IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 25, 2021, accepted July 21, 2021, date of publication July 26, 2021, date of current version August 5, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3100299

Multi-Grained Attention Representation With ALBERT for Aspect-Level Sentiment Classification

YUEZHE CHEN[1], LINGYUN KONG[1,2], YANG WANG[1,2], AND DEZHI KONG[3]
[1]School of Science, Xijing University, Xi'an 710123, China
[2]Artificial Intelligence Laboratory, School of Science, Xijing University, Xi'an 710123, China
[3]School of Mechanical Engineering, Northwestern Polytechnical University, Xi'an 710072, China

Corresponding author: Yuezhe Chen (chen1995086@163.com)

This work was supported in part by the Key Research and Development Program of Shanxi Province under Grant 2019GY-025, in part by the Science and Technology Plan Project of Xi'an under Project 2020KJRC0134, and in part by the Special Fund for High Level Talents of Xijing University under Grant XJ20B07.

ABSTRACT Aspect-level sentiment classification aims to solve the problem, which is to judge the sentiment tendency of each aspect in a sentence with multiple aspects. Previous works mainly employed Long Short-Term Memory (LSTM) and Attention mechanisms to fuse information between aspects and sentences, or to improve large language models such as BERT to adapt aspect-level sentiment classification tasks. The former methods either did not integrate the interactive information of related aspects and sentences, or ignored the feature extraction of sentences. This paper proposes a novel multi-grained attention representation with ALBERT (MGAR-ALBERT). It can learn the representation that contains the relevant information of the sentence and the aspect, while integrating it into the process of sentence modeling with multi granularity, and finally get a comprehensive sentence representation. In Masked LM (MLM) task, in order to avoid the influence of aspect words being masked in the initial stage of the pre-training, the noise linear cosine decay is introduced into $n-gram$. We implemented a series of comparative experiments to verify the effectiveness of the method. The experimental results show that our model can achieve excellent results on Restaurant dataset with numerous number of parameters reduced, and it is not inferior to other models on Laptop dataset.

INDEX TERMS Aspect-level sentiment classification, ALBERT, natural language processing, deep learning.

I. INTRODUCTION

The process of judging the polarity of product reviews is sentiment analysis. Document-level sentiment classification, similar to text classification, is to classify a document or a comment [1], [2]. In fact, the rough classification does not have much value of practical applications. In the real world, a typical comment usually contains opinions on multiple aspects. Therefore, the emotional classification of fine-grained aspects is called one of the important research directions. Aspect-level Sentiment Classification is one of the sub-tasks of Aspect-based Sentiment Analysis (ABSA). Its goal is to classify different aspects of customer reviews (e.g., positive, negative, or neutral). For example, the sentence "*This computer has a big screen, but the battery is too bad.*", there are two aspects in the comment that require polarity classification, namely "*screen*" and "*battery*". At the same time, the opinion words "*big*" and "*bad*" are strongly related to the polar words "*positive*" and "*negative*" respectively.

Many scholars have studied aspect-level sentiment classification models. Machine learning algorithms to build classifiers are often used by researchers under supervised learning [3]–[5]. One of the classic methods is feature-based support vector machine(SVM) [6], [7]. Another excellent representative method is an end-to-end method that uses an attention mechanism to learn better word representations based on neural networks [8]. Because it does not need to deal with tedious feature engineering and has excellent results, this method is favored by many researchers. Among them, in the previous research, LSTM fusion attention mechanism has obtained excellent results [9]–[12]. In addition, Attention over Attention (AOA) solves the problem that the above

The associate editor coordinating the review of this manuscript and approving it for publication was Gang Mei.

method only considers the attention from aspect to text. It can simultaneously generate attention from aspect to text and text to aspect [13]–[15]. It is true that aspect phrases often have multiple words, but their models are too focused on solving the problem of the correlation between context and aspect. This is not conducive to judging logical and ironic comments. Still, the example cited above, "*This computer has a big screen, but the battery is too bad.*", the turning semantics of "*but*" is instructive to the judgment of the polarity of "*battery*". On the other hand, the result of inaccurate sentence representation combined with aspect is also poor. Therefore, we assume that this shortcoming is caused by the inability of word embedding to be encoded within the context.

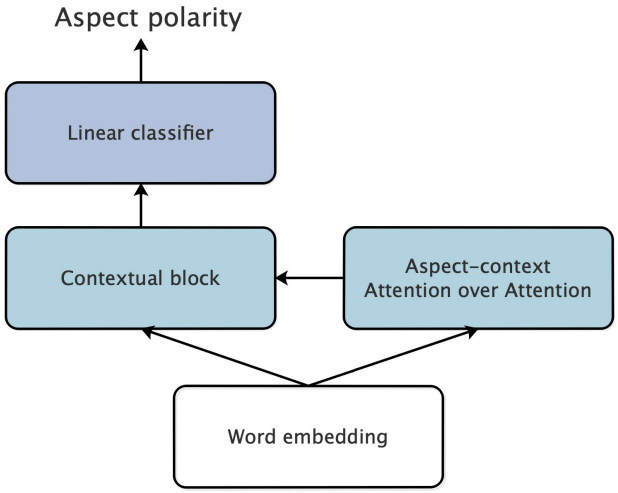

FIGURE 1. The basic architecture of MGAR.

In recent years, a contextual language model that gets rid of the constraints of traditional RNN on data space and time has emerged, namely the Transformer model [16]. Subsequently, Transformer-based language killers followed one after another, for example, OpenAI GPT and BERT applied Transformer encoder and decoder respectively [17]–[20]. Although these models can capture the word features of the same word in different contexts, the number of parameters is too large and requires powerful computing power. In order to ensure that the number of parameters is reduced without reducing the effect, many researchers use knowledge distillation based on the teacher-student model [21]–[23]. Another method takes parameter sharing with a typical example —— ALBERT [24]. We will improve ALBERT to cater to the context of learning of aspect-level sentences.

In this paper, we integrate interactive information about aspects and context into the sentence representation encoded in the context. Our structure mainly includes contextual blocks based on ALBERT processing contextual features and interactive attention modules from aspect to sentence and sentence to aspect, which we call Aspect-Context Attention over Attention (AC-AOA). The structure is shown in Figure 1. In the contextual block, in order to make ALBERT more compatible with ASC tasks, we modified Masked LM task and introduced AC-AOA information in the contextual block of different levels. The purpose is to integrate the aspect-context information into the language model at different learning stages, which helps the language model to understand aspect features. In addition, LSTM is no longer required to model sentences and words by AC-AOA (this part is mainly completed by the context block), and we are inspired by the Transformer model [16], using a multi-head AC-AOA similar to the context block.

