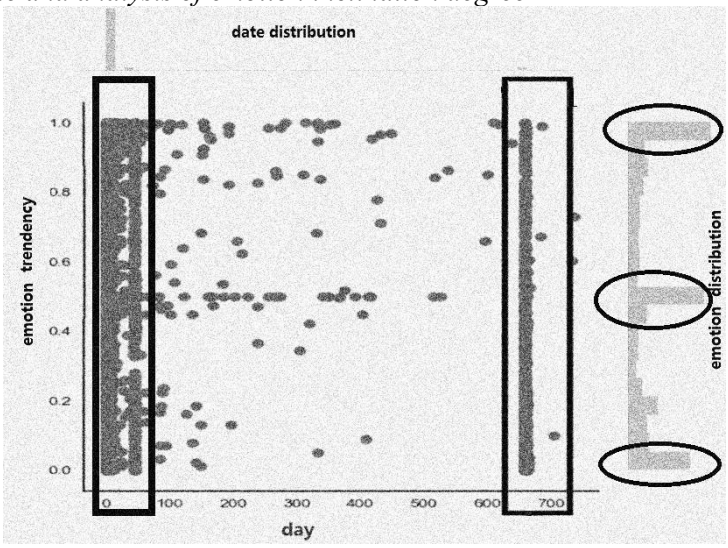

Figure 3. The whole time period comment emotion tendency distribution.

Figure 3 shows the distribution of the emotion inclination degree. The ordinate is the motion inclination degree of each comment, and the larger the value is, the more positive the emotion is, and the closer it is to positive emotion. The abscissa is the number of days since the outbreak of public opinion on August 13.

According to figure 3, the emotional tendency changes greatly with time and comments of different commentators, indicating the characteristics of public opinion on Weibo: Great emotional divergence, changeable, existing explosive and latent periods. By observing the emotional distribution, it can be seen that the emotional distribution has three peaks, which respectively symbolize extremely positive emotion, extremely negative emotion and neutral emotions, indicating that people's emotions fluctuate greatly. From the perspective of time distribution, the emotional tendency tends to become neutral with time after the big divergence in the early stage, but it is the incubation period of public opinion at this time, and occasionally there will be extreme emotional outburst caused by events.

3.4.2. Rule of emotional heat and reason analysis

Calculate the heat according to formula (3) and make the heat -- time curve, as shown in the figure below.

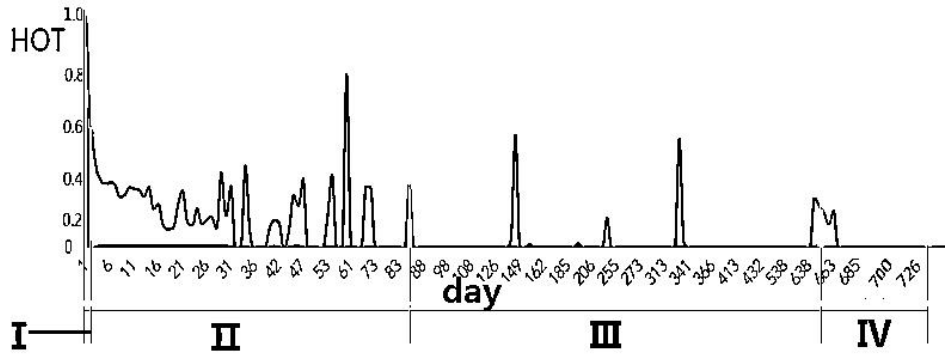

Figure 4. Area diagram of emotional heat changing with time.

According to figure 4, the evolution process of public sentiment can be divided into four stages: hot period, continuous period, incubation period and dying period. The first two days after the outbreak of the accident is regarded as the hot period, and the emotional heat is kept above 0.5.At the beginning of the accident, there are a great deal of comments and people have different emotions towards the accident, information imbalance caused a variety of views among people. From the 3rd to the 85th day after the accident, the emotional heat is stable and high, the information is spread widely, and there are more and more news about the incident. The incubation period of public opinion ranges from 86 to 661 days, netizens' attention is diverted by other things. However, once relevant events occur, the topic will become a hot debate again. Since the 661st days after the incident, the emotional heat maintained at a low value , and it has entered a period of complete extinction.

3.5. Characteristics of the evolution stages

3.5.1. Word cloud in different stages

Figure 5 shows the word cloud of the comments in the four stages. It can be seen that in the whole process of the evolution, the official microblog emphasize firefighters' hero moves and their sacrifices , which aims to guide public opinion in the right way, so that there are words such as "heroes", "pray" in the whole process, but each stage also has its own special words.

