Identification Method of Sensitive Information Based on CNN Model

Ruoxue Bi(✉)

Beijing Normal University - Hong Kong Baptist University United International College, Zhuhai, China
13926937913@163.com

Abstract. E-mail, as one of the important means of modern network communication, is widely used from personal life to workplace. In the era of big data, the phenomenon of e-mail data leakage is also very common. A large number of spam will not only occupy memory, but also lead to the leakage of sensitive information of users and be used by criminals, thus posing a serious threat to the security of enterprises and personal information. This topic will classify the email text, identify the sensitive information in the email and mark the sensitive email, so as to remind users to transfer or protect the sensitive email, improve the protection awareness of sensitive information for individuals and enterprises, reduce the hidden dangers and threats caused by information leakage, improve the efficiency and security of the email system and promote the harmonious development of society.

Keywords: Neural network · Text classification · Word2Vec · text-CNN

1 Introduction

E-mail is an important means of daily communication in modern society, but the security of e-mail and the protection of sensitive information are still very challenging. In recent years, the confidential information identification technology in e-mail has developed rapidly.

At present, there are three main methods to choose from: keyword matching, machine learning and neural network. Using keyword technology [1] to identify sensitive words and organize them can effectively improve the recognition accuracy and easily avoid repeating recognition. However, the effect of this method is not ideal, because even the simplest e-mail may contain keywords in the glossary of sensitive words, and the perceptive words screened manually are also subjective. Using keyword deformation can effectively avoid filtering, which will cause misjudgment of ordinary mail. However, the machine learning method based on statistical feature learning has higher accuracy, but this method tends to ignore the context and misinterpret some texts, resulting in the lack of deep mining of text content.

By using neural network model, we can improve the accuracy of classification because it is able to learn the semantics of text. Compared with the traditional keyword method and machine learning method, this method has obvious advantages. In this

D. Kumar et al. (Eds.): IEIT 2023, AHSSEH 10, pp. 1155–1162, 2023.
https://doi.org/10.2991/978-94-6463-230-9_138

paper, a model based on CNN is proposed to identify sensitive information in e-mail, and the feasibility of the method is checked by comparing with other classification models.

2 Literature Review

In recent years, sensitive information identification has become a hot topic for scholars at home and abroad. Most of the early sensitive information identification methods are based on keyword matching, but this method has some problems such as low efficiency and easy to make mistakes, so it needs more advanced technical means to realize automatic screening and avoidance. After the rapid development of deep learning [2] in recent years, many scholars began to try to use deep learning to optimize and improve the recognition method, and achieved remarkable results.

2.1 Research Status of Sensitive Information Recognition

In recent years, in order to identify and classify sensitive information in e-mail more accurately, there have been two different ways: one is entity identification technology based on naming, and the other is to identify confidential e-mail by using text features. With the continuous progress of technology, more and more scholars really start to use machine learning methods and models to achieve effective text classification.

In China, He Kai [3] introduced the attention mechanism into the classification model, and the classification model is a combination of convolutional neural network and support cross product, which achieved high accuracy of text classification. For example, Yu Hai [4] used the Text-CNN model to study the text detection with sensitive information as a special text classification; Liu Zihao, Zhuang Yi [5], etc. proposed Markov-Gram email feature selection method, which integrated the email structure and the keyword features, so as to improve the detection accuracy of sensitive information. Scholars have also tried the named entity recognition technology based on deep learning. Compared with traditional methods, this method is more efficient and has higher detection accuracy. Chirawan Ronran [6] and others proposed using one-way LSTM for training, and applied it to name entity recognition. Guillaume Lampled [7] and others have established a model based on Bi-LSTM to solve the problem of named entity recognition, which can extract the meaning of words more comprehensively. The model has achieved very good results in Dutch, German and English.

