BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 06/06 pp311–319
Volume 3, Number 4, December 2020
DOI: 10.26599/BDMA.2020.9020024

Multi-Attention Fusion Modeling for Sentiment Analysis of Educational Big Data

Guanlin Zhai, Yan Yang*, Heng Wang, and Shengdong Du

Abstract: As an important branch of natural language processing, sentiment analysis has received increasing attention. In teaching evaluation, sentiment analysis can help educators discover the true feelings of students about the course in a timely manner and adjust the teaching plan accurately and timely to improve the quality of education and teaching. Aiming at the inefficiency and heavy workload of college curriculum evaluation methods, a Multi-Attention Fusion Modeling (Multi-AFM) is proposed, which integrates global attention and local attention through gating unit control to generate a reasonable contextual representation and achieve improved classification results. Experimental results show that the Multi-AFM model performs better than the existing methods in the application of education and other fields.

Key words: educational big data; sentiment analysis; aspect-level; attention

1 Introduction

Educational data mining is an emerging research field[1] in which data mining methods are applied to educational data to provide new insights about learner behaviors and learning approaches and improve learning methods in a data-driven way[2, 3]. For example, Liao et al.[4] predicted student drop-out through Massive Open Online Courses data by using the clustering method to help course organizers improve the course syllabus.

Sentiment Analysis (SA)[5], also known as opinion mining[6], has attracted extensive attention as one of the core tasks in Natural Language Processing (NLP) in recent years. Most SA methods use machine learning methods, such as support vector machine[7], Long Short-Term Memory (LSTM)[8], and attention-based methods[9], to establish a sentiment classifier.

• Guanlin Zhai, Yan Yang, Heng Wang, and Shengdong Du are with the School of Information Science and Technology, National Engineering Laboratory of Integrated Transportation Big Data Application Technology, Southwest Jiaotong University, Chengdu 611756, China. E-mail: ZhaiGL@my.swjtu.edu.cn; yyang@swjtu.edu.cn.

∗ To whom correspondence should be addressed.
Manuscript received: 2020-08-30; accepted: 2020-09-30

In recent years, Aspect-Based Sentiment Analysis (ABSA) has attracted increasing attention as a fine-grained sentiment classification task. It is designed to identify the sentiment polarity of certain aspects rather than focusing on the sentiment polarity of the whole sentence. For example, in the field of education, the review "I find it is difficult to understand the course, even though the teacher is knowledgeable and he is very serious when teaching" includes two aspects: the teacher and the difficulty of the course. The two aspects of the corresponding sentiment polarity are different. How to correctly find an aspect and its corresponding sentiment word is crucial to complete SA tasks.

Although deep neural networks have made considerable improvements compared with traditional machine learning algorithms in various NLP tasks, end-to-end deep learning systems are not flexible in many cases. For example, the neural network usually cannot accurately identify the actual corresponding aspects of the sentiment words when the sentence contains multiple aspects and sentiment words. Therefore, attention-based neural networks have been proposed and have shown excellent results in NLP tasks. In 2016, Wang et al.[10] proposed the embedding of the target aspect behind each word in a sentence to generate a

sentence representation of the embedding aspect and classified the sentences with the embedding aspect information by using neural network and attention mechanism. Ma et al.[11] presented a new model in which contextual attention mechanisms and aspect representation attention mechanisms interact to produce the final global representation. Liu et al.[12] proposed a novel model for performing context-aware user SA. The model involves different forms of semantic relevance and the influence of tweet context information. Han et al.[13] proposed an attention-based neural network framework for uncertain text recognition. In this model, the semantics of the words in the sentence are recognized through the LSTM network, and the convolutional neural network is used to obtain the most important information in the sentence.

Xu et al.[14] established a model of local attention mechanism in 2015, which learns image content automatically and generates corresponding image description. Rush et al.[15] studied a model based on local attention, which generates each word of the summary conditioned on the input sentence. Although the model is simple in structure, it maintains high efficiency in the case of large numbers of training samples and is easily trained end-to-end. Li et al.[16] used a transformation network for target-oriented sentiment classification. They believed that the sentiment polarity of a target is usually determined by some key words in the context. This idea is consistent with the idea of local attention. However, a universal rule for selecting key words in a sentence is lacking because of the complexity of human language. In addition, local attention is likely to overfit and force the network to be too focused on a particular part of the sentence and even ignore the location that provides the key information.

Accordingly, this paper proposes a multi-attention fusion model that considers the contribution of different sentiment resources to different aspects of sentiment polarity and maximizes the advantages of global and local attention.

First, a bidirectional LSTM is used to generate the sentence representation of the embedded aspect, so that the sentence focuses on one specific aspect at a time. Then, the global and local attention corresponding to the embedded sentences are generated respectively, and the gating mechanism is used to control the fusion of global and local attention to obtain a reasonable attention distribution. Finally, the result is fed to the Softmax classifier for final classification.

