Weibo Text Classification Based on Knowledge Graph

Yingjun Liu
East China Normal University, Shanghai, 200062, China
Corresponding author’s e-mail: 51194507010@stu.ecnu.edu.cn

Abstract. As a platform for sharing and communication, Weibo can produce a large amount of short text data every day. Because the Weibo text is short and the text features are sparse, it contains less information, serious colloquialization and weak anti-noise ability, so it is difficult to achieve the desired results by using the traditional text classification method to classify Weibo text. This paper proposes a Weibo text classification method based on knowledge graph, which uses knowledge graph to expand text semantics and enrich text features on the basis of the original Weibo text. Then the TextCNN model is used to classify Weibo texts with extended semantics. The experimental results show that the classification accuracy of this method is better than that of Weibo text classification directly.

1. Introduction
With the rapid development of mobile Internet, more and more users are used to expressing their opinions on social media. Weibo, as the largest social media platform in China, produces data in many fields, and these huge amounts of data often contain a lot of valuable information. However, because Weibo is mostly used for users to publish what they see, think and feel, its content has strong arbitrariness and diversity, and there is a lot of text noise. In addition, Weibo is limited to about 140 words, the content is short, most Weibo has only a dozen or dozens of words, resulting in its sparse features. Due to the irregularity and sparsity of Weibo data, the effect of using traditional classification methods to classify Weibo data is not good.

At present, the methods to solve the sparse features of datasets are basically based on the Bag of Words (BOW) model, including the following two categories: 1) Introduce an external knowledge base. For example, use some common domain knowledge base, encyclopedia entry knowledge base etc. as knowledge sources for extending text. Looking for knowledge matching text keywords from these knowledge bases as extension items to make short text feature richer. Yunjie Fan [1], Qi Jing [2] put forward a classification method of Chinese short text based on Wikipedia, which uses Wikipedia as feature source of short text. This method can’t be restricted by text content, which helps to extend the text sufficiently. However, based on external knowledge base acquisition cost is higher, knowledge quality is unstable, and introducing external knowledge base affects semantic meaning of short text but also limits the solution to short text. 2) Expanding text according to text content. This method uses feature such as word frequency, topic distribution as feature expansion items and expands the text by constructing high frequency word sets, thematic reasoning etc. Xiwei Wang [3] et al. used FP-Growth algorithm to extract association rules of feature items in short text and extract association rules. Xuemin Fu [4] et al. proposed a novel classification method based on high frequency word expansion. Chaozhen Lv [5], Yongjun Hu [6] et al. studied feature extension methods based on thematic distribution. This method does not depend on external knowledge base and extends entirely depending on text features. However, different features need to be constructed according to different problems, which requires higher selection
requirements for feature items, and when text is short enough context information is too few, it can’t solve feature sparse problem properly.

For the non-normative data sets, the usual solution is to dig deeper semantic information to enhance understanding of text semantics and reduce noise interference to short text. In this paper, we use the method based on deep learning to mine deeper semantic information. This method makes up for the deficiency of artificial construction of feature engineering. It uses deep learning model to learn features of short text from massive data concentration thereby enhancing understanding of short text and reducing noise impact. In 2014 Kim[7] firstly applied convolution neural network to sentence grade text classification model and proposed TextCNN model. This model uses trained word vector models as one-dimensional image input so we can capture association between adjacent words using one-dimensional convolution neural network. It performs well when capturing shorter sequence information.

In conclusion, in order to solve the problem of sparse text features in Weibo dataset, this paper introduces knowledge graph as external source to extend Weibo text. After extracting keyword from Weibo text, matching with knowledge graph through entity link, then uses entity disambiguation technology to select entity similarity degree with text similarity as extension text. Aiming at the nonstandard problem of Weibo dataset, text categorization model TextCNN[7] based on deep learning is used to extract deep features of text. Finally, this paper verifies that this method performs well on Weibo data set through experiments.

2. Related work

2.1. Data acquisition

Web crawler technology is to write a program to simulate the browser to surf the Internet, and then let it go to the Internet to grab data. When crawling network information, first get the initial web page URL, to read the information on the current web page through the page structure analysis, and put the read URL into the queue, and then constantly take out the URL from the queue to access and read the information until certain conditions are met. When crawling Weibo datasets, analyze the interface on the Weibo web page “https://m.weibo.cn/” through the browser and get its interface parameters, use the “requests” library to send network requests and obtain response data, use the “lxml” library to parse HTML files, and convert the data to Python data types, which makes file operation easier. In this paper, we crawled Weibo data under ten themes. The topic names and the number of data entries are as follows: Science and technology: 1156, Epidemic: 1075, Huawei: 1154, Tourism: 1196, Parenting: 1052, Food: 1285, Constellation: 1200, Eight Hundred: 1056, Depression: 1056, Fitness: 1230. Each Weibo post, including the serial number, publisher nickname, Weibo body, is stored in txt documents, and the file name is the subject name, such as: “58: Come on, Tang: # Together # Under the epidemic, you and I fight side by side to pay tribute to all the warriors in the fight against the epidemic! # TV series #”, which is located in a file called “Epidemic.txt”.