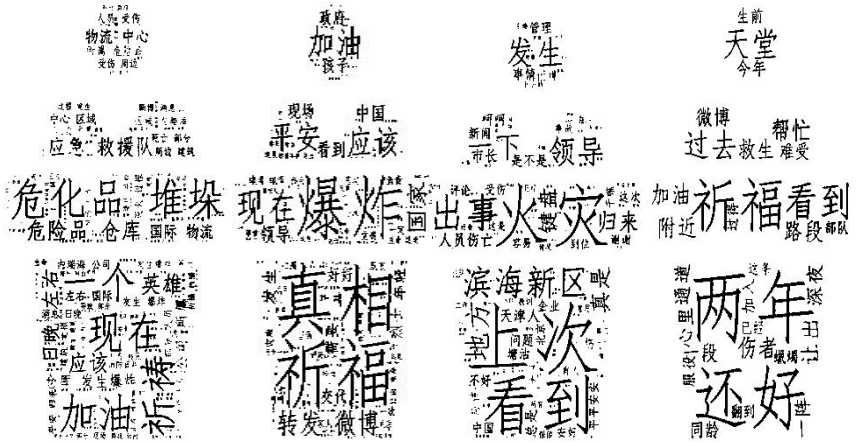

Figure 5. Comment cloud in four stages.

In Stage 1, the comment cloud is mainly a statement of events, such as "warehouse", "hazardous chemicals", "injury", "logistics", etc. It shows people's desire to know and spread the news.

In stage 2, besides positive words such as "pray for good fortune", "peace", "come on", there are words such as "truth" and "leadership", indicating people's different concerns and views on the responsibility of the accident. However, the official media try to be vague about it, which caused the divergence of emotional tendency in public opinion to expand.

In stage 3 ,the words "last time" ,"management" and "leadership" appeared, indicated that people was not satisfied with the leading departments due to the poor management of Tianjin port ,because of the occurrence of another fire in 2016.

In stage 4, the evolution process has been close to the death, some accident witnesses occasionally recall some nostalgic comments in this stage.

3.5.2. Proportion of emotional tendency at each stage

Emotional tendency S was segmented, with 0-0.3 as negative emotion, 0.3-0.6 as neutral emotion and 0.6-1 as positive emotion, and statistical analysis was conducted. The results were shown in table 1.

It can be seen from table 1 that in the first stage, the proportions of the three emotions were equal, which means there were great differences among people's emotions. In stable period, emotions gradually turn into neutral emotions, and negative emotions reduce. During the incubation period, critical events are easy to make it a hot topic again and there are great gaps in people's sentiments. In the dying stage of public sentiments, the proportion of neutral emotion reached a high value, and the evolution process reached dying.

Table 1. Proportion of different emotional types in the four stages.

Period	Proportion		
	Negative	Neutral	Positive
1	0.35	0.31	0.34
2	0.20	0.42	0.38
3	0.57	0.30	0.13
4	0.13	0.63	0.25

3.6. proportion of age groups in different stages

Age groups are divided into five age groups, respectively: [<17], [17-28], [29-37], [38-55] and [>55]. Statistical analysis is conducted on the age of Weibo users at each stage, and the change of the number of different age groups at each stage can be obtained from the curve below, as shown in figure 6.

