Research on Automatic Proofreading Method of Sensitive Information in Content Security

Yonggang Gong*, Yuying Li , Xiaoqin Lian , Junying Fu

School of Computer and Information Engineering, Beijing Technology and Business University, Beijing Key Laboratory of Food Safety Big Data Technology, Beijing, China

*Corresponding author e-mail: gygbenson@126.com

Abstract. Aiming at the problem of automatic proofreading of sensitive information in mass text content, an automatic proofreading method based on the combination of rule and SVM (Support Vector Machine) is proposed. To classify sensitive information based on important sensitive information provided in the “Newly Prohibited Texts and Cautions in Xinhua News Reports” (newest revision) and related central and online texts. According to the different categories, the paper constructs the classification processing rule base, designs the corresponding rules automatic Processing algorithm, and realizes the sensitive information automatic proofreading. At the same time, using the SVM model to analyze the result of the rule processing with emotion, which greatly reduces the false alarm rate. The test result shows that the recall rate of method is 89.98%, the accuracy rate is 98.31%, and 100,000 + text content is processed per second, which solves the key difficult problems in the practical engineering application.

1. Introduction Content

Content security is the requirement of information security in the political, legal and moral level, it is required that the content of information is healthy in politics, and comply with national laws and regulations in law, and conform to the good moral standards of the Chinese nation morally. As the core technology of intelligent information processing in network security, content security provides technical support for advanced network culture Construction and strengthens the network communication of socialist advanced culture, which belongs to the important part of national information security system [1]. Using the filtration methods to monitor and control unhealthy information, to prevent the illegal dissemination of bad information, is widely recognized, there are many specific and filtration-related technologies, however, the complexity of the Internet content flow
determines some single technology is often not enough. In the practical application of these technologies must be organically combined [2-3]. At the same time, a lot of sensitive information covers a wide range of types, if we don't have serious and cautious attitude, solid language skills, high level of policy theory and political sensitivity, it is difficult to find these problems [4]. Based on this, the Xinhua news agency published the "Forbidden words and cautious words in the Press Information report (July 2016 revision)" and gave the uniform norms and standards for proofreading of sensitive words. With the advent and rapid development of the "Internet +" era, new media (Weibo, micro-letter public number, blog, news client, public number, etc.) has become an integral part of people's lives, many news media platform, the daily original press release volume, and the timeliness of the news so that it will be in a short time by the major media widely reproduced forward, and was read by hundreds of millions of netizens. Therefore, if there are typos, political sensitive words and other bad information in the news content, it is possible to have extremely bad influence. Taking today's headline as an example, today's headlines are up to 500,000 dailies, with manual audits impractical. Therefore, it is necessary to use automatic identification technology based on rules and semantic analysis before the manuscript is issued to ensure "timely and accurate" detection of problems, positioning problems and solving problems. The research of automatic proofreading system with high speed processing of mass content sensitive information has important practical significance and wide application field.

This paper takes the sensitive content in the "forbidden words and cautious words" in the news report of the new Edition (latest revision) as reference, on the basis of its specification and relevant central and the important sensitive information provided by the network text, the sensitive information is defined and classified in detail, and the sensitive information proofreading rules base is developed according to the category of sensitive words. The corresponding filtering algorithm is designed for all kinds of rules, and the result of the rule processing is analyzed with the SVM model (Support Vector Machine) to reduce the false alarm rate.

2. Problems description

According to the importance and frequency of errors in sensitive information, it can be divided into the following four categories:

2.1 Errors in Hong Kong, Macao, Taiwan or policy

This error mainly includes relevant laws and regulations, damages the interests of the state, affects national unity or international relations, and involves mistakes in territorial sovereignty, Hong Kong, Macao and Taiwan (especially the Taiwan question) [6]. This kind of error, involving a wide range and a lot of content, which is difficult to implement with a rule method, so for this kind of error, apply range rule, regular expression rule, disable, replace and put quotation marks and other ways, according to different situations, choose easy to operate, the effective way to carry out proofreading. For example, "Hong Kong" and "Macao" cannot be described as countries, nor should they be mentioned in parallel with China [5]. However, in an article published on March 1, 2018, "American Heritage Foundation: Mauritius leading index of economic freedom in Africa", "the top countries in the world are Hong Kong, Singapore, New Zealand, Switzerland, Australia and Finland". That's the wrong statement.
2.2 Information leakage

Information leakage includes: privacy disclosure, military information disclosure and keep their real names secret, etc.

