Journal of Computational Science 25 (2018) 193–200

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

Detecting users' anomalous emotion using social media for business intelligence☆

Xiao Sun[a], Chen Zhang[a], Guoqiang Li[b,*], Daniel Sun[c], Fuji Ren[a,d], Albert Zomaya[e], Rajiv Ranjan[f]

[a] School of Computer and Information, Hefei University of Technology, TunXi Road No. 193, 230009 Anhui, China
[b] School of Software, Shanghai Jiao Tong University, 200240, China
[c] Data61, CSIRO, Australia
[d] Faculty of Engineering, The University of Tokushima, 770-8506 Tokushima, Japan
[e] University of Sydney, Australia
[f] University of Newcastle, Australia

ARTICLE INFO

Article history:
Received 12 March 2017
Received in revised form 7 May 2017
Accepted 31 May 2017
Available online 6 June 2017

MSC:
00–01
99–00

Keywords:
Business intelligence
Sentiment analysis
Anomaly detection
Multivariate Gaussian distribution
Decision making

ABSTRACT

Anomaly detection in sentiment analysis refers to detecting users' abnormal opinions, sentiment patterns or special temporal aspects of such patterns. Users' emotional state extracted from social media contains business information and business value for decision making. Social media platforms, such as Sina Weibo or Twitter, provide a vast source of information, which include user feedbacks, opinions and information on most issues. Many organizations also leverage social media platforms to publish information about events, products, services, policies and other topics frequently, analyzing social media data to identify abnormal events and make decisions in a timely manner is a beneficial topic. This paper adopts the multivariate Gauss distribution with the power-law distribution to model and analyze the users' emotion of micro-blogs and detect abnormal emotion state. With the measure of joint probability density value and the validation of the corpus, anomaly detection accuracy of individual user is 83.49% and of different month is 87.84% by this method. Through the distribution test, the results show that individual users' neutral, happy and sad emotions obey the normal distribution, but the surprised and angry emotions do not. Besides, emotions of micro-blogs released by groups obey power-law distribution, but the individual emotions do not. This paper proposes a quantitative method for abnormal emotion detection on social media, which automatically captures the correlation between different features of the emotions, and saves a certain amount of time by batch calculation of the joint probability density of data sets. The method can help the businesses and government organizations to make decisions according to the user's affective disposition, intervene early or adopt proper strategies if needed.

1. Introduction

According to the 2016 third quarter earnings report [1] released by Sina Weibo, as of September 30, 2016, the monthly number of active users on Weibo has reached 297 million. In September, 2016, the number of active users has reached 132 million, representing an increase of 32% over the same period last year. Micro-blog in the video, travel, sports and other fields have been further developed. In 2016, most of the active users on micro-blog are highly educated, they are the main force of micro-blog, accounting for up to 77.8%, and their emotional states are often characterized by the micro-blogs that they released.

Sina Weibo has a large number of young users, and they are an important part of the main consumer and society. User emotion modeling and anomaly detection on micro-blog is an important field of emotional analysis, which can help the enterprises to make business decisions, help the government to monitor public opinion and public safety through social network, prevent the spread of irrational emotions in social network or even in real world, respond timely to the possible negative incidents to prevent some criminals who attempts to spread rumors [2] through micro-blog

☆ This work was supported by National Natural Science Funds for Distinguished Young Scholar (No. 61203315) and the China Postdoctoral Science Foundation funded project (2015M580532). This work was partially supported by JSPS KAKENHI Grant Number 15H01712. This work was also supported by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR 201407345) and Natural Science Foundation of Anhui Province (1508085QF119).

* Corresponding author.
E-mail address: li.g@sjtu.edu.cn (G. Li).

https://doi.org/10.1016/j.jocs.2017.05.029
1877-7503/

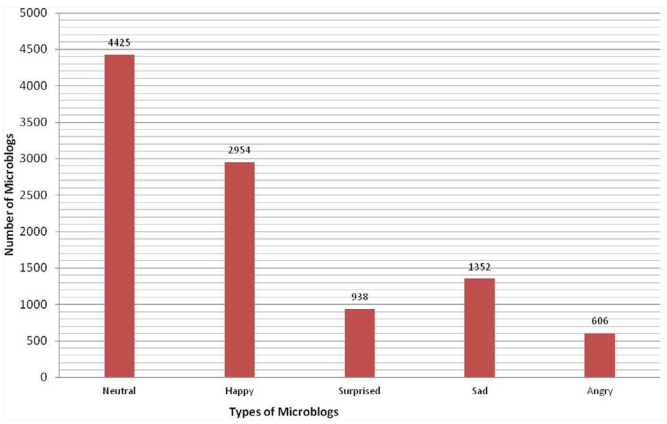


Fig. 1. Five types of micro-blogs.

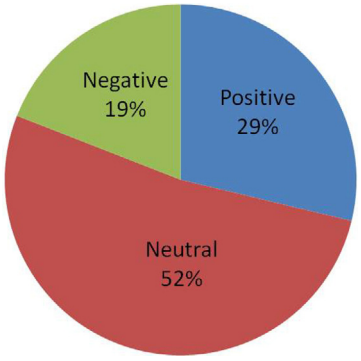


Fig. 2. Micro-blog emotional polarity.

[3]. A considerable amount of data mining research on anomaly detection has been conducted, and this stream has gained considerable interest owing to the realization that anomaly patterns can be detected from large databases through data mining. With the advancement of social media technologies, the ways in which people communicate through their comments, feedback and critiques have dramatically changed. They can post reviews and share their opinions on products, services, policies and other topics through social media platforms. If these emotions or anomalies are undetected or poorly managed, the consequences may be severe, e.g., a business or company whose customers reveal negative sentiments and will no longer support the establishment. Users' emotion and opinion about product in social media is very important for decision making.