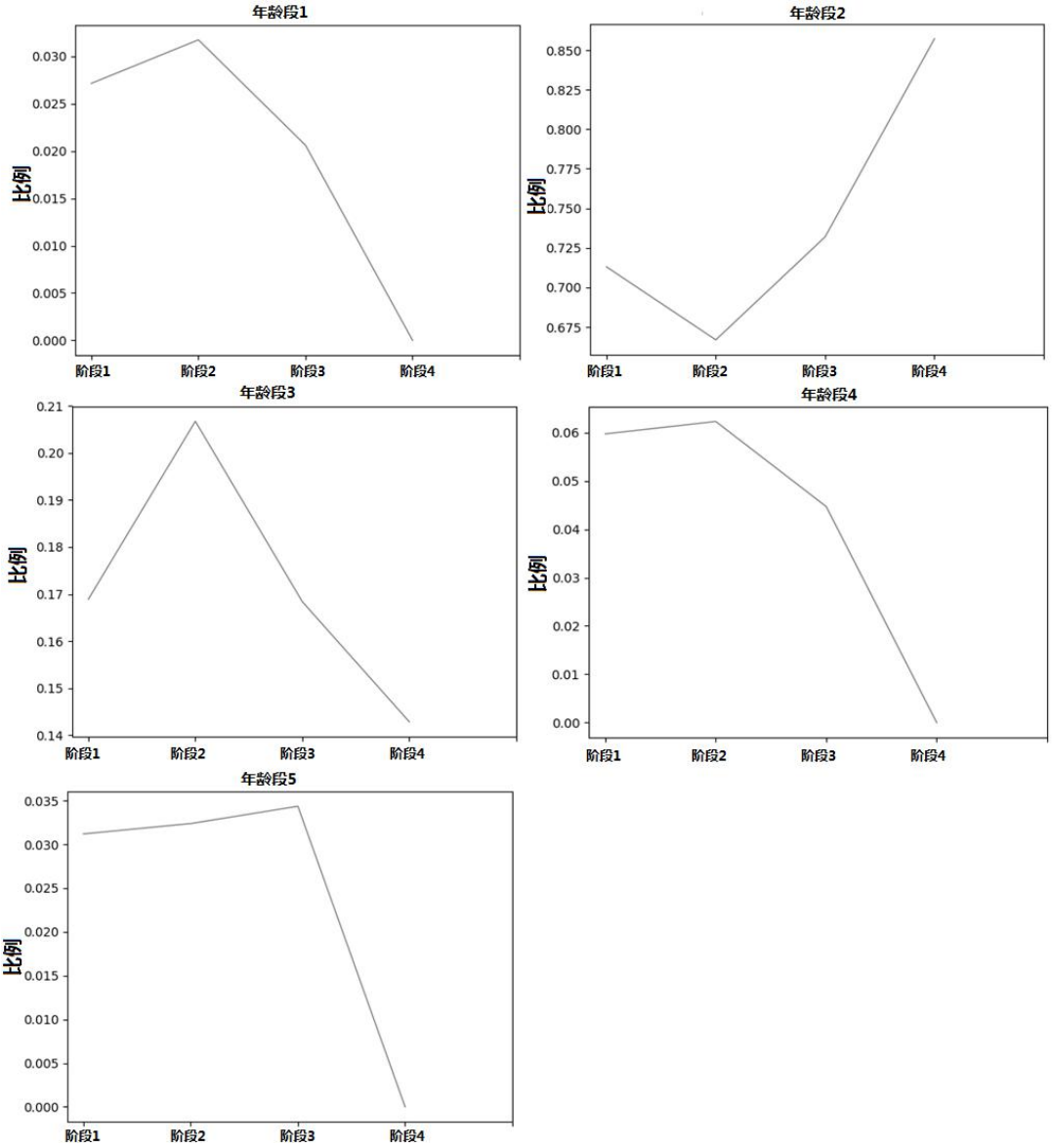

Figure 6. Curve of proportion of different age groups.

As shown in the figure above, group2 (age range [17-28]) and group3 (age range [29-37]) are the dominant groups among the five age groups, accounting for more than 90%. Among them, group2 has the largest proportion, and is always increasing in the four stages. Therefore, our advice for government is that we should focus on young people aged 17-28, especially in the incubation period. Group2 can decide whether the topic will become hot again and when it will die out

4. Tianjin 8•12 accident sentiment prediction model of emotion inclination degree

4.1. Correlation analysis and modelling of emotional propensity

4.1.1. Coding and pre-processing

The gender variable is one-hot coded. Then we use Baidu encyclopedia, sohu news and novels to train Word2Vec model, and the user's personal information, study/work location, nickname and comments release source were transformed into 64-dimensional word vectors with Word2Vec for subsequent analysis and processing.

4.1.2. Correlation coefficient We compute correlation coefficient. To visually display the correlation, a histogram is made, as shown in figure 7.

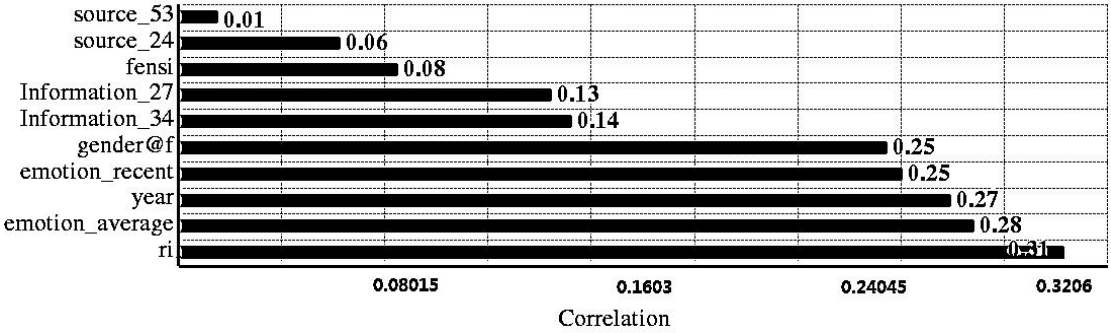

Figure 7. correlation between each variable and emotion inclination degree.

