Research on hot news classification algorithm based on deep learning

Zaiying Wang, Bohao Song
Xi'an University of Science & Technology, Xi'an, Shaanxi, 710054, China
zying_wang@163.com, 136247580@qq.com

Abstract—Nowadays, the Internet development in full swing, has given rise to a variety of network applications, network news as the basic application in the information age, has been more and more customers favor, however, news website to report news type number many, every day the user focus on hot news, so I need to forecast the hot problem in the news, so in view of the study of this problem, this paper studies deep learning application in text categorization, combined with the characteristics of the news text, put forward the double Bi- Gated Recurrent Unit(GRU) + attention deep learning model to predict hotspots, and achieved good results.

Keywords: deep learning; Hot news forecast; Attention; Text classification

I. INTRODUCTION

The Internet is subtly changing the way of life of human beings, which has been transformed from the initial recreation to an indispensable part of People's Daily work and life [1]. Based on the 42nd statistical report on the development of Internet in China in July 2018, as of June 2018, the number of Internet users in China was 802 million, with 29.68 million new Internet users in the first half of the year, 3.8% more than in the end of 2017, and the Internet penetration rate reached 57%. 4.1 percentage points above the global average and 9.2 percentage points above the Asian average. The number of Internet users and the Internet penetration rate in China have maintained a steady growth trend. [2].

With the explosive output of network news, hot spot classification becomes indispensable. Nowadays, many text classification methods based on machine learning, such as naive bayes, Support Vector Machine(SVM), classification tree and vector-based features, provide good interpretability, but the semantics cannot be studied, so that the classification accuracy is not high. However, with the increase of corpora size and increase in the number of subfields, automatic classification has lately become more and more challenging, deep learning based on the text classification technology into people's attention, become very hot, this paper puts forward to assemble a new deep learning model, used in text categorization, namely BiGRU + attention, fully considering the context of the text data and focus problems, the experimental results show that compared with the traditional text categorization, this method has higher precision.

II. RELATED WORK

News prediction, of course, should focus on news report itself before use news prediction, the researchers extracted from known popularity of news history and news popularity of one or more characteristics of high relevance to predict the popularity of the upcoming news, According to existing studies, convolutional neural network (CNN) and long-term and short-term memory neural network (LSTM) are commonly used neural network layers in the field of natural language processing. Since December 2017, Google open-source attention is all your need [3], attention mechanism has also come into people's eyes.

In this paper, from the perspective of deep learning, study and put forward a new method of deep learning text classification, we use the depth study for text classification research, on the basis of using double BiGRU + attention mechanism, can be used to predict the popularity of the news, the first layer of the model using the word embedding layer, is the training sequence into words vector technology, the model output is the conditional probability distribution of words, it is said that the current word in the glossary corresponding to the probability of position words word vector is derived from the parameters of the model of training, Convolutional neural network (CNNs) is modeled according to the structure of the visual cortex, where neurons are not completely connected but are spatially distinct [4]. CNNs has a good effect on the generalization of target classification in image [5]. Recent work has used CNNs for text mining [6]. In the study closely related to the work of this paper, Zhang et al. [7] used CNNs to classify texts and fully connected DNN to provide character-level features. In any case, CNNs requires a large number of training sets. This article USES another basic deep learning architecture is recursive neural networks (RNNs), a variant of the output of a layer connected to its input, more care about semantic logic relationship between text, the next level, attention layer, looking for key words in the text key point, to its dregs, takes its essence, make the accuracy of classification is more, the last layer classification hot news and hot news. This paper introduces how to use deep learning method to classify hot spots. Compared with traditional machine learning and CNN, RNN has a higher accuracy. Deep learning method provides a flexible framework to generate better classification.

III. EXPERIMENTAL DATA

In this study, we use 8687 news from sohu website crawl, from October 20, 2015 to April 25, 2016, a total of fetching hot news and hot news, including financial news, social news, sports news and entertainment news, including hot news from the sohu news list of simplicity, (http://pinglun.sohu.com/), and the list of sohu news buzz buzz list according to the number of comments updated once every 12 hours, ten as hot news, a day

978-1-5386-6243-4/19/$31.00 ©2019 IEEE

before you grab the hot news from the corresponding column of random grab all the news, In order to avoid the imbalance between positive and negative samples, only 10 non-hot news items were captured every day. The captured data included the news title, body, published events and the number of comments within 12 hours. Table 1 counted the number of hot news and non-hot news in different columns. alysis method widely used in system control, decision making, fuzzy mathematics, artificial intelligence, entropy theory and other aspects of social science research [4~6]. Set pair refers to the pair of two set with a certain relationship.

