International Core Journal of Engineering Volume 7 Issue 5, 2021
ISSN: 2414-1895 DOI: 10.6919/ICJE.202105_7(5).0008

Sensitive Text Classification and Detection Method Based on Sentiment Analysis

Yanyan Xu[1,2,*], Yuxiang Li[1,2,a] and Zhiyong Zhang[1,2,b]

[1]School of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China;

[2]Henan International Joint Laboratory of Cyberspace Security Applications, Henan University of Science and Technology, Luoyang 471023 China.

*xyy_smile1012@163.com, [a]liyuxiang@haust.edu.cn, [b]xidianzzy@126.com

Abstract

Nowadays, the popularity and application of social networks are becoming more and more extensive. Sensitive information related to pornography, politics, and terrorism is flooding the Internet. Traditional text-based sensitive information detection methods have inaccurate classification and coarse-grained detection results. For the problem of low accuracy of sentiment analysis detection, this paper proposes a text-sensitive classification detection method based on sentiment analysis. This method first uses Fast Text as a text-sensitive classification model, and improves the accuracy of text-sensitive information classification and detection by introducing text emotional polarity. Experimental results show that the model in this paper is superior to traditional sensitive information in terms of accuracy, recall, and precision. Check the model.

Keywords

Sensitive Information; Emotion Analysis; Classification Detection; Emotional Polarity.

1. Introduction

There are a large number of netizens in China, and the Internet has become the preferred platform for Chinese residents to retrieve information, share knowledge, and obtain services. Online social networks have become a new area for people to socialize with their large number of users, instant information dissemination, and open and shared information resources. The information of social networks presents a trend of diversification, complexity, and quantification. Various sensitive information is flooded in social networks, which seriously affects network safety and health [1-3]. Therefore, how to detect sensitive information efficiently and accurately is an urgent problem to be solved in the current Internet construction.

At present, the research on sensitive text recognition is relatively mature, generally based on the sensitive vocabulary for recognition. The basic idea is to segment and retrieve the text to be detected, and use the keyword matching method to determine that the text contains sensitive words and determine the text as sensitive text, This method is very simple, but the accuracy rate is very low. Sensitive information detection methods based on emotion are still in their infancy and their development is not yet mature. In addition, the fine classification of text-sensitive categories is also an important way to improve the accuracy of sensitive information detection. For example, the literature [4,5] designed corresponding algorithms for sensitive word variants to improve the accuracy of sensitive word review. Literature [6,7] proposed a detection method for hate expression on Twitter, which is used to filter hate speech on the Internet. Literature [8] discussed a simple and effective text classification benchmark. The fast text classifier FastText is usually equivalent to the deep learning

ISSN: 2414-1895 DOI: 10.6919/ICJE.202105_7(5).0008

classifier in terms of accuracy, and the training and evaluation speed has been improved by many orders of magnitude. Literature [9] proposed a network sensitive information classification model, which is based on CNN (Convolutional Neural Network) and latest pre-trained BERT (Bidirectional Encoder Representation from Transformers), is called the BERT-CNN deep learning model. Literature [10] proposed a method to filter bad information based on text analysis and color. Literature [11] uses deep learning methods to detect and recognize sensitive text in pictures. Literature [12] proposed a Twitter spam detection technology based on deep learning. Literature [13] proposed an emotion mining method, which combines the close connection between opinion mining and emotion mining. Literature [14,15] proposed an emotional analysis method based on deep learning to accurately determine the emotional tendency of the text. Literature [16,17] builds a text sentiment analysis model based on machine learning. In terms of sensitive information detection, different emotional characteristics may also have different effects on the judgment of sensitive information. By exploring the internal connections and complementary effects of multi-modal features, the accuracy of sensitive information detection can be enhanced. Through existing analysis, it is found that many results have been achieved in the detection of sensitive information, but the following problems still exist: a) Although the importance of emotional factors to the performance of sensitive information is considered, the effect of emotional polarity and emotional intensity is ignored. The impact of sensitive information judgments. b) Ignoring the problem of text-sensitive classification, the essence of the violation quality inspection is that the classification is different. If only simple two classifications of sensitive and normal are performed, the accuracy rate will be lower.

