Discovering New Sensitive Words Based on Sensitive Information Categorization

Panyu Liu[1,2], Yangyang Li[1](✉), Zhiping Cai[2], and Shuhui Chen[2]

[1] Innovation Center and Mobile Internet Development and Research Center,
China Academy of Electronics and Information Technology, Beijing 100041, China
ssslpy@163.com, yli@csdslab.net

[2] College of Computer, National University of Defense Technology,
Changsha 410073, Hunan, China
zpcai@nudt.edu.cn

Abstract. Sensitive word detection has popped out nowadays as the prosperity of internet technologies emerges. At the same time, some internet users diffuse sensitive contents which contains unhealthy information. But how to improve sensitive information classification accuracy and find new sensitive words has been an urgent demand in the network information security. On the one hand, the sensitive information classification result inaccurate, on the other hand, all the research methods can not find the new sensitive information, in other word, it does not automatically identify new sensitive information. We mainly improved the existing outstanding machine learning classification algorithm, experimental results show that this method can significantly improve the classification accuracy. Beside, by researching word similarity algorithm base on *HowNet* and *CiLin*, we can realize expanding the database of sensitive words continually (i.e., discovery the new sensitive word). Through the methodologies mentioned above, we have got a better accuracy and realized new sensitive word discovery technology which will be analyzed and presented in the paper.

Keywords: Sensitive words · Sensitive information classification · Natural language processing · New word discovery

1 Introduction

With the advent of the Internet era, massive network information resources make it more and more convenient for people to obtain information, life communication, shopping financial management and so on. But at the same time when people get convenience [1], all kinds of pornography, violence, reaction, superstition and other illegal information also follow one another [2], it brought great harm to the people especially the youth, but it also brought great threats to the society [3]. In this regard, researchers engaged in information security have done a lot

This work was supported by CETC Joint Fund under Grant 6141B08020101.

X. Sun et al. (Eds.): ICAIS 2019, LNCS 11632, pp. 338–346, 2019.
https://doi.org/10.1007/978-3-030-24274-9_30

of research and put forward a variety of sensitive information detection technology. Because the Internet features of resource sharing [4], real-time interactive, personalized and virtualization, it shortens the distance between people and promote the social contacting [5] development such as BBS, etc. However, network virtualization leading to people don't have to care about the composition of the conversation object [6]. Internet users can follow one's inclinations to express [7] their views, this phenomenon has led to uneven quality of online speech and even statements that endanger the social thought. Therefore, sensitive word detection technology is critical for the purification of network environment [8], especially for the national security of politic and Ideology.

Definition 1. *Sensitive Information: A word or phrase that has a sensitive political orientation (or anti-ruling party orientation), a violent orientation, an unhealthy color, or an uncivilized language. But some websites according to their own actual situation, set some only applicable to the site's special sensitive words. Sensitive word setting function is widely used in tieba or BBS [9].*

Definition 2. *Sensitive Information: In relevant laws and regulations, network sensitive information is defined in detail. It refers to the information distributed through the Internet which is not in compliance with the law and violates social ethics and morality and has a negative impact on society [10].*

After summarizing, we give our own definition of sensitive information according the need of function to achieved.

Definition 3. *Sensitive Information: In relevant network management regulations, sensitive information is a word or phrase that contain terror content, porn information, antisocialism or anticommunist, which distribute through the internet and has negative impact on society and country.*

Sensitive information is usually words or phrases with sensitive political tendencies, anti-social tendencies, violent tendencies, non-verbal or unhealthy tendencies. Relevant studies show that the deformation of some sensitive words cannot be handled correctly in the case of traditional algorithms [11], and the efficiency of simple text search and replacement is relatively low. In order to avoid filtering sensitive information, publishers usually take some measures. Such as the sensitive word "uniform seduction", it can replace sensitive words directly with Chinese pinyin and so on [12].

Extraction of sensitive information features is the key to study sensitive information filtering system [13]. Text feature extraction methods have been studied in depth, however, if it is directly applied to sensitive information feature extraction, problems such as low accuracy and poor self-adaptability will arise. According to relevant studies [14], currently sensitive information filtering technology mainly has the following key problems to be solved.