Variables with obvious correlation are “emtion_before” which shows the user's emotional tendency in the previous stage, “ri” which implies the number of days since the event occurred, “emotion_average” which is the group's emotion inclination degree at this stage, etc., these variables have strong positive correlation with a user’s emotion inclination degree. In addition, the number of users' followings, gender and personal information are also weakly correlated.

This result shows that emotional tendency changes significantly over time after the event.

For example, there are more negative emotions in the early stage of catastrophic accident, and after that, emotions gradually become neutral over time. So emotion is most affected by time. In addition, due to network empathy and herd mentality, the average emotional inclination of the public also has a positive impact on personal emotional inclination. Besides, even gender can have some influence, and it means woman is more likely to be positive on the issue of emergency.

4.2. Model training

With emotion inclination degree as the prediction target, GBRT was used for regression analysis modeling. Each time, negative gradient of loss function was used to fit the next decision tree. Based on decision integration of multiple regression trees, the prediction model was obtained. The learning rate was set at 0.1, the number of decision trees was set at 100, and the subset of samples accounted for the overall proportion of 0.8.

4.3. Model evaluation

The regression model error can be measured by MSE and MAE.MSE is mean square error, and the calculation expression is:

$$MSE = \frac{1}{N}\sum_{t=1}^{N}(observed_t - predicted_t)^2 \quad (14)$$

MAE refers to the average absolute error. The calculation formula is as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\text{abs}\,(f_i - y_i) \quad (15)$$

We compute MAE and MSE, MAE value was 0.2672 and MSE was 0.1006, which was within the allowable error range, and the model result was considered reliable.

The model still has a lot of room for improvement. More user portrait attributes can be collected in subsequent work to improve prediction accuracy. On this basis, before and after major disasters, GBRT model is used to predict the emotional tendency of Weibo users on the whole network, which can release targeted guidance information to different groups and guide the correct evolution of public opinions.

4.4. Analysis of model results

After preprocessing and coding user’s information, GBRT model is used to predict the emotional tendency of each user in the development of public opinion.

To show the process, information of seven Weibo users in the database were extracted to predict their emotional tendency. To avoid leakage of user information, nickname information is processed here. The sample's personal information is shown in figure 8.

NO.	shijian_review	source	dianzan	gender	guanzhu	fensi	nickname	information	else_info	stage
4	2017/6/26	vivo X5·K歌之王	1	f	239	224	[illegible]洒脱	女 26岁 狮子座 福建	武夷学院	4
10929	2017/6/1	iPhone 6 Plus	2	f	627	179	[illegible]SZ	女 江苏 苏州		3
10857	2017/6/1	柔光自拍vivo X7P	7	f	62	47	[illegible]	女 22岁 水瓶座 四川		3
1877	2015/8/13	三星android智能手	0	f	270	299	[illegible]N	女 处女座 广东 广州		1
4852	2015/8/13	三星android智能手	1	f	270	299	[illegible]MIN	女 处女座 广东 广州		1
10180	2015/9/26	iPhone 6	0	f	48	728	[illegible]thstarc	女 海外 其他		2
10185	2015/9/26	大神手机	0	m	195	7874	[illegible]浪某某	男 巨蟹座 福建 厦门	集美大学诚毅	2

NO.	nickname	text	qinggan	type	predict
4	[illegible]洒脱	['深夜翻自己的微博，又看到你们了。心里一阵难受。在天堂的你们还好吗？']	0.373884	2	0.22
10929	[illegible]_SZ	['管理混乱！天津的官员要脚踏实地工作，少一些虚的空的！']	0.005335	1	0.15
10857	[illegible]懵繆	['上次的教训还没完全吸收，这次又来，天津你是想毁了吗']	0.123978	1	0.09
1877	[illegible]N	['不造谣！不信谣！不传谣！这是我们现在能做的！']	0.905739	1	0.77
4852	[illegible]N	['死了很多人、空气里有毒、政府不作为、媒体在隐瞒，键盘侠们的理想状态']	0.074431	1	0.04
10180	[illegible]thstarc	['英雄一路走好，我们不会忘记你们的']	0.996096	2	0.89
10185	[illegible]某某	['事故责任人应该陪葬！']	0.19648	2	0.2

Figure 8. sample user information table.

