Internet Monitoring Research on Derivative Public Opinions of Emergencies Oriented with Subject Extraction

—Take the Case of Nanny Arson in Hangzhou on "June 22"

Yanchen Cui[a], Peng Zhang[b, *], Lizhi Wu[c]

School of The Chinese People's Armed Police Forces Academy, Langfang 06500, China

[a]2103695856@qq.com, [b]zp_1981@aliyun.com, [c]wulizhi119@sohu.com

Keywords: data mining; subject extraction; emotional analysis; computer simulation

Abstract: This paper aims to adopt the means and methods of emotional analysis and subject mining to analyze and judge the relation between emotional trend of online public opinions and evolution and change of the situation, so as to provide reference to monitoring and handling of public opinions for the government. The emotional dictionary construction for public opinions of fire services is conducted with PMI-IR algorithm in way of domain dictionary construction, and the subject extraction and emotional analysis are conducted for topic comments related to "nanny arson" in micro-blogs and forums in combination with the public opinion features of domestic fire control. The emotional trend and the focus of netizens can be promptly mastered in the time dimension with the methods of emotional dictionary construction and subject extraction adopted in the paper, thus providing certain guidance to prompt monitoring of public opinions and the perspective selection for information issuance.

1. Introduction

Innovation and development for Internet technology facilitate Chinese Internet Society flourished as never before, and the measures for public opinions monitoring and handling adopted by the fire control departments, as Chinese important force of fire extinguishing and emergency rescue cannot be completely suitable to such huge micro-blog user quantity and information mass; thus stories are cooked up and spread around and instigation of group incidents is common, and the public opinions are transformed and upgraded to derivative public opinions for some incidents under the action of multiple factors [1]. It can be found by analyzing the rule of public opinions that the derivative public opinions of emergencies have subjective freedom, carrier diversity and multi-interactivity. So it will provide powerful support for the information issuance opportunity and issuance subject of governments and the related departments to depict and analyze public opinions of micro-blogs, from the perspectives of theme evolution and emotion trend change with more scientific monitoring means of public opinions and scientific data mining methods.

2. Related Theory and Tools

2.1 Digging for online public opinions

The digging degree for public opinions is significant to emotional analysis quality and analysis strategy as the basic link of analysis for online public opinions. The cases where information digging is integrated to industrial domains such as bio-medicine and e-commerce, etc are increasing with the surge of online information quantity and arrival of "big data" age. The research hotspots are mainly demonstrated in two aspects: the first is key technological researches including construction of digging model and improvement of digging algorithm, emotional analysis data and digging faced up with comment information; the second is the specific application of emotional analysis data mining. It mainly reflected in aspects of visualization and information tracking, etc of time and regional information for commentators in the aspect of emotional analysis[2]. The basic

flow can be divided into: data preparation and processing, emotional polarity and calculation for polarity strength, emotional analysis and result visualization when the data mining methods are adopted pertinent to specific cases.

2.2 Emotional analysis of micro-blogs

Emotional analysis is also called opinion digging or advice digging. It refers to analysis of the opinions, feelings, evaluation, attitudes and moods for people of entities and the attributes in texts [3]. The current mainstream research methods can be divided into two kinds for emotional analysis: one is the emotional classification based on supervision, while the other is emotional classification based on non-supervision; it can be further divided into emotional classification with syntax template and webpage retrieval and the emotional classification with the emotional dictionary. It can be found by comparing the two methods that: the effective classification model can be automatically learned in multiple features of supervision learning, but its shortcoming to over rely on marked data is also obvious, and the domain expansibility is bad. However, it does not need manually mark large-scale training data for the method of emotional dictionary, and the errors can be rapidly corrected through correcting rules. To sum up, the emotional analysis of emergent public opinion of micro-blogs is conducted in the paper with the method of the emotional dictionary, and the key analysis objects include emotional word itself, polarity and strength of emotional word with factors of time and micro-blog features integrated.

2.3 Subject extraction oriented with time dimension

The hotspot information can be grasped from the obtained plenty of texts or commentary information for the subject extraction, and the hotspot information is arranged and analyzed in combination with the time dimension for online public opinions, thus the dynamic evolutionary rules for public opinions are mastered. The present common methods are: the method with syntactic relation based on frequency for subject extraction, subject model method based on latent Dirichlet allocation and the method based on supervision learning. The technology for subject extraction with latent Dirichlet allocation algorithm is relatively mature, but the accuracy is greatly reduced when the short-text information represented by micro-blog comments is processed. The attribute extraction algorithm based on frequency is simple and efficient, so the TF-IDF method is adopted in the paper for subject extraction.