There is a lot of public information in the network, such as someone's name, work place, email, telephone number, etc. In some cases, this information can be consolidated and potentially abused, for example, some companies will sell the data as business information. Some fraud groups will use this information to defraud [1]. Army designations, base and numbers are not allowed to be disclosed. For patients with serious infectious diseases, patients with mental illness and AIDS, etc., when these people are involved, the manuscript may use the real surname plus the reference of "some", and the real name shall not appear.

2.3 Pornography

For a large number of harmful contents in violence, eroticism and pornography. If it is rampant on the Internet, it will not only bring instability to the country, but also seriously harm social security.

2.4 Reactionary information

Some lawless elements and cult organizations, the publication of reactionary rhetoric, propaganda, influence the stability of the country and long-term security, this error, in the content of information safety is very rare, but, once it happens, it is a very serious political problem.

3. Overall process of sensitive information processing

The sensitive information proofreading can detect the sensitive information by the given text, and locate and mark the text processing method of the sensitive word [7]. This method realizes the method of proofreading after word segmentation and classification, The method is implemented by using the first word segmentation and then classifying and proofreading. for the suspicious sensitive words after participle, according to the category of sensitive words, apply the corresponding rules for proofreading, and then through emotional analysis to reduce false positives, finally, the sensitive information positioning, highlighting, and give the corresponding error or modify comments [8]. The overall process of sensitive information processing is shown in Figure 1.

![Figure 1. The overall process of sensitive information processing](image_url)

Rules for overall design:

1. The text to be tested is first participle, then the matching tree of AC double group composed of all the sensitive words, and then the sensitive words filtered out, as suspicious sensitive words, record the sensitive word, the position in the text and the type of the sensitive word;
(2) Triggering different sensitive word rules according to the type of sensitive words;
(3) Through the rule proofreading result, will express the wrong sensitive information, carries on the emotion analysis;
(4) The sensitive words after the emotional analysis, as the identified sensitive words, highlighting, and give the corresponding error prompts or modify opinions.

4. Sensitive information processing technology and principles

4.1 Sensitive information proofreading algorithm

In view of the 4 kinds of sensitive information errors proposed in the 2nd section, according to the constructed thesaurus, the word segmentation is done through HanLP, and then the sensitive information proofreading is realized by using the algorithm in the rule base. Due to the space limit, here only the sensitive information about Hong Kong, Macao and Taiwan to proofread the algorithm. Because of the issues of Hong Kong, Macao and Taiwan, which covers a wide range of types. It can’t be uniformly handled with a rule. Therefore, in response to this situation, set up a different rule algorithm, mainly has scope rules and regular expression rules. The scope of rules for example to introduce.

Range rule algorithm

(1) Algorithm process:

![Algorithm flow chart for range rule](image)

The algorithm of the range rule algorithm is shown in Figure 2. The specific algorithm is described as follows:

① Trigger scope rules by suspicious sensitive words;
② Make the information in the include and exclude fields in the rule into an AC binary match tree. For the following use;
③ The scope of suspicious sensitive words is determined by the rule. such as: word, sentence, paragraph, article;
④ The distance between the front and back of the suspicious sensitive word determined by the rule is F, B; If the range of ③ is word, execute ⑤; If the range of ③ is sentence, execute ⑥; If the
range of ③ is paragraph, execute ⑦; If the range of ③ is article, execute ⑧;