Figs. 1 and 2 are 10,275 micro-blogs collected from 100 users from May 2011 to May 2016, including five types of emotions and the proportion of emotional polarity. From Fig. 2, its obvious that the negative emotions in social media account for a quite high proportion, which is worthy of attention and concern.

Currently, the methods used for anomaly detection are mostly unsupervised [4] and nonparametric [5]. Lin [6] proposed a unified hybrid model-a factor graph model combined with Convolutional Neural Network to leverage tweet content and social interaction information for stress detection, which improved the detection performance by 6–9% in F1-score. Guzman [7] proposed a scalable and fast on-line method that used normalized individual frequency signals per term and a windowing variation technique, this method reported keyword bursts which can be composed of single or multiple terms, ranked according to their importance. Niu [8] proposed a rule-based and dictionary-based approach, through the experiment, the language features of emotional expression in micro-blogs were discussed, and Niu provided a basis for the establishment of high-precision emotional analysis system. Zhang [9] built an emotional dictionary based on the emotional words and phrases commonly used of emotional factors to recognize and classify the emotion on micro-blog, which achieved good results. Zhao [10] considered the object of a text to improve the emotional classification accuracy to detect the social anomaly, the Twitter text were chose as the sample for testing, by comparing the proportion of negative emotions to observe anomaly in a day, the conclusions are general and could not accurately analyze the specific abnormal event or user. Li [11], who was based on real-time event monitoring framework and system of micro-blog, proposed a rule-based and statistical method, used time series model to monitor anomaly, which proved more effective than the ordinary model. Yin [12] proposed a micro-blog anomaly ranking detection method based on the lifting coefficient, which effectively prevented the artificial manipulation to improve the ranking of micro-blog. Experiments on the simulation data set showed that the method could effectively identify micro-blog anomaly ranking by micro-blog topology.

Anomaly detection methods mentioned above are mainly based on dictionary, text, neural network, time series, statistic, rule and rank, which require a large number of annotated corpus, but the annotations workload are really heavy. In addition, the current methods tend to classify and analyze all the data on a social platform to detect outbreaks or abnormal events from the time aspect, but there is little research on the detection of abnormal emotion for the individual user.

2. Preparation work

2.1. Data processing

In order to detect the abnormal emotion on micro-blog and model users emotion, this paper is divided into three stages: data processing, abnormal emotion detection and user emotion modeling. Data processing stage is introduced in this chapter in detail, abnormal emotion detection will be claimed in the third chapter, and user emotion modeling will be claimed in the fourth chapter.

In the data process stage, through the Internet crawler technology [13], 10,275 micro-blogs of 100 users from May 2011 to May 2016 are collected. The users include writers, stars, network celebrity, students, ordinary people, etc; the original micro-blog texts are marked with the corresponding user id, release time and other useful information. Micro-blogs of an user during a period of time will be classified into 5 types (five-dimensional vector) based on the preliminary work [14], then the number of "neutral, happy, surprised, sad, angry" emotions of user can be obtained. 5 types of micro-blogs are as the variables related to the user's emotions, and the correlation between the variables and users emotions is researched and modeled. Each type of emotion can be modeled by the single Gaussian model, and the five-dimensional vector of emotions can be modeled by the multivariate Gaussian distribution, through joint probability density (JPD)and a proper threshold [15], the abnormal user or abnormal month can be detected. The corpus annotated with emotions are further processed into two aspects, user and month.

[User: month: emotion category: the number]

[Month: user: emotion category: the number]

From the user aspect is to detect abnormal emotion of a specific user during a period of time; From the month aspect is to see which month appears abnormal users.

2.2. Gaussian distribution

Gaussian distribution is to use the probability density function (normal distribution curve) to accurately quantify things. The Gaus-

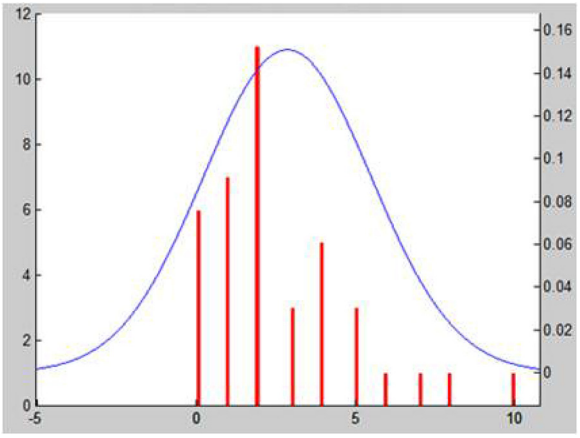


Fig. 3. Fitted graph of normal distribution.

Table 1
Features of emotions on micro-blog.

Features	X1	X2	X3	X4	X5
Emotions	Neutral	Happy	Surprised	Sad	Angry

sian model consists of the single Gaussian model (SGM) and the multivariate Gaussian model (MGM)[16]. In this paper, the single Gaussian distribution is used to model and visualize each emotion of the users, the multivariate Gaussian distribution is used to model the five-dimensional emotions, which can automatically capture the correlation between the different features of variables.