Our main contributions of this paper are summarized as follows: (1) We propose Multi-head Aspect-Context Attention over Attention (Multi AC-AOA), which focuses on interactively extracting features between aspect and context. (2) We contributed a contextual block based on ALBERT. In this part, we modified Masked LM task of ALBERT, which makes it more compatible with aspect-level sentiment classification task and more effective learning sentence representation [24]. (3) Unlike other models that only consider above one of the two points, we integrate different levels of information between aspect and context into sentence representations, forming MGAR-ALBERT. In addition, our experimental results show that compared with other existing methods, the model has better results on Restaurant dataset, also has a certain degree of competitiveness on Laptop dataset from SemEval-2014 Task4 as well as [25].

II. RELATED WORK

A. SENTIMENT CLASSIFICATION

Sentiment classification is one of the common problems in the field of natural language processing. The goal is to detect the polarity of emotions in the text, which is often regarded as a binary or multi-classification problem. The early methods proposed use log-linear regression model to identify whether the semantics of adjectives in a large corpus is positive or negative [26]. Later, a lexicon-based and rule-based method was proposed. The pros and cons of the method largely depended on manual design and prior knowledge [27], [28]. Alternatively, traditional machine learning algorithms, whose core is feature engineering, are used in sentiment classification tasks, such as support vector machines, Naive Bayes, and maximum entropy models [3], [4], [29], [30]. However, in order to obtain high-quality annotation data, feature engineering requires a lot of labor costs.

B. SENTIMENT CLASSIFICATION AT ASPECT-LEVEL WITH NEURAL NETWORK

With the deepening of research, it is found that the same sentence has multiple aspects showing emotional tendencies. The aspect-level sentiment classification is to solve such problems, that is, to distinguish the sentiment polarity of various aspects of the text. In recent years, with the development of neural networks, aspect-level sentiment classification has made enormous progress. Target-Dependent Long Short-Term Memory (TD-LSTM) is an extension of the LSTM model. LSTM is used on both sides of the target

word to better capture context information [31]. Wang *et al.* [32] proposed Attention-based LSTM with Aspect Embedding (ATAE-LSTM), which uses the attention mechanism to integrate into the LSTM network. The attention mechanism can effectively quantify the importance of a word in the sentence, and this mechanism is also used in many other models. Ma *et al.* [11] considered the possibility of multiple words in the target word, and proposed the Interactive Attention Networks (IAN). The model adds a layer of attention operation to calculate the weight of the target word. It applies two LSTMs to calculate the hidden state of the target and the context, respectively, as well as then uses two attention layers to exchange important information about the target and the context. Finally, the fused representation is classified into polarity. Tay *et al.* [10] emphasized not using naive concatenations to model the similarity of word-aspect, so they innovatively introduced the Word-Aspect Fusion Attention Layer, which has two hybrid methods of cyclic correlation and cyclic convolution. Subsequently, the structure named Attention over Attention by Cui *et al.*, an improved version of the attention mechanism, was first applied in the field of cloze-style reading comprehension and achieved excellent results [13]. Then there are many models that borrow AOA to implement aspect-level sentiment analysis tasks [14], [15]. This structure can calculate the attention of both the query and the document at the same time, and can benefit from the mutual information. We were inspired by it and found that there is a similar relationship between aspects and sentences in aspect-level sentiment classification. However, there is a problem with the above methods. Their research focuses on understanding the connection between aspects and sentences, while relatively neglecting the learning of the sentence itself. Recent years, graph neural network also has a strong development in this field. One of the typical models is ASEGCN [33].

With the emergence of more and more machine learning application scenarios, labeling data for each application scenario requires huge labor costs. Tan *et al.* [34] first tried to transfer pre-trained examples labeled in the old domain to unlabeled ones in the new domain. Sentiment analysis at the aspect level also faces a huge shortage of training samples, which largely affects the performance of neural networks. Because LSTM is difficult to parallelize calculations and long-term memory network will bring difficulties to calculating the gradient, many researchers hire transfer learning to solve such problems [35]–[37]. Then, the BERT-based model has become the preferred transfer target for researchers. Sun *et al.* [38] established an auxiliary sentence for the aspect, forming a sentence pair similar to question answering (QA) or natural language inference (NLI), and transformed the ABSA problem into a sentence pair classification task. Xu *et al.* [39] takes an aspect and a review sentence containing the aspect as an input pair, and converts ABSA into a review reading comprehension task (RRC).

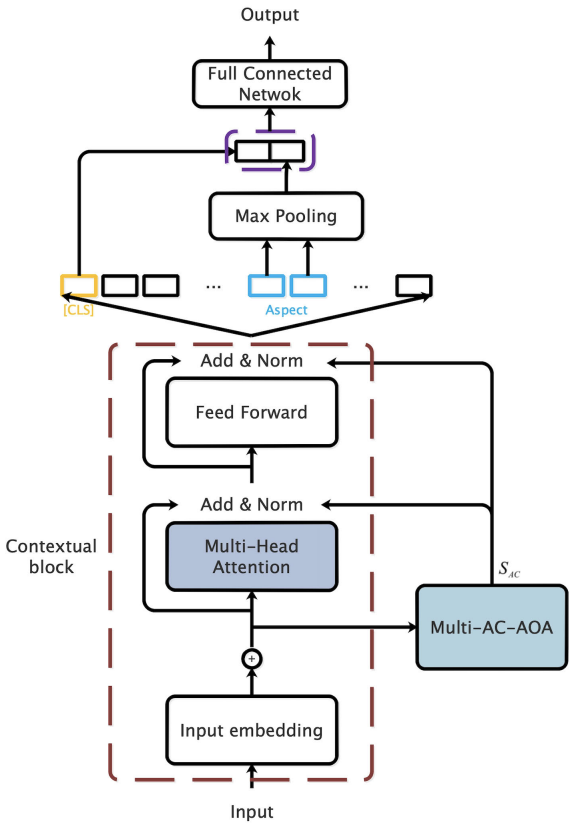

FIGURE 2. The architecture of MGAR-ALBERT.

III. MODEL ARCHITECTURE

The backbone of MGAR-ALBERT is similar to ALBERT, which uses a transformer encoder [16]. Traditional ABSA mainly uses crossed attention and LSTM for classification, and does not distinguish attention within context from attention between aspect and context. Our model includes two parts: Contextual block and Multi-AC-AOA. The architecture of MGAR-ALBERT is shown in Figure 2.

A. TASK DEFINITION

The ABSA task is usually given a tuple (S, A) containing a review sentence and one or more aspects to predict the polarity y of the aspect. The relationship between aspects and sentences is expressed as follows:

$$y_l \propto f(S, A_l) \tag{1}$$

where l represents the l-th aspect.

