DATA SCIENCE
JOURNAL

Hu, Y. 2019. Text Mining and Data Information Analysis for Network Public Opinion. *Data Science Journal*, 18: 7, pp. 1–8. DOI: https://doi.org/10.5334/dsj-2019-007

RESEARCH PAPER

Text Mining and Data Information Analysis for Network Public Opinion

Yan Hu
Sichuan Post and Telecommunication College, Chengdu, Sichuan 610067, CN
hyansc@163.com

Network public opinion information is massive and complex, and it is difficult to make effective use of manual means. In this paper, a method based on pattern matching and machine learning (PMML) was proposed to analyze the emotional tendencies of network public opinion. Firstly, the key words in public opinion were extracted, then the patterns were extracted and matched, and the emotional tendencies of words were calculated to obtain the pattern sequence vectors. Support vector machine (SVM) classifier was used to classify emotional tendencies. The Internet reviews of Meituan hotel were taken as the experimental subject. PMML method was found to have a high classification performance, with a maximum accuracy of 86.75%. It suggested the effectiveness of the proposed method. Then PMML method was used to classify the emotional tendencies of the collected reviews, and the results showed that the negative emotional tendency was greater than the positive tendency, which showed the inadequacy of Meituan hotel. The experiments in this paper provide some basis for the application of PMML in sentiment analysis of Internet public opinion.

Keywords: Network public opinion; text mining; emotional tendency; pattern matching; support vector machine

1. Introduction

With the development of Internet technology, social network has become an important platform for people to obtain and exchange information (Liu & Li 2010). Network public opinion has become the main source of public opinion (Li & Zhang 2014) and the main platform for people to express their opinions on hot issues, social politics, etc. (Tran & Tran 2017). A large number of network public opinion information contains many valuable parts. But effective information cannot be efficiently obtained from mass data using ordinary analysis methods because of the large amount and high complexity of network public opinion information, but the emergence of data mining brings a new direction for analysis of network public opinion information. Network public opinion information can be analyzed and utilized through text mining (Guo et al. 2014). Liu (2010) proposed a method of hot spot detection of network public opinion. First, vector space model was introduced to express the text format, and then K-means algorithm was used to cluster the corpus. Finally, support vector machine (SVM) classifier was used to classify the text. Chen et al. (2016) put forward a collaborative filtering based network public opinion trend prediction method, analyzed its principle, constructed the framework of network public opinion trend prediction, and verified the effectiveness of the method through experiments. Yang (2017) adopted the improved TF-IDF algorithm and introduced the part-of-speech weight coefficient and the position weight of feature words to improve the clustering effect of feature words in network public opinion and better reflect text characteristics. Wang & Tang (2016) put forward a Hidden Markov Model based semantic orientation analysis method to analyze and classify the emotional orientation of network public opinion and found that the algorithm could effectively improve the efficiency and accuracy of network public opinion analysis. Emotional tendencies refer to people's inner preferences and dislikes for certain objective things. Emotional orientation analysis refers to excavating users' psychological attitudes towards characters or events through the analysis and induction of the subjective text content generated by the network. Information of network public opinion usually include people's attitudes towards events. Through clear emotional analysis, people's attitudes towards public opinion can be obtained, which can help guide public opinion and is conducive to social security and stability (Hao et al., 2015). Emotional orientation analysis needs to be based on a large number of data, and the era of big data makes it possible to analyze emotional orientation. Through text mining technology, natural language text can be processed to obtain specific information. In the present study, the emotional tendency in network public opinion was studied, pattern matching was combined with machine learning method, and the emotional inclination in the evaluation of Meituan hotel on the Internet was analyzed using PMML method to verify the effectiveness of the method.

2. Network Public Opinion and Network Security

Public opinion refers to the set of people's attitudes and emotions towards a matter, which is a reflection of people's ideological status. Because of the particularity of the Internet, there are also some characteristics of Internet public opinion.

(1) Universal and anonymous. The spread of information on the Internet is very fast and extensive. The speech made by netizens can be greatly diffused in a very short period of time. Because of the anonymity of the network, freedom of speech is more prominent. The public is unscrupulous when making comments, which also makes the network public opinion more complex.

