A Study of the Chinese spam Classification with Doc2vec and CNN

Hechen Gong, Fucheng You, Shaomei Wang
School of Information Engineering, Beijing Institute of Graphic Communication, Beijing 102600, China

gonghechen123@163.com, youfucheng@bigc.edu.cn, wangshaomei@bigc.edu.cn

Abstract. Convolution neural network is a kind of neural network, which has been proved to be very effective in image recognition and classification. In recent years, convolution neural networks have gradually shifted to the field of natural language processing and become one of the research hotspots. For the construction of word vector text using convolution neural network, only considering the relationship between word granularity level, not considering the relationship between words, nor considering the relationship between semantics, affecting the classification results. In this paper, a method based on Doc2vec and CNN is proposed to classify spam. Firstly, the spam is preprocessed, then the sentence vectors and word vectors of Chinese text are trained by Doc2vec, and finally the trained text vectors are classified by convolution neural network.

1. Introduction
With the development of the Internet and the popularity of e-mail, e-mail has become an increasingly indispensable tool. However, although e-mail has many advantages, it is also full of many hidden dangers. Spam has not yet been defined very strictly. Generally speaking, any e-mail that is forcibly sent to the user's mailbox without the user's permission is spam[1]. The existence of spam not only occupies a large amount of network bandwidth, wastes storage space, affects network transmission and operation speed, causes mail server congestion, reduces network operation efficiency, and seriously affects normal mail services[2]. There are also cheating money, disseminating pornography, reactionary and other content, which has a very bad impact. Therefore, how to filter this kind of spam quickly and effectively has become a hot research topic in recent years, but also become a very important research direction in natural language processing.

Chinese mail spam classification is more difficult than English. The Chinese text will take a variety of forms, and through special treatment, it will evade detection and filtering without changing the original meaning. such as splitting words, adding special symbols between words, Chinese characters, English, Pinyin and other forms of combination, all of which make it more difficult to classify and recognize spam.

2. Journals Reviewed
With the deepening of research in recent years, mail classification technology has also made good progress. From the earliest 1996 Cohen W. W. used rule-based Ripper method[3], but this method cannot adapt to the changes of spam, flexible. The degree is not high, and it will be replaced soon. Similarly, the rule-based method basically has this problem. Starting in 2000, Androutsopoulos et al...
slowly penetrated into the field of deep learning[4], the use of neural networks for spam classification, so that spam has entered the era of machine learning and in-depth learning. Spam categorization is actually a problem of text categorization. Deep learning methods can constantly adapt to changes in spam text representations. Every time it only needs continuous training of new models, it can always conform to the characteristics of spam. Table 1 is a summary of the research on content-based spam filtering technology by scholars at home and abroad. Some of them cite the summary of Lin Jianhong and others[5].

| Research Perspective | Filtration Technology | Literature Sources |
|----------------------|-----------------------|--------------------|
| Rule-based           | Ripper                | Cohen WW (1996) [3] |
|                      | Decision tree         | Carreras X, &Marquez L (2001) [6] |
|                      | rough set             | Yang Liu, Xiaoping Du, Ping Luo (2002) [7] |
| Machine Learning     | neural network        | Androutsopoulos I et.al (2000) [4] |
|                      | SVM+KNN               | Tretyakov K (2004) [8] |
|                      | TF+NB                 | Lu Zhou (2011) [9] |
| Deep Learning        | Improved Neural Network | Jing Zhao (2012) [10] |
|                      | NB                    | Peng Fang (2013) [11] |
|                      | TF+KNN+NB+SVM         | Yu Feng (2013) [12] |
|                      | SVM                   | Changyong Luo (2014) [13] |
|                      | SVM+NB                | FENG W (2016) [14] |
|                      | LDA+Word2vec          | Xiaozhun Kou (2017) [16] |
|                      | LDA+Word2vec+SVM      | Jianhong Lin (2017) [5] |
|                      | CNN+Word2vec          | Meirong Wang (2018) [16] |
|                      | Conv-BiGRU Model      | Yutng Li (2018) [17] |
|                      | Word embedding+GAN    | Qi Wang (2018) [18] |

Through the above research, it is found that spam classification technology has been extended to various fields. At present, the best effect is to use in-depth learning method to classify text. Based on the existing methods, this paper replaces Word2vec word vector training method, and uses Word2vec improved model Doc2vec model to train text feature vectors. In every training process, Doc2vec not only trains words, but also obtains word vectors. At the same time, as a part of the input layer of each training, the shared Paragraph vector will express the theme more accurately, which makes up for the neglect of the word order of the text.

