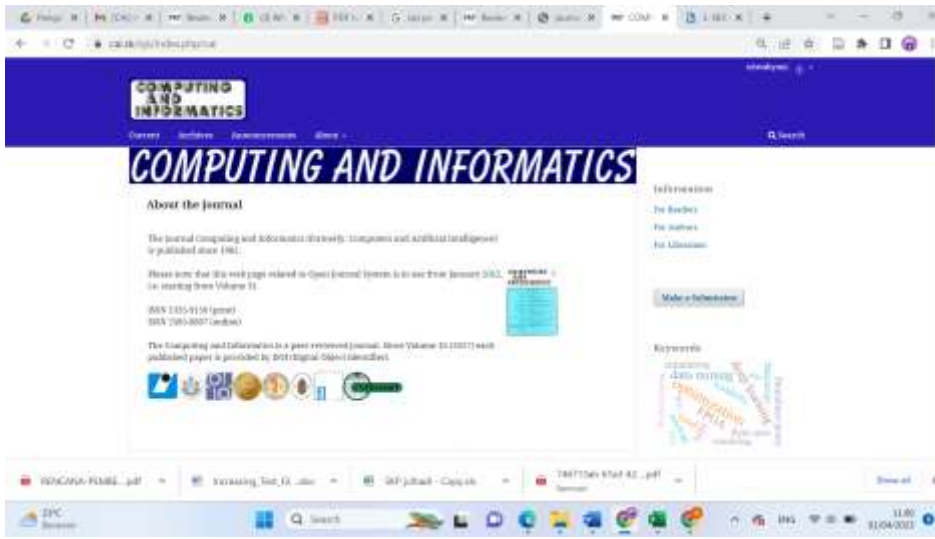
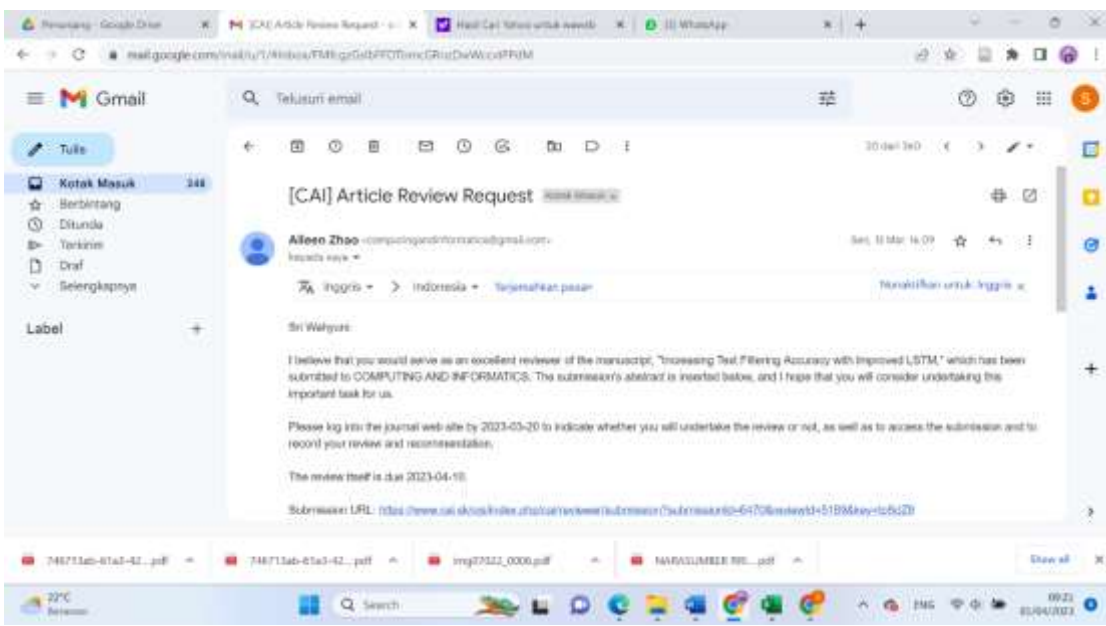
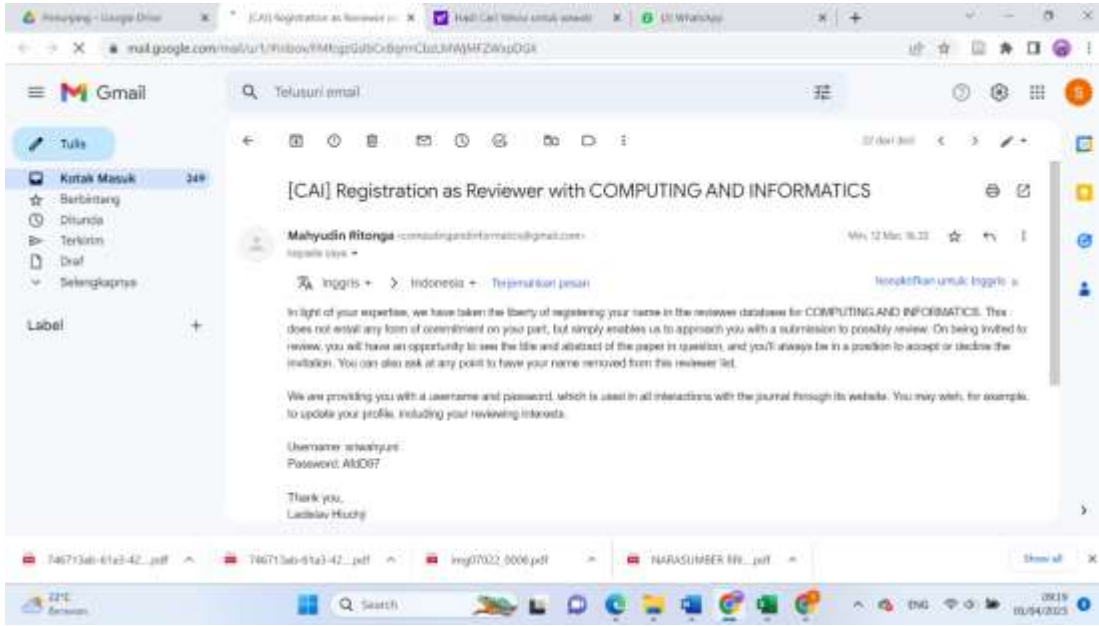

COMPUTING AND INFORMATICS
About the journal


Gmail
Tulis
Kotak Masuk
Berbintang
Ditunda
Terkirim
Draf
Selengkapnya
Label
[CAI] Article Review Request
Aileen Zhao
Sri Wahyuni
I believe that you would serve as an excellent reviewer of the manuscript, "Increasing Text Filtering Accuracy with Improved LSTM," which has been submitted to COMPUTING AND INFORMATICS. The submission's abstract is inserted below, and I hope that you will consider undertaking this important task for us.
Please log into the journal web site by 2023-03-20 to indicate whether you will undertake the review or not, as well as to access the submission and to record your review and recommendation.
The review itself is due 2023-04-10.
Submission URL:



INCREASING TEXT FILTERING ACCURACY WITH IMPROVED LSTM

Wei Dang, Ligao Cai

School of Automation
University of Electronic Science and Technology of China
Chengdu 610054, China
e-mail: wei.dang.cn@gmail.com, 13297087436@163.com

Mingzhe Liu

School of Data Science and Artificial Intelligence
Wenzhou University of Technology
Wenzhou 325000, China
e-mail: liumz@cdut.edu.cn, (Correspondence)

Xiaolu Li

School of Geographical Sciences
Southwest University

Chongqing 400715, China
e-mail: xliswu@swu.edu.cn

Zhengtong Yin

College of Resource and Environment Engineering
Guizhou University
Guiyang, Guizhou 550025, China
e-mail: ztyin@gzu.edu.cn

