ARTICLE IN PRESS

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

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.

© 2017 Elsevier B.V. All rights reserved.

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
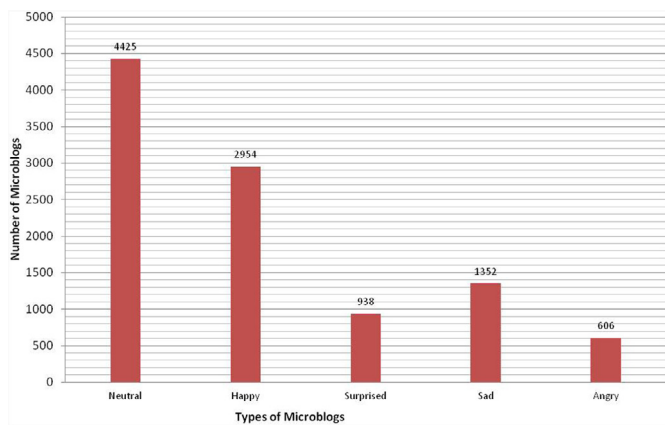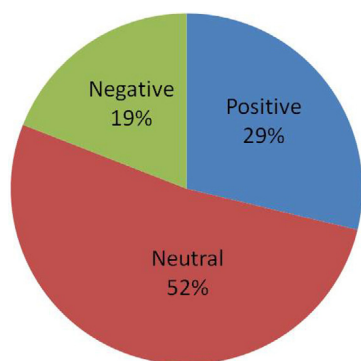
Fig. 1. Five types of micro-blogs.

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-