2.2 Research Status of Word Vector

In the past, the representation of words used statistical information, such as one-hot coding. Bengio [8] proposed a language model based on N-gram neural probability, and more scholars began to study word vectors. Word vector is a text vector representation method based on neural network language model. Mikolov [9] and others put forward a new word vector model. Skip-gram. Its advantage lies in that it greatly reduces the computational complexity, emphasizes grammar and semantics between words, and better expresses the relationship between word vectors. By combining word vector with LSTM

model, Jin Wei [10] and others greatly improved the accuracy of medical data text classification. Bai Heyi [11] used word vector technology to deeply mine the characteristic information in curative health big data in order to better meet the needs of domestic medical services. Hu Wanting [12] and others weighted the word vectors generated by Word2vec to better respond to the requirements of the title and text of news text classification. Xing Xin [13] and other researchers combined the Bi-LSTM technology of attention mechanism with CNN technology, successfully classified the text and effectively extracted the features. By using BERT model, Wen Chaodong [14] and others trained a dynamic word vector, which can preserve the semantic association between words in the text, thus greatly increasing the accuracy and efficiency of classification.

3 Methodology

The detection model in this paper includes input layer, word vector transformation layer, neural network layer and output layer.

3.1 Input Layer

The original text will be input as training data, but the unprocessed original text data contain some disturbing useless information, which will affect the experiment. In order to improve the accuracy of the experiment and the validity of the training data, it is necessary to delete this kind of information, including punctuation marks, semicolons, etc. It is also necessary to remove stop words and stop words, which have no practical meaning when used alone, and can only have meaning in a complete sentence after extra words are combined. Stop words include mood particles, auxiliary words and prepositions. In this study, we will determine the growth of the total amount of text by setting a threshold, and realize the de-duplication of the stopped word list by accumulating the number of occurrences of each word.

3.2 Word Vector Transformation Layer

After the text preprocessing of the input layer, the email text has become more standardized. It is also needed to further transform the email text into a word vector [15] so that the computer can successfully identify it. By using the bag-of-words model, the word vector transformation can be effectively realized. The model divides the text into multiple units and assigns corresponding feature values to them according to the number of occurrences of the units. Thus realizing the transformation. One-hot [16] model is a common word bag model, which uses a series of state registers to convert N states into a set of word vectors. Thus realizing the effective transmission of word vectors. In addition, Wor2vec model is also commonly used in the word bag model. However, one-hot model has an obvious problem: it ignores the relevance between words and assumes that two words are completely independent, so its results often show sparse characteristics, thus falling below the expected results. Word2vec model can transform words into a vector with a long length, thus describing the relationship between words more accurately, thus overcoming the shortcomings of One-hot model.

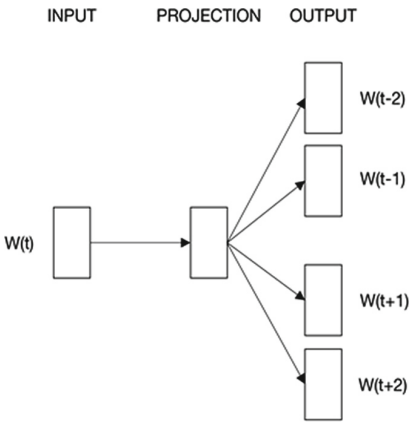

Fig. 1. Skip-gram model

In this paper, the Skip-gram algorithm in Word2Vec is proposed, which has excellent training performance and can significantly shorten the training time. In addition, the training speed of this algorithm is also extremely fast, and the training results are also excellent, Skip-gram model shown as Fig. 1. With the increase of training sample size, the accuracy of Skip-gram model will be improved.

3.3 Neural Network Layer

Convolutional neural network has excellent feedforward performance, can play an exceptional role in large-scale image processing, and can save complex image preprocessing steps, so it has been widely used in the field of image recognition. CNN model usually consists of a convolution layer and pool layer. Because of its superb local feature extraction ability, it can also be used in the field of text classification and recognition. Compared with RNN model, CNN model has higher efficiency.

Text-CNN model is selected in this paper, which includes convolution layer, pooling layer and full connection layer. Generally, when processing image data, the width and height of convolution kernel in convolution layer are the same, but for text-CNN model applied to text classification, the width of convolution kernel is the same as the dimension of word vector, and different convolution kernels can get column vectors with different characteristics. Through 1-Max-pooling, we can gather multiple feature vectors to extract effective information and convert it into a pooled form. The last layer is the fully connected layer, which is mainly responsible for mapping the useful features extracted from the previous layers into the category label space, and integrating the data through the softmax activation function to obtain the classification results. To sum up. This paper proposes a method to identify sensitive information of emails based on CNN model. This method first preprocesses emails, then uses word2vec model to convert emails into word vectors, and then uses CNN model to identify sensitive information of emails. The process of the model for detecting mail sensitive information is shown in Fig. 2.