Firstly, time slicing is conducted for the selected text data, and the data is divided in unit of “day”. Then the lower and upper entropy method is adopted to conduct word frequency statistics for linguistic data in micro-blogs for which the word segmentation is handled with ICTCLAS word segmentation software, and the threshold for lower and upper entropy of words is also set. The words handled in this method contain plenty of predicted subject words, but there are still frequent non-target words such as “yes” and “but”, etc. The TF-IDF method is thus introduced [4], and this method is to rank the importance degree of words in the linguistic data; the details can be seen in Equation (1):

$$TF-IDF(w_i) = freq(w_i) \cdot \log\left(\frac{N}{df(w_i)}\right) \tag{1}$$

The above equation is adopted to respectively calculate TF-IDF value of all words, and the value is compared with the set limit; the words with value higher than this limit will be the candidate set of subject words, and the artificial labeling method is adopted here to manually select the subject words, thus the predicted subject words can be obtained because the quantity of words in the candidate set is small.

2.4 Emotional micro-blogs dictionary construction for emergencies

The emotional dictionary is constructed and divided so as to realize the features such as high accuracy and good domain applicability for the emotional dictionary of emergencies. It is

specifically divided into: universal emotional dictionary, exclusive emotional dictionary and emotional dictionary of Internet words. The literatures can be referred for universal emotional dictionary and emotional dictionary of Internet words [5], and it will be omitted here. The construction for exclusive emotional dictionary of emergencies is mainly elaborated here.

2.4.1 Construction of original word set

The public opinions of micro-blogs for "fire disaster in Daxing, Beijing on November 18", micro-blog topic of "kindergarten explosion in Xuzhou on June 15" and micro-blog themes for "troops review in the 90th anniversary of army foundation", summing to 236 thousand comments are selected as the database so as to ensure that the information source to construct the original word set is scientific and comprehensive; word frequency statistics is conducted for it after the noise reduction, word segmentation and removing the word no longer used, and then the non-emotional words are deleted and the words ranking in the front place are reserved; then the left words are randomly grouped for in-group $PMI-IR$ calculation[6], and then the words ranking in the first three places are kept; then it is combined again for $PMI-IR$ calculation, and finally the words are sequenced and words ranking in the front 20 places are output, thus the exclusive original words for derivative public opinions of emergencies are obtained. The positive and negative original words can be seen in Table 1 and Table 2.

Table 1 Positive original word table

Word	Word property	Word	Word property	Word	Word property	Word	Word property
Sharp	Adjective	Aggressive	Adjective	Sunny	Adjective	Awesome	Adjective
Favor	Verb	Support	Verb	Military spirit	Noun	Proud	Adjective
Strong	Adjective	Cute	Adjective	National power	Noun	Warm	Adjective
Happy	Adjective	Bravo	Adjective	Handsome	Adjective	Powerful	Adjective
Warm	Adjective	Pride	Verb	Self-confident	Adjective	Praise	Verb

Table 2 Negative original word table

Word	Word property	Word	Word property	Word	Word property	Word	Word property
Love dearly	Adjective	Dereliction of duty	Verb	Sad	Adjective	Speechless	Adjective
Tragic	Adjective	At sea	Adjective	Cruel	Adjective	Ashamed	Verb
Helpless	Adjective	Injury	Verb	FALSE	Adjective	Anxious	Adjective
Suspic-ious	Adjective	Bad	Adjective	Hurtful	Adjective	Innocent	Adjective
Furious	Adjective	Punish severely	Verb	Desperate	Adjective	Tragedy	Noun

2.4.2 Selection of target words

The judgment for emotional attribute is conducted for the target words after the original word table is constructed. The semantic similarity between the words and original words is measured by the PMI-IR value, and the calculation result is then sequenced in a descending order; the average of the left values is namely the $PMI-IR$ value of the target word after removing the maximum and minimum, and the specific equation is shown in Equation (2):

$$PMI-IR(element, elementBase) = \sum_{i=1}^{count(elementBase)-1} PMI-IR(element, element_i) \tag{2}$$

$PMI-IR(element, element_i)$ is the PMI-IR value between the target element *element* and the ith

original element $element_i$ in Equation (2), and $count(elementBase)$ indicates the element quantity in the original element set $elementBase$. Then the semantic similarity between the element and the original element set $elementBase$ is finally obtained.

