# Research on patent text classification based on Word2Vec and LSTM

Lizhong Xiao,Guangzhong Wang,Yang Zuo
College of computer science and Information Engineering
Shanghai Institute of Technology
Shanghai, China
Email: isky.imx@gmail.com

***Abstract*—In order to efficiently classify patent texts in the security field, a patent text classification model based on Word2Vec and Long-short term memory (LSTM) was established. Combined with the features of the patent text, first of all, in the text pre-processing process, words frequently appearing in patent documents such as "the invention", "involvement", and "utility model" were added to the stop word list to save storage space and improve efficiency; Secondly, the pre-trained word2vec model was introduced to solve the dimensional disaster caused by the traditional methods. Finally, by training the LSTM classification model, text features were extracted and patent text classification in the security field was performed. 50,000 patent documents were divided into the training set and the test set according to the ratio of 4:1, and the accuracy and ROC curve evaluation model were used to analyze and evaluate the classification results. The results showed that the classification accuracy rate of this method is 93.48%. At the same time, the LSTM classification model, K Nearest Neighbor (KNN) classification model, Convolutional Neural Network (CNN) classification model, and models based on CNN and Word2Vec were further compared. The experimental results showed that this method can better classify the patent texts in the security field, laying the foundation for further research and effective use of patents.**

***Keywords-security field; text categorization; Word2Vec; long and short-term memory networks; stop words***

## I. INTRODUCTION

Coupled with the rapid development of information technology and knowledge economic, the number of patent applications in our country keeps increasing. Patent, which has enormous commercial and research value, considered as a kind of intangible asset also, has become the vital index in the measurement of all countries' comprehensive strength[1]. Currently, how to extract the pioneering and innovative achievements from patent text, transfer them into products, and realize the industrialization eventually, is one of the emphasis of experts and researchers. The classification of patent text, which serves as the fundamental work, plays an important role in patent research, patent mining, strategic decision and other fields[2]. Consequently, the classification of patent text has extremely important research significance and research value.

At present, the patent text pre-processing, text representation and classifier selection have always been the focus and difficulty of patent text classification. Therefore, the auto-classification of patent text received lots of attention from experts and scholars at home and abroad. Marawan[3] et al represented text vector as a Fixed Hierarchy Vectors (FHVs) method. In addition, FHVs indicate multiple levels of a file, which enriches the document representation and sequential classification, improving the classification performance. Jie[4] et al figured out the patent keywords extraction algorithm based on patent classification Skip-gram model, which can increase the efficiency of auto keywords extraction rapidly; Hu Jie[5] et al proposed a text classification model based on convolutional neural network and random forest to realize the patent text classification in mechanical field. Liao Liefa[6] et al proposed another text classification model based on LDA model, implemented the modelling of rare earth patent text corpus to extract the document-topic and topic- feature word matrix from patent text, then realized the purpose of reducing dimensionality and extracting semantic links among documents. Zhai Dongsheng[7] et al performed the patent text classification for speech recognition technology patents, citation patents and acoustic related patents by establishing patent's feature vector with the use of Word2Vec trained word vector model. While, nowadays, the research of patents in security field is deficient. Patents in diverse fields all have their own features, which lead to the fact that no single algorithm can perform the classification for all patents from diverse fields. Moreover, the patent text corpus in the security field is relatively small, which limits the study of patent text classification in security field in a way.

Recurrent neural networks(RNNs) show outstanding performance when dealing with serialized data. Concretely speaking, the network can memorize previous information and apply it into current output calculation, in other word, it remains connection between nodes in hidden layers, which can integrate the information of the front and back positions effectively[8]. While, although the recurrent neural networks can accomplish the task of handling the whole series, it has the deep memory for the last input signal and the relatively shallow memory for the early input signal, which leads to the problem of "gradient disappear". RNNs' long short-term memory(LSTM) model can avoid the "gradient disappear" problem effectively, making good use of the feature information of the context, maintain the sequence information of the text, auto-select the features and classify them in the end.