⑤ Start with the position of the suspicious sensitive word, count F words forward to a0 (the maximum can reach the begin of this sentence), count B words back to b0 (the maximum can reach the end of this sentence). Obtain the text between a0–b0, perform the word segmentation, and put it into the set, then execute ⑨;

⑥ Start with the sentence in which the suspect sensitive word is located, count F sentences forward to a0 (the maximum can reach the begin of this paragraph), count B sentences back to b0 (the maximum can reach the end of this paragraph). Obtain the text between a0–b0, perform the word segmentation, and put it into the set, then execute ⑨;

⑦ Start with the paragraph in which the suspect sensitive word is located, count F paragraphs forward to a0 (the maximum can reach the begin of this article), count B paragraphs back to b0 (the maximum can reach the end of this article). Obtain the text between a0–b0, perform the word segmentation, and put it into the set, then execute ⑨;

⑧ Segment the whole text and put it in the collection to execute ⑨;

⑨ Enter each element in the collection into ① in the form of a stream. While satisfying the conditions of include and exclude, the word is determined as a sensitive word, highlighted, and the corresponding error prompt or modification opinion is given; if one party is not satisfied, it is determined that the word is not a sensitive word;

(2) Instance of scope rule

```
<rule words="nei di" sensitiveType="custom sensitive words" desc="****" business="ALL">
    <action id="1" type="scope" explanation="Suggested to change to: mainland">
        <include distance="0,0" scope="sentence">tai wan, tai shang</include>
        <exclude distance="0,0" scope="sentence">xiang gang, ao men, gang ao, gang shang, "nei di"</exclude>
    </action>
</rule>
```

Sensitive information in the sentence of the situation is very complex, in practice, the need for a variety of sensitive information proofing situation analysis, perfect the rules, and improve the accuracy rate [6]. In the same case, it is necessary to select the more efficient and more appropriate rules from different rules for application.

4.2 emotion analysis

For some sensitive information, such as "Falun Gong", "terrorist organization" and so on, can’t directly prevent its appearing in the report, there are many reports of such bad information is criticized and stopped, can occur. If only by disabling rules to do processing, there will be a large number of misinformation, therefore, Emotional analysis is needed.

Through a large number of experimental results show that part of speech, sensitive words and negative words play a certain role in the detection of sensitive information, and improve the correct rate, but the degree of adverb is reduced slightly and the degree adverb is not suitable to be the characteristic of SVM model. That is: when use of part of speech, sensitive words and negative words of the combination of characteristics have the best effect [9].
5. Experimental result analysis

5.1 Construction of test set
Select the central Ministry of government website 26, provincial government website 32, day active users Tens news site 20, the collected data are divided into: daily data, the whole station collection data and the whole station dead chain data, in order to meet the real-time nature, selected the daily data. Test set selected for nearly 10 days of daily data, total data volume of 819,528, filtering sensitive information, found that contains sensitive information of 11,586 pages, the number of sensitive information is:20646. Processing of the resulting sensitive information and deletion of the prohibited words that occur in specific industries and on specific occasions, as the industry's screening criteria have not been added and, if shown, will affect the accuracy rate.

Because the amount of data after processing is still very large, it is difficult to verify the reliability of the information obtained, and in order to simulate the real error better, this paper chooses 1000 articles containing sensitive information as the test text, which makes up the data set AA. In addition to the sensitive information found in the dataset, other parts of the text are not detected sensitive information, and the dataset is applied as test set data.