In summary, the main contributions of this work are summarized as follows. First, we propose a local attention calculation method that uses a dependency tree to extract the attention score of a sentence in a specific aspect. Second, we propose a gating unit for controlling the weights in the fusion of global and local attention. Third, experimental results show that our model performs better than existing attention models in education and other fields.

The rest of this paper is organized as follows. In Section 2, we discuss previous research about the attention-based methods for SA. In Section 3, we introduce our model for ABSA. In Section 4, we present our different experiments and the results. Finally, Section 5 concludes this work.

2 Related Work

Different from global attention that involves all words, local attention focuses on a subset of words in a sentence. Luong et al.[17] introduced the local attention mechanism to machine translation tasks for the first time. Chen et al.[18] extended local attention with syntax distance constraint by focusing on syntactically related words with the predicted target words. He et al.[19] proposed syntax-based local attention, which performs sentence-level SA. Wang et al.[20] proposed the TMNS network, which can solve the problem that the sentiment polarity is over-dependent on target words in SA. Duan et al.[21] presented an approach that automatically induces target-specific sentence representations. Given their inherent advantages and disadvantages in ABSA, global and local attention could be combined to maximize their advantages. Wang et al.[22] achieved enhanced results in SA tasks by using word-level and clause-level attention.

Many models use a tandem approach to combine information from different sources. However, concatenating local and global attention directly may reduce the performance of the model in some cases. For example, if the local attention result is useful but the global attention result is noisy, the result may be biased toward local attention. Similarly, if global attention is useful but local attention is not, the entire representation may be biased. Therefore, a flexible method is needed to control the degree of information from local and global attention to the final sentence representation. To this end, we use a gating unit to combine different attentions, providing an interpretable way to identify the importance of each word in a sentence to the final prediction.

3 Model

The Multi-Attention Fusion Modeling (Multi-AFM) model is mainly composed of four parts: an aspect-aware sentence representation layer, a memory modeling layer, an attention layer, and a gating layer. The network framework of Multi-AFM is shown in Fig. 1.

The symbols used in this paper are defined as follows for easy understanding: sentence $S = [w_1, w_2, \ldots, w_L]$, where w_i represents each word, which is represented by an embedding vector with dimension k, and L represents the maximum number of words in the sentence. Meanwhile, let a_I denote the I-th aspect of the sentence S, and $0 < I < M < L$, where M represents the largest number of aspects in the sentence.

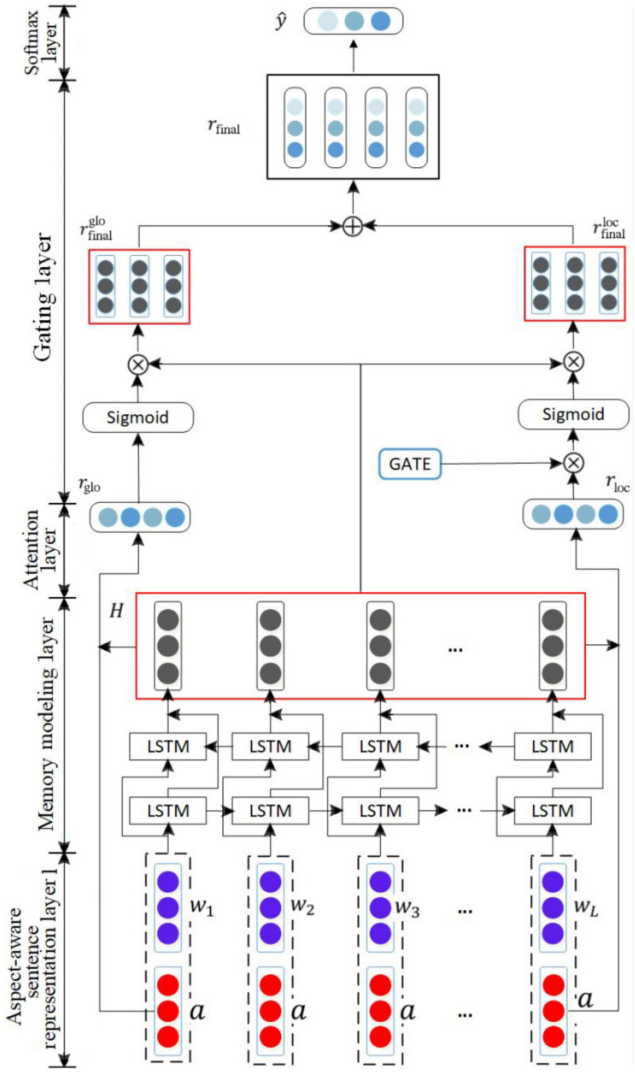

Fig. 1 Network framework of multi-attention fusion modeling.

3.1 Aspect-aware sentence representation

In some sentences, not every word carries sentimental information that can express a particular aspect, and some words do not contain sentimental information, such as stop words. Therefore, we use one sentence to express a specific-aspect sentiment. Multi-AFM first concatenates aspect a_i to every word w_i in the sentence S:

$$S_{a_I} = [[w_1; a_I], [w_2; a_I], \ldots, [w_L; a_I]] \quad (1)$$

where [;] stands for concatenate operation.

3.2 Memory modeling layer

The Multi-AFM model feeds word vectors into a Bidirectional Long Short-Term Memory network (Bi-LSTM) to encode context. The hidden layer size of the Bi-LSTM is D_0.

For forward LSTM, the hidden state at time step $t-1$ is $\overrightarrow{h_{t-1}}$, and the word embedding is S_{a_I}. We can calculate the hidden state $\overrightarrow{h_t}$ at time step t as follows:

$$\overrightarrow{h_t} = \overrightarrow{\text{LSTM}}\left(S_{a_I}, \overrightarrow{h_{t-1}}\right) \quad (2)$$

The backward LSTM is similar to the forward LSTM, except that the input sequence is fed in a reversed way. Multi-AFM concatenates the results of the forward and backward LSTMs, and uses the hyperbolic tangent activation function to process the concatenate results to generate the final hidden state h_t.

$$h_t = \tanh\left(\left[\overrightarrow{h_t}; \overleftarrow{h_t}\right]\right) \quad (3)$$

where $\overrightarrow{h_t}$ represents the output of the forward LSTM at time step t, and $\overleftarrow{h_t}$ represents the output of the backward LSTM at time step t. The final output of the Bi-LSTM is represented as $H = \{h_1, h_2, \ldots, h_L\}$, where $\overrightarrow{h_t} \in \mathbb{R}^{D_0}$, $\overleftarrow{h_t} \in \mathbb{R}^{D_0}$, $h_t \in \mathbb{R}^{2D_0}$.

3.3 Attention layer

The primary purpose of this layer is to use information from global and local views to learn semantic relationships in a specific aspect of a sentence.

3.3.1 Global attention

Given the target aspect vector a_I and the sentence representation $H = \{h_1, h_2, \ldots, h_L\}$, we calculate the attention score α_i for each word representation h_i as

$$m_i = W^{\mathrm{T}}_{\text{att2}} \tanh\left(W_{\text{att1}}\left[h_i; a_I\right] + b_{\text{att1}}\right) \quad (4)$$

$$\alpha_i = \frac{\exp\left(m_i\right)}{\sum_{j=1}^{L} \exp(m_j)} \quad (5)$$

where $W_{\text{att1}} \in \mathbb{R}^{(2D_0+k)\times(2D_0+k)}$ and $W_{\text{att2}} \in \mathbb{R}^{2D_0+k}$ are weight matrices in the training process, $b_{\text{att1}} \in \mathbb{R}^{2D_0+k}$ is bias. Then, global-attention-based representation $r_{\text{glo}} \in \mathbb{R}^{2D_0}$ is formulated as the weighted sum of hidden state h_i with respect to its attention score α_i:

$$r_{\text{glo}} = \sum_{i=1}^{L} h_i \alpha_i \tag{6}$$

3.3.2 Local attention

Since local attention only focuses on a subset of words in a sentence, we first need to select words that are close to the specific target in semantic information. The syntactic dependencies between the target word and its context can be captured by the dependency tree, which contains abundant linguistic information between words.

Therefore, we introduce the syntactic-based distance, which is defined on the dependency tree. Given a sentence S, D is its dependency tree, and each word is a node in D. The distance between two connected nodes is defined as one. We traverse D to calculate the distance from all remaining words to the specific target. A word for local attention is selected based on the syntactic-based distance to the target word. Figure 2 shows an example of the comparison between syntactic-based distance and position-based distance.

Syntactic-based word distance performs better than position-based distance in merging semantic information. We select words within t-step of the grammatical distance to the target word and represent them with LS(t). We then assign attention to the words corresponding to LS(t) as follows:

$$n_i = W_{\text{att4}}^{\text{T}} \tanh(W_{\text{att3}} [h_i; a_I] + b_{\text{att2}}) \tag{7}$$

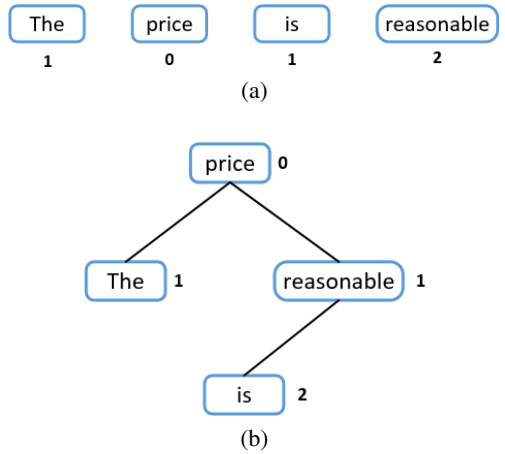

Fig. 2 (a) Position-based distance to target word "price". (b) Syntactics-based distance to target word "price".

$$\beta_i = \frac{\exp(n_i)}{\sum_{j\in \text{LS}(t)} \exp(n_j)} \tag{8}$$

where $i \in \text{LS}(t)$, $W_{\text{att3}} \in \mathbb{R}^{(2D_0+k)\times(2D_0+k)}$, $W_{\text{att4}} \in \mathbb{R}^{(2D_0+k)}$ are weight matrices, and $b_{\text{att2}} \in \mathbb{R}^{2D_0+k}$ is bias.

If the target contains multiple words, the words within the t-step distance of each word in the target are selected. For illustration, we assign a local attention score to each context word:

$$\beta = \begin{cases} 0, & i \notin \text{LS}(t); \\ \frac{\exp(n_i)}{\sum_{j\in \text{LS}(t)} \exp(n_j)}, & i \in \text{LS}(t) \end{cases} \tag{9}$$

The local-attention-based representation is calculated as

$$r_{\text{loc}} = \sum_{i=1}^{L} h_i \beta_i \tag{10}$$

3.4 Gating layer

After obtaining the global and local attention vectors, Multi-AFM uses the gating layer to synthesize specific target information from the local and global attention results. First, the information gate g is calculated by the local and global attention vectors in accordance with Eq. (11):

$$g = \text{sigmoid}(W_{\text{gate}}(r_{\text{glo}} + r_{\text{loc}})) \tag{11}$$

where $g \in \mathbb{R}^{D_0}$, and $W_{\text{gate}} \in \mathbb{R}^{2D_0\times 2D_0}$ is obtained by training. Each dimension in word embedding could reflect different perspectives of word meaning[23]. Therefore, we use a gated unit represented by a vector rather than represented by a scalar. We assume that each dimension of the gating unit controls a different perspective of the attention vector. Afterward, the context is represented as

$$r = r_{\text{glo}} + r_{\text{loc}} \odot g = \sum_{i=1}^{L} h_i \alpha_i + \left(\sum_{i=1}^{L} h_i \beta_i\right) \odot g \tag{12}$$

where $\odot$ represents the Hadamard product.

And we can transform the local attention formula as follows:

$$\tau_{ij} = \beta_i \cdot g_j \tag{13}$$

where g_j represents the j-th dimension of the gating unit g. Due to the role of the hyperbolic tangent function, the value of g_j may be negative. Thus τ_{ij} may be negative, which conflicts with the definition of attention. To make the attention score meaningful, we use Eq. (14) to keep the attention non-negative:

$$t_{ij}^{\text{loc}} = \text{sigmoid}(\beta_i \cdot g_j) \tag{14}$$

Afterward, a normalization function is applied to ensure that the sum of the attention scores of all words in the

j-th dimension is equal to 1:

$$\gamma_{ij}^{\text{loc}} = \frac{t_{ij}^{\text{loc}}}{\sum_{k=1}^{L} t_{kj}^{\text{loc}}} \tag{15}$$

where γ_{ij} is the normalized attention score for w_i on j-th dimension. Finally, the context represented by local attention can be calculated as follows:

$$r_{\text{final}}^{\text{loc}} = \sum_{i=1}^{L} h_i \odot \gamma_i^{\text{loc}} \tag{16}$$

In a similar way, for global attention, $t_{ij}^{\text{glo}} = \text{sigmoid}(\alpha_i)$. We can repeat the above steps to obtain the context of global attention representation:

$$r_{\text{final}}^{\text{glo}} = \sum_{i=1}^{L} h_i \odot \gamma_i^{\text{glo}} \tag{17}$$

Finally, we combine the context representation of the local attention representation with the context representation of the global attention representation to produce the final context representation:

$$r_{\text{final}} = r_{\text{final}}^{\text{loc}} + r_{\text{final}}^{\text{glo}} \tag{18}$$

3.5 Final classification

The representation of the specific target aspect output by the gating layer is fed into the Softmax classifier of size C (C is the number of categories) to output the final sentiment classification result:

$$\hat{y} = \text{softmax}(W_s r_{\text{final}} + b_s) \tag{19}$$

where $\hat{y}$ is the prediction result of sentiment polarity, $W_s \in \mathbb{R}^{2D_0 \times C}$ and $b_s \in \mathbb{R}^C$ are the training parameters in the Softmax layer.

3.6 Training

y stands for the real label and $\hat{y}$ for the predictive label. We use categorical cross entropy with L2-regularizer as loss function to train the network for 30 epochs:

$$L = -\frac{1}{N} \sum_{p=1}^{N} \sum_{q=1}^{C} y_{pq} \log\left(\hat{y}_{pq}\right) + \lambda \|\theta\|_2 \tag{20}$$

where N is the number of samples, p is the index of the sample, q is the classification category, λ is the regularization weight, and θ is the set of parameters that need to be trained in the network. Meanwhile, the Multi-AFM model selects the ADAM algorithm[24] as the optimization algorithm.

4 Experiment

4.1 Experimental parameter and dataset

During the experiment, the values of the hyperparameters are as follows: the dimension of all word embeddings k is 300, the size of the hidden layer D_0 is 300, the initial learning rate lr is 0.001, and the coefficient of L2 regularization λ is 0.0001.

The review data in the field of education are usually not utilitarian. Thus, students usually use concise expressions to describe their feelings. In addition, the expression format is flexible. Hence, the educational data are difficult to use in SA. To test the effectiveness of the Multi-AFM model, we not only conducted experiments on datasets in the field of education but also conducted experiments on datasets in other fields. We used three datasets to evaluate the Multi-AFM model. The label distribution in the datasets is shown in Table 1.

The Education dataset is taken from the course evaluation information of more than 3000 undergraduates from a college in the 2014 – 2017 school year, involving different subjects, grades, and teachers. The dataset focuses on the following four aspects: difficulty, content, practicality, and teacher.

The Course dataset, which comes from the Chinese reviews section of the review set of the education website Coursera (https://www.kaggle.com/septa97/100k-courseras-course-reviews-dataset). The samples in this dataset focus on two aspects: courses and teachers. The Chinese word vectors pre-trained by the Mixed-large comprehensive corpus[25] are used to represent word-level embeddings.

The Restaurant dataset in semeval-2014 ABSA (http://alt.qcri.org/semeval2014/) contains the user's review information. Our goal is to correctly identify the sentiment polarity of the target aspect. The word-level embeddings used in the dataset are obtained using the Glove model[26].

4.2 Model comparison

To conduct a comprehensive evaluation of the Multi-AFM model, we used five models for comparison.

TD-LSTM[27] employs two LSTM networks and uses the target and left and right contexts to model. Finally, the left and right target dependencies are connected to predict the sentiment polarity of the target.

AE-LSTM[10] propagates sentences through the LSTM network and then embeds the word's hidden state

Table 1 Polarity distribution of labels in dataset samples.

Dataset	Positive		Neutral		Negative	
	Train	Test	Train	Test	Train	Test
Education	2481	810	289	60	1065	347
Course	424	104	95	14	41	27
Restaurant	2164	728	633	196	805	196

and aspect into the joint to generate an attention vector. This vector is used to generate the final representation of aspect-level sentiment classification, which is finally sent to the Softmax classifier.

ATAE-LSTM[10] was developed based on AE-LSTM. ATAE-LSTM further enhances the effect of aspect embedding and represents the context by embedding the aspect with each word embedding vector.

IAN[11] inputs the target aspect and its context to two LSTM networks. The output of the hidden layer is the intermediate aspect representation and context representation, respectively. Attention scores are generated from the hidden outputs of the two LSTM networks to generate the final aspect and context representation. The two vectors are connected and input to the Softmax classifier for final classification.

RAM[28] extends MemNet by applying multi-hop attention on the output of Bi-LSTM rather than word embeddings. Moreover, a recurrent function is applied between multiple attentions to model the inner dependencies.

Among the five models, TD-LSTM and AE-LSTM belong to the neural network method. Although ATAE-LSTM and IAN belong to the attention-based method, they do not consider the advantages of local attention. RAM uses multi-attention but does not consider the fusion of local and global attention. Compared with the above models, it can efficiently test the performance of the Multi-AFM model.

4.3 Evaluation indicator and experimental result

In this paper, the classification accuracy rate, precision, recall, and macro-F1 score[20] are used as the evaluation indicators for the three datasets. Each value in the experimental results is the highest value of each indicator in many times experiments.

Taking the experiment of the Multi-AFM model in the Education dataset as an example, we use different hyperparameter values to conduct some comparative experiments, such as the learning rate lr and the coefficient of L2 regularization λ, the experimental results are shown in Table 2.

Table 2 shows that the Multi-AFM model can obtain good results, with minimal difference in the experimental results for different hyperparameter values. Therefore, the performance of the Multi-AFM model is relatively stable, and the setting of hyperparameter values will not significantly affect the experimental results.

In the course of the experiments, similar hyperparameter experiment methods are adopted for the Course and Restaurant datasets to obtain the best performance of the Multi-AFM model in different datasets. The experimental results of different models in the three datasets are shown in Table 3.

Table 2 Performance comparison of Multi-AFM on Education dataset when using different hyperparameter values.

(%)

Hyper parameter (lr, λ)	Accuracy	Precision	Recall	F1 score
$(10^{(-3)}, 10^{(-3)})$	94.1	91.6	92.8	89.9
$(10^{-3}, 5\times10^{-4})$	94.1	**92.2**	93.1	90.0
$(10^{-3}, 10^{-4})$	**94.6**	**92.2**	**93.6**	**90.7**
$(5\times10^{-4}, 10^{-3})$	94.1	91.8	92.9	90.2
$(5\times10^{-4}, 5\times10^{-4})$	**94.6**	92.1	**93.6**	90.2
$(5\times10^{-4}, 10^{-4})$	**94.6**	92.0	**93.6**	**90.7**
$(10^{-4}, 10^{-3})$	93.7	90.7	92.8	90.1
$(10^{-4}, 5\times10^{-4})$	93.8	90.6	93.0	89.7
$(10^{-4}, 10^{-4})$	93.8	90.8	92.7	89.8

From the experimental results in Tables 3 and 4, due to the particularity of the Chinese language, in many cases, the same Chinese word represents a noun, a verb, or an adjective. Therefore, in the case of using a Chinese corpus, the performance of the IAN model is unstable, because it may produce incorrect semantic understanding during the cross-calculation of the attention of the target and the context. Similarly, the RAM model does not perform well in understanding word semantics and finding the distance between each

Table 3 Performance comparison of each model on Education dataset.

(%)

Model	Accuracy	Precision	Recall	F1 score
TD-LSTM	92.4	86.9	88.7	86.0
AE-LSTM	93.8	89.5	90.3	88.2
ATAE-LSTM	**94.6**	91.7	91.1	**91.2**
IAN	89.9	91.8	76.9	74.9
RAM	91.0	80.0	78.7	79.2
Multi-AFM	**94.6**	**92.2**	**93.6**	90.7

Table 4 Performance comparison of each model on Course dataset.

(%)

Model	Accuracy	Precision	Recall	F1 score
TD-LSTM	79.3	73.8	68.0	68.1
AE-LSTM	77.2	72.5	62.4	64.6
ATAE-LSTM	78.6	72.5	60.2	65.0
IAN	78.0	69.9	58.3	57.0
RAM	80.7	83.6	59.9	62.7
Multi-AFM	**81.4**	**86.3**	**72.5**	**69.2**

word and the target. Therefore, the effectiveness of local attention obtained from the dependency tree is shown from the result, and the advantages of attention fusion are better reflected.

The experimental results in Table 5 show that although the Multi-AFM model does not have the highest classification accuracy, it is more stable in terms of precision, recall, and F1 score. RAM improves MemNet by modeling contextual information with bidirectional LSTM and combining features from different attentions non-linearly with a recurrent function. Compared with RAM using a multilevel attention architecture, our model performs better on multiple evaluation indicators, because it can maximize the global information and syntax-based local information instead of allocating attention weights based on word distance to the target. Thus, Multi-AFM can adjust attention weights dynamically. Moreover, the structure of Multi-AFM is simpler than that of RAM, because it does not need to perform multi-hop attention and additional recurrent functions to merge these attention results.

On the basis of the above discussion, Multi-AFM outperforms the baseline models. The Multi-AFM model is improved compared with the method without using local attention, which verifies the effectiveness of deep learning algorithm using local attention.

To further confirm that the fusion of attention improves the classification ability of the model, we remove the local attention and only use global attention in the Multi-AFM model, which is recorded as GAM. We remove the global attention and only use local attention in the Multi-AFM model, denoted as LAM. We use GAM and LAM to perform comparative experiments on the Restaurant and Education datasets, respectively. The experimental results are shown in Table 6.

From the experimental results in Table 6, the Multi-AFM model performs better than the GAM and LAM models. For the GAM model, syntax-based local information that may contain pure opinion modifiers is ignored. Therefore, it may capture irrelevant sentiment words that may affect the final prediction result. The LAM model understands the semantic information of a sentence from a local view. It can extract more accurate information related to the target; however, it will lose a lot of information outside the local view because of the complexity of natural language.

Table 5 Performance comparison of each model on Restaurant dataset.

(%)

Model	Accuracy	Precision	Recall	F1 score
TD-LSTM	75.6	71.0	65.4	67.8
AE-LSTM	76.2	70.0	61.0	63.4
ATAE-LSTM	77.2	63.7	59.1	60.1
IAN	78.6	69.0	68.8	68.9
RAM	**80.2**	74.3	67.8	69.0
Multi-AFM	79.6	**76.8**	**69.0**	**70.3**

Table 6 Performance comparison of each model on different datasets.

(%)

Dataset	Model	Accuracy	Precision	Recall	F1 score
Education	GAM	92.9	86.8	86.0	85.6
	LAM	92.9	89.4	89.1	86.7
	Multi-AFM	**94.6**	**92.2**	**93.6**	**90.7**
Restaurant	GAM	78.2	72.6	66.4	68.7
	LAM	78.0	70.3	65.9	67.3
	Multi-AFM	**79.6**	**76.8**	**69.0**	**70.3**

5 Conclusion

This paper proposes the multi-attention fusion model Multi-AFM for ABSA tasks. This method considers the impact of local attention on contextual representation while using global attention. Multi-AFM controls the weight of global attention and local attention fusion through the gating layer, to achieve improved classification results. Comparative experiments on the datasets in the education field and other similar fields were performed using different methods to test the performance of the Multi-AFM model. Experimental results show that the Multi-AFM model performs better than existing attention-based methods. The proposed Multi-AFM model provides a new method for teaching evaluation in the field of education. In recent years, transformer-based models have been widely used to solve sentiment analysis tasks[29]. For long-input tasks, transformers have huge computational complexity, resulting in slow training speed, and the overall structure of transformers is more complicated than that of LSTM. In addition, the Multi-AFM model demonstrates good overall performance after optimization. In the future, we will try to consider adopting transformer-based related models and using multi-class sentiment analysis[30] to mine fine-grained sentiment polarity, such as happiness, joy, anger, and disgust.

Acknowledgment

The research work was partially supported by the National Natural Science Foundation of China (No. 61976247) and Southwest Jiaotong University Education Reform Project (No. 20201010).

References

[1] N. Bousbia and I. Belamri, Which contribution does EDM provide to computer-based learning environments? in *Educational Data Mining*, A. Peña-Ayala, ed. Springer, 2014, pp. 3–28.

[2] G. Wolfgang and D. Hendrik, Translating learning into numbers: A generic framework for learning analytics, *Educational Technology & Society*, vol. 15, no. 3, pp. 42–57, 2012.

[3] V. Kalakoski, H. Ratilainen, and L. Drupsteen, Enhancing learning at work. How to combine theoretical and data-driven approaches, and multiple levels of data? in *Proc. 23rd European Symp. Artificial Neural Networks, Computational Intelligence and Machine Learning*, Bruges, Belgium, 2015, pp. 331–336.

[4] J. Z. Liao, J. Y. Tang, and X. Zhao, Course drop-out prediction on MOOC platform via clustering and tensor completion, *Tsinghua Science and Technology*, vol. 24, no. 4, pp. 412–422, 2019.

[5] T. Nasukawa and J. Yi, Sentiment analysis: Capturing favorability using natural language processing, in *Proc. 2nd Int. Conf. Knowledge Capture*, Sanibel Island, FL, USA, 2003, pp. 70–77.

[6] B. Liu, Sentiment analysis and opinion mining, *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[7] B. Pang, L. Lee, and S. Vaithyanathan, Thumbs up?: Sentiment classification using machine learning techniques, in *Proc. ACL-02 Conf. Empirical Methods in Natural Language Processing*, Philadelphia, PA, USA, 2002, pp. 79–86.

[8] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[9] X. J. Zhou, X. J. Wan, and J. G. Xiao, Attention-based LSTM network for cross-lingual sentiment classification, in *Proc. 2016 Conf. Empirical Methods in Natural Language Processing*, Austin, TX, USA, 2016, pp. 247–256.

[10] Y. Q. Wang, M. L. Huang, X. Y. Zhu, and L. Zhao, Attention-based LSTM for aspect-level sentiment classification, in *Proc. 2016 Conf. Empirical Methods in Natural Language Processing*, Austin, TX, USA, 2016, pp. 606–615.

[11] D. H. Ma, S. J. Li, X. D. Zhang, and H. F. Wang, Interactive attention networks for aspect-level sentiment classification, in *Proc. 26th Int. Joint Conf. Artificial Intelligence*, Melbourne, Australia, 2017, pp. 4068–4074.

[12] B. Liu, S. J. Tang, X. G. Sun, Q. Y. Chen, J. X. Cao, J. Z. Luo, and S. S. Zhao, Context-aware social media user sentiment analysis, *Tsinghua Science and Technology*, vol. 25, no. 4, pp. 528–541, 2020.

[13] X. Han, B. Y. Li, and Z. R. Wang, An attention-based neural framework for uncertainty identification on social media texts, *Tsinghua Science and Technology*, vol. 25, no. 1, pp. 117–126, 2020.

[14] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in *Proc. 32nd Int. Conf. Machine Learning*, Lille, France, 2015, pp. 2048–2057.

[15] A. M. Rush, S. Chopra, and J. Weston, A neural attention model for abstractive sentence summarization, in *Proc. 2015 Conf. Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 2015, pp. 379–389.

[16] X. Li, L. D. Bing, W. Lam, and B. Shi, Transformation networks for target-oriented sentiment classification, in *Proc. 56th Annu. Meeting of the Association for Computational Linguistics*, Melbourne, Australia, 2018, pp. 946–956

[17] M. T. Luong, H. Pham, and C. D. Manning, Effective approaches to attention-based neural machine translation, in *Proc. 2015 Conf. Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 2015, pp. 1412–1421.

[18] K. H. Chen, R. Wang, M. Utiyama, E. Sumita, and T. J. Zhao, Syntax-directed attention for neural machine translation, in *Proc. 32nd AAAI Conf. Artificial Intelligence*, New Orleans, LA, USA, 2018, pp. 4792–4799.

[19] R. D. He, W. S. Lee, H. T. Ng, and D. Dahlmeier, Effective attention modeling for aspect-level sentiment classification, in *Proc. 27th Int. Conf. Computational Linguistics*, Santa Fe, NM, USA, 2018, pp. 1121–1131.

[20] S. Wang, S. Mazumder, B. Liu, M. W. Zhou, and Y. Chang, Target-sensitive memory networks for aspect sentiment classification, in *Proc. 56th Annu. Meeting of the Association for Computational Linguistics*, Melbourne, Australia, 2018, pp. 957–967.

[21] J. W. Duan, X. Ding, and T. Liu, Learning sentence representations over tree structures for target-dependent classification, in *Proc. 2018 Conf. North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, New Orleans, LA, USA, 2018, pp. 551–560.

[22] J. J. Wang, J. Li, S. S. Li, Y. Y. Kang, M. Zhang, L. Si, and G. D. Zhou, Aspect sentiment classification with both word-level and clause-level attention networks, in *Proc. 27th Int. Joint Conf. Artificial Intelligence*, Stockholm, Sweden, 2018, pp. 4439–4445.

[23] H. Choi, K. Cho, and Y. Bengio, Context-dependent word representation for neural machine translation, *Computer Speech & Language*, vol. 45, pp. 149–160, 2017.

[24] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, in *Proc. 3rd Int. Conf. Learning Representations*, San Diego, CA, USA, 2015.

[25] S. Li, Z. Zhao, R. F. Hu, W. S. Li, T. Liu, and X. Y. Du, Analogical reasoning on chinese morphological and semantic relations, in *Proc. 56th Annu. Meeting of the Association for Computational Linguistics*, Melbourne, Australia, 2018, pp. 138–143.

[26] J. Pennington, R. Socher, and C. Manning, GloVe: Global vectors for word representation, in *Proc. 2014 Conf. Empirical Methods in Natural Language Processing*, Doha, Qatar, 2014, pp. 1532–1543.

[27] D. Y. Tang, B. Qin, X. C. Feng, and T. Liu, Effective LSTMs for target-dependent sentiment classification, in *Proc. 26th Int. Conf. Computational Linguistics*: *Technical Papers*, Osaka, Japan, 2016, pp. 3298–3307.

[28] P. Chen, Z. Q. Sun, L. D. Bing, and W. Yang, Recurrent attention network on memory for aspect sentiment analysis, in *Proc. 2017 Conf. Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 2017, pp. 452–461.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł Kaiser, and I. Polosukhin, Attention is all you need, in *Proc. 31st Int. Conf. Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 6000–6010.

[30] M. Bouazizi and T. Ohtsuki, Multi-class sentiment analysis on twitter: Classification performance and challenges, *Big Data Mining and Analytics*, vol. 2, no. 3, pp. 181–194, 2019.

Guanlin Zhai received the BS degree from Southwest University of Science and Technology, Mianyang, China in 2016 and the MS degree from Southwest Jiaotong University, Chengdu, China in 2020. From 2019 to 2020, he has been an academic visitor in the Centre for Maritime Studies (CMS), National University of Singapore. His current research interests include data mining, natural language processing, and machine learning.

Yan Yang received the BS and MS degrees from Huazhong University of Science and Technology, Wuhan, China in 1984 and 1987, respectively. She received the PhD degree from Southwest Jiaotong University, Chengdu, China in 2007. From 2002 to 2003 and 2004 to 2005, she was a visiting scholar with the University of Waterloo, Waterloo, Canada. She is currently a professor and vice dean with the School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China. Her current research interests include multi-view learning, big data analysis and mining, ensemble learning, semi-supervised learning, and cloud computing.

Heng Wang received the BS degree from Southwest University of Science and Technology, Mianyang, China in 2017 and the MS degree from Southwest Jiaotong University, Chengdu, China in 2020. His current research interests include data mining, multi-view learning, natural language processing, and machine learning.

Shengdong Du received the BS and MS degrees in computer science from Chongqing University in 2004 and 2007, respectively. He received the PhD degree from Southwest Jiaotong University, Chengdu, China in 2020. His research interests include data mining and machine learning.