Computer Life (CPL) ISSN: 1819-4818
DELIVERING QUALITY SCIENCE TO THE WORLD

Research of Chinese Short Text Classification Based on Word2vec

Zhen Li [a], Minmin Duan [b]

Faculty of Economics and Management, Xidian University, Xi'an, China

[a]15114840972@163.com, [b]dmin79856@163.com

Abstract: Short text Because of its less semantic information and high sparsity, traditional text categorization methods do not work properly on short text. Aiming at the above problems, this paper proposes a Chinese short text classification method based on Word2vec. Firstly, the semantic information word vector is realized by Word2vec, and then the weight of the words in each document is calculated by TF-IDF algorithm. The text representation is finally combined with the LIBSVM classification algorithm. Experimental results show that this method can effectively improve the short text classification effect.
Keywords: Text classification, Word2vec, LIBSVM.

1. INTRODUCTION

In recent years, with the rapid development of the Internet and mobile communications, short text has been widely used as a popular text form in short comments, text messages, microblogs, questions and answers, etc. According to China Internet network information center (China Internet network information center, CNNIC) The 42th Statistical Report on Internet Development in China[1]released in June 2018, instant messaging, online news and search engines have become the three most popular information platforms, with 92.1% and 83.1% respectively. And 81.1%. Faced with these large-scale text resources, it is difficult for people to get the information they need. Therefore, short text categorization can help users effectively process and utilize useful information hidden in large documents. Short text technology has become a hot research topic.

Compared with long text, the length of short text is usually controlled at about 140 words, which is characterized by sparsity, real-time, and non-standardity. At present, most traditional classification methods use the Vector Space Model (VSM)[2], but they have high dimensions, sparseness and no consideration of semantic information between texts. One method is to use the search engine to extend the context and Rich text features[3,4]ZELIKOVIT uses the external context corpus for the mark to measure the text similarity[5], Wang Peng et al. use the dependency relationship to extend the short text to achieve effective short text classification[6] The similarity is calculated by analyzing the short text results returned by the search engine. Getting search text snippets from search engines is not an ideal method for most applications because it depends not only on the quality of the search engine, but also on time. Another approach is to use a web database as a source of external knowledge (such as Wikipedia, WordNet, HowNet[7] etc.) to extend the features of short text. Fan Yunjie, Zhao Hui

and others used the Wikipedia knowledge base to extend the short text feature to assist in short text classification[8,9]; Ning Yahui et al. extracted the field high frequency word extended short text from the knowledge base of Knowledge Network[10]; Sheng et al. used the upper and lower position of the network of words to expand the short text.

The external knowledge base only contains a limited range of fields and topics. The vocabulary update speed is slow, and the latest meaning of the words cannot be included in time. There are a considerable number of unregistered words, and the classification effect is very limited. In contrast, we found that these methods still have a lot of room for improvement in short text categorization.

To this end, this paper proposes a short text classification method based on Word2vec[11]model. This method is based on Word2vec training with rich semantic information word vectors. Word vectors can be obtained through large-scale open corpus training. The semantic relationship between the two does not need to rely on an external knowledge base. According to the cosine distance between these word vectors, the TF-IDF[12]algorithm is used to weight the word segmentation in each document. This text representation method can not only solve the problem of semantic deficiency between words, but also effectively reduce the text vector dimension and greatly reduce the training time of the model. Finally, the efficient LIBSVM[13].classification algorithm is used. The experimental results show that compared with the traditional classification method, the method has higher recall rate and accuracy.

2. CHINESE SHORT TEXT CLASSIFICATION PROCESS BASED ON WORD2VEC

The short text classification process proposed in this paper is shown in Figure 1.