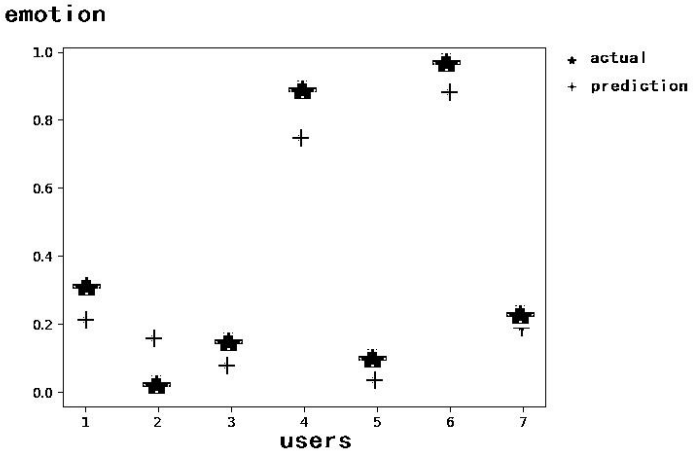

Figure 9. Scatter plots of actual emotion values and predicted values.

Figure 8 shows the value of emotion inclination degree computed by Bayes and the predicted value by GBRT. Figure 9 shows the comparison between them.

The seven predicted values deviated from the actual values, but within the allowable range of error. In the early stage of public opinion, the model can be used to predict the emotional tendency of each person , and government can do some targeted measures according to it. For example, according to word cloud and word frequency in different stages, different age groups, specific guidance should be given to groups with different characteristics and in different stages. Besides, online group identification, offline group guidance can be useful. Identify Internet celebrity users through fan data and microblog data, who could be most the important groups who have the largest influence in the web.

5. Countermeasures and suggestions

According to the analysis results, it is found that the emotional tendency of users is related to the evolution stage of public opinion and the information of user portrait. To some extent, the indicators of user portrait reflect the personality characteristics and knowledge level of users. Therefore, users will have different reactions towards an event. Our model proposed in this paper can be used to predict the emotional tendency of users towards emergencies in advance and guide users' public opinions in a targeted way.

(1) Early warning of public opinion and sentiment.

For example, as shown in section 3.5, user whose ID is 10185 can be predicted in the first stage of the development by model, and combined with the word cloud, the official media can prepare for the potential risks of public opinion in advance.

(2) Public opinion guidance for different user group.

According to the conclusion of correlation analysis in section 3.4, personal emotion is highly correlated with the average public emotion, that is, people will be influenced by others.

Through emotion analysis, positive emotion users are identified and their microblogs are placed at the top of their fans' Weibo page. For users with negative emotions, relevant laws, expert interpretation and other media opinions will be pushed to them to give more reference to him.

(3) Pay attention to the age structure.

In the case of Tianjin 8•12 accident, user group aged 18-37 is the main group of online public opinion. Therefore, after identifying the age characteristics, government should take measures aim to different age group. For example, it is good to adopt more interesting ways for young people. Tik Tok, micro video, kuaishou and other platforms will be useful, as well as the 90's popular cartoon, animation.

(4) Emphasize identity information and Internet celebrity.

Users' information contains variables such as age, gender, educational background, number of fans, number of followings, etc. The number of followers and the number of followings can effectively reflect a person's activeness in Weibo. Among them, users whose number of fans is more than a certain threshold can be identified as Internet celebrity. The emotions of them are often spread through the huge number of fans, which has a significant impact on public opinion. Therefore, the identification of Internet celebrity,targeted emotional analysis and guidance at different stages are conducive to the standardization of cyberspace. In addition, there are school information and workplace information, thus according to it, we can also adopt ways of online identification of the gathering trend of groups and offline guidance.

6. Discussion and conclusion

6.1. Discussion

The research work done in this paper is based on certain assumptions. Based on the following assumptions, the conclusion of this paper is valid:

(1) It is assumed that the number of posts deleted by microblog managers is within the allowable error range and does not significantly affect the analysis of this paper;

(2) It is assumed that the personal information of Weibo users is true and valid;

(3) It is assumed that the comments sample can fully reflect the characteristics of all comments information on Weibo.