Xuan Liu

Computing and Informatics, Vol. 32, 2013, 1001–1025, V 2023-Feb-27
School of Public Affairs and Administration
University of Electronic Science and Technology of China
Chengdu 611731, China
e-mail: liuxuan@uestc.edu.cn

Lirong Yin

Department of Geography and Anthropology Louisiana State University
Baton Rouge, LA 70803, USA e-mail: yin.lyra@gmail.com
(Correspondence)

Wenfeng Zheng

School of Automation
University of Electronic Science and Technology of China
Chengdu 610054, China
e-mail: winfirms@uestc.edu.cn (Correspondence)

Abstract. How to eliminate useless information in the vast network information and retain effective information is a problem that needs to be continuously explored in the field of deep learning. This paper conducts text classification on the network evaluation frequently encountered in daily life, mainly to screen out the meaningless comments published by Internet users, to have access to more useful information. In this paper, a text filtering model was constructed based on word vector and Long Short-Term Memory (LSTM) and improved by adding Deep Averaging Net-works (DAN) and convolutional neural network (CNN). The major improvement of the LSTM & DAN model was to retain the original word vector information and to improve the accuracy of the text classification model without increasing hy-perparameter and model structure complexity. The LSTM & CNN model mainly combines the advantages of convolutional neural network in exploring the deep in-formation of text, which was an improvement over the original LSTM. It was proved by experiments that this improvement is meaningful. Compared with the shallow neural network, the accuracy has been greatly improved.
Keywords: Texting filtering, LSTM, word vector, CBOW, dropout, CNN, DAN
Mathematics Subject Classification 2010: AB-XYZ

1 INTRODUCTION

With the popularity of Internet technology and the enthusiasm of consumers for on-line comments, the Internet has been filled with a large amount of comment data [1]. Internet users suffer from problems such as poor comment quality and information overload when they use these comments to make purchase decisions. While the Internet gives convenience and benefits to people, it has brought some disadvantages such as being unable to contact physical products during consumption, difficulties in remote identification, and possible mismatch between description and physical products [2, 3, 4, 5, 6, 7], so that consumers have to help themselves to make decisions by understanding the comments of other users before making consumption. However, with the rapid increasing of network evaluation quantity and the diversity of evaluation content, it becomes more difficult for users to obtain evaluation infor-mation helpful to them [8, 9, 10]. It is difficult to get really valuable data from a large amount of comment data by manual recognition alone, so how to get the com-puter to automatically screen out the valuable and worthless comments has become an urgent problem to be solved [11, 12, 13, 14]. Therefore, it has important research value for text content filtering [15].

Text automatic classification [16] is a method for attributing similar texts to the same category, using a valuable classification model trained through the existing text collections, and then applying the model to an unclassified dataset to make the same kind of text belong to the same category. This is an efficient text classifica-tion method that makes it more accurate to achieve categorical searches in massive amounts of data.

When classifying the text, it is necessary to convert the text into a form that can be recognized by the machine first, which can ensure the subsequent classifier proceeds effectively. Currently, the mainstream method is to vectorize the text. In recent years, due to the rising of deep learning, it becomes more obvious in the advantage of neural network in text feature extraction, which can be used to excavate the feature information of text at the semantic level, even to visualize the text. The word2vec [32] proposed by Google first shows a method of transforming words into vectors, which has been widely used in the field of natural language processing [18].

In the traditional neural network model citeBashar, the hidden layer connects the input layer and the output layer connects the hidden layer with no connection between the nodes in each layer. It is difficult for a neural network to deal with se-quence problems like natural language. But based on the traditional neural network, the convolutional neural network (CNN) and the deep neural network (DNN) [3] can solve this problem.

N. Widiastuti has raised a series of questions about CNN in the field of NLP [20]. After that, the improved algorithm based on CNN was gradually mined and applied to image processing [?] and other areas. At the same time, deep neural network has also been used in many fields.

Therefore, the network [22] was selected of both short-term and long-term Memory (Long Short Term Memory, referred to as LSTM) in this study. LSTM is a kind

of network model for processing sequence information, which has been widely used in areas such as speech recognition, emotion classification, and instant translation. At the same time, it can also become a more complex network structure together with the other network model, which is one of the hotspots in the field of deep learning. A large number of practices have shown that LSTM has better advantages than other models in the modeling of sequential data. It can capture long-term con-text association in the sequence, which is very powerful in the fitting of non-linear relations and suitable for the modeling of natural language data. The main work of this paper is as follows:

1)Data capture and calibration. 53728 comments of “Wolf Warrior 2” from the Douban movie were grabbed and labeled according to their value.

2)Text feature extraction (also known as text vectorization). According to the specific research content, this paper improves the construction method of text syn-thesis vector in the shallow neural network model. This method considers both the text word frequency vector and the text semantic vector, and then combines both to obtain the text vectorization method with the highest accuracy by experiments for constructing a text filtering model.

3)Construct a text filtering model based on LSTM neural network.

4)Analyze the limitations of the results. These two improved networks have been proposed based on the LSTM network-LSTM & DAN and LSTM & CNN, the classification performances of which have been analyzed.

2 DATASET

2.1 Data scarping

This study selected the online reviews of the movie ”Wolf Warrior 2” from an in-fluential Chinese film website ”Douban Movie” as a data sample, wrote a python crawler code to crawl the reviews, and obtained a dataset of 53728 comments.

2.2 Data Pre-processing

To facilitate classification, the data need to be manually labeled. To avoid bias from working alone, we divided the data into multiple batches, each of which was cali-brated by three staff members. The final value of the calibration is then determined according to

the principle of majority. The specific calculation method is as the following formula (1):

$$\text{brackets} = \left[\frac{\sum_{i=1}^{3} V_i}{2} \right] \quad (1)$$

S is the final calibration value, [. . .] is the floor function, is a certain calibration value given by staff.

To make text data easier for subsequent calculations, the labeled data should be filtered and classified. After filtering out some comment data that has no meaning,Increasing Text Filtering Accuracy with improved LSTM 1005

then select 40,000 data, of which the data calibrated to 1 or 0 were 20,000 respec-tively. Randomly selected 15,000 data from each dataset as the training sample, and the other two sets of 5,000 data each as the test sample.

Since there is no space in the middle like English words, Chinese text needs to be word segmented, this study used stuttering participles (python version) and combines the “stop thesaurus table” (filtering meaningless words) to perform word segmentation processing of the above text data.

3 METHODS

This paper mainly studies the performance of LSTM [23, 24, 25, 26, 27] and its improved network [28, 29] in text value classification. The input of the LSTM network was a word vector, the collected movie comment data was pre-processed to get the word vector in the dataset.

The experimental subjects were the calibrated movie review data sets. The training set was 15,000 positive samples, and 15000 negative samples. 5000 positive samples in the test set, and 5000 negative samples as well. The input was a vector of words trained on the dataset, and the word vector dimension was chosen to be 300.

3.1 Word vector building

To let the computer understand the meaning of each word and dig out words with similar semantics, the first step is to digitize the words. For example, there is a text ”I like basketball”. There are three components in this sentence, respectively ”I”, ”like”, and ”basketball”. A very intuitive method is to use a one-dimensional vector, such as [1, 0, 0, 0, . . . , 0] to represent ”I”, with [0, 1, 0, 0, . . . , 0] to indicate ”like”, use [0, 0, 0, 1, . . . , 0] to represent ”basketball”. If there are 10,000 words in the vocabulary specified in advance, then each word vector is 10000 dimensions, such as the word vector ”I”, except that the first position is 1, and the rest are all zero. This word vector representation method is one-hot coding, which is relatively simple and intuitive. Words with certain connections are not independent of each other. This method has certain problems. For example, the words ”programmer” and ”programming” have a very close relationship, but ”programmer” and ”basketball” are not closely related. At the same time, the dimensions of each vector are too large and sparse, so one-hot coding is not a good representation.

