Research on Automated Detection of Sensitive Information Based on BERT

Meng Ding*, Xing Wang, Changming Wu, Kaixuan Wang, Xue Yang

College of Investigation, Chinese People's Public Security University, Beijing 100038, China

*jinzimin@sina.com

Abstract. With the booming of the Internet, Web public opinion plays an increasingly important role in the stability of the network community. Therefore, the sensitive information hidden on the Internet is likely to lead to unpredictable social impact. This paper focuses on the detection of Chinese sensitive information. First, we build a corpus to train the detection model. Secondly, we apply the Bert method to the detection problem. Then, many popular NLP methods are applied to this problem to show the progress of Bert in a sensitive information detection task. Finally, we got a sensitive information detection model based on BERT with a high F1 score of 97.31.

Keywords: component, formatting, style, styling.

1. Introduction
Nowadays, Internet technology has experienced a flourishing development. The Internet is not only a communication tool, it is often used to collect and send information through various applications and websites, read and publish information from Weibo, Tieba, tweets, community platforms, and BBS by reading and posting the message. The residents' life in the network community is more convenient. In addition, with the development of self-media (or We media), the Internet has become the information publishing center. People of different ages, occupations, and countries voluntarily share and receive information on the Internet. With the rapid development of Internet technology and the gradual expansion of user groups, the interaction of Internet users on social media plays an increasingly important role in online public opinion[1-3]. However, it has attracted the attention of criminals. Illegal information containing sensitive information about terrorism, contraband, pornography, and antisocial will spread on the Internet, which will lead to unpredictable social impact.[4].

Because the Internet is a virtual, hidden, and real-time community, people from all over the world express their views, exchange ideas, and exchange ideas on the Internet every second. The amount of information on the Internet is growing rapidly all the time[5]. Therefore, it is a difficult and impossible task to identify all the information on the Internet manually. How to automatically identify sensitive messages is an urgent problem to be solved.

The detection of sensitive information is essentially a topic recognition problem in natural language processing. Generally speaking, the general method of topic recognition is to classify the isolated information, extract the keywords representing the topic, and identify the topic of the message by
keywords[6]. In the past, scholars focused on the messages written in English[7]. However, with the rapid development of China, the Chinese community has become an important part of the Internet. Therefore, this paper focuses on the detection of Chinese sensitive information. In terms of recognition methods, word2vec has been widely used in natural language processing in the past and has achieved brilliant results[8]. However, the recently proposed Bert method refreshes its record in many NLP problems. On this basis, we choose the best method to solve the problem of Chinese sensitive information recognition[9].

In this paper, our works are summarized as follows:
We built a corpus for the detection of sensitive information written in Chinese.
For the first time, we applied the BERT method on the detection of sensitive information written in Chinese.
Many popular NLP methods such as BiLSTM-CRF, IDCNN-CRF, BiLSTM-Attention-CRF, and so on were applied to this problem. By comparing these results, we proved the advancement of applying BERT on this issue.

2. Related work
As a result of the narrow application range, there are few scholars working on sensitive information recognition problems. Recently, Guxian XU et al. applied the SW-LDA model on the recognition problem of Network sensitive information written in Chinese and got a better result than other LDA based models[10]. And in 2017, M Devi Priya et al. proposed a system for analyzing prohibited English words[11]. This system contains primary sentiment analysis function, Morpheme sentiment analysis function, and prohibited word analysis function. After testing under a practical condition, they got a well-performed detection system[12]. In essence, the problem of detecting sensitive information is a kind of topic recognition problem. The popular method of topic recognition problems such as BiLSTM, BERT, and IDCNN[13], could be applied to the sensitive information detection problem.

3. Method
In the field of natural language processing, there are multiple ways that could help to solve the problem of recognizing sensitive information, such as BERT, Long Short-term Memory(LSTM), and Iterated Dilated Convolutional Neural Network(IDCNN)[14]. In this section, we will focus on the introduction and operating principle of these methods.

3.1. BERT
As a new language representation model, Bert has achieved the latest research results in 11 natural language processing tasks[15]. In BERT, there are two steps, pre-training and fine-tuning. During pre-training, the model is trained unsupervised over different tasks. Then, in fine-tuning, the model is firstly initialized with pre-trained parameters and fine-tuned supervised for a specific task. In this paper, we employed an official pre-trained model in the Keras-BERT project. The details about how BERT works is as follows.

The structure of Bert is a multilayer bidirectional transformer encoder. Bert abandoned the traditional methods, namely recurrent neural network (RNN) and convolutional neural network (CNN), and adopted an attention mechanism to solve the long-term dependence problem[16]. In the input/output representations, BERT uses WordPiece embedding position embedding and segment embedding. WordPiece embedding is used for dividing a word into a limited set of common sub-word units and getting a compromised balance between the validity and the flexibility of the character. Position embedding refers to encoding the position information of a word into a feature vector. And segment embedding stands for distinguishing two different sentences.