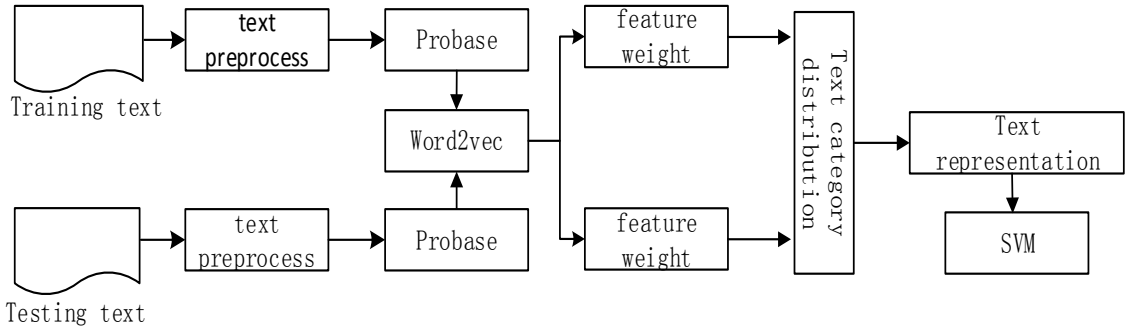

Fig. 1 Chinese short text classification model based on Word2vec

2.1 Text preprocessing

Before training the text in the Word2vec model, the text must be preprocessed first. The text preprocessing is the process of converting the original text into structured data, which can effectively remove the noise in the text, which is more effective representation of the obtained structured data. Text semantics, using the ICTCLAS Chinese lexical analysis system of the Chinese Academy of Sciences for word segmentation and part-of-speech tagging, using the stop word table to match and delete the stop words after the word segmentation, and only retain the feature words [14] that can represent the core part of the sentence.

2.2 Introduction to Word2vec

Word2vec[11] is a word vector training tool that Google opened up in 2013. The text content is simplified into K-dimensional vectors through training processing. Because the model and algorithm fully consider the word-sentence and semantic information during the training process, the vector

generated by the training can be considered to be able to refer to the text and have rich semantic information. There are two types of continuous bog-of-words (CBOW) and Skip_gram models, as shown in Figure 2.

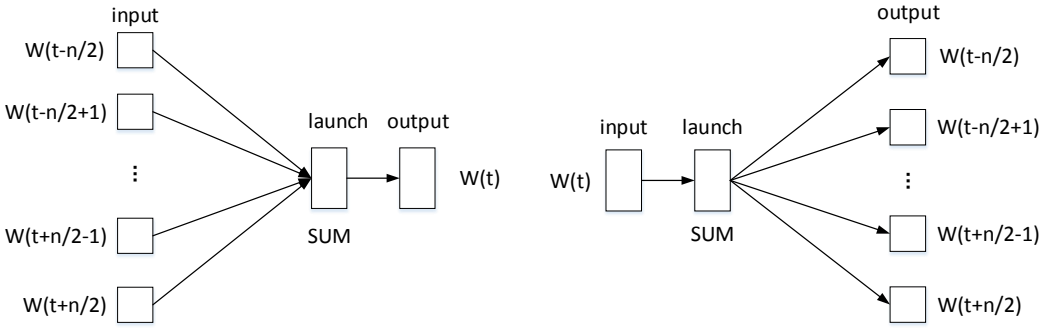

Fig. 1 CBOW and Skip_gram

The CBOW model predicts the current word based on the context. The mathematical representation of CBOW is shown in equation (1):

$$p(w_i|\tau(w_{i-k}, w_{i-k+1}, \cdots . w_{i+k-1}, w_{i+k})) \quad (1)$$

Where is a word in the corpus dictionary that predicts the probability of occurrence of a word by a word with an adjacent context window size of K[15]

Skip_gram predicts the context based on the current word, that is, predicts the vocabulary probability in the adjacent window K by vocabulary. The mathematical representation of Skip_gram is shown in equation (2):

$$p(w_{i-k}, w_{i-k+1}, \cdots . w_{i+k-1}, w_{i+k}|w_i) \quad (2)$$

Compared with the CBOW model, the Skip-gram model is a high-quality distributed vector representation that efficiently captures the relationship between precise grammar and semantic words. The disadvantage is that the training time is longer.

2.3 Word Vector Training

The Word2vec model is used to train the word vector, and the word vector containing the distance between the feature item and other feature items in the document is obtained. At the same time, the word vector representation of all the feature items in the corpus is obtained. The size of the word vector is different from the corpus. The number of items is determined. Word2vec trained a rich semantic information word vector file to calculate words that are semantically similar to a word. For example, if you type the word "Baidu", it will display the words that are closest to "Baidu" and the distance between them. Table 1 shows the five words that are semantically closest to "Baidu".