Table I. The number of data

News type	Hot news	Non-hot news	Total number
Financial news	1228	929	2157
Sports news	1265	967	2232
Social news	1281	969	2250
Enterta inment news	1198	850	2048

Figure 1 and Figure 2 shows the results of the crawl.

Fig.1. HOT news

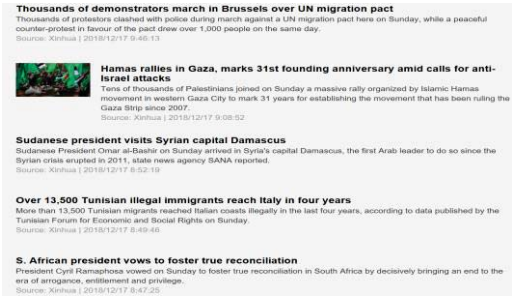

Fig.2..NON-HOT news

When we get these data, we need to preprocess the data Word vector representation refers to low-dimensional real vector to represent words.

Training is used in this article term vectors Word2Vec, this is Google open source training in natural language processing tools, it will all words in the corpus with vector said, this is the distance between the word and the word can be quantitative measurement, through distance measure Word2Vec use Dristributed representation instead of a One - hot model, through the training to map each word to a specified low dimension vector, which can be common statistical methods to study the relationship between the words, The use of vector representation of words in documents can fully consider the semantics and word order of words. Word2Vec is often used for some problems in the field of natural language processing, such as machine translation, relational mining, part-of-speech tagging and so on. It has high computational efficiency and can train a large number of corpus. Word2Vec generally adopts neural network without hidden layer for training, and the input and output are defined as Skip-Gram and CBOW.

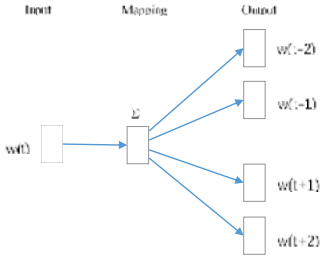

Fig.3. Skip-Gram model structure

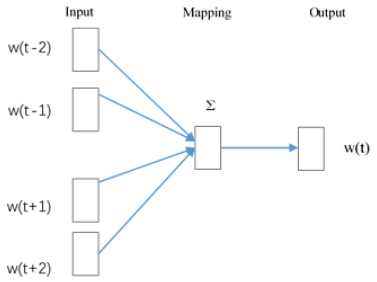

Fig.4. CBOW model structure

IV. BASIC MODEL

CNN: Convolutional Neural Network (Convolutional Neural Network, CNN) in the field of image processing and image recognition has achieved good results, is now also widely used in the field of natural language processing. CNN is a composed of multiple convolution and pooling layer neural network, on the abstraction of the input signal is off work by multiple convolution and pooling layer, finally, in the full connection layer abstract features to the output target mappings, and the difference between ordinary neural network (CNN), This can greatly reduce the number of model parameters and simplify the complexity of the model, thus reducing the risk of overfitting.

RNN: RNN (Recurrent Neural Network) are widely used in the field of natural language processing, RNN is a kind of to setting up the Neural Network modeling of sequence data, it was assumed that a sequence of the current output is not only

has relationship with the current input, output has relationship with before, RNN memory in front of the information and use these information to calculate the current output. However, due to the long text and the long distance between nodes, it is easy for gradient explosion or gradient disappearance to occur in the calculation of RNN, so that RNN cannot be relied on for a long time. LSTM (Long short-term Memory Neural Network) solves the problem of long-term dependence and constantly changes the weight of self-loop by adding input gate, output gate and forgetting gate.and we can seen in Fig5 and Fig6.