3. Relevant Work

3.1 Text Vectorization

The difference between Chinese and English is that before training, words need to be transformed into word vectors. One-hot is a very common technique in classical methods. One-Hot coding mainly uses N-bit state register to encode N states. Each state is encoded by its own register bit, and only one state is valid at any time. One-Hot coding is the representation of classified variables as binary vectors. This first requires mapping classification values to integer values. Then, each integer value is represented as a binary vector, which is zero except for the index of the integer and is marked as 1. For example:

Sequence: ‘red’, ‘red’, ‘green’.

Integer coding: 0, 0, 1

One-Hot Coding: 1 2 3

[1, 0] [1, 0] [0, 1]

But one-hot has one drawback: any two words here are independent and context-free. So Hinto et al[19] proposed a word vector representation method, which mainly maps words distributed to low-
dimensional space, thus solving the problem of vector sparseness. The semantic relationship of text is reflected in the position of word vector in low-dimensional space. Bojanowski P et al[20] proposed that Word2vec regards vocabulary as an atomic object and pays attention to the context of vocabulary. Word2vec uses the co-occurrence degree of words to express the meaning of words through two layers of neural networks to overcome the lack of context links in short texts. However, Word2vec is only based on the dimension of words and does not have the ability of context-based semantic analysis. Therefore, Tomas Mikolov[21] is based on Word2vec model. Doc2vec is proposed.

3.2 Doc2vec
Doc2Vec was put forward on the basis of Word2Vec in 2014. This model not only overcomes the shortcomings of no semantics in the word bag model, but also improves the Word2Vec model. A sentence vector is added to Word2Vec. During the training process, a sentence can also be input into the neural network model as a whole, which not only solves the semantic analysis between the dimensions of words, but also solves the segments. Semantic analysis between falls. Doc2vev, like Word2Vec, has two algorithmic models, namely Distributed Memory Model of Paragraph Vectors (DM) and Distributed Bag of Words version of Paragraph Vector (DBOW). DM model uses paragraph vectors and word vectors to predict the probability of the next word in the context, while DBOW model only uses paragraph vectors to predict the probability distribution of multiple word vectors[21]. The algorithm used in this paper is DM. As shown in Figure.1, it is the schematic diagram of DM.

![Figure 1. DM schematic diagram.](image)

3.3 Convolution Neural Network[22]
Convolution neural network (CNN) is a feedforward neural network. Its artificial neurons can respond to a part of the surrounding units in the coverage area and perform well in large-scale image processing. The core is the convolution operation between input matrix and different convolution kernels, and the result of pooling convolution is the data feature of classification operation. Therefore, convolution neural network is mainly composed of convolution layer, pooling layer and classification layer.
3.3.1 *Input layer*

Convolution neural network is used for data input. In this paper, Doc2vec is used to convert text into vector mode. Through training model, word vector, parameter sum of soft max and paragraph vector/sentence vector are obtained from known training data. Text data is trained to represent a sentence matrix, which is vertically stitched together by word vectors of all words in the sentence, and can be expressed as:

\[ w = w_1 \bigotimes w_2 \bigotimes w_3 \bigotimes \ldots \bigotimes w_n \]  

(1)

Among them:  \( \bigotimes \) is the longitudinal stitching operator,  \( w \) is a matrix representation of sample mail.

3.3.2 *Convolution layer*

The function of convolution layer is to extract the features of sentences. 

Mainly through a  \( h \times k \) dimension convolution kernel  \( w \), the convolution operation is swept up and down on the input layer, and a feature map is obtained by convolution operation. If the height of the convolution core is  \( k \), the dimension is  \( k \), and the eigenvector after convolution is  \( c_i \).

\[ c = (c_1, c_2, \ldots, c_{n-k+1}) \]  

(2)

\[ c_i = f(w \cdot x_{i+k-1} + b) \]  

(3)

Among them:  \( w \) is the weight parameter of convolution kernel,  \( b \) is the bias value,  \( f \) is the activation function, sigmid function or tanh function are commonly used. In this paper, PReLU function is used to accelerate the convergence speed.

\[ f(x) = \max(ax, x) \]  

(4)

3.3.3 *Pooling layer*

The function of pooling layer is to further extract features, extract the most important features, prevent over-fitting, and improve computing performance. This article uses the max-over-time pooling operation.

\[ c_{\text{max}} = \max(c_i) \]  

(5)

Finally, each convolution core corresponds to a value. After the pooling layer is processed, a new feature quantity representing the sentence is obtained.
3.3.4 Fully connected layer
In this paper, the full connection layer can be defined as the classification layer. The pooling layer outputs M new feature quantities and corresponding class combinations in the form of \{(x_{(m)}, y_{(m)})\}.