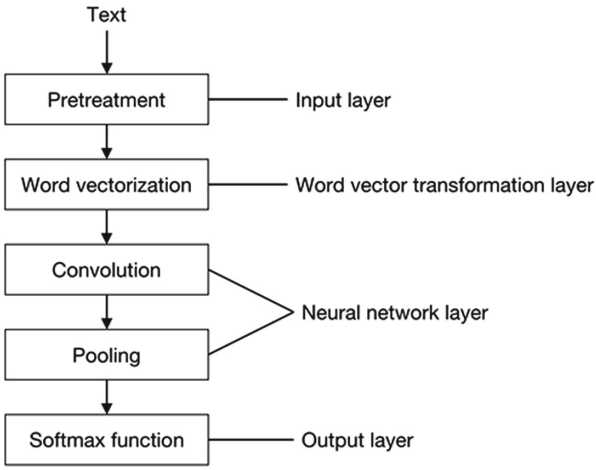

Fig. 2. Process model

4 Experiment

4.1 Data Set

The data set used in this experiment comes from many sources, including social media, social networking sites and e-mail. In order to avoid a large number of illegal characters in the data, we preprocessed all the data. Through data preprocessing, we can convert uppercase letters into lowercase, remove special symbols, ban the use of words, remove related links, and convert words into words that can express meaning. In this sensitive information detection task, I accepted a supervised learning method. We extracted text with sensitive tags from Wikileaks, and obtained 137,901 electronic messages through actual analysis. We use 90% of this information for training and 10% for verification.

4.2 Experimental Environment Setting

The experimental environment is shown in Table 1.

Table 1. Lab environment

Experimental environment	Experimental configuration
Operating system	Ubuntu22.04
Programming language	Python3.8.8
Deep learning framework	Pytroch1.1.0
Graphics card model	RTX309024GB

4.3 Experimental Results and Analysis

In check to see the feasibility of the model, this paper sets up a mail filtering model based on Naive Bayes and text-CNN. Naive Bayes is a traditional machine learning method, text-CNN directly outputs whether the email contains sensitive information. Experiments show that the method proposed in this paper has achieved better performance in identifying sensitive information of emails compared with the other two methods, which prove the practical value of this method.

Through comparative analysis, the accuracy, recall, precision and F1 value are taken as important parameters to measure the effect of this project.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{Tp} + \text{FN}} \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{F1} = \frac{2 \bullet \text{Precision} \bullet \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Among them, TP is the number of texts marked as containing sensitive information in the data set and the model detection result is also containing sensitive information; TN is the number of texts marked as insensitive information in the data set and the model detection result is also insensitive information; FP is the number of texts marked as insensitive information in the data set, but the model detection result is sensitive information; FN is the number of texts marked as sensitive information in the data set, but the model detection result is non-sensitive.

According to the comparison of the performance of the above three models on the test set, the following experimental results shown in Table 2 are obtained.

Depending on the experimental results in the above figure, this paper makes a comparative experiment to evaluate the recognition and classification effects of different models for sensitive information. The experiment shows that when using the model proposed in this paper, its accuracy is improved by about 7 percentage points compared with other benchmark models, which can prove that the text-CNN model is significantly better than the Naive Bayesian model, thanks to deep learning technology, which can capture the deeper semantic features of the text, and at the same time, it can better solve

Table 2. Lab result

Model	Accuracy	Recall	Precision	F1
Naive Bayes	0.785	0.670	0.870	0.757
Text-CNN	0.853	0.832	0.891	0.873

the problems caused by Chinese word segmentation, thus reducing the risk of misjudgment. However, keyword recognition technology alone is not enough to effectively detect the sensitive information in the text. But also needs to consider the context and user's intention and other factors. Therefore, the future research direction should be to comprehensively judge whether the text involves sensitive information by combining various information sources. Only then can we accurately detects the sensitive information in the text.

