A method of filtering bad text in network based on topic sensitive words tagging

Jianying Xiong
Security Technology Department of Jiangxi Police College, Jiangxi Police Institute, Nanchang, 330013, China

Abstract. Bad information filtering is conducive to creating a healthy network. A two-level filtering method based on topic and sensitive words is proposed. In the first stage, the network text is filtered by using thesaurus, by setting the weight of different topics. In the second stage, according to the frequency, position and sensitivity of sensitive words, the value of bad tendency is obtained by weighting the web text. Finally, taking the text set recognition of bad financial publicity content in the network as an example, the result proves that it can improve the efficiency and accuracy of filtering of bad investment information.

1. Introduction
How to filter the bad information on the Internet and help the public get useful information from the multimedia network? There are many network information security strategies, and to prevent the pollution of bad information is one of the problems to be solved in the network analysis and research [1]. At present, the identification of network bad information is mostly analyzed from the perspective of user characteristics and communication characteristics. In fact, sensitive words are very important features of identifying bad information. The analysis of sensitive words of different subject classification helps to improve the discrimination of bad information and curb the spread and spread of bad information. Sensitive topic analysis is an important topic in social network data mining, which has a wide application prospect in many fields such as anti-terrorism, anti-cult, public opinion analysis, market analysis, anti-rumor and anti-defamation [2].

Thesaurus is a collection of words, usually including basic thesaurus and professional thesaurus. The widely used professional thesaurus includes popular thesaurus, professional ontological thesaurus, sensitive thesaurus, emotional thesaurus, etc. Among them, the existing sensitive thesauri mainly include reactionary sensitive thesaurus, violent terrorist sensitive thesaurus, pornographic sensitive thesaurus, spam advertising sensitive thesaurus, etc., which are widely used in various post bar, forum and spam detection. As sensitive topics are often related to both the topic of the text and the speaker's subjective emotions, and a large number of unknown words are often used in texts containing sensitive topics on social networking sites [3]. Compared with general text mining tasks, sensitive topic analysis faces more noise and higher semantic complexity. Therefore, there is a certain gap between the performance of sensitive topic analysis and other text mining problems.
2. Related works
The conventional method of bad information filtering is to use the word orientation judgment based on the network. The filtering method uses less semantic and grammatical elements, and is prone to misjudgment. Bayesian classification algorithm is used to classify the domain according to the weight of ontology elements [4]. The mode matching between the monitored network information and the candidate network information is carried out, and the bad information filtering is realized based on the matching degree obtained. The selection and extraction of sensitive words is a crucial link to affect the effect of topic analysis. Most of the traditional analysis methods are based on keyword filtering technology. Through traversing the content, we can find whether there are words in the sensitive lexicon. If it contains sensitive words, the text is considered to involve sensitive speech. Otherwise, it is considered not to involve sensitive speech [5]. Although it is simple to implement, the filtering of bad information based on sensitive words is affected by the interference of sensitive words such as camouflage. So the effect of this method is limited, and the labor cost is also large. Moreover, the function of this method is limited, so it cannot be regarded as a strict analysis. Because the results of feature extraction of bad information are not clear, the detection effect of bad information is not ideal and the filtering accuracy is low. In order to improve the performance of analysis, many scholars try more text analysis methods. Some researchers proposed that using KNN Text Classification to filter spam information [6]. After extracting the features, the influence degree of frequent spam messages is analyzed by association rules, and the weighted formula is introduced to deal with the complex weighting of spam messages. It is used to distinguish the influence degree of each training sample on the subordinate category, so as to improve the classification decision and complete the filtering of bad information. This method is relatively simple, but the effect of feature extraction of bad information is not very ideal. Others used the improved rake method to filter the bad information. The similarity coefficient of classification tags is obtained by training text. Then, the coefficient matrix of similarity degree of comprehensive labels is obtained by using the self-defined hierarchical relationship of bad information [7]. At the same time, when voting by rake method, the similarity degree of comprehensive tags and the center label are used to re determine the final tag set. The filtering efficiency of the method is high, but the correctness is poor. It also introduces probabilistic topic model into text feature extraction process. The evaluation function of information feature subset is set up [8]. According to the work procedure of quantum evolution method, the process of feature selection of adverse information in multimedia network is constructed to realize the feature selection of bad information and improve the effect of feature extraction and detection of bad information.