Table 1 The five words closest to the semantics of Baidu

Original word	Approximate word	Cosine distance
Baidu	Search for	0.693204
	Web portals	0.587623
	Search engine	0.697913

	Yahoo	0.583331
	Tencent	0.600235

2.4 Improved TF-IDF weight calculation

The most common problem of the most popular TF-IDF weighting algorithm is that the feature weight discrimination of short text is insufficient. The reason is that due to the feature weight calculation problem, an optimized TFIDF algorithm is proposed, which uses a word correction variable. The calculated formula (3) shows that the optimized TFIDF value is also used $\{D_1, D_2, D_3, \cdots, D_m\}$as the weight of the word t in the document. Refer to the setting in[16] to define the difference in the distribution of words under the same category of documents, that is, considering a word or the word representing features, only in a certain category of documents, rarely appear in other In the category of the document.

$$di_{dc}(t_i) = \frac{\sqrt{\sum_{j=1}^{d}(n_{i,j}-\overline{n_{I,J}})^2 / (|d|-1)}}{\overline{n_{I,J}}} \tag{3}$$

The representative word$n_{i,j}$appears $t_i$in the frequency of the documentD_j, and$\overline{n_{I,J}}$ the text $t_i$in the average corpus appearing in each document is trained to obtain the e-dimensional word vector x corresponding to each participle$\mathrm{x} = (\vartheta_1, \vartheta_2, \vartheta_3, \cdots, \vartheta_e)$, wherein if a single word TFIDF value and X are used in the document set Equation (4) to represent the vector of the documentD_i:

$$d_i = \sum_{t \epsilon D_i} x_t \, w(t, D_i) \tag{4}$$

Where is the vector of words $x_t$in the document D_i, which $w(t, D_i)$represents the weight of the words obtained by the improved TFIDF model in the document set. The calculation formula is as shown in (5):

$$w(t, D_i) = \frac{tf(t,D_i)\times idf(t)}{\sqrt{\sum_{t \epsilon D_i}[tf(t,D_i)\times idf(t)]^2}} \times (1 - di_{dc}(t_i)) \tag{5}$$

Then, according to the size of the weight, the K feature items that best represent the text are selected. If the K value is too large, it may cause a dimensional disaster and the amount of calculation increases. If the K value is too small, the text is not well represented.

2.5 classification algorithm

Support vector machine [17,18] is an advantage in the field of text classification, and is widely used in various classification tasks. Any classification problem can be regarded as a two-category problem. The basic idea is to find the hyperplane with the largest edge in a sample data set containing two different classes in the case of linear separability. In the case of linear inseparability, a slack variable is introduced in the constraint of the optimization problem, and the sample in the low-dimensional space is mapped to the high-dimensional space by the nonlinear mapping to make it linear, and the linear algorithm is used to sample the sample in the high-dimensional space. Nonlinear analysis is performed while finding the optimal hyperplane in linear space. The SVM can discover the global minimum of the objective function using known efficient algorithms. Other classification methods use a greedy learning-based strategy to search for hypothesis space. This method generally only

obtains local optimal solutions. The SVM controls the capabilities of the model by maximizing the edges of the decision boundaries. In view of the shortcomings of SVM, this paper adopts the LIBSVM algorithm developed by Dr. Lin Zhiren from Taiwan University. It has the characteristics of flexible application, less parameter setting and easy expansion. It has been adopted by many scholars at home and abroad, and there are many choices of cross-validation parameters of nuclear functions. It is more accurate, so there is a significant improvement in text classification.

3. EXPERIMENTS AND ANALYSIS OF THEIR RESULTS

3.1 Data Preparation

The corpus used in this experiment is a news headline captured from Sina, Sohu, Tencent and other web pages, including 1,000 categories in each of six categories: military, sports, education, technology, entertainment, and economy. The length of each title is about Between 10-20 words. Using the 5-fold crossover experiment method, each type of text was randomly divided into 5 parts, one of which was used as the test text set, and the other four were used as the training text set, and then the five classification results were averaged to obtain the experimental results.

3.2 Performance evaluation method

This paper uses the traditional text classification performance evaluation method. We use the accuracy, recall rate and F1 to evaluate the experimental results.

they are Precision$\mathrm{P} = \frac{TP}{TP+FP}$ Recall$\mathrm{R} = \frac{TP}{TP+FN}$ and $F_1 - measureF_1 = \frac{2PR}{P+R}$

In the above formula, the TP rate shows the percentage of the correct example, which is correctly classified as a positive category, while the TP rate shows the percentage of counterexamples that are misclassified as positive categories, and the FN rate shows the percentage of positive cases that are misclassified as negative categories.

3.3 Analysis of experimental results

The experiment is divided into two groups: Experiment 1 uses traditional short text classification method to classify the corpus, that is, using SVM algorithm, the kernel function of SVM is linear kernel function; Experiment 2 uses the method based on Word2vec proposed in this paper to classify. In the experiment, ICTCLAS was used to segment the words. In the Word2vec word vector training process, the parameters were configured as follows: 1) Select the Skip-gram model; 2) Context sliding window is 5; 3) Word vector dimension is set to 200. The experimental results are shown in Table 2.