3.2. BiLSTM
Bi-directional Long Short-Term Memory(BiLSTM) is a combination of forward LSTM and backward LSTM, which are often used to model context in NLP tasks. LSTM(Long short-term memory) is a
type of Recurrent Neural Network (RNN) that contains a forget gate and memory gate [17]. In LSTM, effective information is selected by forgetting and remembering in the cell state. And at each time step, forgetting, memory, and output are controlled by the forget gate, memory gate, and output gate calculated from the hidden state of the last moment and the current input [18]. For details, assuming that input word at time t is xt, the output of the hidden layer is ht and the cell state is Ct, the output of the forget gate will be described as equation 1.

\[
ft = b(W_f * [ht-i, xt] + bf) \tag{1}
\]

The operating principle of the memory gate is described as equation 2 and equation 3. It refers to the output of the memory gate, and Ct refers to the temporary cell state. In the current cell state, the calculation is shown as equation 4.

\[
it = b(W_i * [ht-i, xt] + bi) \tag{2}
\]

\[
Ct = \tanh(We * [ht-i, xt] + bc) \tag{3}
\]

\[
Ct = ft * Ct-i + it * Ct \tag{4}
\]

The output gate and hidden layer work as equation 5 and equation 6. Ot represents output of the output gate. ht represents output of the hidden layer.

\[
Ot = a(W_o * [ht-i, xt] + bo) \tag{5}
\]

\[
ht = Ot * \tanh(Ct) \tag{6}
\]

Furthermore, we input the different parts of a sentence into LSTM in forward and backward order and concatenate their hidden vectors. Finally, it constitutes Bi-LSTM.

3.3. IDCNN
Iterated Dilated Convolution Neural Network (IDCNN) is the integration of four dilated CNN blocks. And in each block, there is a three-layer dilated convolution structure with a dilation width of 1, 1, and 2. That is why we called it Iterated Dilated CNN [19]. Comparing with the traditional CNN, dilated CNN has a larger receptive field without pooling and could output a wider range of information. It works well in problems that require global information or rely on long sequence information, especially in NLP problems [20].

4. Experiment

4.1. Data set
In this paper, we used the main texts and relevant comments of various reports from People's Daily Online, Sina Weibo, Anti-Terrorism Information Network, and China Anti-Drug Network as the source of data. In detail, a total of 29,285 pieces of news and related comments from December 11, 2015, to April 20, 2020, were collected through crawler technology. Then we selected the texts by whether they contain five types of sensitive words such as terrorism, cult, pornography, contraband, and anti-government. Furthermore, sensitive words are the representative words of a certain topic of sensitive information. For instance, in cult-related messages, sensitive words refer to cult slogans, organization names, names of people, etc. In terrorism-related information, sensitive words refer to weapons, racism, religious titles, etc. In contraband-related information, sensitive words represent obscure expression and cyber word of prohibited items. We classified all kinds of sensitive words contained in the text and build a sensitive word dictionary. Finally, we got a dictionary containing
4743 sensitive words which include 675 for violence and terrorism, 703 for cults, 824 for pornography, 846 for prohibited categories, and 795 for politics.

Due to the lack of relevant public corpora, we process the selected texts and establish corpus data sets with sensitive information. First, we removed URL links, irrelevant spaces, and emoticons, but retained punctuation. Then, we delete modal particles and auxiliary words from the suspension list of the Harbin Institute of technology and obtain 20747 valid corpora. These data were randomly divided into a training set and test set in a ratio of 5:1. Details are shown in Tab.1.

| characters         | tag  | characters | tag  |
|--------------------|------|------------|------|
| Therefore          | O    | In         | O    |
| ,                  | O    | the        | O    |
| we                 | O    | room       | O    |
| particularly       | O    | ,          | O    |
| order              | O    | There      | O    |
| you                | O    | are        | O    |
| to                 | O    | four       | O    |
| quit               | B-antisocial | man     | O    |
| the                | O    | building   | O    |
| Party              | E-antisocial | ice     | B-contraband |
| ,                  | O    | port       | E-contraband |

Tab.1 Details of Dataset

Furthermore, the processed data needed to be labeled before the experiment. Therefore, we tagged the sensitive words in the text data using the BIOES tagging mode referring to the dictionary of sensitive words. In detail, 'B' represents the beginning of a sensitive message, 'I' represents the middle of a sensitive message, 'E' represents the ending of a sensitive message, 's' represents a separated sensitive word, and 'O' represents other words. There are four tags including 'B', 'I', 'E', and 's' in each type of sensitive word. In total, there are 21 kinds of labels to be predicted. An example of label results is shown in Tab.2.
Results and Analysis

In this paper, based on the Bert Keras scheme, the parameters are adjusted and a detection model with a higher F1 score is obtained. The F1 score is a harmonic average of accuracy and recall. What's the meaning of precision and recall? First of all, they are a measure of the performance of the detection model. Precision refers to the proportion of the number of positive samples which was classified correctly to the number of total samples. And recall represents the proportion of the number of positive samples which was classified correctly to the total number of positive samples. Harmonic average these results, and we will get the F1 score finally. In the fine-tuning step, we used the data set mentioned above and the F1 score changed as Fig.1.