This article applies Word2Vec and LSTM to perform the patent text classification in security filed. Compared with the single use of LSTM classification model, Word2Vec model can convert the One-Hot Encoder into continuous value in low dimension, which can avoid the over fitting and resolve the dimensional disaster caused by traditional methods. Meanwhile, the introduction of model can reduce parameters

2473-3547/18/$31.00 ©2018 IEEE
DOI 10.1109/ISCID.2018.00023

of training samples and improve the training efficiency. In the first step of experiment, patent texts were obtainded from the patent website. In the pre-processing of patent text, some words with higher frequency are added to the stop words list according to the traits of patent text. Then, the Chinese data in Wikipedia are trained by Word2Vec to get the word vector with semantic information. The result is used as the input for LSTM model. Finally, LSTM model is trained to execute the text classification.

## II. ALGORITHM PRINCIPLE

### A. Train the word vector based on Word2Vec

This experiment finishes the vectorization of Chinese sample data in Wikipedia and regards it as the input for LSTM model.

Word2Vec, prevalently used in nature language processing (NPL), is a model which can learn semantic knowledge from a large number of text corpus without supervision. It can minimize the space distance among semantically similar words through an embedding space. Word2Vec has two kinds: Skip-Gram(Continuous Skip-Gram Model) and CBOW(Continuous Bag-of-Words Model)[9]. Skip-Gram can forecast context according to current words while CBOW can forecast the current words through context. The structure of Skip-Gram model and CBOW model are showed in Figure 1. They all contain the input layer, projection layer and output layer:

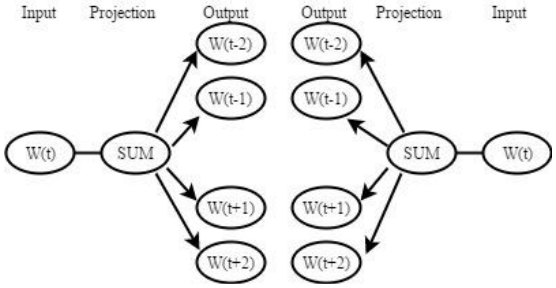


Figure 1. structural diagrams of Skip-Gram model and CBOW model

For statistical model, the maximum likelihood estimation can be used to set the objective function as follows:

$$\prod_{w \in C} p(w \mid Content(w)) \qquad (1)$$

C means corpus, Content(w)means the context of the word w.

### B. LSTM classification algorithm

Although the traditional recurrent neural networks can accomplish the task of dealing the whole time series, it has the deep memory for the last input signal and a relatively shallow memory for early input signals, which leads to the "vanishing gradient" problem. In order to resolve this problem, Hochreiter and Schmidhuber proposed a new special type of RNN, named LSTM. LSTM introduces the CEC(Constant Error Carrousel) unit to deal with the exploding gradient and vanishing gradient problems of back propagation trough time(BPTT) algorithm. In the later period, LSTM undergone the constant improvement, such as the introduction of forget gate, the modification of activation function, the use of memory unit to enhance the connection and so on. The core structure of LSTM is the memory cell, which is showed in the following Figure 2.

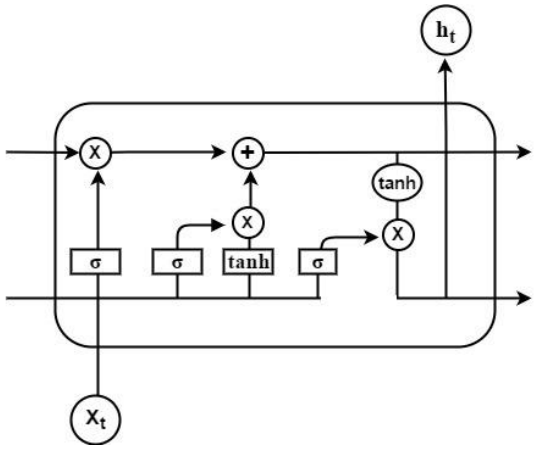


Figure 2. LSTM memory unit structure diagram

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \qquad (2)$$

$$\widetilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \qquad (3)$$

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \qquad (4)$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \qquad (5)$$

$$o_t = \sigma(W_o [h_{t-1}, x_t] + b_o) \qquad (6)$$

Formula (2) represents the input gate; Formula (3) represents the new memory cells of the input gate control; Formula (4) represents the forget gate; Formula (5) represents the state of the current t moment Cell; Formula (6) represents the output gate.