2.2. Data processing

Word segmentation is the first step in Weibo text data processing. This paper uses the “jieba” word segmentation tool to segment sentences into word sets. The “jieba” tool has three word segmentation modes: full mode, precise mode and search engine mode. The full pattern scans out all the words in the sentence as much as possible; the precise pattern tries to cut the sentence most accurately, which is suitable for text analysis; the search engine pattern is to segment the long words again on the basis of the accurate pattern to improve the recall rate. In this paper, the accurate mode is used to segment the Weibo text. Then, according to the stop word list, the stop words are removed from the Weibo text after the segmentation, and the words such as “ah, or, for example” and some meaningless characters are removed after the segmentation. In order to avoid the influence of commonly used words on Weibo, such as map, hypertext, topic, platform, link, retweet, such words are written into the stop word list.

Use TextRank algorithm[8] to get the keywords of Weibo text and the context words of keywords. TextRank algorithm is a text sorting algorithm, which is improved from Google's web page importance
ranking algorithm PageRank algorithm, which is used to generate keywords and abstracts for text. It divides the text into several constituent units (words, sentences) and establishes a graph model, and uses the voting mechanism to sort the important components of the text. The keyword extraction based on TextRank algorithm uses the relationship between the local words of the text itself (co-occurrence window) to sort the subsequent keywords, uses the extraction method of automatic summarization, and extracts the existing keywords and sentences in the document to form a summary, which does not need to learn and train multiple texts in advance, which is concise and effective. When using knowledge graph to expand the feature of Weibo, we need to match the keyword of Weibo text with the knowledge graph, and take the keyword of Weibo text as the name of entity reference. When searching in the knowledge graph according to the entity reference items, many entities with the same name but different background knowledge may be queried. At this time, it is necessary to calculate the similarity between different background knowledge and keyword context words. Select the background knowledge with the greatest similarity as the extended text of the keyword. In the TextRank algorithm, there are two manually set parameters: the length of the co-occurrence window \( L \) and the number of keywords topK. When generating candidate keywords, the co-occurrence relation is used to construct the edge between any two points. There is an edge between the two nodes only if their corresponding words co-occur in a window with a length of \( L \), and \( L \) represents the size of the window, that is, a maximum of \( L \) words. In this paper, the \( L \) value is 6 and the topK value is 5.

Use Word2Vec\textsuperscript{[9]} tool to generate word vector model. Word2Vec is a multi-layer neural network, which obtains the meaning of the word through the context information of the word and vectorizes the word. Compared with the high-dimensional and sparse One-hot coding method used in traditional natural language processing, the word vector trained by Word2Vec is low-dimensional and dense. There are two important models in Word2Vec: the CBOW model and the Skip-Gram model. The CBOW model calculates the probability of the occurrence of the word according to the context of the current word. On the contrary, the Skip-Gram model calculates the probability of several words appearing in its context according to the current word. The corpus used in this paper is “Baidu Encyclopedia”, “Sohu News” and novel text, with a total of 268G. Among them, “Baidu Encyclopedia” has about 8 million pieces of data, 26G; “Sohu News” has about 4 million pieces of data, 13G; the text of the novel has a total of 229G. The CBOW model is used to generate the word vector. The maximum distance between the current word and the target word in the sentence is 5, the dimension size of the output word vector is 128, the hs value is set to 1, and the hierarchical softmax is used. The min-count value is set to 10, which means that words with a frequency less than min-count will be ignored when filtering words.

2.3. Short text expansion
Weibo text extension based on knowledge graph. The keywords of the short text obtained by the TextRank algorithm are used as entity references. N text abstracts of this keyword are obtained from the Chinese general encyclopedia knowledge graph CN-DBpedia\textsuperscript{[10]}.CN-DBpedia is a large-scale general domain structured encyclopedia developed and maintained by the knowledge works laboratory of Fudan University. It mainly extracts information from the plain text pages of Chinese encyclopedia websites (such as Baidu encyclopedia, Interactive encyclopedia, Chinese Wikipedia, etc.). After filtering, fusion, inference and other operations, high-quality structured data is finally formed for machines and people to use. An entity may have many different meanings, and for ease of distinction, each entity in the CN-DBpedia has a short text description that contains the entity's category, topic, or explanation. For example, when querying with the entity denotation item name “Mask” in CN-DBpedia, the set of candidate entities is as follows: ['Sanitary products used to filter air into the mouth and nose', ‘mr. singing songs’].