Previous researchers have proposed many different types of models to estimate the continuous representation of words, including the well-known Latent Semantic Analysis [30] (LSA) and the Latent Dirichlet Allocation [31] (hereinafter referred to as LDA). The word2vec proposed by Mikolov is more concerned with the use of neu-ral networks to learn the distributed representation of vocabulary. Its performance is significantly better than LSA, which can better maintain the linearity between vocabulary; in addition, for big data sets, the calculation of LDA is too heavy.

Tomas Mikolov’s paper [32] published in 2013 proposed a faster and better way to train word vectors, namely the Continuous Bag of Words Model (CBOW). In the same year, the Google team proposed a simple and efficient representation of word vector word2vec based on CBOW. This method can effectively train millions of dictionaries and hundreds of millions of data sets, which would result a very good measure of the similarity between words and words.

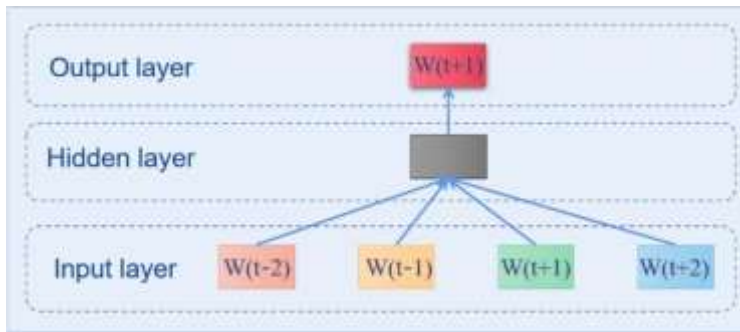


Fig. 1. The strucure of CBOW

This paper used CBOW to train the acquisition of word vectors. The general framework of the continuous word bag model is shown in figure1. The essence is to map the word vector of the original form to one-hot into a low-dimensional dense word vector through a three-layer neural network. The input is a few words in the context of a word, and the output is the word. After repeated iterative training of a large number of samples, a low-dimensional vector representation of each word is finally obtained.

This paper uses the 300-dimensional word vector obtained by CBOW training.

3.2 Construction of neural network models

To find a model with higher accuracy, this paper builds 3 network models based on LSTM: LSTM, LSTM & DAN, and LSTM & CNN. The two networks based on LSTM network - LSTM & DAN, LSTM & CNN were proposed to improve the accuracy of text value classification.

3.2.1 The structure of LSTM

Hochreiter and Schmidhuber first proposed the LSTM [33], a variant of recurrent neural network (RNN). A long-time delay process is added to the network so that the state element can keep error transmission continuously. The traditional model is shown in figure 2. There is only one tanh layer inside each unit.

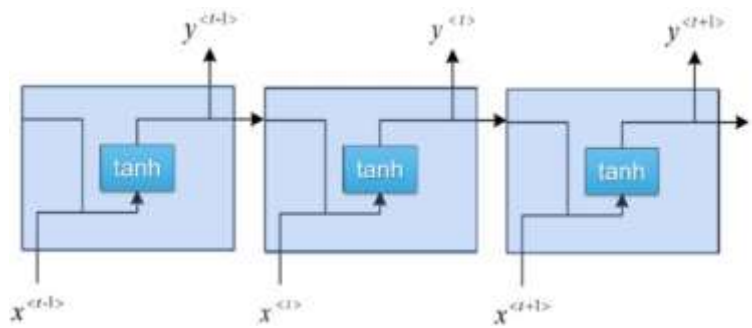


Fig. 2. Basic RNN model

While the overall structure of LSTM is similar to that of RNN, the only difference is that it no longer only uses a single tanh layer but adds three gate control units. As shown in figure 3, it is a minimum unit of LSTM [34].

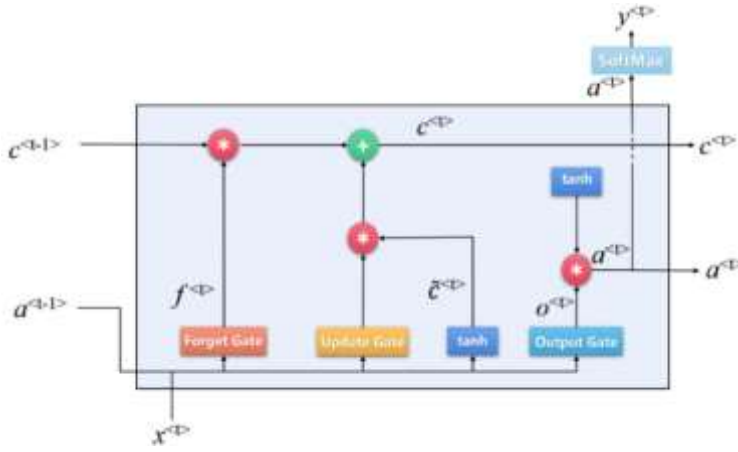


Fig. 3. LSTM unit struture

The relationship between each parameter is calculated by formula 2:

$$
\begin{aligned}
\tilde{c}^{\,t} &= \tanh W_c \; a^{t-1}, x^{t} + b_c \\
i^{t} &= \sigma W_u \; a^{t-1}, x^{t} + b_u \\
f^{t} &= \sigma W_f \; a^{t-1}, x^{t} + b_f \\
o^{t} &= \sigma W_o \; a^{t-1}, x^{t} + b_o \\
c^{t} &= i^{t-1} * \tilde{c}^{\,t} + f^{t} * c^{t-1} \\
a^{t} &= o^{t} * \tanh c^{t}
\end{aligned}
\tag{2}
$$

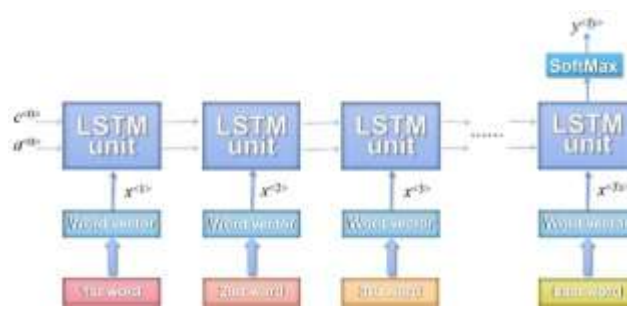
The most important part of LSTM is the control of the state of each unit. The control of long-term state can be regarded as the realization of three control switches. In each unit, there are three gate structures to protect and control information. Respectively for forget gate, update gate and output gate.

The forgotten door [35] is used to determine the retention of the unit state of the t-1 time step transferred to the t time step. Sigmoid is generally selected as the activation function. The input of the forgotten door is the input of the current time step and the output of the hidden node of the previous time step, which means that the forgotten door can implement the control of information.

The update gate is mainly used to control the input information. The input state of the current time step is determined by the output of the previous time step and the input function of the current time step, and the proportion of the newly added information can be controlled by the update gate.

The output gate is used to control the impact of long-term information on the output of the current time step. The output of LSTM is determined according to the output gate and unit state.

Combine the above model with the word vector constructed in the previous part of the paper to form the entire model. The entire model structure is shown below:



3.2.2 The structure of LSTM & DAN

Considering the input of LSTM model with word vector, and DAN model with certain advantages in input when using word vector for text classification, therefore, this section has improved the LSTM model by combining the characteristics of DAN model with the LSTM model.

(1)DAN DAN [36] is an approach proposed by Mohit Iyyer et al. at the ACL, an international top-level conference in 2015.