In response to the above problems, this paper proposes a text-sensitive classification detection method based on sentiment analysis. This method takes into account the context of sensitive words and uses deep learning to classify sensitive text. The emotional features of the text directly reflect the views and emotional tendencies of the text creator, and play an important auxiliary role in the detection of sensitive information. This article will reflect the emotional features of the text from two aspects of text emotional polarity and emotional strength, and use emotion as sensitive information Auxiliary information for detection.

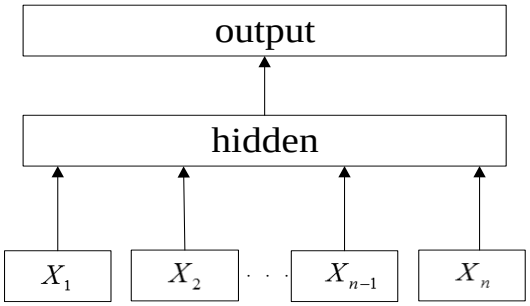

Figure 1. FastText model framework

2. Text-sensitive classification detection model

Text classification is an important task in natural language processing, and linear classifiers are generally considered to be a powerful baseline for text classification problems (Joachims, 1998; McCallum and Nigam, 1998; Fan et al., 2008). Although they are simple, they usually get the most advanced performance if the correct features are used (Wang and Manning, 2012), and they may be extended to very large corpora. This article uses FastText to detect the sensitivity of the text. The text is divided into four categories, including pornographic, political, terrorist and other categories (there may also be marketing, vulgar, fraud, etc., this article only covers the above three types of sensitive information). FastText is a simple and efficient text classification model that integrates word2vec and text classification. The FastText method consists of three parts: model architecture, hierarchical Softmax and N-gram features. The FastText model inputs a word sequence (a piece of text or a

ISSN: 2414-1895 DOI: 10.6919/ICJE.202105_7(5).0008

sentence), and outputs the probability that the word sequence belongs to different categories. The words and phrases in the sequence form a feature vector. The feature vector is mapped to the middle layer through linear transformation, and the middle layer is then mapped to the label. FastText uses a nonlinear activation function when predicting labels, but does not use a nonlinear activation function in the middle layer. The FastText model architecture is shown in Figure 1.

Fasttext model is used to transform each word in the text into vector form and calculates the cosine similarity between two word vectors, as shown in Formula (1).

$$similarity = \cos(\theta) = \frac{x \cdot y}{\|x\|\|y\|} = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \times \sqrt{\sum_{i=1}^{n}(y_i)^2}} \quad (1)$$

The formula suggests that the larger cosine value indicates more similar representation of the two words in higher dimensional space. This finding indicates that the two words are similar in semantics or close in usage. Therefore, the Fasttext model shows excellent performance in word vector representation learning.

3. Text Sensitivity Judgment Method Based on Fine-grained Emotion

Directly use the FastText model to detect and classify text sensitivity, and there will be certain errors. For example, a tweet contains a certain amount of terrorism-related information, but the text context reflects opposition and condemnation of the sensitive information. The text is defined as a sensitive type, which must be biased. It can be seen that the author's subjective emotion in the text has a certain decisive effect on the sensitivity of the text. Therefore, this article introduces the emotional polarity to determine the overall sensitivity of the text.