(2) Freed and controllable. The Internet provides a platform for people to express their opinions. Internet public opinion has a higher degree of freedom than the traditional paper media; there the true and false information is difficult to distinguish. But the network public opinion is not absolute freedom. Certain management and control is very necessary. The correct guidance can also play a certain role in the direction of network public opinion.

(3) Interactive and immediate. The network platform provides great convenience for people's information exchange, not only between people, but also between the public and government and between the public and media. Moreover the immediacy of the network also leads to the immediacy of the network public opinion and the timeliness of speech transmission.

(4) Large influence. Because of the above characteristics, the influence scope and extent of network public opinion are very huge. A hot event can spread to a great extent in a very short period of time under the influence of network public opinion, but the event will become more complex because of different feelings and attitudes. The true or false information will affect the judgment of the public, resulting in confusion.

Complex network public opinion information has brought great challenges to network security. Due to the high freedom and openness of the Internet, the public opinion on the Internet is chaotic. A large number of pornographic, reactionary or violent information is common on the Internet, and the Internet not only makes these garbage and harmful information have a rapid spread of shortcuts, but also makes the vast number of netizens directly and passively face the garbage information.

Besides, the immediacy and high influence of Internet public opinion also poses a great threat to social stability. Faced with hot social and political issues, netizens can make unrestricted speeches on the Internet platform. If people are willing to take advantage of the Internet, they can lead public opinion to an aspect that is not conducive to social stability and stir up public sentiment. If timely and effective control is lack of, it is easy to make the situation develop in a bad direction and form network emergencies, which can bring great negative effects to the country and society.

3. Emotional Tendency Analysis on Network Public Opinion Based on Text Mining

At present, emotional tendency analysis is based on statistics, machine learning and semantics. In this paper, pattern matching was combined with machine learning, and the process of pattern matching and machine learning (PMML) is shown in **Figure 1**.

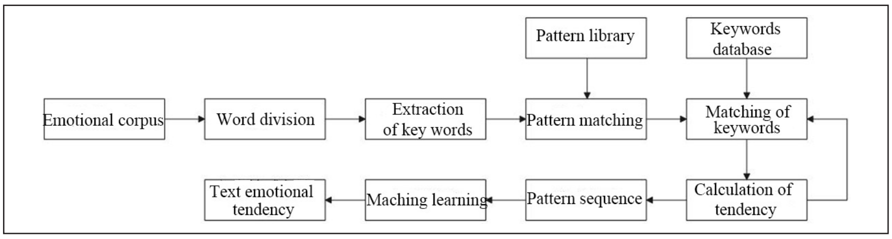

Figure 1: The flow of PMML.

In order to reduce the computational load, emotional tendency is divided into positive and negative regardless of the neutral tendency, which are marked as 1 and – 1, respectively. At the training stage, the text is manually annotated, and the keywords are extracted to establish a keyword library. Then key words which has the largest classification determination degree are used to constitute a pattern library.

In the classification stage, key words are extracted from the words which needed to be classified, and emotional patterns, i.e., pattern characteristics, are generated according to the established pattern. They are matched with the pattern library, and then the emotional tendency is calculated to get the pattern sequence. Finally, the intensity of the emotional tendency of the text is obtained by machine learning such as SVM.

3.1. Extraction of key words

Firstly, a text is divided into nouns (n), verbs (v), adjectives (a), adverbs (d), prepositions (p), conjunctions (c) and others (o) by Institute of Computing Technology, Chinese Lexical Analysis System (ICTCLAS). Each word has a fixed position and function in sentences, which facilitates pattern extraction and matching.

According to the difference between words and expressions, the key words are divided into:

W1: emotionally appraising words (including nouns, verbs, adverbs and adjectives).
W2: negative words (no, none, etc.)
W3: Degree Adverbs (very, quite, particularly, etc.)
W4: transitional conjunctions (but, however, etc.)

3.2. Pattern extraction and matching

Verbs, nouns, adjectives, adverbs can form fixed collocations with adverb of degree and negative words in the expression of emotional tendency, as shown in **Table 1**.