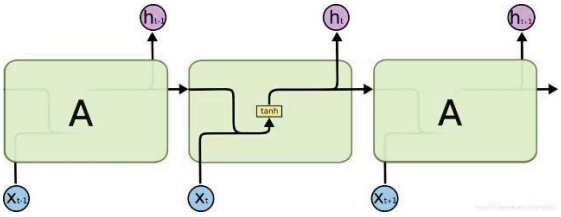

Fig.5. Normal RNN's structure

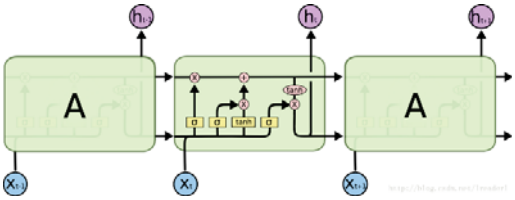

Fig.6. LSTM's structure

According to the Fig 6,The flow of LSTM is as follows: it mainly includes three gates, which are called forgetting gate, input gate and output gate respectively. From the forgotten on the left of the door, then decide what we want from discarded cell state information, the next decision we should state what information is stored in cells, known as the "input layer" to determine what value, we will update the next a tanh layer create candidate vector, the vector will be added to the cell state, in the next step, we will combine the two vectors to create an update value. Finally, we use the "output gate" to determine what we want to output, and only the parts we decide to output.

In this paper, we have chosen a variant of LSTM -- GRU (Gated Recurrent Unit) -- which has maintained the effect of LSTM and simplified its structure. It is a trend that has become more and more popular. The GRU model is as follows. The GRU has only two doors, namely, update door and reset door, Update gate is used to control the degree to which the state information of the previous moment is brought into the current state. The higher the value of update gate is, the more the state information of the previous moment is brought in. Reset doors are used to control the degree to which state information at the previous moment is ignored. The smaller the reset door value, the more ignored it is. The effect of GRU is little different from that of LSTM, but GRU can save a lot of time.

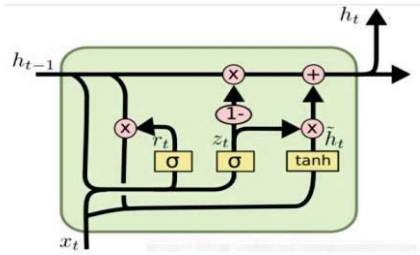

Fig.7.GRU's structure

In addition we want to better solve the semantic understanding, so we adopt the bidirectional RNN, two-way RNN assumes that the current output not only associated with the sequence of before, and it is associated with the sequence of later, just meet our for the text of news, is composed of two RNNs stack up and down together, the output from these two RNNs hidden layer is decided by the state. Similarly, when we superimpose two GRUs, we realize two-way GRU.

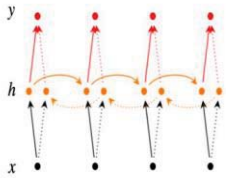

Fig.8. BIRNN

And the final we choose attention mechisam. The Attention mechanism solved the problem of long-distance dependence and let the decoder find the text that needs to be paid Attention to in the input text by itself. Attention model to imitate Attention mechanism of human brain, for example, let's see a picture, although we can see the whole picture of the painting, but when we are thorough careful observation, is really only a small eyes focused, the brain when people mainly focus on this piece of design, that is to say, the brain to the Attention of the whole picture and is not balanced, there is a weight, this is in deep learning Attention model, Attention mechanism is the core idea is used in classifying text context is different, Such considerations are obviously more reasonable.

V. THE EXPERIMENTAL ARRANGEMENT

*1)*Merge CNN Kim et al. achieved the text classification effect by combining two CNN models [9], and applied this result to the prediction of network hot news, built two convolutional neural networks, extracted features respectively and fused the two features, and then used the full connection layer and softmax as the activation function to obtain the probability that the news belongs to hot or non-hot.

*2)*CNN_LSTM Firstly, the text after word segmentation and word stopping is represented by word sequence, and the word vector representation of the text is obtained through embedding layer. Then the convolutional neural network layer and pooling layer are used to abstract the text word vector feature into text feature 1. Then, using the network layer of long and short term memory, text feature 1 was extracted to obtain text feature 2. Finally, using the full connection layer and softmax as the

activation function to get the probability that the news belongs to hot or not hot, then the popularity of the news depends on the corresponding class standard with high probability.