3.1 Fine-grained sentiment analysis

The sentiment polarity of the text is taken into consideration, and fine-grained sentiment analysis is performed according to the sentiment polarity and intensity of the sentiment words in the text. A method for identifying sensitive information based on co-occurrence analysis of sentiment words and sensitive words is proposed. This article uses Dalian University of Technology's emotional vocabulary ontology database to match emotional words in the text. Each emotional word in the lexicon is divided into three types: positive (1), negative (-1), and neutral (0) Emotional polarity. Polarity intensity is a more advanced processing of sentiment analysis. It consists of two parts: sentiment polarity and sentiment intensity. This article sets the value range of sentiment intensity to, the positive and negative values represent the emotional polarity, and the negative value represents the negative direction. Emotion, positive value represents positive emotion, 0 represents neutral attitude, and size represents the intensity of emotion. This article uses Chinese word segmentation to divide the text into multiple words. And through matching with the existing emotional word database and sensitive word database, the emotional word set and the sensitive word set are obtained, and the Cartesian product operation is performed on the two sets. According to whether the elements in the Cartesian product co-occur, the word frequency and individual text emotion polarity can be calculated. The sensitive information recognition method based on the co-occurrence analysis of emotional words and sensitive words. Co-occurrence refers to the co-occurrence of emotional words and sensitive words. According to the principle of closest distance, the smallest distance between the emotional word and the sensitive word in the sentence is taken as the co-occurrence of the two. The formula for calculating the distance between the two is as follows:

$$\mathrm{dis}(w_i, w_j) = \left|index(w_i) - index(w_j)\right| \quad (2)$$

In the above formula, $\mathrm{dis}(w_i, w_j)$ represents the distance between words w_i and w_j, $index(w_i)$ and $index(w_j)$ represent their position subscripts in the phrase after the word segmentation, and the first word subscript is 1, increasing in sequence.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202105_7(5).0008

3.2 Sensitivity Judgment Based on Emotion

This article will reflect the emotional characteristics of the text from the two aspects of the emotional polarity and emotional intensity of the text, and use the emotional characteristics of the text and sensitive information to determine the sensitivity of the text. The calculation method formula is as follows:

$$PositiveSensitiveCount = \sum_{i=1}^{n} Occur\left(Sensitive_{w_i}, \lambda Positive_{w_j}\right), \lambda \in [1,3] \quad (3)$$

$$NegativeSensitiveCount = \sum_{i=1}^{n} Occur\left(Sensitive_{w_i}, \beta Negative_{w_j}\right), \beta \in [-3,-1] \quad (4)$$

$$AllSensitiveCount = PositiveSensitiveCount - NegativeSensitiveCount \quad (5)$$

Where, $PositiveSensitiveCount$ $PositiveSensitiveCount$ represents the positive emotion score of sensitive information, $NegativeSensitiveCount$ represents the negative sentiment score of sensitive information, $AllSensitiveCount$ is the overall emotion score of sensitive information. $Occur\left(Sensitive_{w_i}, Positive_{w_j}\right)$ is the number of co-occurrences of sensitive words w_i and positive emotion words w_j , and $Occur\left(Sensitive_{w_i}, Negative_{w_j}\right)$ is the number of co-occurrences of sensitive words w_i and negative emotion words w_j And n indicates the total number of words after word segmentation. λ is the intensity of positive emotions. β is the intensity of negative emotion.

This article divides emotional polarity into three categories. Based on experience and the results of most researchers, most sensitive words contain negative parts of speech themselves, and the combination of positive emotions indicates that they support or support sensitive words. Acquiescence, therefore, it is easier to draw conclusions about sensitive information in texts of sensitive information that contain positive emotions.

Combining the above calculations, if the overall sentiment score $AllSensitiveCount > 0$, the text is directly judged as the original sensitive category. When $AllSensitiveCount \leq 0$, a second judgment is required to calculate the word frequency of the sensitive words. When the word frequency is greater than the set threshold, it is also directly judged as the original sensitive category. On the contrary, for other categories.