The window length of word collocation is set as 3 considering that the sentence in network public opinion is short. Then for emotional pattern E_i, $i = 1, 2, \dots 10$. Collocation which is suitable for the pattern is selected from key words.

3.3. Calculation of emotional tendency

3.3.1. Calculation of key word tendency

W1 was taken as the subject and HowNet emotional dictionary which is a natural language processing and research software from HowNet was taken as the basis in the calculation of key word tendency. Seed set with evident commendatory and derogatory emotional tendencies is selected, and the emotional tendency of new words is obtained according to the similarity between new words and words in the seed set. For word c_1 and c_2, the similarity is defined as $S(c_1, c_2)$, and the distance between words is defined as D(c_1, c_2), then

$$S(c_1, c_2) = \frac{\partial}{D(c_1, c_2)}, \tag{1}$$

where ∂ stands for a controllable parameter and D is a constant.

Table 1: Matching patterns of phases.

Pattern	Collocations	Example
E1	Adjective + noun	Intimate service
E2	Adjective + adverb	High enough
E3	Adjective + verb	Efficient operation
E4	Noun + adjective	Allocation perfect
E5	Adverb + adjective	Especially like
E6	Negative word + verb	Not satisfactory
E7	Negative word + adjective	Not beautiful
E8	Negative word + adverb + verb	Not very like
E9	Verb + adjective	Consume fast
E10	Adverb + negative word + verb	Not happy

Through the calculation of word similarity, semantic tendency measures can be obtained. The positive emotional paradigm phrase is defined as *cp*, and the negative emotional paradigm phrase is defined as *cn*; the number is M and L respectively. The formula for tendency P of word c is:

$$P(c) = \frac{1}{M}\sum_{m=1}^{M} S(c, cp_m) - \frac{1}{L}\sum_{l=1}^{L} S(c, cn_l). \tag{2}$$

The word is considered as commendatory if the value of P was positive; otherwise it is derogatory.

3.3.2. Calculation of tendency under different patterns

The tendency of different patterns is very different; therefore it should be considered separately.

Adjectives in E1 and E4 indicate emotional tendency, so the tendency of patterns is the tendency of adjectives.

E2, E5, E8 and E10 include adverbs, which will strengthen emotional tendency. The degree of adverbs is divided into three levels, strong (very and especially), medium (relatively and slightly), weak (a little and some), and they are given different emotional weights *R*(*c*).

$$R(c) = \begin{cases} 2, \ strong \\ 1, \ moderate \\ 0.5, \ weak \end{cases} \tag{3}$$

In E3, if the verb is an emotional word and the modifier is an adverb of degree, then it is the same as E5. If the verb is not an emotional word, then the adjective means emotional tendency, the same as E1.

The emotional expression of E9 is more complex, and the same expression may have different emotional tendencies in different contexts. For example, the emotional tendency of "consume fast" is different in "the power of the electroplate consumes fast" and "cupcakes in the store consume fast and usually are sold out before afternoon". Therefore the tendency of E9 needs to be marked manually.

Because the existence of negative words in E6 and E7 will change the polarity of emotional tendency, therefore it is necessary to change the polarity after acquiring emotional tendency. In addition, when the two negatives words appear at the same time, it means double negative.

3.4. SVM based classification method

After the value of each pattern is obtained, the pattern sequence vector can be obtained as well. According to the pattern sequence vector, the tendency of text can be obtained by classification based on machine learning. N-Gram feature extraction is still needed in classification, but pattern matching can greatly reduce the number of features needed and the amount of computation.

For linearly separable problem, there is a training sample set $\{(x_1, y_1)\ (x_2, y_2), \ldots, (x_N, y_N)\}$, in which $x_i \in X, y_i \in Y = \{1, -1\}$. Solve a linear decision function $g(x) = w \cdot x + b$, where w stands for weight vector and b stands for constant. Interval is maximized on the premise of satisfying $y_i[(w \cdot x_i) + b] \geq 0$, $i = 1, 2, \ldots, N$, then

$$\begin{cases} \min \ \frac{\|w\|}{2} \\ st. \quad (y_i(w \cdot x_i) + b) - 1 \geq 0, i = 1, 2, \cdots, N \end{cases}. \tag{4}$$

Solve it using lagrangian multiplier, then

$$w(\alpha) = \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{n} \alpha_i, \tag{5}$$

where α stands for Lagrangian multiplier.