Sensitive information categorization is an important task in Natural Language Processing with many applications. Many researchers have adopted fuzzy matching method, it is useful in sensitive information identification from raw data. Recently, models based on machine learning have become increasingly popular. Up to now, BP neural network algorithm and belief-degree network are used to detect sensitive information. In this paper, we explore ways to scale these baselines to large corpus, in the context of text classification. Inspired by the recent work in feature selection, we show that classification models can train on a large words within five seconds, while achieving performance with the state-of-the-art.

2 Related Work

Because of the potential application value of sensitive information filtering, many scholars have carried out related research. Edel Greevy and others [15] using general text classification method researched on detecting and discovering related information about propaganda of racism on the Internet. Literature research studied the detection and tracking technology about Chinese hot topic in blog field and put forward the similarity calculation method of combing part of speech with word frequency. Tsou and others [16] carried on a classified research on the basis of the study of Chinese Beijing, Hong Kong, Shanghai, Taipei newspapers appraisal on four political figures (Kerrey, Bush, Junichiro Koizumi, Chen Shuibian). In literature [17], polar elements in the text by tagging corpus are first obtained, and then the measurements of the distribution of polar elements, the density of polar element and the intensity of semantic of polar elements are taken for each text, getting the results of texts according to nature of appraisal and degree of strength. Text classification model is highly dependent on the text feature extraction and similarity calculation. As there is no authoritative description of sensitive information features, common practice is to put sensitive words as sensitive information characteristics. In literature [18], it proposes a fusion of different algorithms and attempts to apply more natural language processing techniques to research. The filtering technology based on text content understanding can distinguish the actual meaning of the document dynamically and can achieve better filtering performance, but this is the bottleneck of Chinese word segmentation nowadays.

3 Methodology

3.1 Data Set

In the process of designing the model, we find that there are few text data sets marked as sensitive information in the public data sets currently. Because we employ supervised machine learning algorithm to extract features, if there are too few training sets with target labels, it will lead to inaccurate feature extraction, hence the data set is not meet our experiment requirements. In order to extract sensitive information features accurately, we decided to design a crawler system to crawl data sets which contains sensitive news website. Python, as a simple programming language, has many advantages for the development of crawler. When parsing the HTML source, the beautiful soup library is provided with minimal code to filter the HTML tags and extract the text. Pandas are used for data collection and storage. All the data used in the experiment from the web sites [19]. The total number of sensitive news is 14080. In the stage of data preprocessing, we divide all the sensitive data further. We screened out the news with the political orientation label from crawled sensitive news. Table 1 is description of the total number of crawled news, which is obtained from the Web pages.

Table 1. The total number of sensitive news.

Classification of scientific news	Number of sensitive news
Anti-communist news	1208
Corruption news	4560
Spy news	3750
Malignant events news	250
Civil rights news	2002

Collecting data by crawl technology has been done, we get the data that satisfy the corpus requirements. In order to better utilize this raw sensitive data, we have designed Algorithm 1 to deal with this raw data. The algorithm manifest the process of data screening data tagging data statistics and natural language processing.

3.2 Experiment

Sensitive News Feature Extraction and Model Training. Attention should be paid when extracting feature words in the raw data. Usually, the high-frequency words have higher weight. But the word frequency is not the single factor that affects the weight. In Chinese, the occurrence rates of words are very different. The utilization rates of some words are not high, but they have certain word frequency in a certain text, they are likely to represent the event described by the subject of the text [20]. In the contrary, some high-frequency words have high frequency in all texts; they may not be able to represent the event described by the subject of the text. So we use the TF/IDF algorithm to calculate the weight of words in the text.

$$W_{j,d} = \frac{TF_{j,d} \times \ln\left(N/DF_{j,d}\right)}{\sum_{i=1}^{m}\left[TF_{j,d} \times \ln\left(N/DF_{j,d}\right)\right]^2} \tag{1}$$

Among them, $W_{j,d}$ expresses the weight of feature word j in sentence d, $DF_{j,d}$ expresses the number of sentences in which feature word j appears, N is the total sentences, and $TF_{j,d}$ expresses the occurrences of j in d. Text Categorization, whose core is to build a function from a single text to category, is an important technology of data processing [21], divided into supervised learning, unsupervised learning, semi-supervised learning, enhance learning and deep learning [22].