The input feature \(x_{(m)}\) is the eigenvector obtained by the first two layers of processing, and the text category \(y_{(m)}\) is the text category. For a given test set text vector \(x\), it can be classified by the software Max function:

\[
f(x) = \frac{1}{1 + \exp(-\phi^T x)}
\]

Exp represents the exponential function with \(E\) as the base number and \(\phi\) as the valuation parameter. The value is estimated by the minimum cost function \(J(\phi)\). The formula is as follows:

\[
J(\phi) = \sum_{i=1}^{M} y(i) \log f_{\phi}(x^{(i)})
\]

The return value of the function is the probability value of \(C\) components, and each component corresponds to the probability of an output category, so as to divide the type information of the text and complete the classification.

4. Algorithmic flow and experimental analysis

4.1 Data preprocessing

4.1.1 To stop word processing
As shown in Figure 3, it is an original data in the mail. In the data, we can see similar figures, websites, special symbols, including modal auxiliaries and pronouns in Chinese. These words have no actual meaning in the text. In the mail, the general situation is short text, which contains fewer words. The existence of these words will affect the training effect of the model. This paper uses regular expression technology to filter meaningless characters and numbers such as time, date, website, email address, etc. Reduplicate the "Harbin University of Technology Stopwords List" and "Baidu Stopwords List" to make the mail text cleaner.

4.1.2 Participle
Before Doc2vec trains text data, first of all, We need to segment the text. because different segmentation will lead to different meanings. So a good word segmentation dictionary is very important. HanLP has the characteristics of perfect function, high performance, clear structure, up-to-date and customizable corpus. Figure 4 is a data word segmentation effect.

4.2 Training Text Vector
Here we use DM model, one of the Doc2vec models mentioned in 3.2, to train text vectors, including word sense vectors and semantic vectors. DM distributes sentences and words well in low-dimensional
space, and their position in space reflects their relationship and better reflects the characteristics of the text. In this paper, we use Gensim to implement Doc2vec model and train the text model.

4.3 algorithm flow chart
Through the above research, Chinese spam classification based on Doc2vec and convolutional neural network is mainly divided into the following parts: ① data preprocessing, ② Doc2vec text feature vector training, ③ CNN, ④ softmax classifier, ⑤ classification results and evaluation. The algorithm flow chart is shown in Figure 5.

![Algorithm flowchart](image)

4.4 Experimental result analysis
In order to verify the validity of the model, the experimental data in this paper are from open source data websites. The data set has been labeled as emotional extreme, with 5000 positive comments and 5000 negative comments, totaling 12,000. The whole corpus is divided into two parts. 1000 positive and 1000 negative texts are taken out as test sets, and the rest of the corpus is used as training sets. In order to ensure the comparability of the experimental results, the training corpus for each model is the same as the test corpus.

| Parameter Name           | Value         |
|--------------------------|---------------|
| num_filters              | 128           |
| filer_sizes              | 3, 4, 5       |
| Pooling                  | 1-max pooling |
| batch_size               | 64            |
| dropout_keep_prib        | 0.5           |
| num_epochs               | 200           |
| num_lable                | 2             |
| L2_reg_lambda            | 0.0           |
| learning_rate            | 1e-3          |

As shown in Fig. 6, it is the classification accuracy rate obtained after the text is trained by Doc2vec. In this experiment, 5 is the span value, and the specific parameters are as shown in Table 2 to verify the validity of the model.
In order to verify the effectiveness of the algorithm, we set up a comparative experiment, using the same dimensions and data. The first method is the convolution neural network classification algorithm based on Word2vec, as shown in Figure. 7. The second method is the convolution neural network classification algorithm based on Doc2vec. The classification accuracy of the algorithm is compared. It can be clearly seen from the data in Figure. 6 and Figure. 7 that the classification effect of method 2 is better than that of method 1, which verifies the effectiveness of the method.