Algorithm 1: Text Sensitivity Judgment Method Based on Fine-grained Emotion

Input : Text

Output : Sensitive Type

1) $P_t \leftarrow FastText(text)$; //obtain the sensitive classification probability distribution of the text
2) $SensitiveVocabularySet \leftarrow jieba(Text)$;
3) $EmotionVocabularySet \leftarrow jieba(Text)$;
4) $SensitiveVocabularySet \times EmotionVocabularySet = \left\{\left(Sensitive_{w_i}, Positive_{w_j}\right)\right\}$// Cartesian product
5) $PositiveSensitiveCount \leftarrow \sum_{i=1}^{n} Occur(Sensitive_{w_i}, \lambda Positive_{w_j})$;
6) $NegativeSensitiveCount = \sum_{i=1}^{n} Occur\left(Sensitive_{w_i}, \beta Negative_{w_j}\right)$;
7) $AllSensitiveCount \leftarrow PositiveSensitiveCount - NegativeSensitiveCount$;
8) if $(AllSensitiveCount > 0)$
9) $P_t \leftarrow P_t$;
10) else if $(SensitivityVocabularySet.Count > \gamma)$
11) $P_t \leftarrow P_t$;
12) Return P_t

ISSN: 2414-1895 DOI: 10.6919/ICJE.202105_7(5).0008

4. Experiment and analysis

In order to evaluate and verify the effectiveness of the proposed method, this paper uses a variety of methods to conduct comparative experiments. Next, the specific experimental process will be introduced from four aspects: experimental environment, experimental data set, experimental settings, experimental results and analysis, and visualization.

4.1 Experimental environment

Experimental environment: The development environment used in the experiment is Pycharm, Anaconda3.5, and the development language is python3.5. Libraries such as Tensorflow, Keras, Numpy, Flask, PIL, and Matplotlib were used in the experiment. The final experimental results are displayed using Matplotlib.

Hardware environment: The computer processing model used in the experiment is 3.40GHz Intel Core i5, with 8GB of memory.

4.2 Experimental data set and evaluation criteria

In order to verify the effectiveness of the model, the data sets used in this article are different types of text crawled from the Internet using crawlers. Since sensitive texts are all non-public data sets, the crawled text data sets are based on related sensitive words. Manually processed and marked. The processed text contains four categories: pornographic, political, terrorist and other categories. The final social network text data set is shown in Table 1.

Table 1. Experimental dataset

Sensitive Classes	Text number
Pornography	3643
Politics	2140
Terrorism	2144
Others	2040
Total	9967

4.3 Experimental setting

The experiment uses fastText to train the text word vector model. In the text sensitive information detection model, it includes fast identification of sensitive words, fastText semantic analysis and text classification. The project contains a sensitive vocabulary of about 6w, which can identify common sensitivities such as politics, violence, and pornography. Words, combined with the training results of the fastText model and the text classification library, finally divide the text into four categories, namely: pornographic, political, terrorist, and other categories.

4.4 Experimental results and analysis

In order to verify the effectiveness of the text-sensitive classification model proposed in this paper, a comparative test was carried out with other sensitive information detection methods. Using the text data set of this article, take the number of verification texts k=1000, and use other methods and methods of this article to calculate the accuracy, recall, and precision of each classification. The experimental results are shown in Table 2.

Table 2. Experimental results

Method	Accuracy (%)	Recall (%)	Precision (%)
Our method	86.1	95.8	79.2
Coarse-grained method	78.9	92.3	71.3
Keyword match method	53.3	98.7	65.8

ISSN: 2414-1895 DOI: 10.6919/ICJE.202105_7(5).0008

Table 2 shows that the accuracy of the method used in this article has reached 86.1%, which is significantly improved compared to the other two methods. It can be seen that the introduction of emotional factors has a positive auxiliary effect on the sensitivity detection of text.

At the same time, in order to verify the changes of accuracy and recall rate of the method in this paper under the condition of different verification texts, the number of texts k is respectively taken as 1000, 2000, 3000, and the experimental results are shown in Table 3.

Table 3. The effect of the number of texts on accuracy and recall

Number of texts k	Accuracy (%)	Recall (%)
1000	86.1	95.8
2000	85.9	94.3
3000	87.3	96.7

Table 3 shows that under different verification texts, the fluctuations of accuracy and recall are not large, indicating that the method in this paper has high stability.