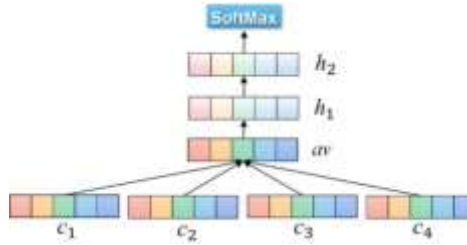


Fig. 5. DAN model structure

DAN is a text classification method proposed by integrating model operation time and accuracy. It is a completely disordered neural network, and the order of words in sentences has no influence on it. Its depth can be arbitrary, which is a feature that also allows it to capture subtle changes in the input. Its model is shown in figure 5. The calculation formula of each item is as follow:

$$av = \sum_{i=1}^{4} \frac{c_i}{4} \tag{3}$$

$$h_1 = f(W_1 \cdot av + b_1)$$

$$h_2 = f(W_2 \cdot h_1 + b_1)$$

Where c_1, c_2, C_3, c_4 is the word vector of each word, av is the average of four word vectors, h_1 is the first hidden layer, h_2 is the second hidden layer, and the number of neurons in each hidden layer is consistent with the dimension of the word vector. W_1, b_1, W_2, b_2 are the parameters to be learned. The last layer is classified by softmax or sigmoid. This is a binary classification problem, so the last layer is classified by sigmoid.

(2)Dropout algorithm

Alex Krizhevsky et al. proposed Dropout [37, 38, 39] algorithm in 2012. The Dropout method can be used to solve the overfitting problem. After using

the berserk method, the original network structure can become ”thin” and the whole model is more adaptable. As shown in figure 6, this is a process of using the Dropout method in neural network:

1010 I. Ja˘s˘sov´a, M. Tak´a˘c, I. Ja˘s˘sov´a, M. Tak´a˘c, I. Ja˘s˘sov´a, M. Tak´a˘c, I. Ja˘s˘sov´a, M. Tak´a˘c, I. Ja˘s˘sov´a, M

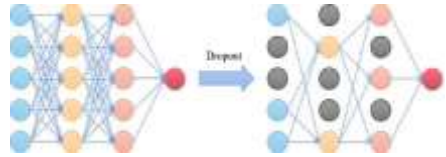


Fig. 6. The Diagrammatic sketch of Dropout Method

The main purpose of this algorithm is to prevent overfitting. Many complex neural networks usually have two disadvantages: time consuming and easy overfit-ting. Overfitting is a common problem of many models. If the trained model is overfitting, the result in application will not be ideal. To solve this problem, pre-vious researchers might train multiple models at the same time, and then use the method of model integration to get the result. However, this kind of method has problems of long learning cycle, slow test, and time consuming, which are hardly to be solved at the same time.

In this section, the Dropout method was applied to the input layer of the DAN model. Before finding the average word vector, the Dropout method was used to invalidate some words in the text randomly, that is, each word vector was set to 0 according to a certain probability p. The model structure after adding the Dropout is shown in figure 7:

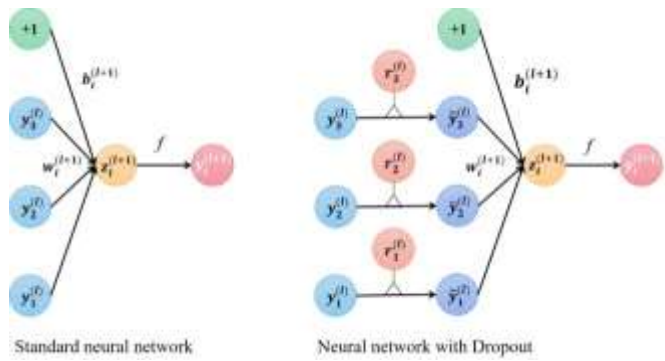


Fig. 7. Dropout comparison chart

For standard neural network:

$$z_i^{(l+1)} = w_i^{(l+1)} y^{(l)} + b_i^{(l+1)}$$

$$y_i^{(l+1)} = f\left(z_i^{(l+1)}\right)$$

For neural network with Dropout:Increasing Text Filtering Accuracy with improved LSTM 1011

$$\begin{aligned} r_j^{(l)} &\sim \text{Bernoulli}(p) \\ y^{(l)} &= r^{(l)} * y^{(l)} \\ z_i^{(l+1)} &= w_i^{(l+1)} y^{(l)} + b_i^{(l+1)} \\ y_i^{(l+1)} &= f\left(z_i^{(l+1)}\right) \end{aligned} \tag{5}$$

The Dropout method may fail some very important word vectors, but it can improve the accuracy of the model. This is because the number of words that are critical to label prediction is often smaller than the number of words that are irrelevant. For example, in the emotional analysis task, neutral words are often the most common. So there's a very high probability of removing extraneous words from the text and reducing the effect of these word vectors on the whole model.

(3)The model of LSTM & DAN

First, the Dropout method was used for all word vectors, with each word inval-idate with a certain probability, and then the average was taken as the last part of the last input.

Second, the output of each LSTM unit was used by the Dropout method to invalidate it with a certain probability, and then the average was taken as the second half of the final input.

The model framework is shown in the following figure 8:

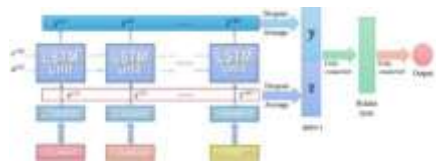


Fig. 8. LSTM & DAN model structure

3.2.3 The structure of LSTM & CNN

When using the traditional LSTM model to handle text classification, only the output of the last unit is kept, and then a full connection layer is connected for classification [40, 41]. For this reason, after a complete LSTM iteration, the output of the last remaining unit inevitably loses the previous part of information, which leads to overfitting of the model on the training set. To

solve this problem, this study proposes an improved model of LSTM with the combination of CNN after considering not only the output of the last unit, but also combining CNN with the feature of preventing overfitting.

The output of each unit of LSTM is kept with the consistency of the dimension of output vector of each unit with the dimension of input vector, and then the
012 I. Ja˘s˘sov´a, M. Tak´a˘c, I. Ja˘s˘sov´a, M. Tak´a˘c, I. Ja˘s˘sov´a, M. Tak´a˘c, I. Ja˘s˘sov´a, M. Tak´a˘c, I. Ja˘s˘sov´a, M

output vector is convolved. The characteristics of CNN model are used to explore the hidden relevance between words, and then to improve the effect of the whole model.

(1)CNN CNN is a special deep learning network structure inspired by biology with its core point of convolution operation. With the deepening of CNN research, problems such as text classification in natural language processing and software defect prediction in software engineering data mining are tried to be solved by using convolutional neural network, which would obtain better prediction results compared with traditional methods and even other deep network models.

Fig. 9. CNN model structure

As shown in the figure below, for CNN, its input data is the original sample form without any artificial processing, followed by numerous operating layers stacked on the input layer. As a whole, these operational layers can be regarded as a complex function f_C NN. The final loss function is composed of data loss and regularization loss of model parameters. The model training is to update the parameters of the model under the driving of the final loss and propagate the error back to each layer of the network.

(2)The model of LSTM & CNN When classifying using LSTM alone, the model overfitted as the number of iterations increases. To solve this problem, this study considers not only retaining the output of the last unit, combined with the charac-teristics of CNNs to prevent overfitting, and proposes a model that combines CNNs to improve LSTM.

Keep the output of each cell of the LSTM, and make the vector output of each cell consistent with the dimension of the input vector, and then convolute the output vector. Use the characteristics of the CNN model to explore the hidden correlation characteristics between words, thereby improving the effectiveness of the entire model.

