IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 18, 2020, accepted February 2, 2020, date of publication February 10, 2020, date of current version February 18, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2972697

Aspect-Context Interactive Attention Representation for Aspect-Level Sentiment Classification

ZHUOJIA WU[1], YANG LI[1], JIAN LIAO[1], DEYU LI[1,2], XIAOLI LI[3], AND SUGE WANG[1,2]

[1]School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China
[2]Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006, China
[3]Institute for Infocomm Research, A*STAR, Singapore 138632

Corresponding author: Suge Wang (wsg@sxu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61632011, Grant 61573231, Grant 61672331, Grant 61906112, and Grant 61603229, in part by the Key Research and Development Program of Shanxi Province through the International Cooperation under Grant 201803D421024 and Grant 201903D421041, in part by the Natural Science Foundation of Shanxi Province under Grant 201901D211174, and in part by the Scientific and Technological Innovation Programs of Higher Education Institutions in Shanxi under Grant 2019L0008.

ABSTRACT Aspect-level sentiment classification aims to determine sentiment polarities of various aspects in reviews, where each review typically contains multiple aspects, that may correspond to different polarities. Aspect-level sentiment classification, unlike document-level sentiment classification, requires different context representations for different aspects. Existing methods normally use Long Short-Term Memory (LSTM) network to model aspects and contexts separately, and they combine attention mechanisms to extract features of a specific aspect in its context. Attention mechanisms are not used for sequence modeling, so aspects are not considered when generating context sequence representations. This study proposes a novel aspect-context interactive representation structure that only relies on an attention mechanism for generating sequence-to-sequence representations in both context and aspect. It is capable of extracting features related to the specific aspect during the process of its context sequence modeling, and generating a high quality aspect representation simultaneously. We have conducted comprehensive experiments to compare with thirteen existing methods. Our experimental results show that the proposed model is able to achieve significantly better performance on Restaurant dataset, as well as very competitive results on Laptop and Twitter datasets.

INDEX TERMS Aspect-level sentiment classification, attention mechanism, interactive representation, natural language processing.

I. INTRODUCTION

Aspect-level sentiment classification aims to determine the reviewer's sentiment tendency of a service or product, based on a given review text [1], which is a sub-task of more comprehensive sentiment analysis [2]. In particular, *aspect*-level sentiment classification is more fine-grained than *document*-level sentiment classification [3], [4]. It is more valuable in many real-world applications, as its sentiment polarity focuses on each specific aspect, instead of a mixed document-level polarity, which usually lacks of detailed sentiment information for different aspects. For example, the sentence "*The menu is limited, but most of the dishes are excellent.*" evaluates the restaurant object based on its two aspects, namely "*menu*" and "*dishes*". In addition, the opinion words "*limited*" and "*excellent*" correspond to negative and positive sentiment polarities of two aspects "*menu*" and "*dishes*", respectively. Typically a review often contains various aspects with their corresponding sentiments that occur in complex context, e.g. Jiang *et al.* [5] reported that 40% of the prediction errors in sentiment classification are caused when the aspects (opinion targets) are not considered. Therefore, the biggest challenge on aspect-level sentiment classification is how to obtain correct sentiment information for a specific aspect when multiple aspects and sentiments are mixed together.

We observe attention-based LSTM network has delivered good results in previous studies [6]–[9]. These meth-

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

ods mainly rely on attention mechanism to select important information associated with the specified aspect in its context, although LSTM fails to identify whether a word in its context is related to the specified aspect. For example, the Interactive Attention Network (IAN) is a typical attention-based LSTM network proposed by Ma *et al.* [7], which first employs LSTM's hidden states to generate representations for aspect and context separately (i.e. averaging all the word states in sequence as aspect/context final representations), subsequently obtains important information contributing to judging sentiment polarity from aspect and its context through an interactive attention network. While interactivity is also one of the important characteristics of our method, there is a key difference, that is our interactive attention can leverage the relationships among all the word-pairs between aspects and contexts (i.e. one word from an aspect while the other word from its context), enabling us to learn better representations by capturing aspect and context relations in fine-grained word-level and explicitly emphasizing those important interactions between keywords in an aspect and its context. Additionally, the time-ordered structure of LSTM can only process text sequences word by word, thus training LSTM for datasets with lengthy texts is time-consuming. Recently Vaswani *et al.* [10] proposed novel Transformer that is the first to rely entirely on the attention mechanism to calculate the representations of sequences without using recurrent neural networks. In the same spirit, We also explore to address the biggest challenge of the aspect-level sentiment classification with a network architecture that only contains attention mechanisms.

It is clear that aspect representation is critical for identifying its sentiment polarity. Traditional methods either use the same LSTM for aspect modeling [7], [8], or simply average its word embeddings [6]. Note aspects may contain multiple words, e.g., "*chicken with black bean sauce*", "*chicken with chili and lemon grass*". Clearly, both aspects contain a common head word "chicken" to indicate that they belong to the same aspect category. We observe aspects that belong to same category, such as chicken, should have common or similar modifiers, for instance, "*chewy*", "*tender*" and "*tough*", no matter what kind of chickens they are. It has been shown that after assigning aspect terms into their corresponding aspect categories, determining the sentiment polarities of high level aspect categories can lead to more accurate results [6], [11]. This result also proves that information among aspects of the same category can be shared. When modeling aspect phrases (an aspect with multiple words), extracting the common category information of different aspect phrases is beneficial for determining sentiment polarities. However, LSTM cannot identify the words in phrases that most effectively represent the category of aspect phrases.