The $PMI-IR$ value for positive and negative original elements is respectively calculated in Equation (3), and the following operation is conducted:

$$PMI-IR(element)=\begin{cases} PMI-IR(element,element_{positive}), \\ PMI-IR\left(element,element_{positive}\right)\geq PMI-IR(element,element_{negative}) \\ \sim PMI-IR(element,element_{negative}), \\ PMI-IR\left(element,element_{positive}\right)< PMI-IR(element,element_{negative}) \end{cases} \tag{3}$$

The calculation of Equation (4) is conducted for the target words judged as positive words:

$$P=\frac{PMI-IR(element,element_{positive})}{PMI-IR(element,element_{negative})} \tag{4}$$

The calculation of Equation (5) is conducted for the target words judged as negative words:

$$P=\frac{PMI-IR(element,element_{negative})}{PMI-IR(element,element_{positivee})} \tag{5}$$

Finally, the P value and the threshold are compared, and the corresponding words with bigger P value then the threshold are determined as emotional words; the threshold is set as 0.63 through multiple verifications of commentary texts.

3. Empirical Analyses

3.1 Data obtaining and cleaning

Crawler is conducted for information of four parts such as commentary contents, commentators, comment time and praise quantity below related subject discussion and news report in the micro-blogs of “nanny arson attack in Hangzhou on June 22” with the cuttlefish collector software, and the crawling time interval is during 12 o’clock on December 11 in 2017 and 12 o’clock on February 15 in 2018. The Ajax time-out is set as 8 seconds so as to ensure accurate and comprehensive information. The obtained data is independently stored in unit of “day”, and then total 274 thousand effective comments are obtained after data cleaning steps including removing spam comments such as no longer used and specific symbols and links, etc; then the comments with the evaluation objects as other involved subjects are screened, thus the final total 143.4 thousand effective comments are obtained. The word segmentation software is used to segment words for the texts after noise reduction, and data preparation is conducted for the calculation of emotional value in the next step.

3.2 Emotional analysis of public opinions

The emotional dictionary of construction in Section 2.4 is adopted, thus the emotional polarity distribution of e-pals for the emergency can be obtained, as shown in Fig. 1:

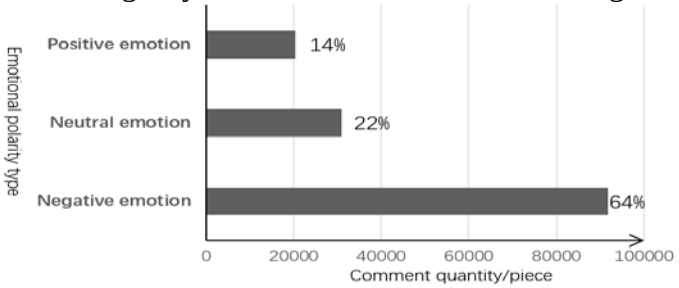

Fig.1. Sentiment analysis of the derived public opinion on the case of "6.22" Hangzhou nanny arson

It can be found from Fig. 1 that the negative emotion occupies the most ratio, reaching 64%, while the neutral emotion ratio is 22% in the second place in the development process for derivative public opinions of the emergency; the ratio for positive emotion is 14%. It indicates that the positive and reasonable emotion for e-pals seriously lacks.

3.3 Subject extraction and text analysis for public opinions

The subject extraction for words in TF-IDF calculation methods is conducted with “day” as the time slice in Section 2.3. The extracted subjects are gathered together with the subjects of officially released information, and then comparison analysis is conducted; the subject distribution and evolution for public opinions in Fig. 2 can be obtained.

As it is shown in Fig.2, the information officially released has extremely difference to influence on public opinions, and the public opinions will be rapidly reduced and present the trend of clearing up, when the subject of information officially released agrees with the focus of the public for actions such as public hearing cases in courts and positive response of fire departments to relatives of the victim, etc; once the subject officially released deviates from the public focus, the e-pals will extremely eager for actual situation on the scene for public opinions of micro-blog, such as the on-site video ranking the first place with the Internet comments as the subject for public opinions on January 15, 2018; but the fire departments did not positively respond to this subject, thus the public opinions were further deteriorated. So related departments should select the angle on which the public are focused for response in combination with actual situation for subject digging.