Assuming that the review sentence

$$S = [w_1, w_2, \ldots, w_n] \tag{2}$$

contains n words and the l-th aspect

$$A_l = [a_1, a_2, \ldots, a_m] \tag{3}$$

contains m words, then there are 3 categories of predicted output.

$$y_l = \{positive, neutral, negative\} \tag{4}$$

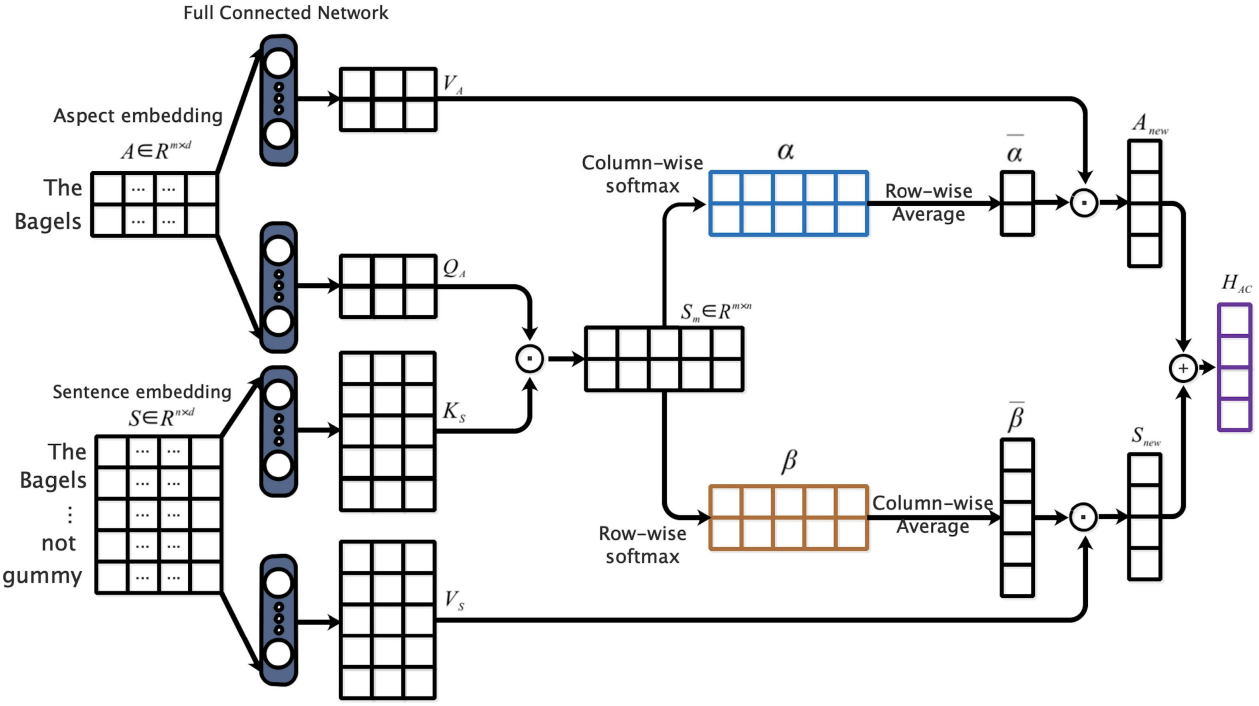

FIGURE 3. The architecture of AC-AOA.

B. ASPECT-CONTEXT ATTENTION OVER ATTENTION

The AC-AOA network mainly integrates Transformer's self-Attention ideas into AOA [13], [16]. The goal of this mechanism is to focus on solving the problem of relevance between the aspect and the context, not to learn the relevance between the aspect and the context and the word weight within the context at the same time. The architecture of AC-AOA is shown in Figure 3.

We first map aspect and sentence respectively to get embedding, which represents the semantic meaning of each word w. Each word vector is given by $e_w \in R^{|V| \times d}$, where $|V|$ is the vocabulary size and d is embedding dimension. Then form the aspect matrix $A = [e_{a_1}, e_{a_2}, \ldots, e_{a_m}]$, $A \in R^{m \times d}$ and the contextual matrix $S = [e_{w_1}, e_{w_2}, \ldots, e_{w_n}]$, where m is the number of aspect words and n is the number of sentence words. A uses a fully connected neural network to map the query vector Q_A and the value vector V_A, and S generates the key vector K_S and the value vector V_S in the same way.

$$Q_A = ReLU(A \cdot W_q + b_q) \tag{5}$$
$$V_A = ReLU(A \cdot W_{va} + b_{va}) \tag{6}$$
$$K_S = ReLU(S \cdot W_k + b_k) \tag{7}$$
$$V_S = ReLU(S \cdot W_{vs} + b_{vs}) \tag{8}$$

where W_q, W_{va}, W_k and W_{vs} represent the weight matrices of the fully connected layers, b_q, b_{va}, b_k and b_{vs} are bias vectors, and $ReLU$ is an activation function [40].

The score matrix is calculated by $S_m = Q_A \cdot K_S$, performs column-wise softmax and row-wise softmax on Sm to convert the value between 0 and 1, and gets the context-to-aspect correlation coefficient matrix $\alpha \in R^{m \times n}$ and the aspect-to-context correlation coefficient matrix $\beta \in R^{m \times n}$. Each column of α represents the weight distribution of the aspect in the context, and each row of β represents the weight distribution of the context in the aspect. In the next step, each row of α is averaged to obtain a vector $\bar{\alpha}$ and each column of β is averaged to obtain a vector $\bar{\beta}$. The formula is described as follows:

$$\bar{\alpha}_c = \frac{1}{n} \sum_{c=1}^{n} \frac{\exp(S_{m_{rc}})}{\sum_{r=1}^{m} \exp(S_{m_{rc}})} \tag{9}$$
$$\bar{\beta}_r = \frac{1}{m} \sum_{r=1}^{m} \frac{\exp(S_{m_{rc}})}{\sum_{c=1}^{n} \exp(S_{m_{rc}})} \tag{10}$$

where r represents the number of the current row, c represents the number of the current column.