## III. RESULTS AND ANALYSIS

In the experiment, the maximum length of every text is set as 200, the dimension of the word vector space is 128. The weights of model are renewed by using random gradient descent method. Futhermore, batchSize is set as 45, epoch is 3. Finally, its performance is compared with the single LSTM model, and the main analyze index in assessing text classification is the accuracy and ROC curve evaluation model. The result is showed in table1 and Figure 3.

TABLE I. COMPARISON OF ALGORITHM RESULTS

| Method | Number of test groups | Correct number of groups | Accuracy | Error |
|---|---|---|---|---|
| LSTM+Word2Vec | 10000 | 9348 | 93.48% | 6.52% |
| LSTM | 10000 | 8576 | 85.76% | 14.24% |

The table 1 shows that, in the classification of patent text in security field, the accuracy of text classification model based on both Word2Vec and LSTM is much higher than

the text classification model based only on LSTM. The reason is that, to introduce the Word2Vec into pre-training is equivalent to increasing the training corpus indirectly. Meanwhile, it can prevent the over fitting, reduce the number of needed parameters in the training and further improve the accuracy. In addition, the area under ROC curve, AUC(Area under the ROC curve), is a method to evaluate the average performance of model. If the curve nears to the top left corner and the area is close to 1, the classification model is better. As shown in the graph, the area of classification model based on both Word2Vec and LSTM is 0.99, while the area of classification model based only on LSTM is 0.98.

At the same time, experiments based on CNN+Word2Vec, CNN, KNN or other classification models were performed separately. CNN has two vital traits: local sensing and weight sharing. Its neurons in the coiling layer are connected only to some neurons in the previous layer, that is, the connections between neurons are not fully connected.

In the experiment, each document is divided into a two-dimensional vector of 100*128. The fixed length of each patent text is 100, and the word vector is 128. In addition, the length of vector is narrowed the 1 convolution layer and pooling layer. Finally, they are conveyed to SoftMax classifier to realize the classification through Dense(all connected layer). K-Nearest Neighbor(KNN)[10] classification algorithm is the simplest machine learning algorithms, it measures the similarity between samples by distance..

TABLE II. COMPARISON OF DIFFERENT ALGORITHMS

| Method | Number of test groups | Correct number of groups | Accuracy | Error |
|---|---|---|---|---|
| CNN+Word2Vec | 10000 | 8118 | 81.18% | 18.82% |
| CNN | 10000 | 8059 | 80.59% | 19.41% |
| KNN | 10000 | 3351 | 33.51% | 66.49% |

The experiment above manifested that the accuracy of text classification model based on Word2Vec and LSTM and the performance on ROC curve are obviously better than other algorithms. Convolution neural network algorithm shows apparent advantages in image processing, while performs not so ideal on text classification. In addition, the KNN algorithm is obviously not suitable for such classification, and LSTM can be better for the processing of the serialized text data.

## CONCLUSION

The classification of patent text, which serves as the fundamental work, plays an important role in patent research, patent mining, strategic decision and other fields. Patents in diverse fields all have their own features, which lead to the fact that no single algorithm can perform the classification for all patents from diverse fields. Moreover, the patent text corpus in the security field is relatively small, which limits the study of patent text classification in security field in a way. In order to resolve the patent text classification problem in security field, Word2Vec and LSTM classification are introduced. The use of pre-trained Word2Vec LSTM model can well solve the dimensional disaster problem caused by traditional methods. This experiment made use of the edge of LSTM neural network, applying 50 thousand patent texts in the training and testing, and the accuracy of the test set reached 93.48%, which is better than the convolution neural network classification model and the K nearest neighbor classification model. Meanwhile, it still remains some deficient in the experiment mainly including the long training time for the model and the low efficiency in the practical application. All in all, how to improve the efficiency of the classification model will be the key point of the next research.