The input of CNN model needs to ensure the same dimension, while 40000 samples measured, the average number of words each sample after word segmentation in 29, considering the length of the sample sizes, lastly, the number of words each comment is 41. When the word number is more than 41, intercept the front only 41 words, and when the word number less than 41, the lack of all parts is filled with zero vector. The model structure is shown in figure 10:

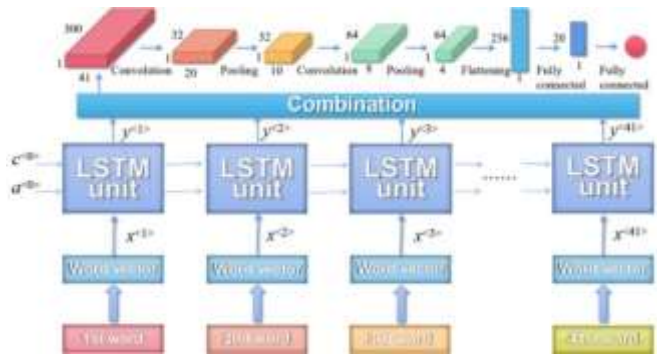


Fig. 10. LSTM & CNN model structure

4 RESULT

4.1 Evaluation

	Comments		Prediction	
			Valid comments	Invalid comment
[ ]			1	0
True	Valid comments	1	VV	IV
	Invalid comments	0	VI	II

Table 1. Experimental data classification table

The evaluation of the experimental results is an important part of the whole experiment. The traditional evaluation methods generally use the accuracy rate, recall rate and F1 value to measure the model effect, but now the machine learning field pays more attention to the overall performance of the model, so in this article, the accuracy is used to measure the experimental results. Assume that the actual classification and prediction

classification results are shown in Table 1. The accuracy is calculated in equation 6:

$$Accuracy = \frac{VV + II}{VV + IV + VI + II} \times 100\% \quad (6)$$

4.2 LSTM-based Classification Model of Text Value

When the LSTM model is used for final classification, the output of the last LSTM unit is only used for prediction. In this experiment, the dimension of the output vector of LSTM unit was selected as 300 and tested on different iterations. The experimental results are as follows:

As can be seen from the above figure, as the number of iterations increases, the loss function shows a downward trend, decreasing until it approaches 0. The

Table 2. The relationship between the number of iterations and the accuracy

Number of iterations	2	4	6	8	10	12	14
Accuracy	84.01%	84.02%	83.48%	83.24%	82.54%	82.20%	82.20%

performance of the model on the training set also gets better as the number of iterations increases, finally approaching almost 100% accuracy. But there is also a very serious problem, and the overall performance of the model on the test set shows a downward trend. This shows that the model has a serious overfitting phenomenon. The reason for this phenomenon may be that the LSTM model only uses the output of the last unit to predict and

classify it. The last unit is theoretically related to all the previous units, but this connection will inevitably lose part of the information.

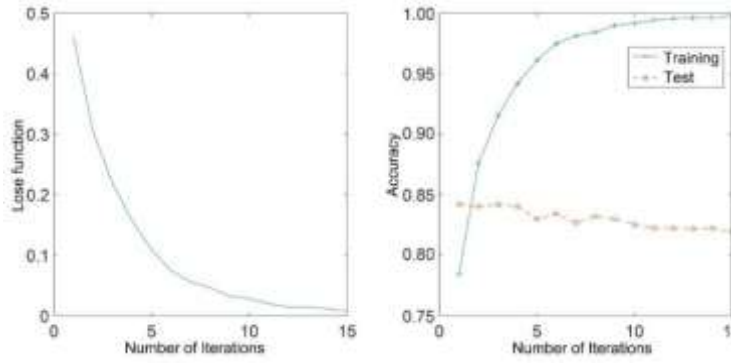


Fig. 11. The relationship between the number of iterations and the accuracy

4.3 LSTM&DAN-based classification model of text value

In the DAN model, the number of hidden layer nodes was recommended as the dimension of the word vector with each hidden layer the same dimension. The experiment in this section also followed the above rules and set the hidden layer node to 300 for three groups of experiments.

The first experiment was to find the best possible value of Dropout. First set the number of hidden layers as 1, and then set the Dropout probability p to 0.0, 0.1, 0.2, 0.3, 0.4 and 0.5, respectively, to observe the performance of each group on the test set. The experimental results are as follows:

It can be seen from the figure above that when the probability of Dropout is 0.3, the maximum accuracy can be obtained on the test set. The reason why the experiment is selected to 0.5 is that when the Dropout probability is greater than 0.5, many important word vectors will be eliminated in a large probability, which

Table 3. The relationship between dropout probability and accuracy of LSTM & DAN model

Dropout	0.0	0.1	0.2	0.3	0.4	0.5
Accuracy	85.17%	86.02%	86.35%	86.61%	86.46%	86.53%

will affect the model effect.

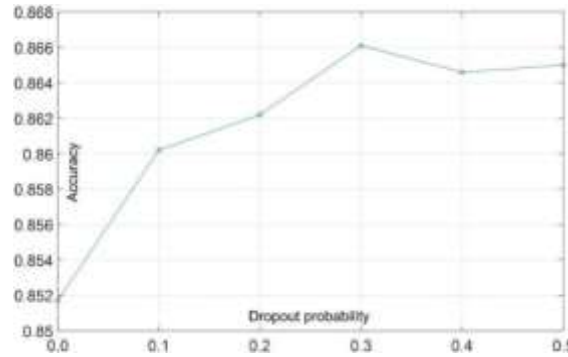


Fig. 12. LSTM & DAN model Dropout probability and accuracy relationship

The second experiment was to find the best hidden layers: The best Dropout probability value has been obtained in the first experiment, so this experiment is conducted on the basis of the Dropout probability value of 0.29. The number of hidden layers is set to 0,1,2,3,4,5,6 in order to observe the performance of each group on the test set. The experimental results are shown as follows:

Table 4. The accuracy of the LSTM & DAN model with different number of hidden layers

Number of hidden layers	0	1	2	3	4	5
Accuracy	84.35%	86.61%	87.86%	87.09%	86.96%	87.03%

It can be seen from figure 13 that the addition of hidden layers greatly improves the experimental results, but it is not the case that the larger the hidden layers is, the better it is. This paper should select different hidden layers according to the specific application scenarios.

The third experiment was to obtain the change of accuracy under the optimal combination of the model. According to the first two experiments, the probability value of Dropout is selected as 0.3 and the number of hidden layers is selected as 2. This experiment is carried out based on these two parameters. The experimental results are shown in table 5 and figure 14:

As can be seen from the above results, although the model can also reach nearly 100% accuracy on the training set, the overfitting problem was avoided on the 1016 I. Ja˘s˘sov´a, M. Tak´a˘c, I. Ja˘s˘sov´a, M. Tak´a˘c, I. Ja˘s˘sov´a, M. Tak´a˘c, I. Ja˘s˘sov´a, M. Tak´a˘c, I. Ja˘s˘sov´a, M

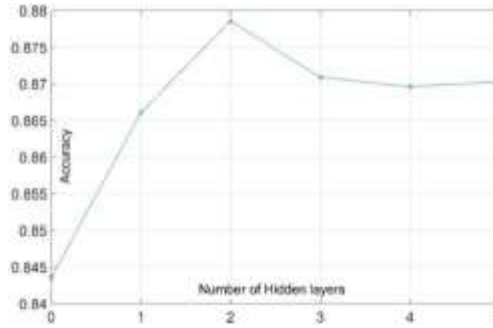


Fig. 13. The relationship between the number of Hidden layers and accuracy of the LSTM & DAN model

Table 5. The accuracy of LSTM & DAN model with different number of iterations

Number of iterat	2	4	6	8	10	12	14

io ns							
	87	87	87	86	87	86	87
Accuracy	.38%	.86%	.14%	.87%	.04%	.82%	.00%

test set. With the increase of iterations, the model can stabilize at a relatively high accuracy rate, without the previous downward trend. The accuracy rate was improved by 3.71% relative to the LSTM model alone.