6.2. Conclusion

In this paper, crawler is written in Python to obtain microblog comment information. On the basis of data cleaning and preprocessing, emotion analysis modeling is conducted by combining word segmentation technology and Naive Bayes classifier to obtain emotion inclination degree. On this basis, according to the number of comments and emotional inclination degree variance, we define the emotional heat, and then the evolution process of public opinion is divided according to emotion heat ,after that, we use statistical methods to study evolution characteristics, and visualize data Finally, taking Tianjin 8•12 accident as an example, we can research the evolution rules of Tianjin 8•12 accident. It has been proved that the model can operate normally, providing theoretical support for the targeted guidance strategies of public opinion in different stages of emergencies.

References

[1] Li H， Wang LT(2018). Micro-blog Hot Topic Discovery Based on Heat term. *China. J.Information Science*, **36(4)**, 45-50.

[2] Chen J, Liu YP, Deng SL(2018). An Analysis on Factors Influencing the Dissemination Effect of Rumor-refuting Information. *China. J.Information Science*, **36(1)**, 91-95.

[3] Chen CY， Huang XL(2018). A Research on the Generation Mechanism and Propagation Law of Microblog Public Opinion. *China. J.Information Science*，**36(4)**, 32-37.

[4] JI XM(2014). Users'Sentiment Mining and Spreading among Chinese Microblogs in the Context of Specific Events. *China. D. Nankai University*.
[5] Zhang P, Lan YX, Li HQ and Wang JY(2016). Analysis of Internet Rumors Classification Based on the Hayshi Quantification Theory. *China. J.Journal of Intelligence*, **35(01)**, 110-115.
[6] Pang T B, Pang B, Lee L(2002). Thumbs up? Sentiment classification using machine learning[J]. *China. J.Empirical Methods in Natural Language Processing*, 79-86.
[7] You YH(2017). The Analysis of Network Public Opinion Development Stage under Big Data Environment. *China. J.Value Engineering*, (35):177-180.
[8] Wu XJ(2018). Analysis of the Theme Evolution of Internet Public Opinion Based On Micro-blog Text —Taking a Fire Case in Blue Qianjiang Community as an example. *China. D. Nanjing University.*
[9] Du ZT, Xie XZ(2013). The Establishment of Public Opinion Forecasting and Early-warning Model with the Methods of Grey Forecasting and Pattern Recognition. *China. J.Library and Information Service*, **57(15)**, 27-33.
[10] Qian AL, Qu BB, Lu YS, Chen PP and Chen GD(2012). Forum Sentiment Trend Prediction Based on Multi Time Series Association Rule Analysis. *China. J.Journal of Nanjing University of Aeronautics & Astronautics*, **44(6)**, 904-910.
[11] Zhang HP, Chen QH(2018). Research on the Prediction of Network Public Opinion Based on Grey Markov Model. *China. J.Information Science*, **36(1)**, 75-79.
[12] Jiang ZY, Ma WR, Zou K and Li L(2018). Research on Emotion Evolution of Network Public Opinion Based on Sentiment Orientation Analysis. *China. J.Journal of Modern Information*, 38(04):50-57.
[13] Pinkerton B(1994). Finding what people want : experiences with the webcrawler. *C.Proc. of the Second International WWW Conference*.
[14] Ahmadi-Abkenari F, Selamat A(2012). An architecture for a focused trend parallel web crawler with the application of clickstream analysis. *J. Information Sciences*, **184(1)**, 266-281.
[15] Li JM , Sun LH, Zhang QR and Zhang CS(2003). Application of native Bayes classifier to text classification. *China. J.*Journal of Harbin Engineering University, **24(1)**, 71-74.
[16] Li H(2012). Statistical learning method [M]. *China. M.*
[17] Sproat R, Emerson T(2003). The first international Chinese word segmentation bakeoff. *C. Sighan Workshop on Chinese Language Processing. Pennsylvania: Association for Computational Linguistics*, 133-143.

Acknowledgments
This work was supported by “National statistical science research key project (no. 2017LZ37)”, “Ministry of Public Security theory and soft science project(No. 2015LLYJWJXY034)”, “Unmanities and social sciences foundation of the Ministry of Education(no.17YJC630214)”, “Hebei province key research and development project(no. 18215601)”, “Hebei statistical research project(no. 2018HY04)”, “Langfang science and technology project (No. 2019013066)”. We are grateful to the anonymous reviewers for their constructive comments that improving the quality of this work.