## ACKNOWLEDGMENT

Shanghai Institute of Technology Collaborative Innovation Fund(number: XTCX2017-17)

## REFERENCES

[1] Guo Peng, Li Honglian, Jiang Jian. The influence of patents on the development of enterprises[J]. Education and Teaching Forum, 2015(5):106-107.

[2] Ma Shuanggang. The Study of Research on Chinese Patent Classification Based on Deep Learning Theory and Method[D]. Jiangsu University, 2016.

[3] Marawan Shalaby, Jan Stutzki, Matthias Schubert, and Stephan Günnemann.An LSTM Approach to Patent Classification based on Fixed Hierarchy Vectors[C].Proceedings of the 2018 SIAM International Conference on Data Mining. 2018, 495-503K. Elissa, “Title of paper if known,” unpublished.

[4] Hu J,Li S,Yao Y.Patent Keyword Extraction Algorithm Based on Distributed Representation for Patent Classification[J]. Entropy, 2018, 20(2):104.

[5] Hu Jie, Li Shaobo, Yulia. Patent text classification model based on convolution neural network and random forest algorithm [J]. science and technology and engineering, 2018 (6).

[6] Liao Lei FA, Le Fu Gang, Zhu Yalan. The Application of LDA Model in Patent Text Classification [J]. modern intelligence, 2017, 37 (3): 35-39.

[7] Zhai Dongsheng, Hu Jinjin, Zhang Jie. Research on Modeling Technology of patent grade classification, [J]. data analysis and knowledge discovery, 2017, 1 (12): 63-73.

[8] Yuan Jie. Research on speech emotion recognition based on the fusion of ANN and GMM [D]. Southeast University, 2016.

[9] Han Songjiang. Research on mining technology of production reviews based on multi-document summarization [D]. Southeast University, 2015.

[10] Zhang Ning, Jia Ziyan, Shi Zhong Zhi. Text categorization with KNN algorithm [J]. Computer Engineering, 2005, 31 (8): 171-172.

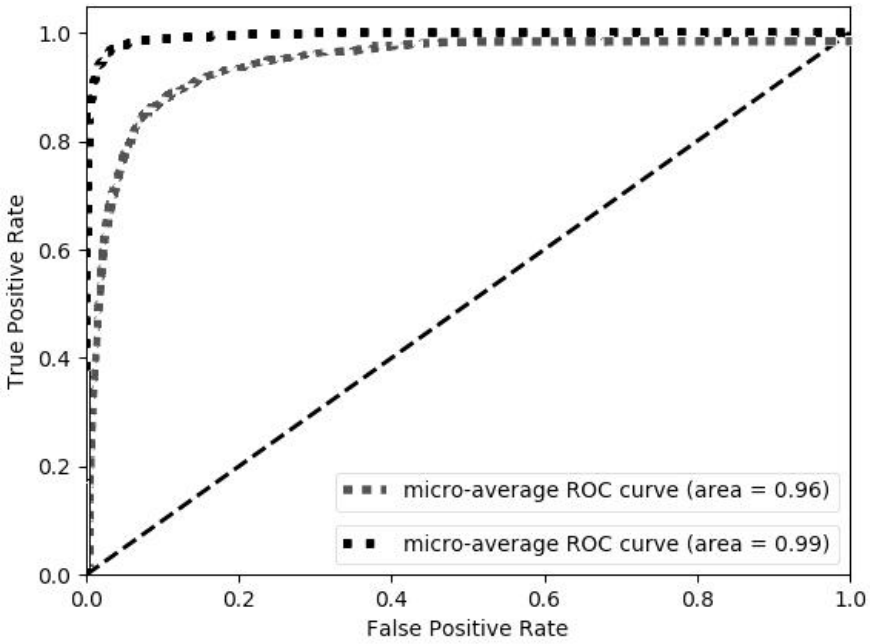


Figure 3. ROC curve of similar algorithms