Next, the mean-valued context-to-aspect attention weight distribution $\bar{\alpha}$ is dot-producted with the value of aspect V_A to obtain a new aspect representation A_{new}. Similarly, a new sentence representation S_{new} can be obtained.

$$A_{new} = \bar{\alpha} \cdot V_A \tag{11}$$
$$S_{new} = \bar{\beta} \cdot V_S \tag{12}$$

The output after a weighted average of A_{new} and S_{new} is used as the output of AC-AOA.

$$H_{AC} = \frac{1}{2}(A_{new} + S_{new}) \tag{13}$$

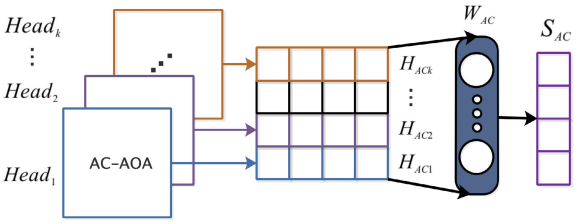

FIGURE 4. The architecture of Multi-AC-AOA.

Finally, the architecture of Multi-AC-AOA is shown in Figure 4, the output H_{ACi} of each head is concatenated and then mapped to obtain a new sentence representation S_{AC}. Its output formula is as follows:

$$S_{AC} = CONCAT(H_{AC1}, \ldots, H_{ACk})W^{AC} \tag{14}$$

where k is the number of heads.

C. CONTEXTUAL BLOCK

1) ARCHITECTURE

The encoder structure of the popular Transformer are applied by the contextual block [16]. As shown in Figure 2, the main framework is similar to ALBERT [24], which contains two sub-layers, namely multi-head self-attention mechanism and position-wise fully connected feed-forward network. In order to solve the degradation problem caused by the increase in the number of network layers, the residual network is also applied [41]. The main difference between the contextual block and ALBERT is that the output of Multi-AC-AOA, which contains the sentence representation of the information between the aspect and the context, is introduced, so that the contextual block focuses on the calculation of the context's own attention. The network layer will be described next.

Since this model uses parallel computing to solve the timing problem, position coding is required. The formula for position coding is as follows:

$$P(i, 2j) = \sin(i/10000^{\frac{2j}{d}}) \tag{15}$$

$$P(i, 2j+1) = \cos(i/10000^{\frac{2j}{d}}) \tag{16}$$

where i is the position of the word in the sentence, j is the position of the word vector, and d is the dimension of the word vector.

Then, using the scaled dot product attention, the query Q, the key K and the value V are calculated as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V \tag{17}$$

where queries Q, keys K and values V are the mapping matrix of input embedding, d is the number of columns of Q, K and V.

Multi-Head Attention is equivalent to the integration of multiple different self-attentions, the formula is as follows:

$$\begin{aligned} MULTIHEAD(H_E) &= CONCAT(head_1, \ldots, head_k)W^O \\ where\ head_i &= Attention(H_E W_i^Q, H_E W_i^K, H_E W_i^V) \end{aligned} \tag{18}$$

where H_E is the input embedding after position encoding, and k is the number of heads.

In the propagation process, in order to improve the problem of gradient disappearance, a residual layer is added and layer normalization is performed [41], [42]. And in this step, the output of Multi-AC-AOA is merged for the first time.

$$\begin{aligned} H_{MultiHead}(H_E, S_{AC}) = LayerNorm(&\frac{1}{2}(H_E + S_{AC}) \\ &+ MULTIHEAD(H_E)) \end{aligned} \tag{19}$$

where S_{AC} is the output of Multi-AC-AOA, which is a sentence representation that integrates the interactive information between aspects and sentences.

The other sublayer is a fully connected feed-forward network, which mainly contains two linear transformations, as well as there is a *ReLU* function between the two. Similarly, as with multi-head attention, layer normalization is also adopted by this sublayer. This is the second time to fuse with the output of Multi-AC-AOA.

$$\begin{aligned} FFN(H_{MultiHead}) = \max(0, H_{MultiHead}W_1 + b_1)W_2 \\ + b_2 \end{aligned} \tag{20}$$

$$\begin{aligned} H_{FFN}(H_{MultiHead}, S_{AC}) = LayerNorm(&\frac{1}{2}(H_{MultiHead} + S_{AC}) \\ &+ FFN(H_{MultiHead})) \end{aligned} \tag{21}$$

In order to comprehensively consider the internal learning of the sentence and the learning of the correlation between the aspect word and the sentence, in formulas 19 and 21, we incorporate the output of Multi-AC-AOA into the two stages of the contextual block. Our inspiration is mainly that human beings have different understandings of a certain thing at different stages of learning.

2) PRE-TRAINING TASK

ALBERT is an excellent language model. In order to reduce the number of parameters, it uses two strategies — factorized embedding parameterization and cross-layer parameter sharing. We also use this strategy to reduce the number of parameters from $O(V \times H)$ to $O(V \times E + E \times H)$, and make each contextual block use the same parameter matrix. However, because the masked object of ALBERT's Masked LM (MLM) task is not specifically considered for aspect-level sentiment classification tasks, it will affect the judgment of the sentence on the aspect if the aspect-related keywords are masked in the early stage of training. We use the n-gram mask to generate a mask for the MLM target. The input needs to determine whether it is related to the aspect. If it is related to it, the probability of masking the word is reduced. The length of each n-gram mask is randomly selected. The probabilities of length n are listed as follows:

$$P_{MLM}(n) = \begin{cases} \dfrac{1/n}{\rho \sum_{k=1}^{N} 1/k} & if \approx aspect \\ \dfrac{1/n}{\sum_{k=1}^{N} 1/k} & otherwise \end{cases} \tag{22}$$

TABLE 1. Statistics of the datasets.

Dataset	Train	Test	Positive		Neural		Negative	
			Train	Test	Train	Test	Train	Test
Restaurant	3608	1120	2164	728	637	196	807	196
Laptop	2328	638	994	341	464	169	870	128

where $\rho = \rho_{base} \cdot \tau$, ρ_{base} is the original coefficient, and τ is the noise linear cosine decay rate. And the maximum length of n-gram is set to 3.

D. OUTPUT LAYER AND MODEL TRAINING

The final representation $S_f \in R^{(n+1)\times d}$ of the sentence is provided by the hidden state of the last layer.

$$S_f = [w_{f_0}, w_{f_1}, w_{f_2}, \ldots, w_{f_n}] \tag{23}$$

where n is the length of the sentence, d is the dimension of the word vector, and w_{f_0} is the vector of the classification mark [CLS] of the sentence. The final representation of the aspect word $A_f \in R^{m\times d}$ is the sub-matrix of S_f.

$$A_f = [a_{f_i}, a_{fi+1}, a_{f_{i+2}}, \ldots, a_{f_{i+m-1}}] \tag{24}$$

where m represents the length of the aspect phrase. A_f is selected to run max pooling. This output f and the representation of [CLS] are concatenated as the final representation of the classification [43]. Finally, the classification representation is sent to the fully connected network layer and the softmax layer for classification.

$$O_f = Softmax(CONCAT(w_{f_0}, \max\{0, A_f\})W_f + b_f) \tag{25}$$

For the loss function, cross entropy with L2 regularization is applied to training. The given training data (S_i, A_i, y_i) comes from corpus C, where S_i is the i-th sentence, A_i is the i-th aspect in the sentence, and y_i is the true sentiment polarity in S_i correspond to A_i. The Adam algorithm is used to minimize the error between the predicted value $\widehat{y}$ and the true value y_i [44]. The equation of the loss function is as follows:

$$L(\theta) = E_{(S_i,A_i,y_i)\sim C}[-\log P_\theta(y_i|f_i)] + \frac{\delta}{2}||\theta||_2^2 \tag{26}$$

where θ denotes all trainable parameters and δ is a L2 regularization.