Fig. 3 is the distribution of the number of specific emotion on micro-blog from a user during 10 months. As shown in Fig. 3, the fitted graph does not exactly match the Gaussian distribution, because the number (N) of data set is small. The fact is that as N increases, it converges quickly to the Gaussian distribution. The single Gauss model can detect abnormal data in a set of data in a certain extent, but it is complex for it to solve the problem with multiple variables.

In Table 1, X1–X5 are 5 features that are considered to be influential for anomaly detection of micro-blog, X1 = "neutral", X2 = "happy", X3 = "surprised", X4 = "sad", X5 = "angry". For the original model, these features need to be computed one by one, in this paper, it is assumed that the user's multidimensional emotion follows the multivariate normal distribution, the association between these features can be captured by the multivariate Gaussian model and the anomalies can be detected by the computing the JPD and setting the appropriate threshold.

3. Model

3.1. Anomaly detection

Anomaly detection is the detection of abnormal samples from the data set. Anomaly detection in sentiment analysis refers to detecting users' abnormal opinions, sentiment patterns or special temporal aspects of such patterns. Users' emotional state extracted from social media contains business information and business value for decision making. The anomalies detection may be due to sudden sentiment changes hidden in large amounts of text. If these anomalies are neglected or poorly managed, the consequences may be severe. Abnormal cases are very few in anomaly detection. There are three concrete methods for anomaly detection, model-based approach [17], proximity-based approach [18], and density-based approach [19]. Professor Andrew Ng has taken the

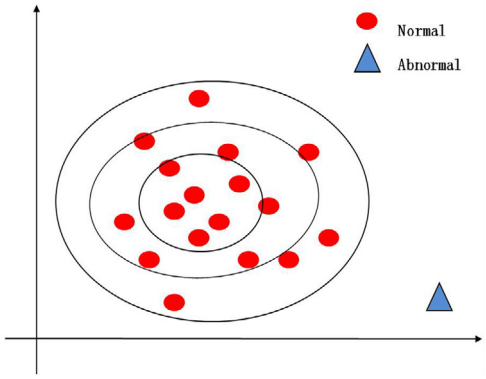


Fig. 4. Anomaly detection of aircraft engine.

anomaly detection of aircraft engine [20] as an example to explain the basic principle of density-based anomaly detection:

A variety of factors can influence the aircraft engine and cause the anomalies. Assuming that X = engine-generated heat, Y = engine vibration intensity, given a set of data $D = (D1, D2, \ldots, Dn)$. Since the engine anomaly detection is based on two variables to determine, according to the (X, Y) value, these data points can be plotted on the graph. As shown in Fig. 4, the density of oval points are large, and they are marked as normal points; the rightmost triangle data point are significantly deviated from the normal data group, whose density is much smaller than the oval points, then it is marked as anomaly.

The principle of anomaly detection based on density is that data points that are farther away from the neighboring points (low density) will be marked as outliers [18]. In the experiment, data in low probability density will be considered as anomalies, and then determine whether the user and month that corresponding to the data are abnormal or not.

3.2. Anomaly detection on micro-blog

At present, most of the studies focus on the single Gauss model or the two-dimensional Gauss model, and there are few models or methods that use 3 or more variables in the anomalies detection. In this paper, users' micro-blog emotions were divided into 5 categories, so that the users' emotions can be characterized by 5 dimensional variables. The single Gauss distribution can be used for the visualization of each emotional type of micro-blog, and the 5-dimensional Gauss model will be used to solve the problem of abnormal emotion detection.

Given the training set $\{X_j^{(1)}, X_j^{(2)}, \ldots, X_j^{(m)}\}$, which is a matrix of $m * n$; a model $p(x)$ can be fitted by setting first:

$$\mu_j = \frac{1}{m}\sum_{i=1}^{m} X_j^{(i)} \tag{1}$$

$$\sum_j = \frac{1}{m}\sum_{i=1}^{m}[(X_j^{(i)} - \mu)(X_j^{(i)} - \mu)^T] \tag{2}$$

where m is the number of samples; "j" is from 1 to n, and n is the number of variables. μ (n-dimensional) is the mean of each column vector, and $\sum$ is the covariance matrix. Given an example $x^{(i)}$ (5-dimensional variable), the JPD of x can be calculated as follows:

$$p(x) = \frac{1}{(2\pi)^{\frac{n}{2}}|\sum|^{\frac{1}{2}}}\exp[-\frac{1}{2}(x^{(i)} - \mu)^T\sum^{-1}(x^{(i)} - \mu)] \tag{3}$$

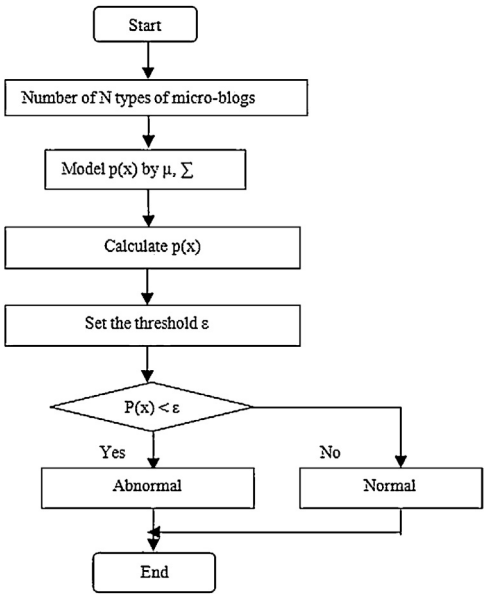


Fig. 5. Micro-blog abnormal emotion detection model.