Table 2 Comparison of experimental results

category	Word2vec+SVM			Word2vec+TF-IDF+LIBSVM		
	precision	recall	F1	precision	recall	F1
Military	78.15	75.39	76.13	86.13	83.24	83.79
Sport	63.67	60.28	56.89	71.45	70.93	68.24
Education	62.53	67.42	62.73	72.36	79.86	76.29
Technology	75.33	71.25	75.71	84.90	81.25	83.74
Entertainment	66.82	72.78	69.90	75.22	78.47	75.16
Economic	68.39	60.45	62.67	76.84	77.26	78.36
Average	69.15	67.93	67.34	77.82	78.50	77.60

It can be seen from the experimental results in Table 3 that the short text classification results based on Word2vec proposed in this paper are improved in all aspects compared with the traditional classification method. It can be explained that the word vector representation of the feature item is obtained by Word2vec training, which well represents the semantic relationship between the feature items, and also reduces the vector dimension of the feature item. At the same time, the word vector is clustered, and the semantic information and distance information distinguish the similar texts very well, which greatly reduces the time of model training. The experimental results also prove that the proposed method has certain advantages in text classification.

4. CONCLUSION

This paper is based on Word2vec model for Chinese short text classification. This method is based on Word2vec training of word vectors with rich semantic information. It handles the sparseness of text representation and the lack of semantics between words. TF-IDF weighted calculation document vector information research in the corpus shows that the short text classification method proposed in this paper can improve the classification effect.

REFERENCES

[1] The 42th Statistical Report on the Development of China's Internet. http:// www. cnnic. net. cn/hlwfzyj/
[2] Salton G. A vector space model for automatic indexing[J]. Communications of the Acm, 1974, 18(11):613-620.
[3] Banerjee S, Ramanathan K, Gupta A. Clustering short texts using wikipedia[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2007:787-788.
[4] Hu X, Sun N, Zhang C, et al. Exploiting internal and external semantics for the clustering of short texts using world knowledge[C]// ACM Conference on Information and Knowledge Management. ACM, 2009:919-928.
[5] Zelikovitz S, Hirsh H. Improving Short-Text Classification Using Unlabeled Background Knowledge to Assess Document Similarity[C]// 2000:1183--1190.
[6] WANG Peng, FAN Xing-hua. Experimental study on the use of dependencies in Chinese text classification[J]. Computer Engineering and Applications, 2010, 46(3): 131-133.
[7] Liu Z, Yu W, Chen W, et al. Short Text Feature Selection for Micro-Blog Mining[C]// International Conference on Computational Intelligence and Software Engineering. IEEE, 2010:1-4.
[8] Fan Yunjie, Liu Huailiang. Research on Chinese Short Text Classification Based on Wikipedia[J]. Modern Library Information Technology, 2012(3): 47-52.
[9] Zhao Hui, Liu Huailiang. A Chinese Short Text Classification Algorithm Based on Wikipedia[J]. Library and Information Service, 2013, 57(11): 120-124.
[10] Ning Yahui, Fan Xinghua, Wu Wei. Short Text Classification Based on Domain Word Ontology[J]. Computer Science, 2009, 36(3): 142-145.
[11] Tomas Mikolov. Word2vec project [EB/OL]. [2014-09-18].
[12] Wu H C, Luk R W P, Wong K F, et al. Interpreting TF-IDF term weights as making relevance decisions[J]. Acm Transactions on Information Systems, 2008, 26(3):55-59.
[13] Chang C C, Lin C J. LIBSVM: A library for support vector machines[J]. 2011.
[14] Wang Yuanzhen, Qian Tieyun, Feng Xiaonian. Automatic Classification of Chinese Text Based on Association Rules Mining[J]. Journal of Chinese Computer Systems, 2005, 26(8): 1380-1383.
[15] Zhang Q, Gao Z, Liu J. Research of Weibo Short Text Classification Based on Word2vec[J]. Netinfo Security, 2017.
[16] Huang X, Wu Q. Micro-blog commercial word extraction based on improved TF-IDF algorithm[C]// Tencon 2013-2013 IEEE Region 10 Conference. IEEE, 2013:1-5.
[17] Joachims T. Text categorization with Support Vector Machines: Learning with many relevant features[M]// Machine Learning: ECML-98. Springer Berlin Heidelberg, 1998:137-142.
[18] Isa D, Lee L H, Kallimani V P, et al. Text Document Preprocessing with the Bayes Formula for

Classification Using the Support Vector Machine[J]. IEEE Transactions on Knowledge & Data Engineering, 2008, 20(9):1264-1272.