5. Conclusions

Aiming at the traditional two-class classification of text sensitive information detection, low accuracy of coarse-grained sentiment analysis and low detection efficiency, etc. This paper proposes a text-sensitive classification detection method based on sentiment analysis. This method takes into account the context of sensitive words and uses a fastText classification method to classify sensitive text. Then combined with the influence of emotional polarity and intensity on sensitive information, the sensitivity of the content can be judged more accurately. Experiments prove that this method has higher detection accuracy. Most of the texts on social networks are short texts. In order to express the true mood of the moment, users often use expressions to express them more vividly. The emotions expressed in facial expressions also contain important emotional information, which also has a certain impact on the sensitivity of text detection. The detection of text sensitivity by incorporating multi-feature emotions is also an important research direction.

Acknowledgments

The work was supported by National Natural Science Foundation of China Grant No.61972133, Project of Leading Talents in Science and Technology Innovation for Thousands of People Plan in Henan Province Grant No.204200510021, and Program for Henan Province Key Science and Technology No.212102210383. We thank reviewers and editors for their valuable suggestions, comments and helps.

References

[1] H. Chen, Y. Wu, and D. J. Atkin, Third person effect and Internet pornography in China, Telematics Informat, vol. 32, no. 4, pp. 823–833,Nov. 2015

[2] Nardina OV. Keeping Minors Safe in Cyberspace: Extremist and Terrorist Threats[C]// International Scientific Conference "Far East Con" (ISCFEC 2020). Atlantis Press, 2020: 1640-1646

[3] Saurwein F, Spencer-Smith C. Combating disinformation on social media: Multilevel governance and distributed accountability in Europe[J]. Digital Journalism, 2020, 8(6): 820-841

[4] Fu Y, Yu Y, Wu X. A sensitive word detection method based on variants recognition[C]//2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI). IEEE, 2019: 47-52

[5] Liu W, Chen C, Wong K Y. Char-net: A character-aware neural network for distorted scene text recognition[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1)

ISSN: 2414-1895 DOI: 10.6919/ICJE.202105_7(5).0008

[6] H. Watanabe, M. Bouazizi and T. Ohtsuki, Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection, in IEEE Access, vol. 6, pp. 13825-13835, 2018, doi: 10.1109/ACCESS.2018.2806394.

[7] Davidson T, Warmsley D, Macy M, et al. Automated Hate Speech Detection and the Problem of Offensive Language[J]. 2017.

[8] Joulin A, Grave E, Bojanowski P, et al. Bag of Tricks for Efficient Text Classification[J]. 2016

[9] Wang Y, Shen X, Yang Y. The Classification of Chinese Sensitive Information Based on BERT-CNN[M]// Artificial Intelligence Algorithms and Applications. 2020.

[10] Xiong J Y, Yao L Y. An Information Filtering Algorithm Based on Text and Complexion Detecting [C]// 2008 ISECS International Colloquium on Computing, Communication, Control, and Management. IEEE, 2008, 1: 308-311

[11] Cao X, Wai T O P. Research on Text Detection in Network Advertisement Picture Based on Depth Learning[C]// 2018 International Conference on Information Systems and Computer Aided Education (ICISCAE). IEEE, 2018: 169-174.

[12] Wu T, Liu S, Zhang J, et al. Twitter spam detection based on deep learning[C]// Proceedings of the australasian computer science week multiconference. 2017: 1-8.

[13] Yadollahi A, Shahraki A G, Zaane O R. Current State of Text Sentiment Analysis from Opinion to Emotion Mining[J]. ACM Computing Surveys (CSUR), 2017.

[14] Shi S, Zhao M, Guan J, et al. Multi-Features Group Emotion Analysis Based on CNN for Weibo Events[J].

[15] Dos Santos C, Gatti M (2014) Deep convolutional neural networks for sentiment analysis of short texts. In: Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers, pp 69–78

[16] Moh M, Gajjala A, Gangireddy SCR, Moh T-S (2015) On multi-tier sentiment analysis using supervised machine learning. In: 2015 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology (WI-IAT), vol 1. IEEE, pp 341–344

[17] Rout J K, Choo K K R, Dash A K, et al. A model for sentiment and emotion analysis of unstructured social media text[J]. Electronic Commerce Research, 2018, 18(1):1-19.