The final decision function is

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i (x_i \cdot x) + b\right). \tag{6}$$

For nonlinear problem, linear classification after non-linear conversion can be realized after slack variable ζ is introduced. The objective function is:

$$\begin{cases} \min \quad \frac{\|w\|^2}{2} + C\sum_{i=1}^{N}\zeta_i \\ y_i\left[(w \cdot x_i) + b\right] \geq 1 - \zeta \quad i = 1, 2, \ldots, N \\ \zeta \geq 0 \end{cases}, \tag{7}$$

where C stands for penalty factor.

Through solution based on Lagrangian multiplier, the final classification function is:

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{n}\alpha_i y_i K(x_i \cdot x) + b\right). \tag{8}$$

4. Instance Analysis

4.1. Data source

Taking "Meituan hotel" as the key word, search on the micro-blog platform. Comment texts were collected using Bazhuayu data collector from 0:00 on June 15, 2018 to 24:00 on June 18, 2018. A total of 7,892 comments were obtained, and 200 positive comments were selected (such as "The hotel has a good environment and the service is very good) and 200 negative comments (such as "sound insulation and hygiene are too bad, not sleeping at all"). The 400 comments were regarded as classification subjects. Tag of positive and positive emotion were marked.

4.2. Data processing

The comments were segmented using ICTCLAS. Then key words were extracted according to section 4.1. Characteristics phases which were successfully matched were obtained using the method described in section 4.2 to form pattern sequence vector.

Seven hundred and seventy-four words were selected from HowNet emotional dictionary as the seed set, and it included 452 positive emotional words and 322 negative emotional words. Positive emotional words included clean, comfort, beautiful, quiet, enthusiastic, tender, etc., and negative emotional words included sick, bad, excessive, disappointed, etc.

4.3. Evaluation indicators

For the evaluation of text classification effect, the commonly used indicators are precision rate and recall rate. Precision rate refers to the proportion of correctly identified samples, and recall rate refers to the ratio of number of the correctly identified samples, which is used for representing the completeness of classification. The precision rate and recall rate of text classification is shown in **Table 2**.

The precision rate and recall rate were calculated according to **Table 2**:

Precision: $P = \frac{w}{w + x}$

Recall rate: $R = \frac{w}{w + y}$.

Table 2: The classification situation.

	Actual number of texts which belong to the class	**Actual number of texts which do not belong to the class**
Number of texts which are identified as the class by the classifier	w	x
Number of texts which are identified not as the class by the classifier	y	z

4.4. Experimental results

Firstly, N-gram feature selection method and SVM classifier were used to analyze the emotional tendency of the data. Three feature presentation methods, UniGrams, BiGrams and TriGrams, were used. 70% of the 400 comments were used for SVM classifier based training and 30% for experiment. The evaluation indexes were PP (accuracy rate of commendatory comments), PR (recall rate of commendatory comments), NP (precision rate of derogatory comments) and NR (recall rate of derogatory comments). The results are shown in **Table 3**.

Table 3 shows that the classification results of the three feature representation methods were not very good. Generally speaking, BiGrams' classification performance was slightly better, and its precision and recall rates were slightly higher than those of the other two methods, which showed that BiGrams was a little more accurate and comprehensive in text classification.

Then the PMML method combined with feature selection was used to analyze the emotional tendency.

The comparison of **Tables 4** and 3 shows that the precision rate and recall rate of each item were obviously higher than those shown in **Table 3** when PMML method was used, which indicated that that the PMML method in this paper is effective and can classify the emotional inclination of comment text accurately and comprehensively. It could be noted from **Table 4** that the BiGrams feature representation method had a high precision rate and recall rate, and the precision rate of commendatory comments was 86.75%, which was similar to **Table 3**. It indicated that BiGrams feature representation method had better performance in text classification.