4.4 LSTM & CNN-based Classification Model of Text Value

The word vector dimension in the model selects 300 dimensions, and the output of each LSTM unit also selects 300 dimensions, that is, in the figure above are all 1 by 300 vectors. There are 41 LSTM units in total. In this experiment, the output of all LSTM units is retained and spliced into a vector. For the first convolution, set the convolution kernel size to 3, the step size to 2, and the number of convolution

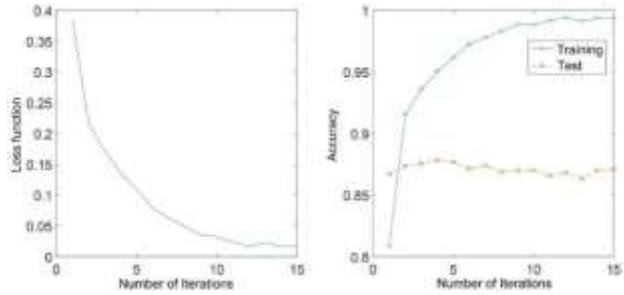


Fig. 14. The relationship between the number of iterations and accuracy relationship of the LSTM & DAN model

cores to 32. For the first pooling, set the size of the pool kernel to 2, the step size to 2, and the type of pooling to maximum pooling. For the second convolution, set the convolution kernel size to 3, the step size to 1, and the number of convolution cores to 64. For the second pooling, set the size of the pool kernel to 2, the step size to 2, and the type of pooling to maximum pooling. It is then flattened, and fully connected to a hidden layer with 20 nodes, and finally output. The model experiment results are shown in table 6 and figure 15:

Table 6. The accuracy of LSTM & CNN model with different number of iterations

Number of iterations	2	4	6	8	10	12	14
Accuracy	88.59%	87.10%	87.68%	87.64%	87.01%	87.84%	87.64%

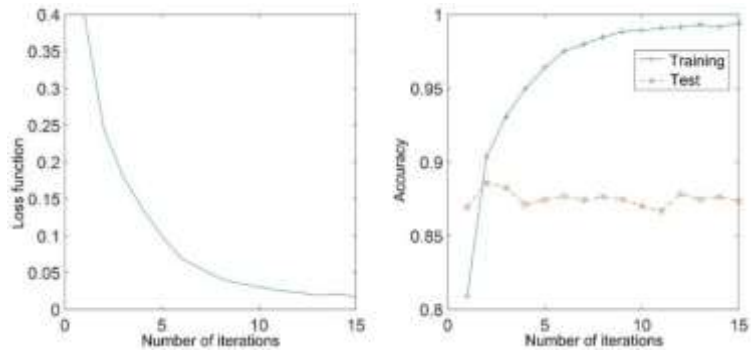


Fig. 15. The relationship of number of iterations and accuracy relationship of LSTM & CNN model

As can be seen from the above data, the model can also achieve almost 100% accuracy on the training set, but on the test set, it is a 0.73% improvement compared to the LSTM & DAN model, which also proves that the CNN can mine out more text features.

4.5 Performance of three LSTM models

The performance of the three LSTM models is shown in figure 16. As can be clearly seen in the figure above, the two improved models have significantly improved accuracy compared with the original LSTM model, and both of them have solved the problem of overfitting. The LSTM & CNN model generally performs better than the LSTM & DAN model.

As can be seen in the figure above, the two improved models have significantly improved in accuracy compared with the original LSTM model, and both solve the

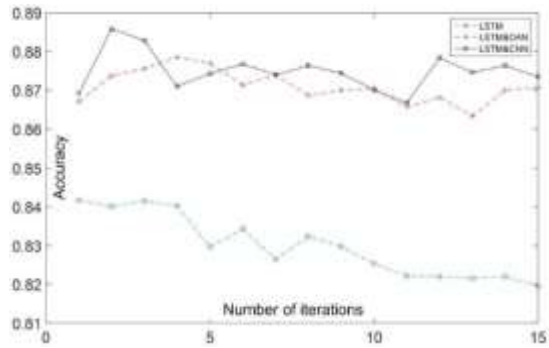


Fig. 16. Comparison of three LSTM models

model overfitting problem. The LSTM & CNN model overall performed better than the LSTM & DAN model.

5 DISCUSSION AND CONCLUSION

Firstly, the text data obtained is vectorized in this paper. Then the LSTM model is introduced to carry out the text classification experiment by the LSTM model.

In the experimental results of the LSTM model, it can be seen that the loss function presents a downward trend as the number of iterations increases. The performance of the model on the training set also gets better and better with the increase of the number of iterations with the accuracy rate of nearly 100% in the end. But there is also a very serious phenomenon - the performance of the model on the test set as a whole shows a downward trend. This indicates that there is a serious over-fitting phenomenon in the model, which may be caused by the fact that LSTM model only uses the output of the last unit for prediction and classification. Theoretically, the last unit is related to all the previous units, but some information is inevitably lost in this connection.

To solve the overfitting problem, this paper proposes the improvement schemes LSTM combining DAN model and CNN model respectively.

Three experiments conducted under the LSTM & DAN model show that the model can also achieve nearly 100% accuracy in the training set, furthermore, the overfitting problem is avoided in the test set. With the increase of the number of iterations, the

model can stabilize at a relatively high accuracy without any previous downward trend. Compared with LSTM model alone, the accuracy is improved by 3.71%.

Experiments conducted under the LSTM & CNN model show that the model can achieve almost 100% accuracy in the training set, while in the test set, it can improve by 0.73% compared with the LSTM & DAN model, which also proves that CNN can dig out more text features. LSTM & CNN method is the most time-consuming

model among all models, but it is also the model with the highest accuracy, which improves 4.42% compared with the LSTM model.

Although this article has done certain research and experiments on text feature extraction and text classification models, there are still many areas that need to be improved and adjusted urgently:

(1)The selection of the corpus needs to be improved. Due to the limitation of computer performance, the data set has not been selected too large. Later, more large-scale training can be carried out on the cloud server. At the same time, in the sample calibration process, there will be more or less human factors. Later, unsupervised learning methods can be considered for text classification.

(2)Chinese text classification needs to be developed. Now, there are not many researches on processing at the character level. If there is a good learning model or method, character-based processing can retain more text information, which can be further improved in theory. The follow-up will be studied at the character level to increase the accuracy of the model.

6 FUNDING

Supported by the Sichuan Science and Technology Program (2021YFQ0003)

7 AVAILABILITY OF DATA AND MATERIALS

Data available in a publicly accessible repository that does not issue DOIs Publicly available.

Dataset was analyzed in this study.

Data URLs:(https://movie.douban.com/subject/26363254/comments?status=P)

REFERENCES

[1] Cheng, M.; Jin, X., What do Airbnb users care about? An analysis of online re-view comments. International Journal of Hospitality Management 2019, 76, 58-70. https://doi.org/10.1016/j.ijhm.2018.04.004.

[2] Tang, Y.; Liu, S.; Deng, Y.; Zhang, Y.; Yin, L.; Zheng, W., Con-struction of force haptic reappearance system based on Geomagic Touch hap-tic device. Computer methods programs in biomedicine 2020, 190, 105344. https://doi.org/10.1016/j.cmpb.2020.105344.

[3] Ding, Y.; Tian, X.; Yin, L.; Chen, X.; Liu, S.; Yang, B.; Zheng, W., Multi-scale relation network for few-shot learning based on meta-learning, In International Conference on Computer Vision Systems, Springer: 2019; pp 343-352. https://doi.org/10.1007/978-3-030-34995-031.

[4] Ni, X.; Yin, L.; Chen, X.; Liu, S.; Yang, B.; Zheng, W., Semantic rep-resentation for visual reasoning, In MATEC web of conferences, EDP Sciences: 2019; p 02006. https://doi.org/10.1051/matecconf/201927702006.

[6] Zhao, Y.; Wang, L.; Tang, H.; Zhang, Y., Electronic word-of-mouth and consumer purchase intentions in social e-commerce. Electronic Commerce Research and Appli-cations 2020, 41, 100980. https://doi.org/10.1016/j.elerap.2020.100980.

[7] Xiao, L.; Guo, F.; Yu, F.; Liu, S., The Effects of Online Shopping Context Cues on Consumers' Purchase Intention for Cross-Border E-Commerce Sustainability. Sus-tainability 2019, 11 (10). https://doi.org/10.3390/su11102777.