5.2 Result analysis
The data set AA is tested by the automatic proofreading method of the sensitive information designed in this paper, and then the artificial check is performed to obtain the final experimental result, as shown in Table 1. The formula for the recall rate $R$ and the accuracy $P$ in the table is as shown in (1) and (2) below. Among them, the number of correctly detected results is represented by $TP$, the number of results of the error detection is represented by $FP$, and the number of results that should be detected but not detected is represented by $FN$:

Recall rate: 
$$R = \frac{TP}{TP + FN} \times 100\% \quad (1)$$

Accuracy rate: 
$$P = \frac{TP}{TP + FP} \times 100\% \quad (2)$$

For the proofreading method, in general, when the recall rate is increased, the accuracy rate is lowered; conversely, when the accuracy rate is increased, the recall rate is lowered. In order to control the recall rate and accuracy to an acceptable range, use equation (3) to represent $F$:

$$F = \frac{2 \times R \times P}{R + P} \quad (3)$$

| Table 1: implementation results |
|-------------------------------|
| Number of articles | Total number of errors found $(TP + FP)$ | Total number of errors in text $(TP + FN)$ | number of correctly detected results $TP$ | recall rate (%) | accuracy rate (%) | False alarm rate (%) | $F$ |
|-----------------------|------------------------------------------|--------------------------------------------|------------------------------------------|---------------|------------------|---------------------|---|
| Text with sensitive information | 1000 | 1773 | 1937 | 1743 | 89.98 | 98.31 | 11.69 | 93.96 |

For sensitive information involving semantics, it is more complex and can’t be implemented by
rules, which is not dealt with in this paper. And there may be a small amount of errors in the manual checking part. Excluding these reasons, as can be seen from the results of table 1, the recall rate, accuracy and F value are very prominent. In order to reduce the omission, the rules that we start to write broader and adapt wildly, then through a large number of data validation, statistics and analysis of each rule of false alarms, then change the rules, and the methods of affective analysis are less false alarms. This method excludes some of the sensitive information that cannot be validated. To ensure the recall rate, through the manual audit to improve the rules of improving the accuracy rate, and the effect is very obvious. It is suitable for the current content security sensitive information automatic proofreading, has been applied in the actual environment, to help the relevant functional departments early detect some sensitive information, which should be dealt with as soon as possible. So, it can’t cause serious impact.

6. Prospect

For the current proofreading results, some words appear on specific occasions, will be sensitive information. For example, Hong Kong and China shall not be mentioned together. So the term "zhongGang" is not allowed to appear, but "zhongGang food" is not sensitive information, and there may be false alarms for such situations, and some of them are defective in the rules. All need to complement the rule base. In addition, the sensitive words detected by the system, especially the semantic aspects, can’t ensure the accuracy of 100%, but also need manual check. The system will be perfected and improved by further study of deep neural network in the future.

References

[1] Zhang Huanguo, Han Wenbao, Lai Xuejia, Lin Dongdai, Ma Jianfeng, Li Jianhua. Overview of cyberspace security, J. Chinese Science: Information Science, 2016, 46(02): 125-164.
[2] Fang Binxing, Guo Yunchuan, Zhou Yuan. ICCON control model and evaluation of Internet information content security, J. Chinese Science (F: Information Science), 2009, 39(09): 951 - 965.
[3] Liu Qi, Li Jianhua. Research on Network Content Security Administration System and Its Key Technologies, J. Computer Engineering, 2003(02): 287-289.
[4] Cai Yanjing, Cheng Xiaohong, Cheng Xianyi. Research on Algorithm of Network Sensitive Information Features Extracting, J. Journal of Changzhou University (Natural Science Edition), 2014, 26(04): 80-85.
[5] Xinhua News Agency. Forbidden words and cautious words in news reports of Xinhua News Agency, 2016—07.
[6] Zhang Yangsen, Tang Anjie, Zhang Zewei. Chinese Text Proofreading for Political News Field, J. Journal of Chinese Information Processing, 2014, 28(06): 79-84+128.
[7] He Shan. Research on Sensitive Words Detection Based on Hadoop in Network Security Audit. Donghua University, 2015.
[8] Wen Durina. Research on an Ontology-based Sensitive Word Filtering Method. Jilin University, 2014.
[9] Li Tingting, Ji Donghong. Sentiment analysis of micro-blog based on SVM and CRF using various combinations of features, J. Application Research of Computers, 2015, 32(04): 978-981.