The pre-trained Word2Vec model is used to transform keywords and their contextual words into word vector sets. For these n text abstracts, “jieba” word segmentation is used to deactive word processing, and then Word2Vec model is used to transform them into word vector sets. The similarity between the keyword context words and the n text abstracts is calculated according to the generated word vector set. Finally, the text summary with the highest similarity is selected as the extended text of
the Weibo text. Generate a short text extension matrix. Splice Weibo text segmentation, Weibo text keywords, the most similar target entity summary participle as the final Weibo extended text.

2.4. Text classification model
Traditional language models and text classification tasks regard text data as a time series with only one dimension, so cyclic neural networks are used to represent such data. In fact, the text data can also be regarded as an one-dimensional image, thus one-dimensional convolution neural network can be used to capture the association between adjacent words. The text classification model TextCNN based on convolution neural network is a classical text classification method based on deep learning. In this model, the features of the text are trained through the convolution neural network, and then the features of the text are input into the classifier to get the classification results. TextCNN model mainly uses one-dimensional convolution layer and timing maximization pool layer. When using TextCNN calculation, we first define multiple one-dimensional convolution kernels, and then use these convolution kernels to do convolution operations on the input respectively. Convolution kernels with different widths may capture the correlation of different numbers of adjacent words. All the output channels are sequentially pooled, and then the pooled output values of these channels are connected as vectors. Through the full connection layer, the linked vector is transformed into various kinds of output. The TextCNN model used in this paper is: Input layer, Convolution Pooling layer * 3, Splicing layer, Full Connection layer, Dropout, Full Connection layer, Softmax Classifier layer. The purpose of adding Dropout is to reduce the sensitivity to noise and prevent over-fitting caused by noise. In this paper, the parameters of the TextCNN model are set as follows: dimension of the word vector entered: 50; convolution kernel size: 3, 4, 5; convolution kernels: 256; activation function: ReLU; dropout_rate: 0.2; batch_size: 128; epochs: 100; the learning rate of every 50 epoch, is reduced to 0.1 of the original.

3. Experiment and analysis
In this paper, precision, recall, F1-measure and accuracy are selected to evaluate the classification results in each category. Macroscopic average score and weighted average score are selected to evaluate the comprehensive results of short text classification. In the classification results, T represents the number of samples of correct classification, F represents the number of samples of misclassification, P represents the number of positive samples, and N represents the number of negative samples, then there are four kinds of cases:

- True Positive (TP): Refers to the positive tuple that is correctly classified by the classifier.
- True Negative (TN): Refers to a negative tuple that is correctly classified by the classifier.
- False Positive (FP): Refers to a negative tuple that is mistakenly marked as a positive tuple.
- False Negative (FN): Refers to a positive tuple that is mistakenly marked as a negative tuple.

According to the above four situations, precision, recall, F1-measure, accuracy, macro average and weighted average are defined as follows:

- The precision rate (precision): It can be seen as a measure of precision, that is, the percentage that tuples marked as positive classes are actually positive classes. As shown in formula 3-1.
  \[
  \text{precision} = \frac{TP}{TP + FP} \quad (3-1)
  \]

- The recall rate (recall): It is a measure of completeness, that is, the percentage of positive tuples marked as positive. As shown in formula 3-2.
  \[
  \text{recall} = \frac{TP}{TP + FN} = \frac{TP}{P} \quad (3-2)
  \]

- The F1-Measure (F1-measure): It is used to comprehensively measure the recall rate and precision rate of text classification results, using the harmonic mean of precision and recall rate, as shown in formula 3-3.
  \[
  F1 - \text{measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (3-3)
  \]

- The accuracy (accuracy): It is the percentage of tuples correctly classified by the classifier. As shown in formula 3-4.
  \[
  \text{accuracy} = \frac{TP + TN}{P + N} \quad (3-4)
  \]
The macro average (macro avg): It averages the precision rate, recall rate and F1-measure of each category. The weighted average (weighted avg): It is an improvement of macro average, taking into account the proportion of the number of samples in each category in the total sample.

Weibo texts that do not use knowledge graph to expand the semantics are segmented by “jieba” and removed the stopped words and then put into the TextCNN model for training. 10% of the data is randomly selected as the test set, the classification accuracy of the test set is 0.62. Other evaluation metrics are shown in Table 1.