4.5. Classification of comment texts

According to the experiment shown in the last section, PMML + BiGrams method shows better performance in categorizing text emotional tendencies. This method was used to classify the emotional tendency of 7,892 hotel comments. The results are shown in **Table 5**.

The classification results suggested that the proportion of comments with negative emotional tendency was greater than that with positive emotional tendency. It was because that most of the users tend to complain on the micro-blog when they had a bad living experience or in a bad mood, and it also showed that Meituan hotel failed to provide a good living environment for users because of many problems. In order

Table 3: The comparison of classification performance of N-Gram.

Feature	UniGrams	BiGrams	TriGrams
PP	62.72%	67.62%	65.85%
PR	54.27%	47.55%	43.72%
NP	60.38%	63.82%	59.32%
NR	72.54%	76.41%	64.78%

Table 4: The comparison of classification performance of PMML.

Feature	PMML + UniGrams	PMML + BiGrams	PMML + TriGrams
PP	82.15%	86.75%	78.25%
PR	80.53%	83.49%	76.43%
NP	82.78%	84.64%	80.59%
NR	79.21%	78.21%	77.67%

Table 5: The classification of emotional tendency of comment texts.

Category	Number (n)	Percentage (%)
Positive (commendatory) emotional tendency	2764	35.02
Negative (derogatory) emotional tendency	5099	64.61
Not classified	29	0.37

to get considerable development, attract new customers and retain old customers, Meituan hotel need to strengthen the management of hotels and improve quality of service.

In addition, 29 comment texts were not categorized, which showed the deficiency of emotional tendency analysis. For example, there is a comment, "the staff told me that there was WiFi in the hotel when I booked, but it is not true." This comment clearly expresses a derogatory emotional tendency, but neither pattern nor machine learning can identify and analyze it. Because of the variety of language expression forms in the network public opinion, especially the use of non-standard language and the existence of network vocabulary has brought great challenges to text mining technology.

5. Discussion and Conclusion

With the development of science and technology, the Internet has gradually become the main cradle of public opinion (Song et al. 2014). With the development and popularization of the Internet, network public opinion is playing an increasingly important role in society. More and more network public opinions has become social public opinion problems through the Internet. It is found that network public opinion contributes to more than half of social events, and positive public opinion can raise social concern and promote the solution of events. However, it is difficult to distinguish the true from the false, and there is harmful information in public opinions. Because of the high openness and freedom of the network, harmful information is difficult to be completely eradicated, which brings about a lot of network security problems. The endless harmful information exerts a subtle influence on people's mood and thinking through various means and encourages the dissemination of improper speech. With its growth, its influence extends from the network to the reality, causing social panic and turbulence. Therefore, the influence of network public opinion can not be ignored, and the massive public opinion data information needs to be processed and analyzed using effective methods. Text mining technology is of great help to the analysis of network public opinion information. Network public opinion analysis includes sensitive problem mining, hot spot tracking, trend control and emotional orientation analysis (Wang & Liu 2011). Through the analysis of emotional tendencies, we can grasp the people's real feelings and attitudes toward people or events, understand the people's position in current politics, and make decisions that conform to public opinion (Luo 2014). Analysis of emotional tendency of hot news can guide public opinion, stabilize social order, and prevent bad comments.

In order to effectively analyze public opinion information on the Internet, this study proposed PMML. Firstly, a pattern library was established by extracting keywords. Then, keyword orientation and pattern orientation were calculated. Finally, emotional orientation classification was realized by SVM. In order to verify the validity of the method, this study classified the emotional orientation on the evaluation on "Meituan Hotel" on the Internet. As to feature representation methods, UniGrams, BiGrams and TriGrams were compared, and it was found that BiGrams method in combination with PMML had the best classification performance, with a forward accuracy rate of 86.75%. Hence BiGrams was used for text classification. The results showed that 35.02% of the 7892 reviews was positive, 64.61% was negative, and 29 texts (0.37%) were not classified. The results showed that the method had good performance in classifying emotional tendencies of the hotel reviews and the negative tendency was greater than the positive tendency. The findings is an alarm for hotel managers. They are urged to enhance hotel management and provide better services. The 29 unclassified texts indicated that there were some shortcomings in the method, which needs to be further studied, and also showed that the text mining was difficult.