[8] Dwidienawati, D.; Tjahjana, D.; Abdinagoro, S. B.; Gandasari, D.; Munawaroh, Customer review or influencer endorsement: which one influences purchase intention more? Heliyon 2020, 6 (11), e05543. https://doi.org/https://doi.org/10.1016/j.heliyon.2020.e05543.

[9] Huang, Y.; Wang, N.-n.; Zhang, H.; Wang, J., A novel product recommendation model consolidating price, trust and online reviews. Kybernetes 2019, 48 (6), 1355-1372. https://doi.org/10.1108/K-03-2018-0143.

[10] Yang, X., Influence of informational factors on purchase intention in so-cial recommender systems. Online Information Review 2020, 44 (2), 417-431. https://doi.org/10.1108/OIR-12-2016-0360.

[11] Liu, S.; Xiao, Z.; You, X.; Su, R., Multistrategy boosted multicolony whale vir-tual parallel optimization approaches. Knowledge-Based Systems 2022, 242, 108341. https://doi.org/10.1016/j.knosys.2022.108341.

[12] Su, R.; Gu, Q.; Wen, T., Optimization of High-Speed Train Control Strategy for Traction Energy Saving Using an Improved Genetic Algorithm. Journal of Applied Mathematics 2014, 2014, 507308. https://doi.org/10.1155/2014/507308.

[13] Huang, W.; Li, Y.; Zhang, K.; Hou, X.; Xu, J.; Su, R.; Xu, H., An Efficient Multi-Scale Focusing Attention Network for Person Re-Identification. Applied Sci-ences 2021, 11 (5). https://doi.org/10.3390/app11052010.

[14] Zhang, K.; Huang, W.; Hou, X.; Xu, J.; Su, R.; Xu, H., A Fault Diagnosis and Visualization Method for High-Speed Train Based on Edge and Cloud

Collaboration. Applied Sciences 2021, 11 (3). https://doi.org/10.3390/app11031251.

[15] Reyes-Menendez, A.; Saura, J. R.; Filipe, F., The importance of behavioral data to identify online fake reviews for tourism businesses: A systematic review. PeerJ Computer Science 2019, 5, e219. https://doi.org/10.7717/peerj-cs.219.

[16] Hayati, H.; Khalidi Idrissi, M.; Bennani, S., Automatic classification for cognitive engagement in online discussion forums: text mining and machine learning approach, In International Conference on Artificial Intelligence in Education, Springer: 2020; pp 114-118. https://doi.org/10.1007/978-3-030-52240-721.

[17] Mikolov, T.; Chen, K.; Corrado, G.; Dean, J., Efficient estimation of word representations in vector space. arXiv preprint arXiv 2013. https://doi.org/10.48550/arXiv.1301.3781.

[19] Bashar, A., Survey on evolving deep learning neural network architectures. Journal of Artificial Intelligence 2019, 1 (02), 73-82. https://doi.org/10.36548/jaicn.2019.2.003.

[24] Smagulova, K.; James, A. P., A survey on LSTM memristive neural network architectures and applications. The European Physical Journal Special Topics 2019, 228 (10), 2313-2324. https://doi.org/10.1140/epjst/e2019-900046-x.

[25] Yenter, A.; Verma, A., Deep CNN-LSTM with combined kernels from multiple branches for IMDb review sentiment analysis, In 2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON), 19-21 Oct. 2017; 2017; pp 540-546. https://doi.org/10.1109/UEMCON.2017.8249013.

[26] Behera, R. K.; Jena, M.; Rath, S. K.; Misra, S., Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data. Information Processing & Management 2021, 58 (1), 102435. https://doi.org/10.1016/j.ipm.2020.102435.

[27] Bhuvaneshwari, P.; Rao, A. N.; Robinson, Y. H., Spam review detection using self attention based CNN and bi-directional LSTM. Multimedia Tools and Applications 2021, 80 (12), 18107-18124. https://doi.org/10.1007/s11042-021-10602-y.

[28] Dankwa, S.; Zheng, W., Twin-delayed ddpg: A deep reinforcement learning technique to model a continuous movement of an intelligent robot agent, In Pro-ceedings of the 3rd International Conference on Vision, Image and Signal Processing, 2019; pp 1-5. https://doi.org/10.1145/3387168.3387199.

[29] Dankwa, S.; Zheng, W., Modeling a Continuous Locomotion Behavior of an Intelligent Agent Using Deep Reinforcement Technique, In 2019 IEEE 2nd International Conference on Computer and Communication

Engineering Technology (CCET), Beijing, China, IEEE: Beijing, China, 2019; pp 172-175. https://doi.org/10.1109/CCET48361.2019.8989177.

[30] Maulud, D. H.; Zeebaree, S. R.; Jacksi, K.; Sadeeq, M. A. M.; Sharif, K. H., State of art for semantic analysis of natural language processing. Qubahan Academic Journal 2021, 1 (2), 21-28. https://doi.org/10.48161/qaj.v1n2a44.

[31] Jelodar, H.; Wang, Y.; Yuan, C.; Feng, X.; Jiang, X.; Li, Y.; Zhao, L., Latent Dirichlet allocation (LDA) and topic modeling: models, ap-plications, a survey. Multimedia Tools Applications 2019, 78 (11), 15169-15211. https://doi.org/10.1007/s11042-018-6894-4.

[32] Mikolov, T.; Le, Q. V.; Sutskever, I., Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168 2013. https://doi.org/10.48550/arXiv.1309.4168.

[33] Hochreiter, S.; Schmidhuber, J., Long Short-Term Memory. Neural Computation 1997, 9 (8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735.

[34] Dankwa, S.; Zheng, W., Special issue on using machine learning algo-rithms in the prediction of kyphosis disease: a comparative study. Applied Sciences 2019, 9 (16), 3322. https://doi.org/10.3390/app9163322.

[35] Gers, F. A.; Schmidhuber, J.; Cummins, F., Learning to Forget: Con-tinual Prediction with LSTM. Neural Computation 2000, 12 (10), 2451-2471. https://doi.org/10.1162/089976600300015015.

[36] Iyyer, M.; Manjunatha, V.; Boyd-Graber, J.; Daum´e Iii, H., Deep unordered com-position rivals syntactic methods for text classification, In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th In-ternational Joint Conference on Natural Language Processing, Beijing, China, July 26-31; https://doi.org/10.3115/v1/P15-1162.

[37] Hinton, G. E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. R., Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv 2012. https://doi.org/10.48550/arXiv.1207.0580.

[38] Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhut-dinov, R., Dropout: a simple way to prevent neural networks from over-fitting. The journal of machine learning research 2014, 15 (1), 1929-1958. https://doi.org/10.5555/2627435.2670313.

[39] Rosenfeld, A.; Tsotsos, J. K., Incremental learning through deep adaptation. IEEE transactions on pattern analysis machine intelligence 2018, 42 (3), 651-663. https://doi.org/10.1109/TPAMI.2018.2884462.

[40] Albawi, S.; Mohammed, T. A.; Al-Zawi, S., Understanding of a convolutional neural network, In 2017 international confer-ence on engineering and technology (ICET), Ieee: 2017; pp 1-6. https://doi.org/10.1109/ICEngTechnol.2017.8308186.

[41] Xie, Y.; Liang, R.; Liang, Z.; Huang, C.; Zou, C.; Schuller, B., Speech Emotion Classification Using Attention-Based LSTM. IEEE/ACM Transac-tions on Audio, Speech, and Language Processing 2019, 27 (11), 1675-1685. https://doi.org/10.1109/TASLP.2019.2925934.

Wei Dang graduated from Central South University in 2022. She is now a master student in the School of Automation, University of Electronic Science and Technology of China.

Ligao Cai received the master's degree in Control Science and Control En-gineering from University of Electronic Science and Technology of China in 2018, respectively. He is currently working as an software development engineer at the Huawei Technologies CO.LTD. His research areas include software engineering, deep learning, Natural language processing, and social network analysis.