Table 1. The result of text classification with unexpanded semantics.

| Category           | Precision | Recall | F1-measure |
|--------------------|-----------|--------|------------|
| Science and technology | 0.71      | 0.64   | 0.67       |
| Epidemic           | 0.74      | 0.73   | 0.73       |
| Huawei             | 0.78      | 0.39   | 0.52       |
| Eight Hundred      | 0.83      | 0.67   | 0.74       |
| Constellation      | 0.86      | 0.67   | 0.75       |
| Tourism            | 0.70      | 0.47   | 0.56       |
| Parenting          | 0.74      | 0.56   | 0.64       |
| Food               | 0.72      | 0.63   | 0.67       |
| Depression         | 0.87      | 0.63   | 0.73       |
| Fitness            | 0.30      | 0.80   | 0.44       |
| macro avg          | 0.72      | 0.62   | 0.65       |
| weighted avg       | 0.72      | 0.62   | 0.64       |

The extended semantic text of Weibo texts is trained in the TextCNN model using knowledge graph. 10% of the data is randomly selected as the test set, and the classification accuracy of the test set is 0.74. Other evaluation metrics are shown in Table 2.

Table 2. The result of text classification with extended semantics by using knowledge graph.

| Category           | Precision | Recall | F1-measure |
|--------------------|-----------|--------|------------|
| Science and technology | 0.72      | 0.69   | 0.70       |
| Epidemic           | 0.85      | 0.80   | 0.82       |
| Huawei             | 0.86      | 0.58   | 0.69       |
| Eight Hundred      | 0.87      | 0.74   | 0.80       |
| Constellation      | 0.83      | 0.88   | 0.85       |
| Tourism            | 0.91      | 0.69   | 0.78       |
| Parenting          | 0.76      | 0.67   | 0.71       |
| Food               | 0.84      | 0.82   | 0.83       |
| Depression         | 0.93      | 0.70   | 0.80       |
| Fitness            | 0.38      | 0.81   | 0.52       |
| macro avg          | 0.79      | 0.74   | 0.75       |
| weighted avg       | 0.79      | 0.74   | 0.75       |

According to the above table, after the feature extension of Weibo text by using knowledge graph, the classification accuracy is improved by 0.12 and the weighted average of F1-measure is improved by 0.11. The experimental results show that the accuracy of Weibo text classification using knowledge graph is better than that of Weibo text classification directly.

4. Conclusion

In view of the lack of and non-standard text features of Weibo, this paper expands its text with the knowledge graph, and use the text classification model TextCNN based on deep learning to mine the deeper semantic information of the text. First of all, a total of 11311 Weibo texts under ten themes were crawled by web crawler technology. After that, we use the “jieba” word segmentation tool to segment the text, remove the stopped words, and use the TextRank algorithm to extract the keywords of the text. The keyword is matched with the knowledge graph as an entity reference item, and the entity that best...
matches the keyword is obtained by calculating the similarity. Finally, the text summary of the entity is taken as the extended text of Weibo text. The experimental results show that the method is effective in classifying Weibo short text data sets. However, this paper does not take into account that there are a large number of colloquial expressions, word abbreviations, misspellings and so on in Weibo text. Moreover, the Word2Vec pre-training model used in this paper has limited ability to understand the semantics of these texts. The next step is to make improvements on this basis.

References

[1] Yunjie Fan, Huailiang Liu. Research on Chinese short text Classification based on Wikipedia [J]. Modern library and information technology, 2012(03):47-52.
[2] Qi Jing, Liguo Duan, Aiping Li, Qian Zhao. Short text relevance calculation based on Wikipedia [J]. Computer engineering, 2018, 44(02):197-202.
[3] Xiwei Wang, Xinghua Fan, Jun Zhao. A Chinese short text classification method based on feature extension [J]. Computer application, 2009, 29(03):843-845.
[4] Xuemin Fu, Xudong Chen. Research on short text classification based on high-frequency word expansion [J]. Journal of Suzhou University, 2018, 33(07):90-92+96.
[5] Chaozhen Lu, Donghong Ji, Feifei Wu. Short text classification based on LDA feature extension [J]. Computer engineering and application, 2015, 51(04):123-127.
[6] Yongjun Hu, Jiaxin Jiang, Huiyou Chang. Chinese short text classification based on LDA high frequency word extension [J]. Modern library and information technology, 2013(06):42-48.
[7] Kim Y. Convolutional Neural Networks for Sentence Classification[C]. EMNLP. 2014:1746-1751.
[8] Mihalcea, Rada, Tarau, Paul. TextRank: Bringing Order into Texts[C]. Proc Conference on Empirical Methods in Natural Language Processing. 2004.
[9] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer ence, 2013.
[10] Bo Xu, Yong Xu, Jiaqing Liang, Chenhao Xie, Bin Liang, Wanyun Cui, and Yanghua Xiao. CN-DBpedia: A Never-Ending Chinese Knowledge Extraction System. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, pp. 428-438. Springer, Cham, 2017.