Emotional tendency analysis of network public opinion information is of great significance to maintain network security and promote social stability. In this paper, a PMML based method was proposed to classify the emotional tendency of network public opinion. The experiments showed that the accuracy of PMML in combination with BiGrams feature representation method was 86.75%, which indicated that the method was feasible, has broad application prospects, and strengthen network harmony and security.

Competing Interests

The author has no competing interests to declare.

References

Chen, X, Xia, M, Cheng, J, Tang, X and **Zhang, J.** 2016. Trend prediction of internet public opinion based on collaborative filtering. In: *International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, Changsha, China*, 13–15 August 2016, 583–588. DOI: https://doi.org/10.1109/FSKD.2016.7603238

Guo, K, Shi, L, Ye, W and **Li, X.** 2014. A survey of Internet public opinion mining. In: *International Conference on Progress in Informatics and Computing, Shanghai, China*, 16–18 May 2014, 173–179. DOI: https://doi.org/10.1109/PIC.2014.6972319

Hao, X, An, H, Zhang, L, Li, H and **Wei, G.** 2015. Sentiment Diffusion of Public Opinions about Hot Events: Based on Complex Network. *Plos One*, 10(10): e0140027. DOI: https://doi.org/10.1371/journal.pone.0140027

Li, YZ and **Zhang, MS.** 2014. Design and Implementation of Internet Public Opinion Monitoring System. *Advanced Materials Research*, 926–930: 1902–1905. DOI: https://doi.org/10.4028/www.scientific.net/AMR.926-930.1902

Liu, H. 2010. Internet Public Opinion Hotspot Detection and Analysis Based on Kmeans and SVM Algorithm. In: *2010 International Conference of Information Science and Management Engineering, Xi'an, China*, 7–8 August 2010, 257–261. DOI: https://doi.org/10.1109/ISME.2010.207

Liu, H and **Li, X.** 2010. Internet Public Opinion Hotspot Detection Research Based on K-means Algorithm. In: *Advances in Swarm Intelligence, First International Conference, Beijing, China*, 12–15 June 2010, 594–602. DOI: https://doi.org/10.1007/978-3-642-13498-2_78

Luo, Y. 2014. The Internet and Agenda Setting in China: The Influence of Online Public Opinion on Media Coverage and Government Policy. *International Journal of Communication*, 8(2): 1289–1312.

Song, B, Zhu, JM and **Huang, QF.** 2014. The Internet public opinion grooming model based on cluster dynamics and evolutionary game theory. *Xitong Gongcheng Lilun Yu Shijian/System Engineering Theory & Practice*, 34(11): 2984–2994.

Tran, YH and **Tran, QN.** 2017. Estimating public opinion in social media content using aspect-based opinion mining. In: *International Conference on Mobile Networks and Management*, 101–115.

Wang, W and **Tang, Y.** 2016. Network public sentiment orientation analysis based on HMM Model. In: *International Conference on Wireless Communications, Signal Processing and NETWORKING, Chennai, India*, 23–25 March 2016, 2269–2273. DOI: https://doi.org/10.1109/WiSPNET.2016.7566546

Wang, X and **Liu, Q.** 2011. Topic semantic orientation compute based on sentiment words Ontology. *Computer Engineering & Applications*, 47(27): 147–151.

Yang, Y. 2017. Research and Realization of Internet Public Opinion Analysis Based on Improved TF – IDF Algorithm. *International Symposium on Distributed Computing and Applications To Business, Engineering and Science*, 80–83. DOI: https://doi.org/10.1109/DCABES.2017.24

How to cite this article: Hu, Y. 2019. Text Mining and Data Information Analysis for Network Public Opinion. *Data Science Journal*, 18: 7, pp. 1–8. DOI: https://doi.org/10.5334/dsj-2019-007

Submitted: 02 November 2018 **Accepted:** 04 January 2019 **Published:** 23 January 2019

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS