Sentiment Analysis Model Based on Self-Attention and Character-Level Embedding

HONGBIN XIA,1,2, CHENHUI DING1, AND YUAN LIU1,2
1School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China
2Jiangsu Key Laboratory of Media Design and Software Technology, Jiangnan University, Wuxi 214122, China
Corresponding author: Hongbin Xia (hbxia@jiangnan.edu.cn)
This work was supported in part by the National Natural Science Foundation of China Project.

ABSTRACT Aiming at the problem of insufficient sentiment word extraction ability in existing text sentiment analysis methods and OOV (out-of-vocabulary) problem of pre-training word vectors, a neural network model combining multi-head self-attention and character-level embedding is proposed. An encoder-decoder (Seq2Seq) model is used for sentiment analysis of text. By adding character-level embedding to the word embedding layer, the OOV problem of pre-trained word vectors is solved. At the same time, the self-attention mechanism is used to perform attention calculations within the input sequence to find intra-sequence connections. The model uses bidirectional LSTM (Long Short-Term Memory) independently encodes the context semantic information of the input sequence to obtain deeper emotional features, to more effectively identify the emotional polarity of short text. The effectiveness of our method was verified on the public Twitter dataset and the movie review dataset published by ACL. Experimental results show that the accuracy of the model on the three categories of Twitter, movie reviews and IMDB datasets reaches 66.45%, 79.48% and 90.34%, respectively, which verifies that the model has excellent performance in datasets in different fields.

INDEX TERMS Sentiment analysis, seq2seq, character-level, bidirectional long short-term memory, attention mechanism.

I. INTRODUCTION
In recent years, with the emergence of social networks such as twitter and microblog, people can easily and conveniently share diversified information about individuals on social platforms such as texts, pictures, videos. Understanding emotions is an important aspect of personal development and growth, and as such, it is a key tile for the emulation of human intelligence. Besides being important for the advancement of AI, emotion processing is also important for the closely related task of polarity detection [1]. Through the extraction and analysis of these texts with emotional information, we can better understand the user’s behavior, and find out the user’s inclination to products and the degree of attention to hot events [2].

Sentiment analysis is a field of natural language processing (NLP), and it study and mine the emotion expressed in the text. Since Pang and Lee [3] put forward the work of sentiment analysis in 2007, with the development of deep learning and more and more scholars’ research, sentiment analysis has been greatly developed. Kim [4] used the pre-trained word vector to apply the convolutional neural network (CNN) in the sentence classification task for the first time. However, due to the insufficient scale and depth of the network, the ability to extract features is limited. Amir Hussain et al. explore the potential of a novel semi-supervised learning model based on the combined use of random projection scaling as part of a vector space model, and support vector machines to perform reasoning on a knowledge base [5]. Wang et al. proposes a stacked residual LSTM model to predict sentiment intensity for a given text. By investigating the performances of shallow and deep architectures, and introduce a residual connection to every few LSTM layers to construct an 8-layer neural network. Experimental results show that the proposed method outperforms lexicon-, regression-, and conventional NN-based methods proposed in previous studies [6].

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
Kalchbener et al. [7] proposed a dynamic K-MAX pooling method to solve the polarity judgment problem of twitter short text. The network can process input sentences of different lengths, and generate a feature map on the sentence, which can clearly capture the short-term and long-term relationship, and achieved good results. In 2011, Mikolov et al. [8] proposed to use RNN (recurrent neural network) to model, making full use of context information to build language model, to obtain better results. Tang et al. [9] established a text level cyclic neural network model by using cyclic neural network, which has higher advantages than the standard cyclic neural network model, and has made progress in emotional classification tasks.

In recent years, with the wide application of attention mechanism in NLP, more and more scholars begin to use attention mechanism to solve NLP problems. AOA_LSTM model is proposed by Huang et al. [10] in 2018. Two-way LSTM is used to construct the attribute eigenvector matrix of a sentence, and then the attribute attention matrix is used to calculate the emotional eigenvector of a specific attribute. The model can pay more attention to the important information in the attribute sequence and dig out the deeper emotional feature information through context coding and attention calculation. In the senseval-2017 competition, Baziotis et al. [11] and others used deep LSTM combined with attention model to assign weight to key words in short text through attention mechanism, enhanced the influence of keywords on sentence emotion, and achieved good results in twitter emotion classification. In 2018, Tian et al. [12] and others used two-way GRU (gated recurrent Unit) combined with attention mechanism, which has achieved good results in the short text classification of e-commerce reviews. These methods prove that deep learning combined with attention mechanism can achieve better results in short text sentiment classification. In 2018, Tan et al. [13] and others applied self-attention mechanism to semantic role tagging task (SRL). In the paper, the author regards SRL as a sequential tagging problem, and then proposes to use deep attention network. In the model, self-attention can learn tag potential dependency information in the top layer of the model. The model has achieved good results in the SRL datasets of conll-2005 and conll-2012.

The attention models used in the above literature are common basic models based on probability distribution, which cannot get the relationship between words better. The model released by Google in the paper “Attention Is All You Need”[14] shows excellent performance in machine translation tasks. So we will use the Multi-Head Self-attention model in the above combined with deep learning. We use the multi-head self-attention model to weight the important words in short text, and pay more attention to the relationship between the internal text sequences, so that the important emotional words can get higher weight. We propose a twitter short text sentiment analysis model (Self-AT-LSTM) based on the multi head self attention mechanism and LSTM. The main contributions of this paper are as follows:

1. By using character level embedding and pre-training word vectors, the model input can deal with spelling errors and rare words elastically, so as to effectively solve the OOV (out of vocabulary) problem.

2. The twitter sentiment analysis method based on the combination of multi head self-attention mechanism and bidirectional LSTM is proposed. The multi head attention mechanism is introduced into Twitter text sentiment analysis, and the advanced characteristics of the above models in emotional analysis tasks are verified by our experiments.

3. Post attention LSTM is added after the attention layer to simulate the decoder function and extract the predicted features of the model.

II. SELF ATTENTION SENTIMENT ANALYSIS MODEL

This paper proposes a network model (self multi-head attention-LSTM, Self-AT-LSTM) which integrates multi head self attention and bidirectional LSTM. As shown in Figure 1, the network model is mainly composed of embedding layer, bidirectional LSTM layer, self multi-head attention layer, post attention layer and SoftMax layer. The network model draws on the design idea of neural network translation model [15], adopts the network structure of encoder decoder based on LSTM. As shown in Figure 1. It independently encodes the context through two-layer bidirectional LSTM, captures the association between the current word and other input words in the coding stage, to learn the emotional feature information more fully. After that, it use a multi-head self-attention mechanism to calculate the weight corresponding to the hidden state of the LSTM at each time. In addition, it obtains the background variables, to find the internal relations between the model input sequences. Some of the above models realize the encoder function. After the attention layer, the post attention LSTM layer is added to simulate the decoder function to extract model features and better classify text emotion.

A. INPUT LAYER

According to the existing research [16], using convolutional neural network can extract morphological information (such as prefix and suffix of words) in word characters, and can use character embedding as an extension of word vectors. In addition, can provide additional information for words without word vectors. Therefore, this paper uses character-level word embedding vector and word vector containing position as the input of short text coding model. The input short text word sequence is represented as \( \{x_1, x_2, \ldots, x_n\} \), \( x_i \) means the ith word in the sentence, where \( x_i \) words contain L characters, \( c_j \) is the word \( x_i \) Each character in I embeds a vector, and each character represents its corresponding feature. As shown in the figure 2, we use a standard convolution neural network to process the character sequence in each word, and train the character level vector E of the word \( e_{ij}^{Wc} \), the calculation
formula is as follows:

\[ e^W_i = \max_{1 \leq j \leq L} \left( \begin{array}{c} W^T_{CNN} \begin{bmatrix} e^c \left( \frac{c \cdot ke - 1}{2} \right) \\ \vdots \\ e^c \left( \frac{c \cdot ke - 1}{2} \right) \end{bmatrix} + b_{CNN} \\ W^T_{CNN} \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \end{array} \right) \]

where \( W_{CNN} \) and \( b_{CNN} \) is the training parameter, \( c_j \) is the embedding vector of each character in the word \( \chi \), \( ke \) represents the convolution kernel size, and \( \max \) represents the maximum pooling operation.

We splice word vector and character vector:

\[ E_i = \text{Concat} (e^W_i, e^W_i) \] (2)

In which, \( e^W_i \) is the word vector, and \( e^W_i \) is the character-level vector of the word obtained by training.

**B. BI-LSTM LAYER**

Self-AT-LSTM network model uses two independent bidirectional LSTM networks, which can better encode contextual emotional information. The forward LSTM reads the input sequence in the normal order (\( x_1 \sim x_t \)). After that, the reverse LSTM reads the input sequence in reverse order (\( x_t \sim x_1 \)), input vector \( x_i \) at each time \( t \). Calculation of long-term and short-term memory gates:

\[
 f_t = \sigma (W_f \cdot [h_{t-1}, x_t] + b_f) \\
 i_t = \sigma (W_i \cdot [h_{t-1}, x_t] + b_i) \\
 C'_t = \tanh (W_c \cdot [h_{t-1}, x_t] + b_C) \\
 C_t = f_t \cdot C_{t-1} + i_t \cdot C'_t \\
 o_t = \sigma (W_o \cdot [h_{t-1}, x_t] + b_o) \\
 h_t = o_t \cdot \tanh (C_t) 
\]

(3)-(8)

Calculate the forward hidden state of each time \( (\tilde{h}_1, \tilde{h}_2, \ldots, \tilde{h}_t) \) and reverse hidden state \( (\bar{h}_1, \bar{h}_2, \ldots, \bar{h}_t) \). Finally, concatenate the forward hidden state \( \tilde{h}_j \) and the backward hidden state \( \bar{h}_j \) in series, the hidden state of the moment is obtained:

\[ h_j = [\tilde{h}_j | \bar{h}_j]^T \] (9)

**C. ATTENTION LAYER**

The role of attention layer is to weight words input in each time step according to attention calculation, so that important
emotional words can get higher weight. The general idea of attention is a coding sequence scheme, which can be understood as a sequence coding layer. Attention calculation is defined as:

\[
{\text{Attention}}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V
\]  

Q, K and V are the abbreviations of query, key and value respectively. Key means corresponding to value and used to calculate similarity with query as the basis for attention selection. Query means the query when attention is executed at one time, and value means the data that is noticed and selected.

\[
Q \in \mathbb{R}^{n \times d_k}, K \in \mathbb{R}^{m \times d_k}, V \in \mathbb{R}^{m \times d_v},
\]

The value of the V dimension. By calculating the dot product of query and each key, divide by \( \sqrt{d_k} \) normalized, then calculated as weight by softmax activation function, and finally multiplied by value. The Multi-head Attention model is shown in Figure 3.

The calculation is defined as follows:

\[
\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)
\]

\[
\text{MultiHead}(Q, K, V) = \text{Concat} (\text{head}_1, \ldots, \text{head}_h)
\]  

Using \( h \) different linear transformations map the dimension \( key \), \( value \) and \( query \) of \( d_{\text{model}} \) to \( d_k \) dimension, \( d_v \) dimension and \( d_q \) dimension, then substituting into equation (8) to generate a total of \( h \times d_v \). The final output is obtained by a linear transformation.

This paper uses the multi-head self-attention mechanism, which is defined as follows:

\[
a_t = \text{MultiHead} [h_i, h_i, h_i]
\]  

In the above formula, \( h_i \) is the hidden state of the input sequence, that is, the output of the LSTM layer. The purpose of it is to calculate the attention within the input sequence and find the internal connection of the sequence.

Network model calculates attention weight through attention layer \( a_t \). Hidden state \( h_t \) at time \( t \) output by LSTM weighted average, is transformed into background variable:

\[
c_{t'} = \sum_{i=1}^{T} d_t h_t
\]  

D. POST-ATTENTION-LSTM LAYER (DECODER)

The background variable \( c \) output by the encoder encodes the input sequence \( x_1, x_2, \ldots, x_t \), and then use post attention LSTM as the decoder. At the time step \( t' \) of the output sequence, the decoder takes the output \( y_{t'-1} \) of the previous time step and the background variable \( c \) as input. It transforms them with the hidden state \( s_{t'-1} \) of the previous time step into the hidden state \( s_{t'} \) of the current time step. Therefore, we can use function \( g \) to express the transformation of the hidden layer of the decoder:

\[
s_{t'} = g (y_{t'-1}, c_t, s_{t'-1})
\]

E. OUTPUT LAYER

With the hidden state \( s_{t'} \) of the decoder, we can use the custom output layer and softmax operation to calculate \( P (y_{t'} | y_1, \ldots, y_{t'-1}, c) \), and obtain the probability distribution of the output \( y_{t'} \) the current time step. Finally, the classification result is calculated through the Dense layer and Softmax layer.

F. MODEL TRAINING

Text sentiment analysis is essentially a classification problem. This paper does three-class and two-class processing for twitter data set, and two-class processing for IMDB data set and Movie Review data set. In this paper, back propagation algorithm is used for network model training, and L2 regularization is introduced to avoid over fitting of network model. By minimizing the cross entropy loss function, the network model is optimized and the emotion classification task is completed. The cross entropy loss function is as follows:

\[
\text{loss} = - \sum_{i=1}^{D} \sum_{j=1}^{C} y_{ij} \log y_{ij} + \lambda \parallel \theta \parallel^2
\]
TABLE 1. Datasets.

| Datasets       | docs  | Class | |\[|\text{train}| |\]| |\[|\text{test}| |\]| |
|----------------|-------|-------|-------------|-------------|
| Twitter        | 49570 | 3     | 44613       | 4957        |
| Movie Review   | 9596  | 2     | 8262        | 1334        |
| Twitter        | 25000 | 2     | 22500       | 2500        |
| IBDM           | 50000 | 2     | 25000       | 2500        |

where D is the size of training set, C is the number of categories, and y is the prediction category, y′ is the actual category, λ \| θ \|² is the regular item.

III. EXPERIMENTAL DATA AND EXPERIMENTAL SETTINGS

A. DATASET

The method proposed in this paper is used in two open datasets of different fields to solve the task of short text sentiment analysis.

(1) Semeval2017 data set is the twitter data set of task 4 of the international semantic evaluation competition. The short text of the data set contains three emotional polarity: positive, neutral and negative.

(2) The twitter dichotomous data set published by the kaggle competition contains positive and negative emotional polarities.

(3) the movie review data set published ACL (Association for Computational Linguistics) International Computer Language Association and IMDB movie review data set, each review contains positive and negative. Table 1 illustrates the statistics of the data used in the experiments in this paper.

B. COMPARED METHODS

We compare our proposed model with the following methods:

(1) CNN. Based on the convolution neural network model proposed by Kim [4], it is a more basic convolution neural network.

(2) AT-CNN(Attention-based CNN) literature [17] proposed the convolution neural network based on word attention, adding the attention layer after the word embedding layer, which can obtain more important local features and achieve better results.

(3) CNNs-LSTMs According to the model of CNN and LSTM proposed by Cliche [18], we use the unlabeled data set to train the word vector, use the labeled data set to fine tune the parameters in the training, and finally achieve excellent results in the twitter emotion classification task.

(4) AT-LSTM(Attention-based LSTM) The Bi-LSTM network based on attention mechanism proposed in [11] has achieved better classification effect than the traditional LSTM network in the emotional classification of twitter.

(5) AT-GRU(Attention-based GRU) The Bi-GRU network based on attention mechanism proposed in [12] has achieved good results in data sets such as movie reviews and hotel reviews.

C. EXPERIMENTAL PARAMETER SETTINGS

The word vector dimension of the Embedding layer in the model is 300, and the output dimension of the two-way LSTM is 300. Embedding layer: Gaussian noise σ value is 0.3 and dropout value is 0.3. The two-way LSTM dropout value is 0.5, the L2 regularization weight is 0.0001, and the learning rate is 0.001. The model is trained by 50 samples in each batch.

D. RESULTS

In this paper, accuracy, macro average recall and F1 score are used to measure and evaluate the effect of short text sentiment analysis. The definition is as follows:

\[
\text{Accuracy} = \frac{TP + TN}{P + N} \tag{17}
\]
\[
\text{recall} = \frac{TP}{TP + FN} \tag{18}
\]
\[
\text{macro-recall} = \frac{1}{n} \sum_{i=1}^{n} \text{recall}_i \tag{19}
\]
\[
F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{20}
\]

where: P is all positive samples and N is all negative samples. True positives (TP): the sample is positive and the prediction result is positive; false positives (FP): the sample is negative and the prediction result is positive; true negatives (TN): the sample is negative and the prediction result is negative; false negatives (FN): the sample is positive and the prediction result is negative. Accuracy: TP / (TP + FP). The correct prediction is the positive proportion of all predictions.

The experimental results of this paper are shown in Table 2 and Table 3.

In order to verify the effectiveness of the self-attention mechanism and character level embedding layer in this model, the above two parts of the model are removed.

TABLE 2. Results of twitter three categories experiment.

| Model            | Accuracy | Recall | F1    |
|------------------|----------|--------|-------|
| CNN              | 0.6145   | 0.6398 | 0.6259|
| AT-CNN           | 0.6488   | 0.6442 | 0.6491|
| CNNs-LSTMs       | 0.6536   | 0.6785 | 0.6810|
| AT-GRU           | 0.6649   | 0.6643 | 0.6553|
| AT-LSTM          | 0.6593   | 0.6735 | 0.6678|
| Self-AT-LSTM     | 0.6645   | 0.6831 | 0.6761|
TABLE 3. Accuracy of compared models on different datasets.

| Model    | Movie Review | IMDB     | Twitter |
|----------|--------------|----------|---------|
| CNN      | 0.7566       | 0.8429   | 0.6501  |
| AT-CNN   | 0.7738       | 0.8856   | 0.7044  |
| CNNs-LSTM | 0.7684       | 0.8657   | 0.6795  |
| AT-LSTM  | 0.7715       | 0.8932   | 0.6850  |
| AT-GRU   | 0.7768       | 0.8825   | 0.6789  |
| Self-AT-LSTM | 0.7948     | 0.9034   | 0.6985  |

TABLE 4. The impact of self-attention mechanism and character level embedding layer on model (accuracy).

| Model       | Twitter | Movie Review | IMDB     |
|-------------|---------|--------------|----------|
| LSTM        | 0.6402  | 0.7597       | 0.8523   |
| NAT-LSTM    | 0.6419  | 0.7642       | 0.8576   |
| NChar-LSTM  | 0.6586  | 0.7874       | 0.8841   |
| Self-AT-LSTM | 0.6645 | 0.7948       | 0.9034   |

respectively, as shown in Table 4. LSTM is an experiment that removes both the attention layer and the character-level embedding layer, and NAT-LSTM is an experiment to remove the attention layer, and NChar-LSTM is an experiment for removing the character-level embedding layer. The experimental results in each data set are shown in Table 4.

It can be seen from Table 2 and Table 3 that the method proposed in this paper has achieved better results in two and three classifications of data sets in different fields. In the twitter three-class dataset experiment of SemEval 2017 task 4, the recall rate, accuracy rate and F1 value of the algorithm in this paper are all the best, and the recall rate of the algorithm in this paper is 0.2% higher than 0.681 of the best known model [17]. At the same time, the accuracy rates of 0.7948 and 0.9034 were achieved in the Movie Review and IMDB datasets, which are the highest of all models. while the accuracy of twitter two-class data sets is 0.6985, which is only 0.6% lower than the highest accuracy of 0.7044. The experimental results show that our model can better solve text sentiment analysis tasks in different fields.

From the above experimental results, we can see that the classification effect of the model AT-CNN is better than that of the CNN model in all aspects. Because CNN treats all words equally and extracts local features of each place, it is unable to distinguish the correlation between the characteristic words of the input text and the emotion, and it has no ability to identify the key words, so it performs generally in the emotion analysis. The At-CNN, which integrates attention mechanism, greatly improves the feature selection ability of the network, and make it more effective in emotion classification task of text. In the three and two categories of twitter, compared with CNN model, the accuracy increased by 3.4% and 4.7% respectively. In the movie review and IMDB data sets, the accuracy increased by 1.7% and 4.2% respectively. The validity of attention mechanism is verified.

AT-LSTM combined with general attention mechanism has a good performance, while TD-LSTM model is not good for classification. AT-LSTM has a better improvement compared with TD-LSTM, especially in IMDB data set, with an accuracy of 5.8%, which shows that attention mechanism can pay high attention to the feature information of specific targets during model training, so that the network can better recognize the text Emotional polarity.

To verify the impact and effectiveness of self-attention and character-level embedding layers on the model, the self-attention layer and character-level embedding layer are removed in ablation experiment, as shown in Table 4. When the above two layers are removed respectively, the classification accuracy of the model decreases to different degrees, especially in the NAT-LSTM layer the accuracy of model classification decreases more obviously. This is because without the calculation weighting of attention layer to the hidden state of the previous layer’s LSTM, post attention LSTM layer without background variable cannot play the role of decoder better. Finally, it can only be used as a common LSTM layer in the model, so the accuracy has a significant decline. After the Char-Embedding layer is removed, the accuracy of the model also drops to a certain extent. This is because the OOV problem occurs when the model only uses ordinary pre-training word vectors as input, which can verify that Char-Embedding can solve the OOV problem of pre-trained words Vector.

Comparing the Self-AT-LSTM model and AT-LSTM model proposed in this paper, we can see that the data sets in four fields have better results. In the three categories and two categories of twitter, the accuracy increases by 0.5% and 1.3%, respectively. In the movie review and IMDB data sets, the accuracy also increases by 2.3% and 1%. Because of the introduction of character level embedding as input, the model in this paper can alleviate the vocabulary problems encountered in model input and deal with spelling errors and rare words more flexibly than the common model with pre-training word vector as input. In terms of output, since the vocabulary of the character-level model is small, the computational cost is lower and the training speed is faster. The Self-attention model, to a certain extent, can solve the problem that RNN lacks hierarchical information and relies on a long distance. For example, RNN can’t deal with two words which are far away from each other, and can’t express the hierarchical information well. After adding self-attention mechanism, we can do attention calculation in the input sequence and add higher attention score to the key words, to obtain the connection between the sequences. At the same time, the multi head part uses different linear transformations to learn different word relations for each head calculated by self-attention. Therefore, the model proposed in this paper has certain advantages in solving the problem of emotion analysis of different types of sentences, and achieves better results in the task of emotion classification.
IV. CONCLUSION
For short text sentiment analysis, this paper proposes a Self-AT-LSTM model, which integrates character level embedding and multi-head self-attention mechanism. Through self-attention mechanism, our model can pay more attention to the relationship between words in each sentence, mining deeper emotional information in the sentence. At the same time, the model can independently encode context information through double-layer bidirectional LSTM. The experimental results of four different areas of evaluation data sets show the feasibility and effectiveness of this method, which can better find the emotional tendency of text information. Most of the current research is to combine the attention mechanism with the LSTM network. Most of these models based on recurrent neural network need higher time cost to train. We will improve the self-attention neural network for this problem next in the future.

REFERENCES
[1] E. Cambria, “Affective computing and sentiment analysis,” IEEE Intell. Syst., vol. 31, no. 2, pp. 102–107, Mar. 2016.
[2] D. Xiao and L. Ting, “Mining user consumption intention from social media using domain adaptive convolutional neural network,” in Proc. Assoc. Adv. Artif. Intell., 2015, pp. 2389–2395.
[3] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” Found. Trends Inf. Retr., vol. 2, nos. 1–2, pp. 1–135, 2008.
[4] Y. Kim, “Convolutional neural networks for sentence classification,” 2014, arXiv:1408.5882. [Online]. Available: http://arxiv.org/abs/1408.5882
[5] A. Hussain and E. Cambria, “Semi-supervised learning for big social data analysis,” Neurocomputing, vol. 275, pp. 1662–1673, Jan. 2018.
[6] J. Wang, B. Peng, and X. Zhang, “Using a stacked residual LSTM model for sentiment intensity prediction,” Neurocomputing, vol. 322, pp. 93–101, Dec. 2018.
[7] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, “A convolutional neural network for modelling sentences,” in Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics, 2014, 655-665.
[8] T. Mikolov, S. Kombrink, L. Burget, J. Cernocky, and S. Khudanpur, “Extensions of recurrent neural network language model,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2011, pp. 1–5.
[9] D. Tang, B. Qin, and T. Liu, “Document modeling with gated recurrent neural network for sentiment classification,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2015, pp. 1422–1432.
[10] B. Huang, Y. Ou, and K. M. Carley, “Aspect level sentiment classification with attention-over-attention neural networks,” in Proc. Int. Conf. Social Comput., Behav.-Cultural Modeling. Cham, Switzerland: Springer, 2018, pp. 197–206.
[11] C. Baziotis, N. Pelekas, and C. Doukeridis, “DataStories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis,” in Proc. 11th Int. Workshop Semantic Eval., 2017, pp. 747–754.
[12] Q. Yu, H. Zhao, and Z. Wang, “Attention-based bidirectional gated recurrent unit neural networks for sentiment analysis,” in Proc. 2nd Int. Conf. Intell. Pattern Recognit. Cham, Switzerland: Springer, 2019, pp. 67–78.
[13] Z. Tan, “Deep semantic role labeling with self-attention,” in Proc. AAAI, 2018, pp. 1–5.
[14] A. Vaswani, “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5998–6008.
[15] I. Sutskever, O. Vinyals, and V. Q. Le, “Sequence to sequence learning with neural networks,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 3104–3112.
[16] J. Lee, K. Cho, and T. Hofmann, “Fully character-level neural machine translation without explicit segmentation,” Trans. Assoc. Comput. Linguistics, vol. 5, pp. 365–378, Dec. 2017.
[17] W. Shengyu, Z. Bijing, S. Qi, and H. Xuli, “Word attention-based convolutional neural networks for sentiment analysis,” J. Chin. Inf. Process., vol. 32, no. 9, pp. 123–131, 2018.
[18] M. Cliche, “BB_twtr at SemEval-2017 task 4: Twitter sentiment analysis with CNNs and LSTMs,” 2017, arXiv:1704.06125. [Online]. Available: http://arxiv.org/abs/1704.06125

HONGBIN XIA was born in Hefei, China, in 1972. He received the master’s and Ph.D. degrees in computer science from Jiangnan University, Wuxi, China, in 2003 and 2008, respectively. He is currently an Associate Professor and the Vice Dean of the Department of Artificial Intelligence and Computer Science, Jiangnan University. His research interests include natural language processing and computational intelligence. He once won the Grand Prize of Science and Technology Award of the China Federation of Commerce.

CHENHUI DING was born in Bengbu, China, in 1994. He received the bachelor’s degree in mechanical engineering from the Guilin University of Technology, China, in 2016. He is currently pursuing the master’s degree in software engineering with Jiangnan University, China. His research interests include natural language processing and deep learning.

YUAN LIU is currently a Professor and a Doctoral Supervisor with the School of Artificial Intelligence and Computer Science, Jiangnan University, China. He is also the Principal Investigator of many provincial research projects and has successfully completed them over the years. He has published more than 40 academic articles (including more than ten SCI and EI articles) and books in the authoritative and core journals. His main researches focused on the software development of network information systems, network security, and digital media applications. His current research interests include network traffic measurement, social networks, and digital media.

He is a member of the 863 Expert Panel in the field of information security technology, Ministry of Science and Technology, China, a Senior Member of the China Computer Federation (CCF), and a member of the Cyber Security Association of China (CSAC).

* * *