Algorithm 1. Preprocessing of raw sensitive data for training data

```
$sensitivelabel = set()$
$length = Constant$
$Equalcorpus = \varnothing$
for each $new \in corpus$ do
    if $new \in$ corpus then
        if Length $t \geq 1$ then
            $new \Leftarrow [: Constant]$
            $sensitivelabel \Leftarrow new[lable]$
            $Equalcorpus \Leftarrow new$
        else
            $continue$
        end if
    else
        $continue$
    end if
end for
$T \Leftarrow T \cup [t]$
return $T$
for each $item \in Equalcorpus$ do
    $result = item.strip()$
    $result = segment$
end for
return $result$
```

Algorithm 2. Sensitive News Feature Extraction

```
$MAX\_SEQUENCE\_LENGTH = Constant1$
$EMBEDDING\_DIM = Constant2$
$TEST\_SPLIT = Constant3$
$TRAIN\_TEXTS = OPEN\_FILE1()$
$TEST\_TEXTS = OPEN\_FILE2()$
$ALL\_TEXTS = TRAIN\_TEST + TEST\_TEST$
$CountVector = Initialize(ALL\_TEXTS)$
$TFIDFTRANSFORMER = Initialize()$
$COUNT\_TRAIN = CountVector.fit\_transform(TRAIN\_TEXTS)$
$COUNT\_TEST = CountVector.fit\_transform(TEST\_TEXTS)$
$TRAIN\_DATA = TFIDFTRANSFORMER.fit(COUNT\_TRAIN).transform(COUNT\_TRAIN)$
$TEST\_DATA = TFIDFTRANSFORMER.fit(COUNT\_TEST).transform(COUNT\_TEST)$
$X\_TRAIN = TRAIN\_DATA$
$Y\_TRAIN = TRAIN\_LABELS$
$X\_TEST = TEST\_DATA$
$Y\_TEST = TEST\_LABELS$
```

This algorithm for title classification of scientific news is an algorithm of Chinese text categorization and bases on title of scientific news. we respectively designed algorithm to extract sensitive news feature (Algorithm 2)and to train classification model and prediction (Algorithm 3)

Algorithm 3. Classification Model Training and Prediction

```
$TRAIN\_ALGORITHMS \Leftarrow ML\_NN\_ALGORITHMS\_SETS$
$x\_TRAIN = TRAIN\_DATA$
$y\_TRAIN = TRAIN\_LABELS$
$x\_TEST = TEST\_DATA$
$y\_TEST = TEST\_LABELS$
for each $ALGORITHM \in TRAIN\_ALGORITHMS$ do
    $MODEL\_ALGORITHM = ALGORITHM.initialize()$
    $MODEL\_ALGORITHM.FIT(x\_TRAIN, y\_TRAIN)$
    $MODEL\_SAVE = JOBLIB.DUMP(MODEL\_ALGORITHM, FILE\_PATH)$
    return $MODEL\_SAVE$
end for
$MAX\_SEQUENCE\_LENGTH = CONSTANT$
$ISOMETRIC\_LIST = \varnothing$
$TEST\_LIST = \varnothing$
$TEST\_CONTENT \Leftarrow file\_open(test\_corpus)$
$ISOMETRIC \Leftarrow TEST\_CONTENT[: MAX\_SEQUNECE\_LENGTH]$
for each $new \in ISOMETRIC$ do
    $new = new.strip()$
    $OUTSTR = news.segment()$
    return $OUTSTR$
end for
$MODEL\_SETS = \varnothing$
for each $model \in read().MODEL\_SAVE$ do
    $MODEL\_SETS \leftarrow model$
end for
for each $MODEL \in MODEL\_SETS$ do
    $PREDICT\_MODEL \Leftarrow MODEL$
    $PREDICT\_MODEL.TRANSFORM(OUTSTR)$
    $TFIDFTRANSFORMER \Leftarrow INITIALIZE()$
    $TEST\_DATA = TFIDFTRANSFORMER.fit\_transform()$
    $PRES = PREDICT\_MODEL(TEST\_DATA)$
    $NUM, SUM = ZERO$
    $PREDS = PRES.TOLIST()$
end for
for each $I \in PREDS$ do
    if Value $I \geq 1$ then
        $SUM = SUM + 1$
    else
        $continue$
    end if
end for
```