3. Topic sensitive words recognition model
3.1. Model framework
The model adopts two-level matching and filtering mode of topic classification and weighted matching of sensitive words. the filtration process is as shown in Figure 1. The feature terms will be obtained after text is processed, and the text expression feature matrix will be formed. Then using thesaurus and Bayesian theorem to classify the network information. Different topics will have different weights. The weight of document information feature items in the vector space model is recalculated by topic and sensitive words. Finally, by calculating the joint probability of bad information, the whole document tendency can be calculated to determine whether it is suspicious bad information. When dealing with the bad information, the sentence is taken as the basic processing unit, and three choices are made according to the tendency intensity of the bad text: a) delete the text containing the bad sentence directly; b) delete the bad sentence only; c) submit it to the background management interface for manual follow-up processing.
Suppose there are m documents, each of which has n statements and M feature words, involving K topics. Each document has its own topic. Set document to \((d_1, d_2, \ldots, d_m)\). The statement in each document indicates that \((S_1, S_2, \ldots, S_m)\). The subject weight is \((SW_1, SW_2, \ldots, SW_m)\). According to the sensitive word list, the weight of sensitive words hit by each document is \((MW_1, MW_2, \ldots, MW_m)\). Then the calculation method of adverse information \(TQ\) is as follows.

\[
TQ = SW \sum_{i=0}^{M} MW_i
\]  

(1)

### 3.2. Topic classification based on LDA

Bad information is often concentrated on a few sensitive topics. We are more concerned about such as food safety, child assistance or loss, and etc.. However, it can be divided into several categories, including junk information, illegal information, illegal vulgar information, reactionary information and so on. LDA topic model is also namely potential Dirichlet distribution model [9]. It is a document topic generation model and a three-tier Bayesian probability model with three-tier structure of words, topics and documents.

![Figure 2. Two stage filtering mode.](image)

The topic distribution is a polynomial distribution, and the parameters of the polynomial distribution obey the Dirichlet distribution [10]. As shown in Figure 2, for the nth word in a document \(D\), first a topic is used from the topic distribution of the document, and then a word is used in the word distribution corresponding to the topic. The operation is repeated until all the M documents have completed the above process.

In Python, LDA analysis steps mainly include reading data and word segmentation; removing stop words; building TF/IDF matrix. Each line represents a test document, and each column represents the
TF/IDF value of a word. LDA analysis (fit and transform), output the results. At the critical moment of training, assuming that the number of topics is M and the number of iterations is 1500, the model can be initialized in the following ways: lda.LDA (n_topics = M, n_iter = 1500, random_state = 1).

3.3. Construction of sensitive vocabulary based on illegal finance

Traditional bad information detection methods have the problem of low accuracy, which is based on keyword character matching and coarse-grained sentiment analysis. A keyword extraction method based on ontology model is proposed. The weighted calculation method of word frequency factor, position factor and sensitivity level are used to extract keywords [11]. \( MW_i \) is the weight of a sensitive word, it can be computed as formula (2).

\[
MW_i = loc_i \times fre_i \times lev_i
\]  

(2)

In the above formula, \( loc_i \) is the position factor, which distinguishes two positions \{1,2\}, one is the title and the other is the body. \( fre_i \) is the word frequency factor; \( lev_i \) is the level factor of sensitive words. Generally, sensitive words are divided into three levels \{1,2,3\}, which respectively represent absolute prohibition, general prohibition and need to be reviewed. These three levels are all manually divided.

The word frequency factor \( fre_i \) is computed as formula (2), \( f_i \) is the word frequency of a word.

\[
fre_i = f_i / (1 + f_i)
\]  

(3)

Taking financial illegal information as an example, in recent years, financial risk events have occurred one after another, which seriously infringe on the legitimate rights and interests of financial consumers. From the regulatory practice, illegal financial activities often release a large number of illegal marketing information on the network in the early stage to lure financial consumers. The common suspected illegal and false financial propaganda texts mainly involve financial related words such as credit investigation, loan and interest rate. The sensitive lexicon is constructed from the aspects of investment income, investment guarantee and investment subject. The sensitive vocabulary and its sensitive level are shown in the table I. The common neutral nouns are relatively low, and the words of seductive combination are more sensitive. In practice, we can use Python program to detect sensitive words.

| Serial number | sensitive word       | sensitive value |
|---------------|----------------------|-----------------|
| 1             | investment           | 1               |
| 2             | Credit investigation | 2               |
| 3             | recommend            | 2               |
| 3             | No credit is required| 3               |
| 4             | Black door second    | 3               |
| ...           |                      |                 |

Table 1. Sensitive word list.

4. Experimental analysis

4.1. Experimental description

Taking illegal finance as an example, suspected illegal and false financial advertisements generally have six characteristics. 20 suspicious information is manually screened out by grabbing the content of Social Forum on the network, and mixed with 200 ordinary texts. The suspicious false propaganda texts are detected by model. In the experiment, precision ratio \( p \) and recall ratio \( R \) are used as experimental evaluation indexes, and comprehensive evaluation index \( F \) value is selected as the harmonic average of \( P \) and \( R \) to measure the accuracy of evaluation and analysis. These indicators are defined as follows: the number of suspicious texts judged as correct is \( nprE \), the number of suspicious
texts judged by the system is Nall, and the number of suspicious texts should be judged as nori. Then the calculation formula of each index is as follows:

$$P = \frac{N_{pre}}{N_{all}} \quad (4)$$

$$R = \frac{N_{pre}}{N_{ori}} \quad (5)$$

$$F = \frac{2 \times P \times R}{(P + R)} \quad (6)$$

4.2. Experimental results
The samples selected in the experiment are suspicious financial illegal publicity text information. There are six characteristics of suspected illegal and false financial advertisements. These characteristics mainly include: 1). Commitment that "the return on investment and financial management is significantly higher than the benchmark loan interest rate over the same period"; 2). Commitment to "100% principal and interest protection"; 3). Promise "no credit report, as long as one ID card, you can make 24-hour loan", "no credit reference seconds", "black door second"; 4). The words "investment is risky" are not marked; 5). The word "advertisement" was not marked prominently; 6).Use the names or images of regulatory agencies, academic institutions, industry associations, professionals and beneficiaries for recommendation and certification. Because there are some obvious sensitive words in the investment publicity text, such as investment and financing return, significantly higher than, 100% principal and interest protection, no credit report and other sensitive words. But neutral words such as principal and interest protection, credit report and investment may also appear in the general financial management text, and certain risks may be identified in the labeling of sensitive words.

The sensitivity tendency value of each network text is calculated by the filtering model in this research. The calculation results are analyzed and described by basic statistics. As shown in Table II, the calculation results show that the average value and the quartile value of the calculation results are low because most of the texts do not involve sensitive information. And most of the text tendency values are low. The maximum value of the calculation result is 16.66667, and the range of the data is large, which also shows that the tendency of suspicious text has a good discrimination here.

| count | 220 |
|-------|-----|
| mean  | 1.693636 |
| std   | 2.780644 |
| min   | 0 |
| 25%   | 0 |
| 50%   | 0.833333 |
| 75%   | 1.75 |
| max   | 16.66667 |

Table 2. Sensitive word list.

The calculation results of the content sensitivity tendency of all 220 web texts are shown in the scatter diagram Fig.3. The text with high sensitivity will get a relatively high degree of bad text tendency. If we set a corresponding threshold, we can divide these texts into certain categories.

In order to distinguish the bad text, we set 3.5 as the threshold for the weighted tendency value. The total number of suspicious texts identified is 27. Of these identified texts, 18 are the real target text, 9 were recognition error text, and 2 target text were not recognized. The calculated values of P, R and F are as Fig.4. The experimental results show that it has a good recognition, which is conducive to the realization of bad financial information filtering. But it is easy to identify the common investment text. In the risk identification scenario, even if some mistakes are made in the propaganda texts at the edge, it can also guide them to further standardize their own propaganda.
5. Conclusions

In this study, through analyzing the content characteristics of bad financial investment publicity text, the model combines with the existing research results of bad information monitoring, and proposes a two-level network text filtering method based on topic and sensitive words tagging. After processing the filtered text, we use thesaurus and sensitive thesaurus to obtain the adverse tendency vector of each sensitive word on the text by weighting, and select the text operation of the bad sentence according to the tendency intensity of the bad text. The experimental results show that the recall of the algorithm is not less than 90% in the case of network bad financial propaganda content text. But the precision rate is relatively low. However, from the perspective of financial risk prevention and control, it helps to promote these enterprises who published misjudged texts to further standardize publicity. In order to improve the efficiency of applying to different topics, the next step of our work is to improve the semantic structure of sensitive words.

Acknowledgement

This work was supported by Ministry of education's humanities and Social Sciences youth fund project supports "Research on social media communication governance based on trust and content risk perception (20 YJC 860032).

References

[1] Wechsler H. Cyberspace Security Using Adversarial Learning and Conformal Prediction. Intelligent Information Management, 2015, vol.07, no.4, pp.195-222.
[2] Kasper A, Krasznay C. Towards Pollution-Control in Cyberspace: Problem Structure and Institutional Design in International Cybersecurity. International and Comparative Law Review, 2019, vol.19, no.2, pp.76-96.
[3] Fang Q, Sang J, Xu C, et al. Topic-Sensitive Influencer Mining in Interest-Based Social Media Networks via Hypergraph Learning. IEEE Transactions on Multimedia, 2014, vol.16, no.3, pp.796-812.
[4] Hu, Yuhuang , et al. "Learning sufficient representation for spatio-temporal deep network using information filter." , 2014, pp.655-658.
[5] Yi-Gui D , Yu-Ying W U . Information Filtering Algorithm of Text Content-based Sensitive Words Decision Tree. Computer Engineering, 2014, vol.40, no.9, pp.300-304.
[6] Zhang, Shichao, Li, et al. Efficient kNN Classification With Different Numbers of Nearest Neighbors. IEEE Transactions on Neural Networks & Learning Systems, 2018, vol.29, no.5, pp.1774-1785.
[7] Bing Z, Yiliang S, Cong L. Research on Improvement of Mass Multimedia Information Filtering Technology Under Multi - Data Inflow// 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS). 2019.
[8] Feng G, Li S , Sun T, et al. A Probabilistic Model Derived Term Weighting Scheme for Text Classification. Pattern Recognition Letters, 2018, vol.110, pp.23-29.
[9] Fu R, Qin B, Liu T. Open-categorical text classification based on multi-LDA models. Soft Computing, 2015, vol.19, no.1, pp.29-38.