Fig. 5 is the emotion anomaly detection on micro-blog model proposed in this paper, which can be concluded as the following steps,

- Number of user's/month's N types of micro-blogs $\{X_j^{(1)}, X_j^{(2)}, \ldots, X_j^{(m)}\}$.
- Model $p(x)$ by μ, $\sum$.
- Calculate $p(x)$.
- Set the threshold ε.
- Flag an anomaly if $p(x) < \varepsilon$.

The JPD value shows the density of the sample, in order to detect abnormal samples quantitatively and accurately, a threshold will be set. When the JPD value of the sample is less than the threshold, which means that the density of the sample is too low, and it will be marked as abnormal.

3.3. Threshold selection

The range of JDP value is used to evaluate whether the data set is abnormal or not, so a proper threshold is essential for this work to be determined. In this paper, the threshold selection is realized by the following steps:

- Calculating the JPD values for all the data sets.
- Dividing all the data sets into two parts, the cross validation set (CVS) [21] and the test set [22].
- Comparing the accuracy by selecting different thresholds on the cross validation set to obtain the optimal threshold.
- Using the optimal threshold to get the accuracy on test set.

The JPD value of each data set is calculated by the multivariate Gauss model $p(x)$, and $p(x)$ can be calculated according to the formula (3). The process of the threshold selection will be presented in Section 4.2.

4. Experiments and results

In the experiment part, a total of 10,275 micro-blogs of 100 users are randomly selected and classified into 5 types according to the user and month. Since the data will be very sparse and not easy to analyze if counted in a day or in a week, data is collected and computed monthly in this paper. After the semi-automatic annotation of the micro-blogs, a five dimensional data set:(number of neutral emotion, number of happiness, number of surprise, number of sadness, number of anger) will be obtained. Through batch calculation of the JPD of the data, a threshold is used on the test set, if the JPD value of a sample on the test set is smaller than the threshold, then it will be marked as abnormal. The threshold selection and accuracy of the detecting method are two main parts of the experiments. Finally, the K-S test [23] is used to detect the distribution of the group emotion on micro-blog, which models the users' social emotion from the groups aspect.

Table 2
JPD of users (user12).

User12	X1	X2	X3	X4	X5	JPD
Dec.2015	2	0	0	0	0	8.15E−04
Nov.2015	1	1	0	0	0	7.70E−03
Oct.2015	1	0	0	0	2	3.79E−04
Sep.2015	1	2	0	0	0	7.38E−03
Aug.2015	1	4	1	0	1	5.01E−04
Jul.2015	1	0	1	2	0	2.78E−04
Jun.2015	1	2	0	0	0	7.38E−03
May.2015	0	0	0	2	0	6.92E−04
Apr.2015	0	1	0	1	0	5.40E−03
Mar.2015	0	0	1	0	0	9.16E−04
Oct.2014	0	1	0	0	0	9.18E−03
Jul.2014	0	0	1	0	1	9.96E−04
Jun.2014	0	1	0	0	0	9.18E−03
May.2014	0	0	0	0	1	6.22E−03
Feb.2014	2	0	0	0	1	1.54E−03
Jan.2014	0	1	0	1	1	3.10E−03
Sep.2013	1	1	0	0	0	7.70E−03
Aug.2013	1	2	0	0	0	7.38E−03
Jul.2013	0	3	0	1	0	7.61E−04
Jun.2013	3	2	1	3	2	1.06E−04
May.2013	*10*	*11*	*2*	*3*	*1*	*3.40E−06*

4.1. Joint probability density

The data in Table 2 is collected from user12 among the 100 users, the column 2 to the column 6 is the number of five micro-blog emotions of this user. It can be seen from the table that most JPD values (the last column) of each month are from 1E−03 to 1E−04, but the JPD of the data in May 2013 is 3.40E−06, which is significantly smaller than that of other groups, which is marked as the abnormal case.

Fig. 6 is the original micro-blogs of the corresponding user of Table 2, it can be seen that in May 2013, the user's mood does appear "tired, insomnia" and other abnormal states.

In Table 3, most of the JPD values are also from 1E−03 to 1E−04, but there are 4 shadowed data that are smaller than the others. It is obvious that these data far less than the other values will be labeled as suspected anomalies.

Fig. 7 is the original micro-blogs of the corresponding month of Table 3, in May of 2016, the user 19 and user20 appeared "sad, poor, forget the past, give up" and other abnormal emotional states.

4.2. Threshold selection and accuracy

For further verification of this method and improvement of accuracy, total 10,275 micro-blogs are processed into 1700 data sets, in which 500 data sets are as cross validation set, 1200 data sets are as test set; by observation and comparison, 3 suitable thresholds are chose on the cross validation set. Different thresholds will get different accuracies, and the threshold related to the highest accuracy will be selected for the test set.

Fig. 6. Verification of abnormal emotional text (user12).

Table 3
JPD of months (May, 2016).

May.2016	X1	X2	X3	X4	X5	JPD
User1	*14*	*0*	*0*	*0*	*0*	*1.41E−05*
User2	1	2	0	0	1	6.50E−04
User3	0	0	0	1	0	3.21E−03
User4	0	0	0	1	0	3.21E−03
User5	0	2	1	1	1	2.71E−03
User6	1	2	0	1	0	3.01E−03
User7	0	1	0	0	0	4.72E−03
User8	*1*	*0*	*2*	*1*	*1*	*2.59E−05*
User9	1	1	1	1	0	2.09E−03
User10	0	0	0	1	0	3.21E−03
User11	7	1	0	0	0	3.21E−03
User12	0	0	0	1	0	3.21E−03
User13	0	3	1	0	1	9.26E−04
User14	0	1	0	0	0	4.72E−03
User15	0	1	0	0	0	4.72E−03
User16	2	0	0	0	0	7.15E−03
User17	1	0	0	0	0	6.61E−03
User18	1	0	0	0	0	6.61E−03
User19	*5*	*1*	*1*	*0*	*2*	*5.45E−05*
User20	*8*	*10*	*2*	*2*	*0*	*2.13E−06*
User21	2	0	0	0	0	7.15E−03

Table 4
Cross validation results based on users.