IV. EXPERIMENTS

A. EXPERIMENT SETTING

The datasets we used are named Laptop and Restaurants from SemEval-2014 Task4 in the experiments [25]. The aspect terms and polarities of the dataset reviews are labeled by professional annotators and divided into three categories: positive, neutral and negative. In addition, the current popular accuracy and Macro-F1 are used as our standard for evaluating models. The distribution of each dataset classified by emotional polarity is listed in Table 1.

In our experiments, we applied the pre-trained ALBERT-base parameters to initialize the contextual block and fine-tune it. ALBERT-based model is configured with a 12-layer network and we use a word embedding size $d = 128$. Multi-AC- AOA with the number of heads set to 6 is the focus of model training. All weights in this part are initialized with normal distribution and the input embedding is shared with the contextual block. The batch size on Restaurant and Laptop dataset is set to 128, and the Adam optimizer with a learning rate of 0.001 is selected [44]. The main hyper-parameter setting is shown in Table 2.

We format our inputs as "$[CLS], w_1, w_2, \ldots, A, \ldots, w_n$", where $A = [a_1, \ldots, a_m]$, n is the number of sentence words and m is the number of aspect words. The specific example is shown in Figure 5.

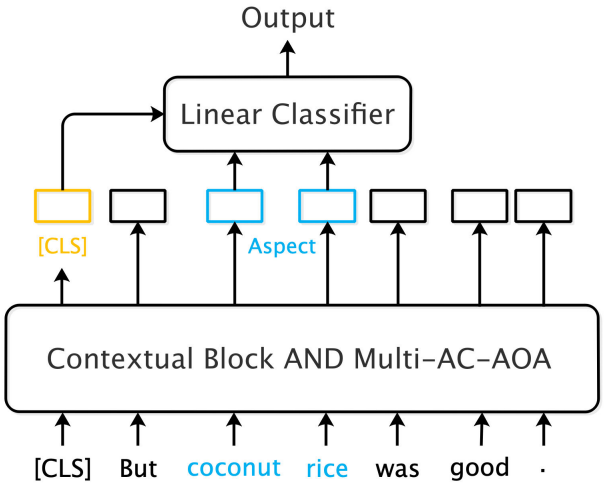

FIGURE 5. Input/output example of the architecture of MGAR-ALBERT.

B. COMPARED METHODS

To verify the effectiveness of our model, 11 representative models were selected for comparison. The compared models are summarized below.

LSTM is a panacea model in the field of natural language processing. It is based on the one-way sequence model of the recurrent neural network model, which generates a hidden state for each word, and the last hidden state is used as a vector for sentiment classification [9].

TD-LSTM uses LSTM on both sides of the target word to model the context information, and the hidden states on both sides of the target word are classified as the final representation [31].

ATAE-LSTM, an extension of the AT-LSTM network, is an attention-based LSTM architecture. After the

TABLE 2. The final value of the main hyper-parameters for the datasets. L.R, B.S, H.D, E.S, L2, MSL represent the learning rate, batch size, hidden dimension, the number of embedding sizes, L2 regularization and maximum sequence length, respectively.

Dataset	L.R	B.S	H.D	Layers	E.S	Heads	L2	DR	Classifier DR	MSL
Restaurant	le-3	128	768	12	128	6	le-3	0.2	0.1	512
Laptop	le-3	128	768	12	128	6	le-4	0.2	0.1	512

combination of word embedding and aspect embedding is passed through LSTM, the hidden states are obtained and the attention vector is calculated after concatenating the hidden states with the learned aspect vector. The final sentence representation is the sum of the weights of the hidden vector [32].

MemNet employs a multi-hop attention mechanism for contextual word embedding. The goal is to obtain the degree of relevance of contextual words and finally capture the sentence representation [45].

IAN firstly considers that aspect terms may have multiple words. The input embeddings containing sentence representation and aspect embedding are respectively input into two LSTMs, and all hidden states are averaged to obtain the hidden state representations of aspect words and sentences, which are used as the subsequent interactive information of the two. According to the word weight, the representations of the aspect words and sentences are recalculated, and then the two are spliced as the final representation to send to softmax for sentiment classification [11].

RAM uses bidirectional LSTM (BiLSTM) to generate memory from the input, and weights memory slices to targets, according to their relative positions, so that different targets in the same sentence have their own customized memory. After that, multiple attention is performed on the position-weighted memory, and the attention results are non-linearly combined with recurrent networks (i.e. GRUs). Finally, softmax is performed on the output of GRU to predict the emotion of the target [46].

AOA-LSTM uses BiLSTM to convert sentences and aspect embeddings into hidden states respectively, and uses the AOA mechanism to interactively merge the hidden states of the two to calculate the comprehensive weights with sentence to aspect and aspect to sentence. Finally, the final sentence representation is calculated through the weight and the hidden state of the sentence [14].

The main improvement of **MGAN** is to propose a fine-grained attention mechanism and a multi-grained attention mechanism to comprehensively consider the relevance of sentences and aspects, and finally integrate multi-grained attention as a classification vector [12].

DMMN-SDCM is based on a memory network and introduces semantic analysis information to guide the attention mechanism to build a Deep Mask Memory Network (DMMK) as well as effectively learn the information provided by other non-target aspects. At the same time, the context moment proposed by the model is embedded in the sentiment classification of the entire sentence and is designed to provide background information for the target aspect [47].

BERT-PT proposes Review Reading Comprehension (RRC) inspired by machine reading comprehension (MRC). The model uses BERT as the basic model, and proposes a joint post-training approach to enhance the representation of knowledge related to aspect-level sentiment classification tasks [39].

AEN-BERT uses the attention mechanism to model the context and targets based on the trained BERT. This paper raised the problem of label unreliability, and the model added label smoothing and regularization. The goal of this approach is to encourage the model to reduce the consistency of fuzzy labels [48].