5 Conclusion

In recent years, with the development of science and technology, the identification and classification of sensitive information in e-mail has become the focus of domestic and foreign scholars' research, but in contrast, the research on the protection and identification and detection of sensitive information is still relatively backward. Although the related technology of text recognition has been relatively flawless, the recognition of email content for specific scenes still faces challenges. However, after the introduction of new models and mechanisms, the efficiency of the models can be enhanced, making the identification of sensitive mail information more efficient and accurate, such as neural network and CNN model. By introducing the word vector and text-CNN model, it organically integrates various deep learning technologies, greatly improving the accuracy and reliability of sensitive email information. By establishing a sensitive information detection model of e-mail based on the CNN model, and comparing with other traditional models, the superiority of this model is proved. The experimental results show that the model has been greatly improved compared with the traditional text classification model and email text recognition method, and the accuracy and training speed has also been optimized and improved.

In view of the challenges in the retrieval and identification of sensitive information, the future work will focus on improving the accuracy of test data, at the same time, strengthen the exploration of context semantics, and combine with other advanced technical means to continuously promote the progress of mail sensitive information identification technology, in order to grasp the content and real use of mail more clearly, and then greatly improve the retrieval accuracy of sensitive information. According to the changing environment, a more targeted and efficient mail sensitive information detection technology is developed.

References

1. Li Yang, Pan Quan, Yang Tao. Recognition of sensitive information based on short text sentiment analysis [J]. Journal of Xi'an Jiaotong University, 2016, 50(9): 80-84.
2. R D Seeja, A Suresh. Deep learning based skin lesion segmentation and classification of melanoma using support vector machine (SVM), 2019, 20(5): 1555-1561.
3. He Kai, Guan Youqing, Gong Rui. Text classification model based on deep learning and support vector machine [J]. Computer Technology and Development, 2022,32(07):22–27.
4. Yu Hai. Design and implementation of unstructured text sensitive information detection system based on convolutional neural network [D]. Beijing University of Posts and Telecommunications, 2019.

5. Liu Zihao, Zhuang Yi. An algorithm for detecting sensitive information in e-mail [C]//Communication Branch of Chinese Institute of Electronics, Beijing Institute of Electronic Technology Application. Proceedings of the 11th Annual Conference of Hunan Computer Society, the 8th National Symposium on Information Hiding and Multimedia Security. Science Press, 2009: 5.
6. Ronran C, Lee S. Effect of character and word features in bidirectional lstm-crf for new [C]//2020 IEEE International Conference on Big Data and Smart Computing (BigComp). IEEE, 2020: 613–616.
7. Lample G, Denoyer L, Ranzato A M . Facebook Exploring unsupervised machine translation [J]. Robot industry, 2017, No.17(06): 36–38. https://doi.org/10.19609/j.cnki.cn10-1324/tp.2017.06.005.
8. Bengio Y, Ducharme R, Vincent P. A neural probabilistic language model[J]. Advances in Neural Information Processing Systems, 2000, 13.
9. Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space [J]. Computer Science, 2013.
10. Jin Wei, Meng Jun, Huang Yufei, et al. Medical posture recognition method based on CNN and high-speed communication technology [J]. Microcomputer application, 2022, 38(07): 20–22+26.
11. Bai Heyi. Research on Intelligent Analysis Method of Health Big Data Based on Convolutional Neural Network [J]. Electronic Design Engineering, 2021, 29 (10): 10–14. https://doi.org/10.14022/j.issn1674-6236.2021.10.003.
12. Hu Wanting, Jia Zhen. News Text Classification Based on Weighted Word Vector and Convolutional Neural Network [J]. Computer System Application, 2020, 29 (05): 275–279. https://doi.org/10.15888/j.cnki.CsA.007391.
13. Xing Xin, Sun Guozi. ACRNN Text Classification Based on Dual-channel Word Vector [J]. Computer Application Research, 2021, 38 (04): 1033–1037. https://doi.org/10.19734/j.ISSN.1001-3695.2020.05.0127.
14. Zhang Haifeng, Ceng Cheng, Pan Lie, et al. News topic text classification method combining BERT and feature projection network [J]. Computer Application, 2022, 42(04):1116–1124.
15. Jing Dongsheng, Xue Jinsong, Feng Renjun. Spam text classification method based on deep Q network [J]. Computer and Modernization, 2020 (6): 89–94.
16. HARRIS D, HARRIS S. Digital Design and Computer Architecture [M]. Morgan Kaufmann, 2010.