Deep Text Information Type Detection Model Based on Statistical Learning in Engineering Application

Chengfu Wei*
School of intelligent equipment, Shandong University of Science and Technology, Taian, Shandong 271019

*Corresponding author: chengfuwei@sdust.edu.cn

Abstract. In response to the problem of financial news data in the traditional neural network model, an improved text classification model is proposed. The model optimizes the neural network model based on Word-Embedding + LSTM, which has achieved better text classification in the experiment.

Keywords: Deep learning, text classification, neural network, LSTM recurrent neural network, word-embedding

1. Introduction

With the rapid development of deep learning sectors in recent years, artificial neural networks have gradually showed its huge potential in the field of natural language treatment [1]. Recent studies have shown that circulating neural network (RNN) is very good in terms of text feature vector extraction [2].

In order to further improve and enhance the accuracy of the text classification model for text data, the paper proposes a new model structure. When constructing the Word-Embedding word vector, according to the index of the word and the part of speech corresponding to the word, combined with the article title and article content comprehensively constructs the Multiword-Embedding word vector, and then the constructed word vector is extracted through a single-layer Bi-LSTM cyclic neural network and the final text feature vector is spliced, and finally the text data is multi-classified through a fully-linked FC layer Processing, better test results were obtained in the experiment.

2. Related theories

2.1. Word-Embedding word vector

The generation of feature vectors is an indispensable step to transform human-understandable data into machine-understandable data [7]. The vector generation method is based on the experience of researchers using "0" or "1" to manually define the one-hot real number vector. Although such vector data has certain characteristic expressiveness, it often causes dimensional disasters, while Word-Embedding is A series of low-latitude floating-point vectors, which do not rely on artificially customized features but learn the features of the corresponding data through the model [8]. Experiments have shown that this method can learn better feature distribution. The formula has the following expression:

\[ \text{Vec}_i = \text{Emb}(\text{Index}(W_i)) \]
2.2. LSTM circular neural network
The LSTM recurrent neural network is a variant of the RNN network. It improves the calculation method in Cell on the basis of RNN [9], making it have long-term dependence in the training and inference process, effectively reducing the disappearance of gradients. There is no need to save lengthy context information during the learning process. LSTM neural network is particularly outstanding in solving temporal modeling problems, and is now widely used in the field of natural language processing [10]. The biggest change of the LSTM recurrent neural network is the improvement of the Cell of RNN and the introduction of the MemoryCell unit [11], which includes four elements: an InputGate, a cyclic self-connected neuron, a ForwardGate and an OutputGate. The introduction of this unit can better tune the neuron state transmission of RNN.

![LSTMMemoryCell structure](image)

Figure. 1 LSTMMemoryCell structure

As shown in Figure 1, LSTM is restricted by three kinds of "gates" in the process of its parameter transfer, namely, "input gate", "output gate" and "forgotten gate". These three kinds of gates are in their respective parameter matrices. Some parameters are randomly discarded during the calculation process, so that most of the parameters that do not work can be discarded. According to the formula, it can be seen that when training to obtain the gradient, the gradient can be forced to change to "1", which effectively avoids the gradient disappears or the gradient explodes due to continuous multiplication when deriving.

3. Improved text classification model

3.1. Word vector construction based on MultiWord-Embedding
The word-embedding word vector construction method is proposed, so that researchers no longer need to manually extract the feature vector to construct the on-hot word vector. The word-embedding word vector is based on its controllable word vector dimension and dense floating-point data. The characteristic effectively solves the dimensional disaster that often appears in the on-hot vector and the excessive sparseness of the vector. When constructing a most basic word-embedding word vector, it is only necessary to obtain the vector composed of the corresponding variables in the embedding variable matrix according to the index of the word, and then the word vector of the word can be formed. And in the process of model training, each element in the vector is used as an optimizable parameter, and the model is trained and tuned together to find the best feature expression of the word.

Based on the Word-Embedding technology and combined with the characteristics of news data, this paper proposes a multi-dimensional construction method of Multi Word-Embedding word vector constructing Word-Embedding. First of all, the word vector of the paper is generated by using the dictionary meaning of the word and the corresponding part of speech feature of the word, so that the word vector that composes a word has more dimensional characteristics. The specific method is to construct the dictionary index of the words in the news article, and then construct the corresponding part-of-speech index according to each word, and then construct the independent word-embedding word matrix of the two indexes respectively, when generating the word-embedding of a word, it is necessary
at the same time, the corresponding vectors are taken from different word matrices and spliced. The construction method of the internal word vector of MultiWord-Embedding can be expressed by the following formula:

$$
Vec_i = \text{Emb}_d(\text{IndexD}(W_i)) \parallel \text{Emb}_p(\text{IndexP(Property}(W_i)))
$$

3.2. Bi-LSTM+MultiWord-Embedding text classification model

Based on the Word-Embedding and LSTM models, this paper proposes a text classification model of MultiWord-Embedding+Bi-LSTM. According to the characteristics of news industry articles, the overall structure of the MultiWord-Embedding word vector combined with the title is processed as follows:

![Figure 2 Bi-LSTM+MultiWord-Embedding text classification model](image)

3.3. Algorithm process description

The algorithm process of the text classification model based on Bi-LSTM and MultiWord-Embedding word vector can be described by the following steps:

1) Input a text string;
2) Perform word segmentation and part-of-speech tagging on the text string to obtain a word segmentation list. Each element in the list is a two-tuple of a word and the part of speech corresponding to the word;
3) Convert the words and the marked part of speech in the word segmentation list into the corresponding index of the word dictionary and the part of speech dictionary, and obtain a two-dimensional matrix of Nx2, where N represents the number of words in the sentence;

4. Related experiments

4.1. Experimental design

The data set used in the experiment of the thesis is financial industry-related news materials, including 10,000 training sets and 2,000 test sets. The data set contains 7 categories, namely Internet financing, online lending, financial policy, financial technology, Crowdfunding, banking, financial insurance. When constructing the training set and the test set, the paper randomly shuffles these articles and guarantees the proportion of subcategories.
All the neural network model codes in the experiment of the thesis are implemented based on Google's TensorFlow. The CPU of the machine used in the experiment is Intel Core i5 processor, the main frequency is 3.80 GHz, and the operating system is Windows 7, 64 bits. The total time for model training is 1.5 hours.

4.2. Experimental data
The paper will carry out the classification experiment of the model according to the methods and details described in the experimental design. The specific distribution of the data obtained after the training set and the test set are scrambled and randomly selected is shown in Table 1.

| category       | Financial management | loan | policy | Technology | Crowdfunding |
|---------------|----------------------|------|--------|------------|--------------|
| Training set  | 20000                | 15000| 10000  | 25000      | 10000        |
| Test set      | 4000                 | 3000 | 2000   | 5000       | 3000         |

The training set obtained by the optimal parameter model trained by the four neural network models and the correct rate data of the test set are shown in Table 2 below. The main data comparison basis in Table 2 is the test set and training set of each neural network model. According to the percentage of correctness, it can be seen that the model proposed in the paper increases the correctness to 93% after obtaining the optimal parameters.

| model            | Correct rate of training set | Test set correct rate |
|------------------|------------------------------|-----------------------|
| LR+WORD          | 0.99                         | 0.89                  |
| CNN+WORD         | 0.99                         | 0.89                  |
| LSTM+WORD        | 0.99                         | 0.89                  |
| Bi-LSTM+MWORD    | 0.99                         | 0.89                  |

4.3. Experiment analysis
Judging from the experimental data in the above table, the Bi-LSTM+Multiword-Embedding model has obtained relatively excellent experimental results in news text classification, and the accuracy rate of the test set reaches 93%, which is better than other comparison models. There are two main reasons for this. First, Word-Embedding expands the feature expression of words by constructing a combined dictionary and part-of-speech construction. The combination of news content and headline word vectors makes the Word-Embedding feature more obvious, and secondly, Bi-LSTM solves the relationship expression of the sequence of words to a certain extent. Compared with LR, CNN and one-way LSTM, it provides more word position information. In summary, the model of Bi-LSTM + MultiWord-Embedding provides better text classification results in news data.

5. Conclusions
Based on the LSTM recurrent neural network and Word-Embedding word vector, the paper proposes a text classification model based on the Bi-LSTM recurrent neural network plus dictionary and part of speech MultiWord-Embedding word vector. The thesis has carried out in-depth research on the relevant theory of the model, and carried out a detailed design of the overall structure of the model, and carried out a comparison experiment on the accuracy of text classification with the logistic regression model, the convolutional neural network model and the one-way LSTM model. The training set and test set data used in the experiment in the paper come from public historical financial news data obtained by major financial websites. Through the comparison of the classification accuracy of the test set in the experiment, it can be seen that the improved text classification model in the financial field has achieved 93% The classification accuracy of the test set is higher than the classification accuracy of the other
three neural network models, which also proves that the model structure proposed in the paper has better feature expression than the traditional neural network. This model also provides a new idea for the study of text classification methods based on neural networks.

References

[1] Learning Word Vectors for Sentiment Analysis [C]// Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2011.

[2] Shamseera S P, Sreekanth E S. Learning Word Vectors To Dual Sentiment Analysis.

[3] Yuan W, Li Z, Jie L, et al. Word Vector Modeling for Sentiment Analysis of Product Reviews [C]// Communications in Computer & Information Science. Springer Berlin Heidelberg, 2014: 168-180.

[4] Liu J, Zheng S, Xu G, et al. Cross-domain sentiment aware word embeddings for review sentiment analysis [J]. International Journal of Machine Learning and Cybernetics, 2020: 1-12.

[5] Lai K P, Lam W, Ho. Domain-Aware Recurrent Neural Network for Cross-Domain Sentiment Classification [C]// DSIT 2020: 2020 3rd International Conference on Data Science and Information Technology. 2020.

[6] Y Zhang, Miao D, J Wang. Hierarchical Attention Generative Adversarial Networks for Cross-domain Sentiment Classification [J]. 2019.