C. MAIN RESULTS

The experimental results of the compared models are reported in Table 3. First, we observe that our proposed MGAR-ALBERT outperforms all the compared models on Restaurant dataset. In fact, its accuracy and Macro-F1 surpass AEN-BERT by 2%-5% respectively. On Laptop dataset, the effect of our model is basically the same as BERT-PT, which is also a good performance. Such results are sufficient to prove the effectiveness of the model.

Because LSTM simulates the human memory process of a certain thing well and its storage space is still very small, it is always favored by researchers. LSTM pushes the experimental results to a new benchmark. It can be seen from the table that the attention mechanism is indeed very popular due to the fact that there are different degrees of association between words. Many models based on the attention mechanism, which surpass LSTM-based models, have been proposed. In general, Attention-based models are better than LSTM-based models. In addition, LSTM has achieved unexpected results under the blessing of AOA. Since the probability of multiple aspect words in a sentence on Restaurant dataset is greater than that on Laptop dataset, MGAR-ALBERT that integrates the interactive information between aspect words and sentences is more prominent on Restaurant dataset. Therefore, the connection between aspect words and sentences cannot be ignored, especially on Restaurant dataset. In summary, whether it is LSTM-based models or Attention-based models, MGAR-ALBERT surpasses them when it integrates interactive information.

Next, we analyze the performance of our model on Laptop dataset. By comparing the methodology with other models, we summarized some neglected factors. First of all,

TABLE 3. Comparison of accuracy and Macro-F1 with different models. Conflict data removed from SemEval-2014 datasets. The results with "♯" are copied from the DMMN-SDCM paper and those with symbol "♭" are our reimplemented as well as others results are retrieved from published papers. '-' means not reported. The best and second best scores in each column are shown in bold and underlined fonts respectively.

	Model	Laptop		Restaurant	
		ACC	Macro-F1	ACC	Macro-F1
LSTM-baselines	LSTM	65.82♭	64.02♭	74.61♭	63.56♭
	TD-LSTM	71.83♯	68.43♯	78.00♯	66.73♯
Attention-baselines	ATAE-LSTM	68.65♯	62.45♯	77.23♯	64.95♯
	MemNet	70.33♯	64.09♯	78.16♯	65.83♯
	IAN	72.10♯	67.48♯	77.95♯	67.90♯
	RAM	75.01♯	70.51♯	79.79♯	68.86♯
	MGAN	75.39	72.47	81.25	71.94
	DMMN-SDCM	77.59♯	73.61♯	81.16♯	71.50♯
	AOA-LSTM	74.5	-	81.2	-
BERT-baselines	BERT-PT	<u>78.07</u>	75.08	<u>84.95</u>	<u>76.96</u>
	AEN-BERT	**79.93**	**76.31**	83.12	73.76
Our Model	MGAR-ALBERT no AC-AOA	75.45	71.31	82.57	72.23
	MGAR-ALBERT	77.98	<u>75.85</u>	**85.13**	**77.68**

DMMN-SDCM introduces semantic dependency information to guide the attention mechanism to construct a deep mask memory network, and it is more special that it can effectively consider information provided by other non-target aspects. Thus, its performance on Laptop dataset is more prominent. Our model ignores the above-mentioned semantic dependency and non-target aspect information. Additionally, with the advent of excellent BERT, many BERT-based sentiment classification models have been produced, typically AEN-BERT, whose results greatly exceed the previous methods. AEN-BERT adds label smoothing and regularization to improve the problem of label unreliability. This also leads to a relatively better performance of the model on Laptop dataset. Compared with AEN-BERT, although our model is slightly inferior on Laptop dataset, our model can still achieve this result when the number of parameters is greatly reduced. Now, with the explosive growth of the number of model parameters, we believe that the reduction of the number of parameters is particularly important. In order to reflect the effect of our method, we conducted ablation experiments. According to the paper method, we conducted the same experiment using ALBERT. MGAR-ALBERT model is improved by 2% to 4% with Multi-AC-AOA than without Multi-AC-AOA. This also fully proves that Multi-AC-AOA makes up for many losses due to the reduction of the number of parameters. In summary, the experimental results in the table show that its inherent ability to learn a large number of features makes it surpass many past models. Objectively, it is slightly inferior to AEN-BERT on Laptop dataset. However, on Restaurant dataset, our proposed MGAR-ALBERT achieves the best performance, and the results on both datasets are better than ALBERT.

To conclude, we verified the effectiveness of MGAR-ALBER through comparative experiments. The advantage of this method lies in the use of a powerful language extractor to ensure the understanding of the semantic meaning of the sentence, and at the same time special learning for the relevance of aspects and sentences.

D. CASE STUDY

The contribution of MGAR-ALBERT lies in the targeted learning of the dependence between words as well as between aspects and sentences. This section will give examples to analyze the effectiveness of each part.

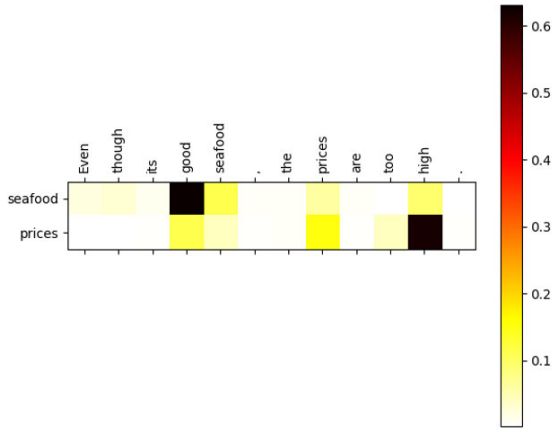

FIGURE 6. Example of weight assignment from aspect to context. The degree of importance is represented by the shade of the color.

An example is given separately from the Restaurant and Laptop datasets. The weight distribution of aspects and sentences is visualized, and the degree of association between words is represented by the depth of color. Black represents the greatest relevance and white represents the smallest. As shown in Figure 6, for example, "*Even though its good seafood, the prices are too high.*". This example contains two aspects: "*seafood*" and "*prices*". The weight ratio of "*good*" in the sentence is the largest for the aspect word

"*seafood*". When the aspect word is "*prices*", "*high*" gets the highest weight in the sentence. First of all, the visualization results show that the polarity of "*seafood*" is positive, the polarity of "*prices*" is opposite, and the transition word such as "*Even though*" is also assigned a part of the weight. This shows that AC-AOA has the ability to extract key information for the opposite polarity. In addition, the model has a certain shielding ability for characters that are not related to polarity, such as "*the*", "*are*" and ".", which are shown in light colors in the figure. Unexpectedly, when there are different aspects of the same sentence, not only the core words related to the polarity of this aspect are given a higher weight, but other related words also have a certain proportion. From this point, we speculate that the model also has certain contextual analysis capabilities.