New Sensitive Word Discovery. We get sensitive news after the Chinese word segmentation. The first step is to filter out all the words that contains one word in the segmented sensitive news, because it is difficult for single words to become new sensitive words. Secondly, we designed an algorithm to combine each word in sensitive news with every word in the database of sensitive words, in order to reduce the number of total words, this paper further filter out common stop words. The second step is to combine each word from sensitive words library and each word from segmented news. For example, suppose the number of sensitive lexicon words is m and the number of pre-processed sensitive news words is n, then the total number after pairing is $m * n$. We combine the $HowNet$ and the $CiLin$ to calculate the similarity of each pair of word. However, due to the difference in the structure and nature of $HowNet$ and $CiLin$, it's inaccurate if calculate the similarity of each pair word simply. In order to achieve our goals better. We design two weights between the two words. The process are described in Algorithm 4 as follow.

Algorithm 4. Discovering new sensitive word

```
$sensitivewords \Leftarrow sensitvedictionary$
$HowNet = HowNetdict$
$CiLin = CiLindict$
$weight1 = Constant1$
$weight2 = Constant2$
$wordgroup = \varnothing$
for each $text \in corpus$ do
    if $new \in$ corpus then
        if Length $t \geq 2$ then
            for each $word \in sensitivewords$ do
                $wordgroup \Leftarrow pair(new.word, word)$
                $similarityvalue1 \Leftarrow HowNet(wordgroup)$
                $similarityvalue2 \Leftarrow CiLin(wordgroup)$
                $pip \Leftarrow similarity1 \times weight1 + similarity \times weight2$
            end for
        else
            $continue$
        end if
    else
        $continue$
    end if
end for
if value $pip \geq threshold$ then
    return $wordgroup$
end if
```

4 Result and Discussion

The analysis of the experiment results in Table 2 shows that other methods such as Naive Bayes are not very effective in classification of sensitive information. The classification effect of decision tree is better than that of naive bayes. The classification of sensitive news is different from the classification of general news. Sensitive news classification does not have a large amount of corpus for classifier learning and training. Therefore, naive bayes algorithm has no ideal decision tree for sensitive news classification. The establishment of decision tree is a top-down induction process. The decision tree is generated according to the continuous division of each feature. The similarity calculation method based on *HowNet* and the similarity calculation method based on *CiLin* can complement each other in natural language processing applications. Because of the different application goals of *HowNet* and *CiLin*, there is a big difference in the entries. Please refer to literature for details. Despite the continuous optimization and expansion of the *HowNet* and *CiLin*, there are still great differences between them in many aspects due to the differences in structure and properties. Therefore, we combine the *HowNet* and *CiLin*, it can effectively improve the accuracy of new sensitive words discovery. The pearson correlation coefficient calculated without any preprocessing is 0.825. Compared to other word similarity calculations, the new sensitive discovery algorithm calculations are relatively good.

Table 2. Contrast effect of different classification.

Classifier	Precision	Recall	F-measure
Decision tree	0.98	0.96	0.97
Naive Bayes	0.88	0.86	0.85
SVM	0.75	0.74	0.74
KNN	0.60	0.56	0.53
LSTM	0.72	0.70	0.73

5 Conclusion

Sensitive information feature extraction and classification is the foundation and key of the sensitive information filtering model. It is varies according to dataset. This article improve the accurate of classification sensitive information base on unique supervised categorization machine learning and deep learning technology. It also proposes a method to detect sensitive information based on web information extraction. This paper mainly utilize the existing outstanding decision tree classification algorithm. By experimental results analysis, this method can obviously increase the sensitive news classification accuracy. In addition, we apply word similarity calculation algorithm combine *HowNet* with *CiLin*, we can expand the lexicon of sensitive words continually, in other word, through the methodologies mentioned above, this method have got a better accuracy and realized new sensitive word discovery technology.