5. Conclusion
In the era of information explosion, the popularity and convenience of e-mail has made spam more and more rampant, which may affect people's lives at any time. So detecting and filtering spam becomes especially important. Based on the predecessors, this paper proposes an improved convolution neural network algorithm, which effectively combines Doc2vec and CNN, so that there are not only semantic meanings but also semantic relations in the original features. The effectiveness of the method is verified by comparison experiments. Sex. CNN originally had excellent effects in image recognition. How to identify image garbage in spam is a problem that needs to be further studied in the next step.

Acknowledgement
This work was partially supported by Joint Funding Project of Beijing Municipal Commission of Education and Beijing Natural Science Fund Committee (KZ201710015010), Project of National Scientific Found (No.61370188), Project of Beijing Municipal College Improvement Plan (PXM2017_014223_000063) and New Project of Green Printing and Publishing Technology by Cooperative Creating Center (PXM_014223_000025) and BIGC Project (Ec201803 Ed201802 Ea201806)

References
[1] Zhimin L 2015 Research on Spam Identification and Processing Technology Beijing Institute of Technology Press
[2] He H, Xiaoyuan J, Xiwei D and Fei W 2019 The text message classification model of CNNs based on Skip-gram Computer Technology and Development pp 1-6.

[3] Cohen W W 1996 Learning rules that classify e-mail AAAI spring symposium on machine learning in information access

[4] Androutsopoulos I, Koutsias J and Chandrinos KV, et al 2000 An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages Proceedings of the 23rd annual international ACM SIGIR conference on Research and development In information retrieval

[5] Jianhong L, Jiantong Z and Jing X 2017 Research on spam filtering method based on LDA and Word2vector Network security technology and application pp73-75.

[6] Carreras X and Marquez L 2001 Boosting trees for anti-spam email filtering arXiv preprint cs/0109015

[7] Yang L, Xiaoping D and Equality L 2002 Intelligent Analysis, Filtering and Rough Set Discussion of Spam The 12th China Computer Society Network and Data Communication Conference (Wuhan)

[8] Tretyakov K 2004 Machine learning techniques in spam filtering Data Mining Problem-oriented Seminar MTAT

[9] Wei Z 2011 Research on content-based spam filtering system (Jilin: Jilin University) p 22

[10] Jing Z 2012 Research on Key Technologies of Spam Filtering Based on Content Feature Analysis (Shandong: Shandong Normal University) p 34

[11] Peng F 2013 Design and implementation of spam filtering technology based on content analysis (Chendu: University of Electronic Science and Technology of China) p 46

[12] Wei F 2013 Research on feature-based dimension reduction algorithm for spam detection (Zhejiang: Zhejiang University) p 27

[13] Changyong L 2014 Research on content-based spam detection method (Zhejiang: Zhejiang University) p 45

[14] FENG W, SUN J and ZHANG L, et al 2017 A support vector machine based naive Bayes algorithm for spam filtering 2016 Performance Computing and Communications Conference (New Jersey: IEEE Press) pp 1-8

[15] Xiaohuai Y and Hua C 2017 Design and Implementation of Spam Filtering System Based on Topic Model Telecommunications Science pp 73-82

[16] Meirong W 2018 Text Classification Algorithm Based on Convolution Neural Network Journal of Jiamusi University(NaturalScience Edition) pp 354-357.

[17] Yuting L 2018 Spam text classification method based on deep learning (Shanxi: North University) p 32

[18] Qi W, Zhongyang W, Chenrong H and Lei Pan 2018 Spam Message Classification Algorithm Based on Word Embedding and Generating Against Network Journal of Nanjing Institute of Technology(Natural Science) pp 20-27.

[19] Hinton G E 1986 Learning distributed representations of concepts Proc of the 8th Annual Conference of Cognitive Science Society p 1-12

[20] Bojanowski P, Grave E and Joulin A, et al 2016 Enriching Word Vectors with Subword Information EMNLP 2016 pp 26-27.

[21] LE Q, MIKOLOV T 2014 Distributed representations of sentences and documents proceedings of the 31st International Conference on Machine Learning pp 1188-1196.

[22] Xiaochuan Z, Linfeng Y, Ruiting S and Yihao Z 2018 Study on Short Text Classification Based on CNN and LDA Software Engineering p 17-21.