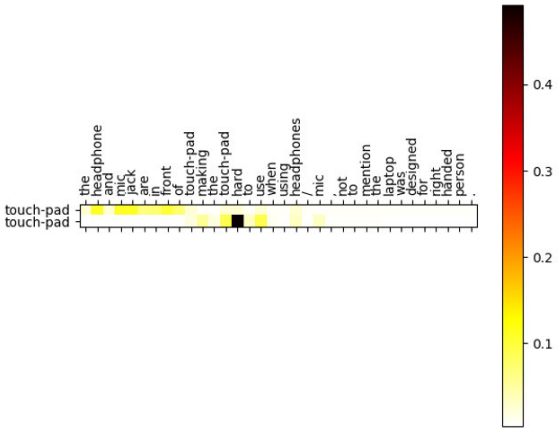

FIGURE 7. Example of weight assignment from multiple aspects to context. The degree of importance is represented by the intensity of the color.

As shown in Figure 7, an example of the Laptop dataset is " *The headphone and mic jack are in front of touch-pad making the touch-pad hard to use when using headphones / mic, not to mention the laptop was designed for right handed person.*". This sample contains multiple aspects, namely "*mic jack*", "*touch-pad*", "*headphones*" and "*mic*", where "*touch-pad*" appears twice. This part mainly focuses on case analysis that contains multiple same aspects in the sentence. For convenience, only the parts of the same aspect are extracted from weight visualization. In Figure 7, the "*touch-pad*" in the first and second rows is the weight distribution of the word in the sentence for the first and second occurrences. When the "*touch-pad*" first appeared, the important part of the attention weight is assigned to the first half sentence and the weight values are basically equal. Its polarity is judged to be "neutral". The "*hard*" in the second line of the visualization has the highest proportion and the polarity is "negative". This shows that AC-AOA also has a better ability to distinguish when there are multiple identical aspects in a sentence.

Figure 8 reveals the visualization results of the final representation of the multi-head attention mechanism. The sample

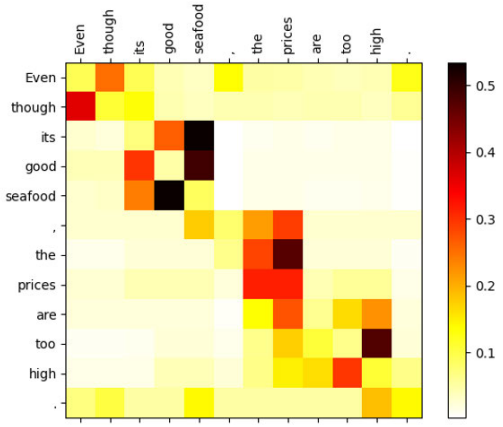

FIGURE 8. An example of the weight distribution of the self-attention mechanism. The degree of importance is represented by the intensity of the color.

used in this section is consistent with the restaurant reviews mentioned above. The contextual block mainly focuses on the learning of the association between words in the sentence. On the whole, the area with large weight distribution forms a diagonal line from upper left to lower right. This is in line with the basic common sense that the distance between words is inversely proportional to the degree of association between the two. Secondly, the weight between "*Even*" and "*though*" is relatively large. On the one hand, the word distance between the phrases is very close. On the other hand, the phrase is very common in life and can be easily used as data for training. Finally, characters of no actual semantic meaning, such as commas and periods, bear relatively weak weight in sentences, but they have a slightly larger weight on transition words, which shows that punctuation still has a certain guiding significance for understanding transition word.

In general, both Multi-AC-AOA and context blocks have excellent performance in their respective modeling tasks.

V. CONCLUSION

BERT's excellent text representation ability performs well on many NLP tasks, including aspect-level sentiment classification. However, it is not enough to understand the sentence itself. The study of the relationship between key information and sentences is also an indispensable part of sentiment analysis. But we found that the design of special models for different tasks is limited by the excessive number of original model parameters. In this paper, we proposed a novel neural network model called Multi-Grained Attention Representation with ALBERT for Aspect-Level Sentiment Classification to handle aspect-level sentiment analysis tasks. Our contribution is mainly to design a special network to integrate various aspects of information, and to simulate the human learning process to optimize the pre-training task under the reduced number of basic model parameters. The special design network not only makes up for the loss of

effect due to the reduction of the number of parameters, but also brings better results. We believe that our special design network can contribute to fields where emotional orientation needs to be considered.

In future work, because the judgment of emotional tendency is multidimensional, we need to design more functional networks to realize a model that can fully understand emotion. We may further research on knowledge distillation to find higher-quality basic models. For additional special networks, we will also consider adding some important elements that have been overlooked, such as semantic dependency information and aspect extraction.