*3)*Double layer BiGRU +attetntion Model, the diagram below is divided into two parts, respectively when the sentence modeling and modeling of document, the aforementioned model in modeling of a sentence, through the study of the characteristics of the words in the sentence combination, form a sentence vector, from sentence chapter modeling at the next higher level, this paper will be done through two-way GRU helped first term vectors encoding, combining the output of hidden layers and attention mechanism, the formation of the characteristics of the sentence, said after each sentence is equivalent to one word, repeat the previous step to sentence the modeling process, complete the sentences to document modeling, One of the biggest benefits of attention is that it gives an intuitive sense of how important sentences and words are to categorizing. Finally, softmax layer was connected to get the popularity of the news category.

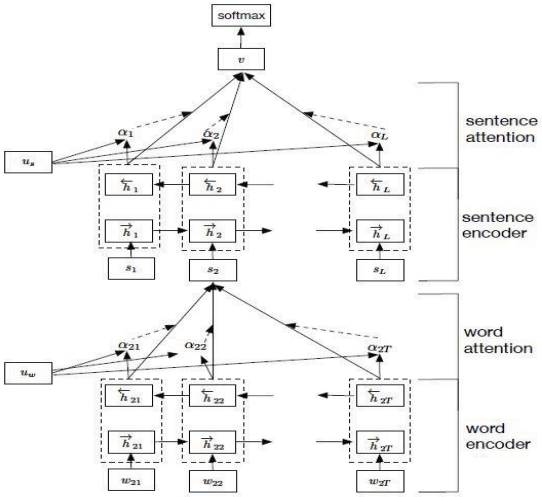

Fig .11. BiGRU +attention model

DESIGN AND RESULT ANALYSIS

The evaluation index

Use 10-fold cross-validation. The evaluation indexes are accuracy rate (P1), recall rate (R1), F value (F1) of hot news, accuracy rate (P0), recall rate (R0), F value (F0), macro average F value (F) and overall prediction accuracy (Acc) of non-hot news.

Remember *TP* refers to the number of hot news correctly predicted, *TN* refers to the number of non-hot news correctly predicted, *FP* refers to the number of news actually non-hot but predicted to be hot, and *FN* refers to the number of news actually hot but predicted to be non-hot.

1) Accuracy

The prediction accuracy refers to the ratio between the number of accurately predicted news and the number of news in the test set, which can comprehensively evaluate the prediction effect. The calculation method is as follows

$$\text{Acc}=\frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

2) Precision

The accuracy rate is for the prediction results. It represents the proportion of hot news (non-hot news) correctly predicted to hot news (non-hot news) in the prediction results, and measures the accuracy rate of the prediction results. The formula for calculating the accurate rate P1 of hot news is as follows:

$$\mathrm{P}_1=\frac{TP}{TP+FP} \tag{2}$$

The numerator is the number of hot news correctly predicted, and the denominator is the number of hot news in the predicted result, P1 is the accurate rate of hot news. Similarly, the accuracy rate P0 of non-hot news is:

$$\mathrm{P}_0=\frac{TN}{TN+FN} \tag{3}$$

the overall average precision rate is:

$$\mathrm{P}=(\mathrm{P}_0+\mathrm{P}_1)/2 \tag{4}$$

3)Recall

Recall rate refers to the proportion of correctly predicted hot news (non-hot news) in the original sample, which measures the recall rate of the predicted results. The calculation formula of recall rate R1 of hot news is as follows:

$$\mathrm{R}_1=\frac{TP}{TP+FN} \tag{5}$$

The numerator is the number of hot news correctly predicted, the denominator is the number of hot news in the test set, and R1 is the recall rate of hot news. Similarly, the recall rate of non-hot news R0 is:

$$\mathrm{R}_0=\frac{TN}{TN+FP} \tag{6}$$

The overall average recall rate is:

$$\mathrm{R}=(\mathrm{R}_0+\mathrm{R}_1)/2 \tag{7}$$

4) f-measure

In order to balance the accuracy rate and recall rate, F value is used to measure the predicted results. F value F1 of hot news is calculated as follows. Similarly, F value F0 of non-hot news can be obtained.

$$\mathrm{F}_1=\frac{2\mathrm{P}_1\mathrm{R}_1}{\mathrm{P}_1+\mathrm{R}_1} \tag{8}$$