References

1. Wu, S.: Research on Synthesis Governance of Internet Harmful Information. Beijing University of Posts and Telecommunications, Beijing (2011)
2. Greevy, E., Alan, F.S.: Classifying racist texts using a support vector machine. In: Proceedings of the 27th Annual International ACM SIGIR Conference, New York, USA, pp. 468–469 (2004)
3. Wu, O., Hu, W.: Web sensitive text filtering by combining semantics and statistics. In: Proceedings of IEEE NLP-KE 2005, pp. 215–259. IEEE Press (2005)
4. Guo, X., He, T.: Survey about research on information extraction. Comput. Sci. **42**, 14–17 (2015)

5. Xia, Y., Wong, K.F., Li, W.: A phonetic-based approach to Chinese chat text normalization. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, pp. 993–1000 (2006)
6. Shen, B., Zhao, Y.-S.: Optimization and application of OPTICS algorithm on text clustering. J. Convergence Inf. Technol. (JCIT) **8**(11), 375–383 (2013)
7. Wang, S., Wang, L.: An implementation of FP-growth algorithm based on high level data structures of Weka-JUNG framework. J. Convergence Inf. Technol. (JCIT) **5**(9), 287–294 (2010)
8. Che, W., Li, Z., Liu, T.: A Chinese language technology platform. In: COLING: Demonstration Volume, Beijing, China, pp. 13–16 (2010)
9. Pan, L., Zhu, Q.: An identification method of news scientific intelligence based on TF-IDF. In: International Symposium on Distributed Computing and Applications for Business Engineering and Science, pp. 501–504 (2015)
10. Epochtimes homepage. http://www.epochtimes.com/gb/ncid277.htm. Accessed 4 Oct 2018
11. He, W., Guozhong, W., Liliang, L.: Fast automatic elimination of vertical parallax of multiview images. In: IEEE 10th International Conference on Signal Processing Proceedings, pp. 1004–1007 (2010)
12. Metwally, A., Agrawal, D., El Abbadi, A.: Efficient computation of frequent and top-k elements in data streams. In: Eiter, T., Libkin, L. (eds.) ICDT 2005. LNCS, vol. 3363, pp. 398–412. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30570-5_27
13. Lim, E.-P., Nguyue, V.-A., Jindal, N., et al.: Detecting product review spammers using rating behaviors. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Managment (CIKM 2010) (2010)
14. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. **34**(1), 147 (2002)
15. Liu, T.Y., Yang, Y., Wan, H., Zeng, H.J., Chen, Z., Ma, W.Y.: Support vector machines classification with a very large-scale taxonomy. SIGKDD Explor. Newsl. **7**(1), 36–43 (2005)
16. Brank, J., Grobelnik, M.: Training text classifiers with SVM on very few positive examples. Technical report, MSR-TR-2003-34, Redmond: Microsoft Research (2003)
17. Tang, Y., Lian, H., Zhao, Z., Yan, X.: A proxy re-encryption with keyword search scheme in cloud computing. CMC: Comput. Mater. Continua **56**(2), 339–352 (2018)
18. Cui, J., Zhang, Y., Cai, Z., Liu, A., Li, Y.: Securing display path for security-sensitive applications on mobile devices. CMC: Comput. Mater. Continua **55**(1), 017–035 (2018)
19. Liu, W.Y., Song, N.: A fuzzy approach to classification of text documents. J. Comput. Sci. Technol. **18**(5), 640–647 (2003)
20. Ng, V., Dasgupta, S., Arifin, S.M.N.: Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In: Proceeding of the COLING/ACL Poster Sessions (2006)
21. Blitzer, J., Dredze, M., Pereira, F., et al.: Boom-boxes and blenders: domain adaptation for sentiment classification. In: Proceedings of the Association for Computation Linguistic (ACL) (2007)
22. Asur, S., Huberman, B.A., Szab, G., Wang, C.: Trends in social media: persistence and decay. In: Adamic, L.A., Baeza-Yates, R.A., Counts, S. (eds.) Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, pp. 434–437. The AAAI Press, Menlo Park (2011)