REFERENCES

[1] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *Proc. ICML*, 2011, pp. 1–8.
[2] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, nos. 1–2, pp. 1–135, 2008.
[3] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Rep., Stanford*, vol. 1, no. 12, p. 2009, 2009. [Online]. Available: https://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf
[4] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," 2002, *arXiv:cs/0205070*. [Online]. Available: https://arxiv.org/abs/cs/0205070
[5] A. Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," *Comput. Intell.*, vol. 22, no. 2, pp. 110–125, 2010.
[6] S. M. Mohammad, S. Kiritchenko, and X. Zhu, "NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets," 2013, *arXiv:1308.6242*. [Online]. Available: http://arxiv.org/abs/1308.6242
[7] S. Liu, F. Li, F. Li, X. Cheng, and H. Shen, "Adaptive co-training SVM for sentiment classification on tweets," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2013, pp. 2079–2088.
[8] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier, "Effective attention modeling for aspect-level sentiment classification," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 1121–1131.
[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[10] Y. Tay, L. A. Tuan, and S. C. Hui, "Learning to attend via word-aspect associative fusion for aspect-based sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–8.
[11] D. Ma, S. Li, X. Zhang, and H. Wang, "Interactive attention networks for aspect-level sentiment classification," 2017, *arXiv:1709.00893*. [Online]. Available: http://arxiv.org/abs/1709.00893
[12] F. Fan, Y. Feng, and D. Zhao, "Multi-grained attention network for aspect-level sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 3433–3442.
[13] Y. Cui, Z. Chen, S. Wei, S. Wang, T. Liu, and G. Hu, "Attention-over-attention neural networks for reading comprehension," 2016, *arXiv:1607.04423*. [Online]. Available: http://arxiv.org/abs/1607.04423
[14] 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 Predict. Behav. Represent. Modeling Simulation*. Cham, Switzerland: Springer, 2018, pp. 197–206.
[15] Z. Wu, Y. Li, J. Liao, D. Li, X. Li, and S. Wang, "Aspect-context interactive attention representation for aspect-level sentiment classification," *IEEE Access*, vol. 8, pp. 29238–29248, 2020.
[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*. [Online]. Available: http://arxiv.org/abs/1706.03762
[17] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018. [Online]. Available: https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf
[18] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.
[19] T. B. Brown *et al.*, "Language models are few-shot learners," 2020, *arXiv:2005.14165*. [Online]. Available: http://arxiv.org/abs/2005.14165
[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805
[21] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*. [Online]. Available: http://arxiv.org/abs/1503.02531
[22] R. Tang, Y. Lu, L. Liu, L. Mou, O. Vechtomova, and J. Lin, "Distilling task-specific knowledge from BERT into simple neural networks," 2019, *arXiv:1903.12136*. [Online]. Available: http://arxiv.org/abs/1903.12136
[23] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, "MobileBERT: A compact task-agnostic BERT for resource-limited devices," 2020, *arXiv:2004.02984*. [Online]. Available: http://arxiv.org/abs/2004.02984
[24] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," 2019, *arXiv:1909.11942*. [Online]. Available: http://arxiv.org/abs/1909.11942
[25] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. Al-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. de Clercq, V. Hoste, M. Apidianaki, X. Tannier, N. Loukachevitch, E. Kotelnikov, N. Bel, S. M. Jiménez-Zafra, and G. Eryigit, "SemEval-2016 task 5: Aspect based sentiment analysis," in *Proc. Int. Workshop Semantic Eval.*, 2016, pp. 19–30.
[26] V. Hatzivassiloglou and K. McKeown, "Predicting the semantic orientation of adjectives," in *Proc. 35th Annu. Meeting Assoc. Comput. Linguistics, 8th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 1997, pp. 174–181.
[27] N. Kaji and M. Kitsuregawa, "Building lexicon for sentiment analysis from massive collection of html documents," in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn. (EMNLP-CoNLL)*, 2007, pp. 1075–1083.
[28] D. Rao and D. Ravichandran, "Semi-supervised polarity lexicon induction," in *Proc. 12th Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL)*, 2009, pp. 675–682.
[29] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing," *Comput. Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.
[30] R. Moraes, J. F. Valiati, and W. P. G. Neto, "Document-level sentiment classification: An empirical comparison between SVM and ANN," *Expert Syst. Appl.*, vol. 40, no. 2, pp. 621–633, Feb. 2013.
[31] D. Tang, B. Qin, X. Feng, and T. Liu, "Effective LSTMs for target-dependent sentiment classification," 2015, *arXiv:1512.01100*. [Online]. Available: http://arxiv.org/abs/1512.01100
[32] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based LSTM for aspect-level sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 606–615.
[33] Y. Xiao and G. Zhou, "Syntactic edge-enhanced graph convolutional networks for aspect-level sentiment classification with interactive attention," *IEEE Access*, vol. 8, pp. 157068–157080, 2020.
[34] S. Tan, G. Wu, H. Tang, and X. Cheng, "A novel scheme for domain-transfer problem in the context of sentiment analysis," in *Proc. 16th ACM Conf. Conf. Inf. Knowl. Manage. (CIKM)*, 2007, pp. 979–982.
[35] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier, "Exploiting document knowledge for aspect-level sentiment classification," 2018, *arXiv:1806.04346*. [Online]. Available: http://arxiv.org/abs/1806.04346
[36] Y. A. Winatmoko, A. A. Septiandri, and A. P. Sutiono, "Aspect and opinion term extraction for hotel reviews using transfer learning and auxiliary labels," 2019, *arXiv:1909.11879*. [Online]. Available: http://arxiv.org/abs/1909.11879
[37] Y. Liang, F. Meng, J. Zhang, J. Xu, Y. Chen, and J. Zhou, "An iterative knowledge transfer network with routing for aspect-based sentiment analysis," 2020, *arXiv:2004.01935*. [Online]. Available: http://arxiv.org/abs/2004.01935
[38] C. Sun, L. Huang, and X. Qiu, "Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence," 2019, *arXiv:1903.09588*. [Online]. Available: http://arxiv.org/abs/1903.09588
[39] H. Xu, B. Liu, L. Shu, and P. S. Yu, "BERT post-training for review reading comprehension and aspect-based sentiment analysis," 2019, *arXiv:1904.02232*. [Online]. Available: http://arxiv.org/abs/1904.02232
[40] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[42] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*. [Online]. Available: http://arxiv.org/abs/1607.06450
[43] Z. Gao, A. Feng, X. Song, and X. Wu, "Target-dependent sentiment classification with BERT," *IEEE Access*, vol. 7, pp. 154290–154299, 2019.
[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[45] D. Tang, B. Qin, and T. Liu, "Aspect level sentiment classification with deep memory network," 2016, *arXiv:1605.08900*. [Online]. Available: http://arxiv.org/abs/1605.08900
[46] P. Chen, Z. Sun, L. Bing, and W. Yang, "Recurrent attention network on memory for aspect sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 452–461.
[47] P. Lin, M. Yang, and J. Lai, "Deep mask memory network with semantic dependency and context moment for aspect level sentiment classification," in *Proc. IJCAI*, Aug. 2019, pp. 5088–5094.
[48] Y. Song, J. Wang, T. Jiang, Z. Liu, and Y. Rao, "Attentional encoder network for targeted sentiment classification," 2019, *arXiv:1902.09314*. [Online]. Available: http://arxiv.org/abs/1902.09314

YUEZHE CHEN is currently pursuing the M.S. degree with the School of Science, Xijing University, China. His research interests include sentiment analysis, natural language processing, and deep learning.

LINGYUN KONG received the B.S. and M.S. degrees from Air Force Engineering University, China, in 1988 and 2001, respectively, and the Ph.D. degree from Northwestern Polytechnical University, China, in 2011. He is currently a Professor with the School of Science, Xijing University, China. His research interests include robots engineering and novel control methods.

YANG WANG received the B.S., M.S., and Ph.D. degrees from Northwestern Polytechnical University, China, in 2011, 2014, and 2018, respectively. His research interests include nonlinear control system theory and robust control.

DEZHI KONG received the B.S. degree from Northeastern University, China, in 2014, and the M.S. degree from The University of Sydney, Australia, in 2017. He is currently pursuing the Ph.D. degree with Northwestern Polytechnical University, China. His research interests include robot control and nonlinear control theory.