Threshold	CVS	TP	FP	Accuracy
1E−04	36	29	7	80.56%
4E−05	36	32	4	88.89%
1E−05	36	27	9	75%

Table 5
Cross validation results based on months.

Threshold	CVS	TP	FP	Accuracy
1E−05	36	30	6	83.33%
1E−06	36	32	4	88.89%
1E−07	36	31	5	86.11%

"TP" is true positive (marked as abnormal, the actual data is abnormal), "FP" is false positive (marked as abnormal, the actual data points is normal). From Table 4, 1E−04, 4E−05, 1E−05 are chose for the cross validation set based on users, when the threshold is 4E−05, the accuracy of anomaly detection is 88.89%, which is the highest among the 3 thresholds. From Table 5, 1E−05, 1E−06, 1E−07 are chose for the cross validation set based on months, when the threshold is 1E−06, the accuracy of anomaly detection is 88.89%, which is the highest among the 3 thresholds. Through the experiment on the cross validation set, the threshold of different user is set to 4E−05; the threshold of different month is set to 1E−06.

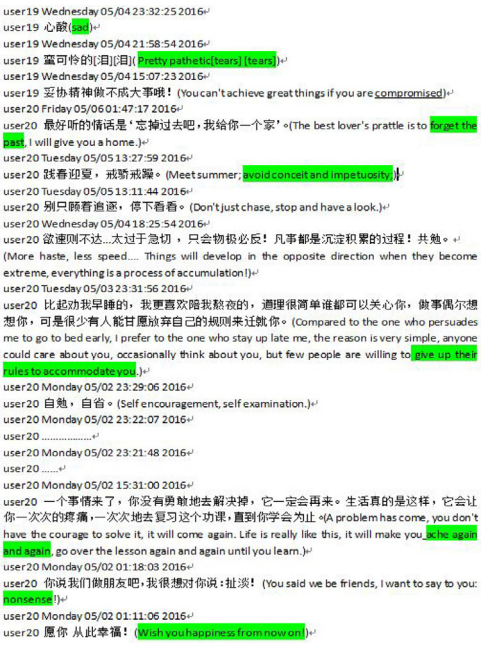


Fig. 7. Verification of abnormal emotional text (May, 2016).

Table 6
Accuracy based on users.

Threshold	Test data	TP	FP	Accuracy
4E−05	109	91	18	83.49%

Table 7
Accuracy based on months.

Threshold	Test data	TP	FP	Accuracy
1E−06	74	65	9	87.84%

Table 6 is the accuracy obtained from the user's aspect, 109 data sets are marked as abnormal, after the comparison with the original micro-blog, whose emotions are semi-automatic labeled, of which 91 data sets are true positive, 18 data sets are false positive, and the accuracy is 83.49%. Table 7 is the accuracy based on different month, 74 data sets are marked as abnormal, after the comparison with the original micro-blog, of which 65 data sets are true positive, 9 data sets are false positive, and the accuracy is 87.84%. This method has a better performance in detecting abnormal emotion users in a period of time.

In order to evaluate the method proposed, the method is compared with the other three existing methods. Fig. 8 shows the anomaly detection accuracy of existing three methods and the proposed method, NMF (nonnegative matrix factorization) gets the low accuracy, about 51%. The real-time monitoring [11] method gets the accuracy of 73.33%, whose accuracy is not high, but the advantage of this method is that the detection is dynamic and real-time. SSDM [24] gets the accuracy of about 85.2%, compared with the former methods, it is greatly improved, but the method is for the Twitter spam detection, Chinese micro-blog is more complicated to deal with than the Twitter text, and the anomalies are more difficult to detect due to the diversity and concealment of Chinese. In this paper, the JPD value is introduced to quantify the

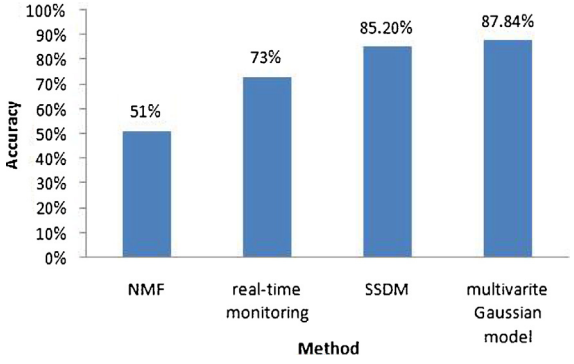


Fig. 8. Comparison the accuracy of four methods.

Table 8
K-S test of individual user.

Kolmogorov–Smirnov						
Item		Var1	Var2	Var3	Var4	Var5
N		20	20	20	20	20
Normal(a, b)	Mean	1.4	1.8	0.4	0.8	0.5
	Standard	2.5	1.4	0.7	0.9	0.7
Most	Absolute	0.3	0.2	0.4	0.3	0.4
Extreme	Positive	0.3	0.2	0.4	0.3	0.4
Difference	Negative	−0.3	−0.2	−0.3	−0.2	−0.2
Kolmogorov–Smirnov Z		1.3	0.7	1.9	1.2	1.6
Asymp. Sig. (bilateral)		0.1	0.7	0	0.1	0

Table 9
K-S test of group.