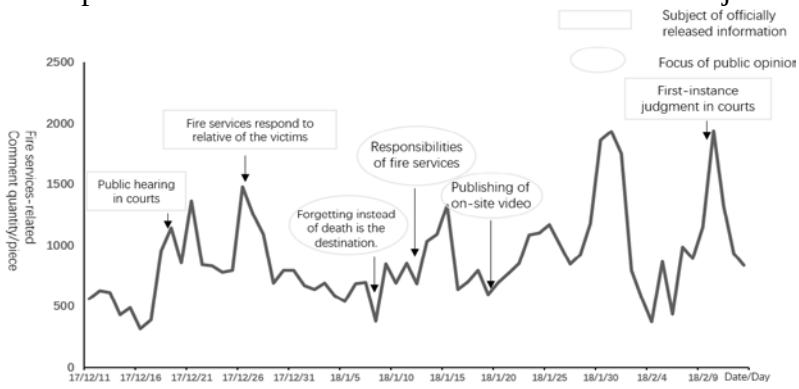

Fig.2. Subject evolution of public sentiment about the case of"6.22" Hangzhou nanny arson

4. Conclusion and prospects

4.1 Discussion

(1) The method in Literature [4] is referred for threshold setting in PMI-IR for the construction of emotional dictionary of fire control, and the progressive increase for thresholds is controlled through experiments; the best filtering effect can be obtained when the threshold is 0.63 after measurement, so it is adopted.

(2) The subject comments not related to fire control are screened in the paper to prevent the commentary information that e-pals hold negative emotions towards criminal suspects from affecting the judgment and analysis of the overall public opinions, when the subject digging and data pre-processing are conducted, thus the analysis target will be clearer.

5. Conclusion

The emotional dictionary applicable to public opinions of micro-blogs for emergencies is established in the paper with algorithms such as TF-IDF and PMI-IR, etc by integration the thought of time sequence. The emotional analysis and subject extraction are conducted for the pre-processed data, and then the change of emotional distribution and attention subject of e-pals for “nanny arson attack in Hangzhou on June 22”is obtained, thus realizing the purpose of monitoring and analysis for public opinions and providing reference for corresponding governmental departments to select

the response time and angle. The design for emotional strength has not been conducted for exclusive emotional dictionary of emergencies for the construction of the dictionary, so the traversal search method can be used to assign the emotional value in the next step, thus realizing the target of monitoring real-time emotional value of e-pals. Meanwhile, the alarm threshold for public opinions will be set in the monitoring system for consideration to realize the function of automatic alarm.

Acknowledgements

Fund Programmes: Emergencies-oriented study on the early warning and countermeasure of the Internet rumor risk" of Humanities and Social Science Fund of Ministry of Education(Number: 17YJC630214); "Research on the risk modeling and countermeasures of network public opinion" of the National Social Science Fund of China(Number:15CXW015);"Research on key technologies of information mining and decision making in emergencies under new media environment" of Science and Technology Programs of Hebei Provinces(Number: 17455610);"Research on public opinion information mining and decision support in emergencies under big data environment" of the Soft Science Research Plan Item of Langfang (Number: 2017029034)

References

[1] Antonella Ianni, Valentina Corradi, The dynamics of public opinion under majorityrules, J. Review of Economic Design (2008) 555-575.

[2] Ding Z Y, Jia Y, Zhou B. Survey of data mining for Microblogs. Journal of Computer Research and Development ,51(2014) 691-706.

[3] Sentiment Analysis: Ming Opinions, Sentiments and Emotions. Bing Liu, China Machine Press, Beijng,2017.

[4] Turney P D, Littman M L. Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus [R]. National Research Council of Canada, Tech. Rep.:EGB 1094,(2002).

[5] Wang X T. Sentiment Analysis of Popular Event Based on Chinese Microblog Network. P.h.D. Dissertation. Beijing: Beijing University of Posts and Telecommunikcations,2014.

[6] Turney P D. Mining the Web for Synonyms: PMI-IR Versus LSA on TOEFL[C]. In Proceedings of the 12th European Conference on Machine Learning ,Freiburg, Germany(2001) 491-502