![Graph showing variation of F1 score in this model](image)

**Fig. 1** Variation of F1 score in this model

In this figure, at the beginning of fine-tuning, the F1 score increased rapidly. With the progress of training, the change of F1 score tended to be flat. Finally, we got a detection model with a high F1 score of 97.31.

To prove the advancement of BERT using in the detection of sensitive information, we selected four popular method combinations training on our data set. The results are shown in Tab.3.

| quit | B-antisocial | to | O |
|------|--------------|----|---|
| the  | O            | take | B-contraband |
| communist | E-antisocial | drugs | E-contraband |

**Tab.2** Sample of Label Results
| Models                | Precision | Recall | F1 score |
|----------------------|-----------|--------|----------|
| BERT-fine-tuning     | 97.07     | 97.57  | 97.31    |
| BiLSTM-CRF           | 95.01     | 94.23  | 94.49    |
| IDCNN-CRF            | 98.03     | 93.36  | 94.04    |
| BiLSTM-Attention-CRF | 90.33     | 90.14  | 90.10    |
| BERT-BiLSTM-CRF      | 96.69     | 97.23  | 96.90    |

Tab.3 Details of the dataset

In this table, BERT-fine-tuning represents the result of our fine-tuned model based on the project of BERT-Keras. The results show that compared with other traditional methods, the fine-tuning Bert model has better performance. Especially, when comparing the result of BiLSTM-CRF with the BERT-BiLSTM-CRF, we can see the advantages of Bert in dealing with sensitive information detection.

References
[1] Zhang R, Lee H and Radev D 2016 Dependency Sensitive Convolutional Neural Networks for Modeling Sentences and Documents Proc. of NAACL-HLT 2016 (San Diego: Association for Computational Linguistics) pp 1512-1521
[2] Vaswani A, Shazeer N and Parmar N 2017 Attention is all you need J. arXiv
[3] Wu Y, Schuster M and Chen Z 2016 Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation J. arXiv
[4] Lu Y, Zhang Y and Ji D 2016 Multi-prototype Chinese character embedding LREC (Berlin: Springer)
[5] Kalchbrenner N, Grefenstette E and Blunsom P 2014 A convolutional neural network for modelling Sentences J. Eprint Arxiv
[6] Wang R, Li Z and Cao J 2019 Convolutional Recurrent Neural Networks for Text Classification Int. Joint Conf. on Neural Networks (IJCNN) (Budapest: IEEE) pp 1-6
[7] Devlin J, M W Chang and K Lee 2018 BERT: pre-training of deep bidirectional transformers for language understanding J. arXiv
[8] H Sak, Senior A and Beaufays F 2014 Long short-term memory recurrent neural network architectures for large scale acoustic modeling J. computer science
[9] Li X and X Wu 2015 Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition J. arXiv
[10] Xu G, X Wu, H Yao, F Li and Z Yu 2019 Research on Topic Recognition of Network Sensitive Information Based on SW-LDA Model IEEE Access (Beijing) (Budapest: IEEE) 7 pp 21527-21538
[11] Niu X and Y Hou 2017 Hierarchical Attention BLSTM for Modeling Sentences and Documents Int. Conf. on Neural Information Proc vol 10635 (Switzerland: Springer Chamin) pp 167-177
[12] Xie Z 2017 Closed-Set Chinese Word Segmentation Based on Convolutional Neural Network Model China National Conf. on Chinese Computational Linguistics Int. Symp. on Natural
Language Proc. Based on Naturally Annotated Big Data vol 10565(Switzerland: Springer Cham) pp 24-36

[13] Kalchbrenner N, Grefenstette E and Blunsom P 2014 A convolutional neural network for modelling sentences J. Eprint Arxiv

[14] Rios A and Kavuluru R 2015 Convolutional neural networks for biomedical text classification: application in indexing biomedical articles J.arXiv

[15] Niu X, Hou Y and Wang P 2017 Bi-Directional LSTM with Quantum Attention Mechanism for Sentence Modeling Int. Conf. on Neural Information Proc vol 10635(Switzerland: Springer Cham) pp 178-188

[16] Kavasidis I , Palazzo S and Spampinato C 2018 A saliency-based convolutional neural network for table and chart detection in digitized documents J.arXiv

[17] Yin W, Schütze, Hinrich and Xiang B 2015 ABCNN: attention-based convolutional neural network for modeling sentence pairs J. Computer ence

[18] Ma M, Huang L and Zhou B 2015 Dependency-based Convolutional Neural Networks for Sentence Embedding Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Int. Joint Conf. on Natural Language Proc Vol 2(Beijing:Association for Computational Linguistics) pp 15-20

[19] S Ha, J M Yun and S Choi 2016 Multi-modal Convolutional Neural Networks for Activity Recognition IEEE Int. Conf. on Systems, Man, and Cybernetics(Kowloon: IEEE) pp 3017-3022

[20] Zhang Y and Wallace B 2015 A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification J. Computer ence