Kolmogorov–Smirnov						
Item		Var1	Var2	Var3	Var4	Var5
N		100	100	100	100	100
Normal (a, b)	Mean	2.4	3.7	0.7	1.2	0.4
	Standard	2.6	4.4	1	1.3	0.7
Most	Absolute	0.3	0.3	0.3	0.4	0.4
Extreme	Positive	0.3	0.2	0.3	0.4	0.4
Difference	Negative	−0.3	−0.3	−0.2	−0.2	−0.3
Kolmogorov–Smirnov Z		3.1	2.6	3.1	3.7	4
Asymp. Sig. (bilateral)		0	0	0	0	0

anomaly detection, through the experiments, the accuracy rate of 87.84% is achieved. As the texts on social media are very diverse and the features of anomaly are not obvious, the accuracy of anomaly detection are not high at present and need further study.

4.3. User emotion modeling

User's five categories of micro-blogs of each month can be seen as a five-dimensional matrix, the data of each dimension is tested and verified whether they are subject to Gaussian distribution or not. According to K-S test, as shown in Tables 8 and 9, VAR1–VAR5 represent "neutral, happy, surprised, sad, angry" five variables. Asymp. Sig. (bilateral) is the *P*-value, which is generally set to a threshold value as 0.05, when *P*-value is larger than 0.05, the data can be assumed to be a normal distribution; otherwise, the original hypothesis (the data is subject to the normal distribution) is rejected. For the individual user and group users, the K-S test is separately used and obtained the results as the following tables:

Table 8 is the K-S test of an individual user, as the table shows, *P*-value of "neutral, happy and sad" types are bigger than 0.05, consistent with previous assumption, but "surprised" and "angry" emotions are not subject to the normal distribution, because these two emotions have the characteristic of outburst. Table 9 is the K-S test results of group (100 users), for each type of micro-blogs, finding that the *P*-value is less than 0.05, indicating that the emotion of group on micro-blog is not subject to the normal distribution.

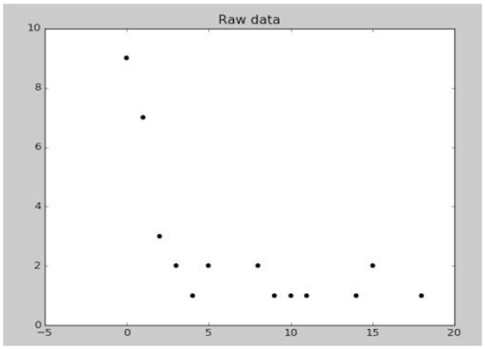


Fig. 9. Distribution of original data.

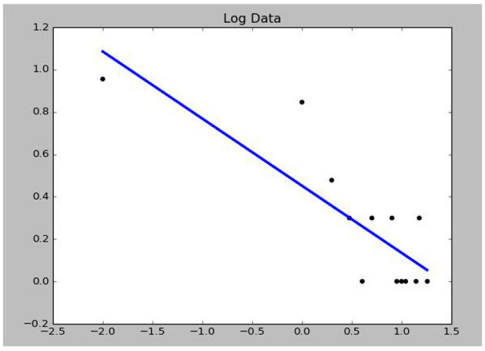


Fig. 10. Distribution of logarithmic data.

In fact, the micro-blog sentiment of the group tends to satisfy another exponential distribution: power law distribution [25], which is a famous and widespread social phenomenon, also known as long-tail distribution. This paper makes an experiment and verification for this inference: setting *x* as the abscissa, *x* is the number of each type of micro-blogs of all users in a month, and setting the frequency of *x* as the vertical axis, these data can be fitted by matlab.

Fig. 9 shows the original distribution of a set of data. The row data is close to the long-tailed distribution shape, after taking the logarithm of row data, the log data is obtained, which approximates a straight line as shown in Fig. 10. Finally, the residual sum [26] is used to evaluate the reasonableness of the distribution. The smaller the residual sum, the more likely the power-law distribution of the set of data can be considered. The distributions of all users' micro-blogs of every month are tested in this experiment, which proves that the distribution of group's emotion on micro-blog is subject to the power-law distribution.

5. Discussion

In the beginning of this paper, we assumed that the micro-blog emotions of users obey normal distribution, and a model for user emotion and anomaly detection on social media was researched in this paper. Through experiments and analysis of real data, some inferences are also obtained as the following:

- Emotions such as "neutral, happy, sad" of micro-blog are subject to the normal distribution.
- Emotions such as "surprised, angry" of micro-blog are explosive and not subject to the normal distribution.
- Group emotions on micro-blog are not subject to the normal distribution, they tend to satisfy another index distribution: power-law distribution.

The innovative point of the Gauss model is that the emotion on micro-blog is 5-dimensional in this paper, this method can automatically capture the correlation between different features of variables and model the multiple emotions well. Besides, through the batch calculation of JPD of the samples, the anomaly detection can be detected quantitatively.

There are also two aspects need to be improved in this paper, one is due to the sparseness of micro-blog data of each user, the abnormal emotion of individual user can only be detected in a month or in a week at present, if the data is enough, the method can model emotions and detect anomaly on micro-blog in a day, but the abnormal user in a certain period on micro-blog can be detected timely. Another is that the time performance need improvement, the threshold selection is still a time-consuming process, we compensate for this by batch computing the JPD of the data sets, thus saving a part of the time.