F value is one of the most common measurement methods of comprehensive measurement accuracy and recall rate, but it can only represent the classification effect of a certain class in the classification process. In the prediction of network hot

news in this paper, we should not only pay attention to the recognition effect of hot news, but also pay attention to the recognition effect of non-hot news. Only when the recognition effect of both kinds of news is achieved, can the prediction of network hot news have higher application value. Therefore, using the macro average F value to measure the overall prediction effect of the popularity of network news:

$$F=\frac{F_0+F_1}{2} \tag{9}$$

In order to evaluate the prediction performance more objectively, we use the ten-fold cross-validation method. The so-called 10-fold cross-validation means that the sample data set is divided into 10 parts on average and denoted as, which will be used as the test set, and the remaining 9 parts as the training set to obtain the evaluation index. Then select any data set other than the test set and the rest as the training set to obtain the evaluation index; At this time, each data set is used as a test set to test the classification results, and 10 groups of evaluation indicators are obtained. The average value of these results is taken as the final evaluation index of classifier performance.

Experimental design and result

BiGRU+attention mentioned above was compared with CNN-LSTM, mergeCNN and deep learning models through experiments. Some traditional deep learning models were implemented with Keras toolkit. Tensorflow was selected as the underlying library in this paper.

Table II .The result of experimental

model	P1	R1	F1	P0	R0	F0	F	Acc
CNN-LSTM	0.813	0.74	0.771	0.693	0.765	0.723	0.747	0.751
mergeCNN	0.688	0.723	0.699	0.788	0.749	0.764	0.731	0.737
doubleBiGru+attention	0.799	0.816	0.805	0.751	0.721	0.730	0.768	0.775
doubleCNN	0.697	0.622	0.636	0.748	0.777	0.751	0.694	0.712
doubleLSTM	0.452	0.532	0.488	0.727	0.783	0.734	0.611	0.674

The bar chart of experimental results is as follows:

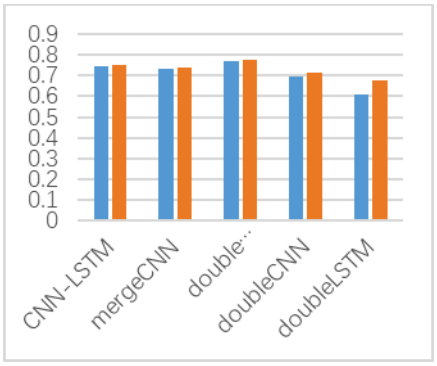

Fig .12 . The result of image

VI. CONCULSIONS

In this paper, we summarized the development background and current situation of the hot news network, grasping system design and implementation of network news, and draw lessons from the inside of the machine translation knowledge, using BiGRU + attention deep learning model with the traditional common deep learning model experiment, the experimental results show that the predicted results of our model is superior to the traditional model of deep learning model was applied to prediction of network hot news through a simple adjustment to reach the effect of close to 80, for the use of deep learning predicted popularity laid the foundation of network news. In the following work, we can try to use Bert word vector to design more complex deep learning network to achieve better prediction effect.

ACKNOWLEDGMENT

This work was supported by the Key Research and kernel is 5, the number of convolution kernel is 400, the Development Project of Shaanxi Province, China No. 2017GY-157.

REFERENCES

[1] R.J. Ellison, D.A. Fisher, and R.C. Linger, Survivable Network System: An Emerging Discipline. Technical Report, CMU/SEI-97-TR-013, Carnegie Mellon University, 1997.

[2] R.J. Ellison, D.A. Fisher, and R.C. Linger, “Survivability: Protectiong Your Critical Systems,” IEEE Internet Computing, vol. 3, issue 6, pp. 55-63, 1999.

[3] D. Dan, Y.Q. Zhang, “Research on Definition of Network Survivability,” Journal of Computer Research and Development, vol. 43(Suppl.), pp. 525-529, 2006.

[4] K.Q. Zhao, Set pair analysis and its preliminary applicatio. Hangzhou, Zhejiang Science and Technology Press, 2000.

[5] K.Q. Zhao, A.L. Xuan, “Set Pair Theory - A New Theory Method of Non-Define and Its Applications,” System Engineering, vol. 14, issue 1, pp. 18-23, 1996.

Y.L. Jiang, C.F. Xu, “Advances in Set Pair Analysis Theory and its Applications,” Computer Science, vol. 33, issue 1, pp. 205-2