Mingzhe Liu received a B.S. degree in computer application from the Chengdu University of Technology, Sichuan, China, in 1994, and M.S. and Ph.D. degrees in computer science from Massey University, New Zealand, in 2006 and 2010, respec-tively. He is currently a professor at the College of Computer Science and Cyber Security, Chengdu University of Technology. His research interests include image and video processing, machine learning and deep learning, data mining, and big data. He has supervised more than 50 postgraduate students to completion. He has published over 150 peer-reviewed papers.

Xiaolu Li received the Ph.D. degree in detection technology and automatic equipment from the University of Electronic Science and Technology of China, Chengdu, China, in 2015. She is currently a Lecturer of geographic information science with the School of Geographical Sciences, Southwest University, Chongqing, China. Her research interests include the spatio-temporal information mining and coupled human and natural system.

Zhengtong Yin is an associate professor at the College of Resource and Environment Engineering, Guizhou University since 2009. He received the P.h.D in Earth Exploration and Information Technology from the Chengdu University of Technol-ogy in 2009. The focused research interests involve Artificial Intelligent and machine learning. She has published more than 40 papers. Research Interests: AI/ML, Com-plex Dynamics, Pattern Recognition, Visual Reasoning, Visual Question Answer-ing, NLP, Surgical Robot, Geospatial AI, GIS/RS, Image Fusion, Surgical Vision, 3D Visualization, Artificial Neural Network, ComputerGraphics, Image Processing, Machine Vision, 3D Reconstruction, Medical Imaging, Data Mining, Earth Surface Process, Cloud Computing, Geography and Environmental Science.

Xuan Liu is an Professor at the University of Electronic Science and Technology of China. She earned her Ph.D in urban studies in National University of Singapore and Bachlor's degree in urban and regional studies from Peking University. She was a visiting scholar in the Department of Architecture, Swiss Federal Institute of Technology Zurich (ETHZ) and in Belk College of Business, University of North Carolina at Charlotte in 2013 and 2020. Her research focuses on patterns and mechanism of land use in urban and rural China. By applying machine learning algorithms on urban big data, she reviews and models the changes of land use patterns and space value. She also adapt various theories to explain the changes.

Lirong Yin is a Ph.D. student in the Department of Geography and Anthro-pology at Louisiana state university. with a study interest in remote sensing, server weather and climate change, Coastal environment, natural hazard, and coupled hu-man and natural dynamic system. Acquired the Master of Science in Geography from Louisiana State University and the Bachelor of Science in Geography Informa-tion Science from the University of Iowa, she has experienced the artificial intelli-gence studies and machine learning techniques, geo-data processing and information analysis skills. She has published more than 70 papers.

Wenfeng Zheng is an associate professor at the School of Automation of the University of Electronic Science and Technology of China since 2008. He received the P.h.D from the Chengdu University of Technology in 2008. The focused research interests involve Artificial Intelligent and machine learning. He has published more than 160 papers, and authorized more than 50 Chinese national invention patents. He is a member of Association for Computing Machinery, a member of IEEE. a member of America Association Geographer, a member of American Geophysical Union, and a membership of China Association of Inventions. He is ranked among the 2022 world's top 2% scientists list of Stanford University.

Reviewer Files

Scientific level	: good
Originality of results	: Good
Topicality	: Excellent
Suitability for the journal	: Excellent
Style	: good
Overall appraisal of manuscript	: 7 - publication recommended

Other important notes **for the Editorial Board:**

always consistent and continuous in managing and publishing journals according to the theme, in order to make a real contribution as a solution to various problems in the industrial world

Brief content of the paper indicating the contribution made **by the author(s):**

In this paper, a text filtering model was constructed based on word vector and Long Short-Term Memory (LSTM) and improved by adding Deep Averaging Net-works (DAN) and convolutional neural network (CNN). The major improvement of the LSTM & DAN model was to retain the original word vector information and to improve the accuracy of the text classification model without increasing hy-perparameter and model structure complexity. The LSTM & CNN model mainly combines the advantages of convolutional neural network in exploring the deep in-formation of text, which was an improvement over the original LSTM. It was proved by experiments that this improvement is meaningful. Compared with the shallow neural network, the accuracy has been greatly improved.

Comments and suggestions for the author:

This topic is interesting to study, because until now a program is needed that is able to store and store data for a long time.

Need to uncover knowledge gaps that have not been discovered by

researcher

Data collection is sufficient

Need to add the latest references related to this topic

The issues raised are in accordance with reality.

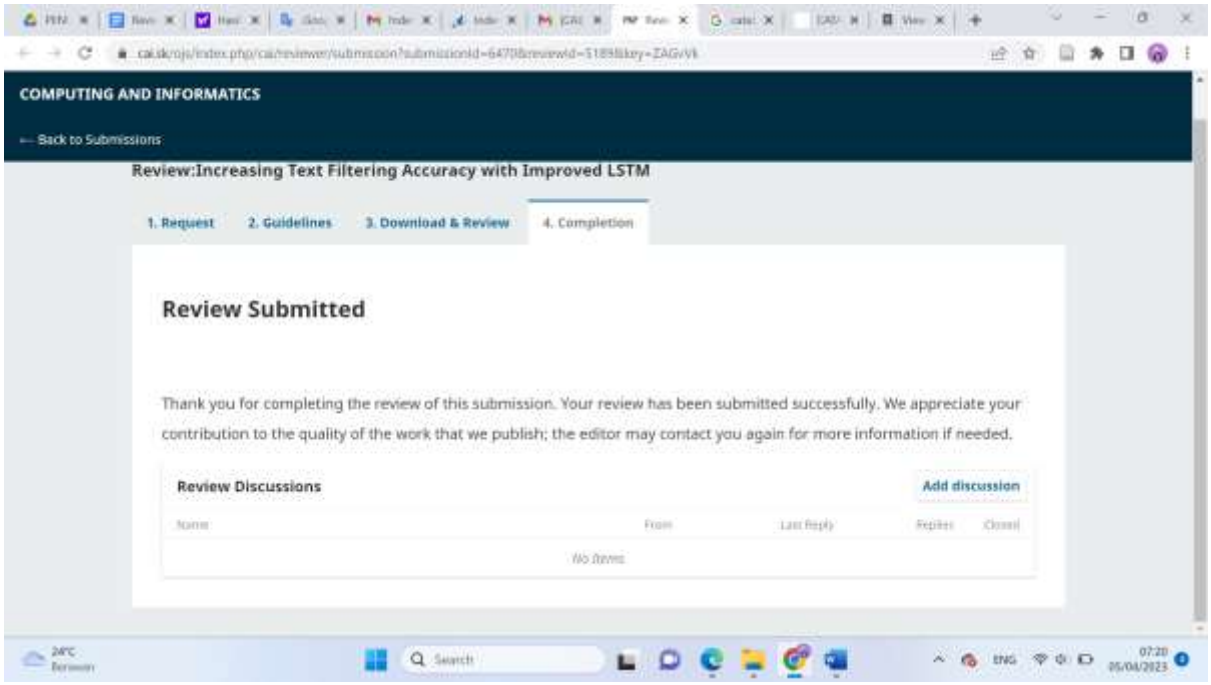
Concepts that have been poured are reinforced with the latest theories or

research results.

Novelty found

footnote is omitted, just follow the standard rules of journal writing. References are recommended using citation management such as mandeley, zotero, or endnote. Add references from reputable journals

this research can be improved on a better model, namely a model that can accommodate a larger data capacity.


COMPUTING AND INFORMATICS
← Back to Submissions
Review:Increasing Text Filtering Accuracy with Improved LSTM
1. Request
2. Guidelines
3. Download & Review
4. Completion
Review Submitted
Thank you for completing the review of this submission. Your review has been submitted successfully. We appreciate your contribution to the quality of the work that we publish; the editor may contact you again for more information if needed.
Review Discussions
Add discussion
Name
From
Last Reply
Replies
Closed
No items