6. Conclusion

This paper makes a study on the users emotion modeling and abnormal emotion detection on social media. The multivariate Gaussian model and joint probability density are introduced to detect abnormal users emotions of user on micro-blog. Results show that the accuracies of abnormal detection are 83.49% and 87.84% according to the user and month respectively. The experiment also shows that the "neutral, happy, sad" emotions of the individual user are subject to the normal distribution through the K-S test, while the "surprised, angry" emotions are not, and the emotion of micro-blogs released by group is subject to power-law distribution, while the individual user is not. This paper combines multivariate Gaussian model with joint probability density to detect anomalies on social media, and proposes a relatively comprehensive approach to model user and group emotions, which are meaningful to detect the abnormal emotion, monitor the public security and help the enterprise to make the reasonable business decision.

References

[1] Micro-blog user development report in 2016 (2017). URL http://www.useit.com.cn/thread-14392-1-1.html.

[2] F. Yang, Y. Liu, X. Yu, M. Yang, Automatic Detection of Rumor on Sina Weibo, 2012, pp. 1–7, http://dx.doi.org/10.1145/2350190.2350203.

[3] Y. He, W. Deng, D. Zhang, B. School, S. University, Study on sentiments recognition and classification of Chinese micro-blog, J. Intell. 33 (3) (2014) 136–139.

[4] G. Huang, S. Song, J.N. Gupta, C. Wu, Semi-supervised and unsupervised extreme learning machines, IEEE Trans. Cybern. 44 (12) (2014) 2405–2417, http://dx.doi.org/10.1002/9781118557693.ch4.

[5] X. Dang, R. Serfling, Nonparametric depth-based multivariate outlier identifiers, and masking robustness properties, J. Stat. Plan. Inference 140 (1) (2010) 198–213, http://dx.doi.org/10.1016/j.jspi.2009.07.004.

[6] H. Lin, J. Jia, J. Qiu, Y. Zhang, G. Shen, L. Xie, J. Tang, L. Feng, T.S. Chua, Detecting stress based on social interactions in social networks, IEEE Trans. Knowl. Data Eng. PP (99) (2017) 1.

[7] J. Guzman, B. Poblete, ACM SIGKDD Workshop on Outlier Detection and Description, in: On-line relevant anomaly detection in the twitter stream: an efficient Bursty keyword detection model, 2013, pp. 31–39, http://dx.doi.org/10.1145/2500853.2500860.

[8] Y. Niu, M.H. Pan, O. Wei, X.Y. Cai, Emotion analysis of Chinese microblogs using lexicon-based approach, Comput. Sci. 41 (9) (2014) 253–257.

[9] Z. Jing, Z. Bo, L. Liang, H. Min, Y. Teng, Recognition and classification of emotions in the Chinese microblog based on emotional factor, Beijing Daxue Xuebao Ziran Kexue Ban/acta Scientiarum Naturalium Universitatis Pekinensis 50 (1) (2014) 79–84, http://dx.doi.org/10.13209/j.0479-8023.2014.016.

[10] Z. Wang, V. Joo, C. Tong, X. Xin, H.C. Chin, IEEE International Conference on Cloud Computing Technology and Science, in: Anomaly detection through enhanced sentiment analysis on social media data, 2014, pp. 917–922, http://dx.doi.org/10.1109/cloudcom.2014.69.

[11] L.I. Ling-Yun, A.O. Ji, Z. Qiao, L.I. Jian, Research on security event real-time monitoring framework based on micro-blog, Netinfo Secur. (2015), http://dx.doi.org/10.3969/j.issn.1671-1122.2015.01.004.

[12] G. Yin, Y. Zhang, Y. Dong, W. Yuan, H. Dong, A boost factor based detection method for abnormal rank of microblogging, J. Harbin Eng. Univ. 34 (4) (2013) 488–493, http://dx.doi.org/10.3969/j.issn.1006-7043.201206039.

[13] Z.H. Zhou, H.R. Zhang, J. Xie, Data crawler for Sina Weibo based on python, J. Comput. Appl. 34 (11) (2014) 3131–3134, http://dx.doi.org/10.11772/j.issn.1001-9081.2014.11.3131.

[14] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (3) (2007) 27, http://dx.doi.org/10.1145/1961189.1961199, article 27.

[15] P.U. Diehl, D. Neil, J. Binas, M. Cook, International Joint Conference on Neural Networks, in: Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing, 2015, pp. 1–8, http://dx.doi.org/10.1109/ijcnn.2015.7280696.

[16] S.F. Yuan, S.T. Wang, Multi-classification method applied to face recognition based on mixed Gaussian distribution, Appl. Res. Comput. 30 (9) (2013) 2868–2871.

[17] J. Liang, R. Du, Model-based fault detection and diagnosis of HVAC systems using support vector machine method, Int. J. Refrig. 30 (6) (2007) 1104–1114, http://dx.doi.org/10.1016/j.ijrefrig.2006.12.012.

[18] T. Id, A.C. Lozano, N. Abe, Y. Liu, Proximity-based anomaly detection using sparse structure learning, SDM (2009) 97–108, http://dx.doi.org/10.1137/1.9781611972795.9.

[19] S.H. Ma, J.K. Wang, Z.G. Liu, H.Y. Jiang, Density-based distributed elliptical anomaly detection in wireless sensor networks, Appl. Mech. Mater. 249-250 (2012) 226–230, http://dx.doi.org/10.4028/www.scientific.net/amm.249-250.226.

[20] Machine learning, week9, anomaly detection (2017). URL https://www.coursera.org/learn/machine-learning.

[21] K. Chen, J. Lei, Network cross-validation for determining the number of communities in network data, Br. J. Psychiatry 178 (5) (2014) 410, http://dx.doi.org/10.1080/01621459.2016.1246365.

[22] H. Yu, J. Yang, J. Han, X. Li, Making SVMs scalable to large data sets using hierarchical cluster indexing, Data Mining Knowl. Discov. 11 (3) (2005) 295–321, http://dx.doi.org/10.1007/s10618-005-0005-7.

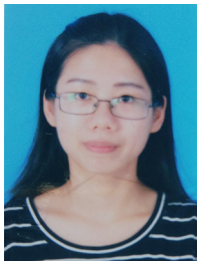
[23] H.J. Eghbali, K-S test for detecting changes from Landsat imagery data, IEEE Trans. Syst. Man Cybern. 9 (1) (1979) 17–23, http://dx.doi.org/10.1109/tsmc.1979.4310069.

[24] X. Hu, J. Tang, Y. Zhang, H. Liu, International Joint Conference on Artificial Intelligence, in: Social spammer detection in microblogging, 2013, pp. 2633–2639.

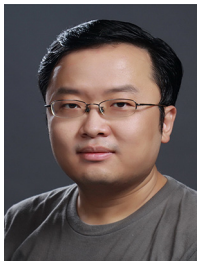
[25] A. Clauset, C.R. Shalizi, M.E.J. Newman, Power-law distributions in empirical data, SIAM Rev. 51 (4) (2009) 661–703, http://dx.doi.org/10.1137/070710111.

[26] D.M. Rosen, M. Kaess, J.J. Leonard, Rise: an incremental trust-region method for robust online sparse least-squares estimation, IEEE Trans. Robot. 30 (5) (2014) 1091–1108, http://dx.doi.org/10.1109/tro.2014.2321852.

Xiao Sun received the M.E. degree in 2004 from the Department of Computer Sciences and Engineering at Dalian University of Technology, and got his doctor's degrees in Dalian University of Technology (2010) of China and the University of Tokushima (2009) of Japan. He is now an associate professor in AnHui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine at Hefei University of Technology. His research interests include Affective Computing, Natural Language Processing, Machine Learning, and Human-Machine Interaction.



Chen Zhang received his Master degree in 2016 from School of Science, Anhui University of Science and Technology, Huainan, China. He is currently a Ph.D. student at School of Computer and Information, Hefei University of Technology. His research interest includes Natural Language Processing, Sentiment Analysis, and psychology computing.



Guoqiang Li received the B.S., M.S., and Ph.D. degrees from Taiyuan University of Technology, Shanghai Jiao Tong University, and Japan Advanced Institute of Science and Technology in 2001, 2005, and 2008, respectively. He worked as a postdoctoral research fellow in the graduate school of information science, Nagoya University, Japan, during 2008–2009, as an assistant professor in the school of software, Shanghai Jiao Tong University, during 2009–2013, and as an academic visitor in the department of computer science, University of Oxford during 2015–2016. He is now an associate professor in school of software, Shanghai Jiao Tong University.

Daniel Sun is a research scientist in Data61, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia. He is also a conjoint lecturer in School of Computer Science and Engineering, the University of New South Wales, Australia. He received his Ph.D. in Information Science from Japan Advanced Institute of Science and Technology (JAIST) in 2008. From 2008 to 2012, he was an assistant research manager in NEC central laboratories in Japan. From 2013 to 2016, he was a researcher in National ICT Australia (NICTA).

Fuji Ren received the B.E., M.E. from Beijing University of Posts and Telecommunications, Beijing, China, in 1982 and 1985, respectively. He received the Ph.D. Degree in 1991 from Faculty of Engineering, Hokkaido University, Japan. He worked at CSK, Japan, where he was a chief researcher of NLP from 1991. From 1994 to 2000, he was an associate professor in the Faculty of Information Sciences, Hiroshima City University. He became a professor in the faculty of engineering, the University of Tokushima in 2001. His research interests include Natural Language Processing, Artificial Intelligence, Language Understanding and Communication, and Affective Computing. He is a member of the IEICE, CAAI, IEEJ, IPSJ, JSAI, AAMT and a senior member of IEEE. He is a Fellow of The Japan Federation of Engineering Societies. He is the President of International Advanced Information Institute. Computing; Taiwan Association of Cloud Computing. He is vice chair of IEEE TCSC and IEEE senior member.

Albert Y. Zomaya received the Ph.D. degree from Sheffield University. He is currently the Chair Professor of High Performance Computing & Networking and Australian Research Council Professorial Fellow in the School of Information Technologies, The University of Sydney. He is also the Director of the Centre for Distributed and High Performance Computing which was established in late 2009. His research interests include algorithms, complex systems, parallel and distributed systems.

Rajiv Ranjan received the Ph.D. degree. He has been a reader (associate professor) of computing science at Newcastle University since September 1, 2015. He is an internationally renowned researcher in the areas of cloud computing, Internet of Things (IoT), and big data. By applying ground-breaking combination of well-founded formal models from four domains (Operations Research, Computational Statistics, Peer-to-Peer Networking, and Performance Engineering) of computer science, he has developed novel algorithmic techniques and distributed system architectures that facilitate service level agreement (SLA) driven autonomic provisioning of multimedia (e.g., content delivery networks), eScience (e.g., scientific work-flows), and IoT big data applications (e.g., remote sensing, smart homes, smart cities, etc.) applications over multiple private and public cloud data centres.