More specifically, our interactive structure has two advantages over the separate LSTM structure. Firstly, it can take advantage of the interactive information between an aspect and its context when modeling text sequences. The weight differences generated by attention mechanism cause our model to focus on the features that are conducive to the specified aspect polarity discrimination (these differences exist in each word of the aspect phrases and its contexts). Secondly, it works well for aspect representations, which can automatically identify the most representative words in the aspect phrases. To the best of our knowledge, this is the first work leveraging the interaction among word-pairs between contexts and aspects to better presentation aspects.

This paper proposes a novel attention-based sequence modeling method for aspect-context interactive representation. In summary, our main contributions are can be summarized as follows:

- The representation structure of **Multi**-layer **A**spect-**C**ontext **I**nteractive **A**ttention (MultiACIA) can generate a sequence representation of contexts by focusing on those words associated with a specific aspect. It is more suitable than LSTM for handling mixed information of multiple aspects in a context. Furthermore, MultiACIA can be parallelized to effectively reduce the training time.
- The proposed interactive structure can model aspect phrases simultaneously. Particularly, when aspects contain multiple words, MultiACIA can focus on the representative words such as "*chicken*" and assign lower weights to auxiliary words "*with black bean sauce*", which is critical for sentimental classification, as our model can focus on those core aspect words by ignoring less irrelevant words.
- Our experimental results show that the proposed model, comparing with thirteen existing methods, is able to achieve significantly better performance on Restaurant dataset, as well as very competitive results on Laptop and Twitter datasets.

II. RELATED WORK

Aspect-level sentiment classification, which is also known as target-oriented sentiment classification [12], was originally performed using rule-based and statistic-based methods [5], [13]. In particular, matching is performed according to some predefined rules to determine the sentiment polarity of an aspect [14], [15]. Alternatively, by employing a bag-of-words model and designing a set of features such as a sentiment lexicon, a traditional machine learning method, such as the SVM or naive Bayesian model, is subsequently employed to perform sentiment classification task [16]–[19]. These methods and corresponding results, however, heavily depend on the quality of manually designed rules and their feature engineering outcomes.

In recent years, neural network methods such as recurrent neural networks, recursive neural networks, convolutional neural networks, and gated neural networks, have been widely used for aspect-level sentiment classification [20]–[24]. Attention mechanisms were first proposed to solve problems in the field of visual images [25] and then applied in natural language processing (NLP) to handle machine translation problem [26]. Attention mechanisms can assign different weights to different words, for focusing on important parts of a sentence. Therefore, they can effectively be employed to improve the accuracy of aspect-level sentiment classification [6], [7], [27].

To utilize aspect information properly, Wang *et al.* [6] incorporated it when modeling contexts using LSTM and focused on words related to sentiment polarities in a given context using attention mechanisms. However, it is not efficient to consider aspect information based on word embedding stitching. Ma *et al.* [7] emphasized the importance of aspects and proposed an approach based on aspect and context separation representations by modeling aspects using LSTM. This interactive attention mechanism is capable of generating context representations based on aspects and aspect representations based on context. However, this method ignores the interactions between words in aspects and contexts when calculating attention weights because pooling operations are performed after the LSTM outputs are obtained. Huang *et al.* [8] applied the attention-over-attention (AOA) mechanism on aspect-level sentiment classification. The AOA mechanism was first proposed by Cui *et al.* [28] to learn the relationships between words in contexts and problems for question answering. The AOA mechanism not only highlights the words in a sentence that contribute to sentiment expression of a specific aspect, but also assigns different weights to words in a specific aspect phrase. Although various forms of attention mechanisms have been used to model the relationships between aspects and contexts, most studies have used attention-based LSTM architectures. The AOA-LSTM and TNet-LF/AS models delivered state-of-the-art performance on two SemEval-2014 datasets.

When annotating aspect-level sentiment classification training data, all aspects in the review and their corresponding sentiment polarities need to be labeled. Due to the difficulty of annotating, existing public datasets for aspect-level sentiment classification are relatively small, which greatly limits the effectiveness of neural models [29]. Some studies try to transfer knowledge from document-level to aspect-level [29], [30]. He *et al.* [29] used pre-training and multi-task learning to achieve the purpose of transferring knowledge. Zhuang et al. use a capsule network that mainly includes aspect routing and dynamic routing methods to transfer knowledge [30]. Other studies include various pre-trained language models represented by BERT [31]–[33]. These models are constantly obtaining the state-of-the-art results on many NLP tasks. On aspect-level sentiment classification, there are also some methods that combine BERT with this task and obtain strong results [34]–[37]. Nevertheless, their performances largely depend on BERT's powerful language feature representation capabilities, and training large BERT model requires massive datasets, powerful GPU clusters and considerable training time. For our study, the focus is not on how to more effectively combine pre-trained models with aspect-level sentiment classification. Instead, we focus on how to better handle the essential and more complicated problem, that is the mixing of multiple aspects and sentiments in a review [8], [29], [38].

III. THE PROPOSED MODEL

A. PROBLEM DEFINITION

Assume that a review S contains multiple aspects. We use T_i to denote the i-th aspect. A review $Sentence = [w_1, w_2, \ldots, w_n]$ consists of n words, and an aspect $Target_i = [w_j, \ldots, w_{j+m-1}]$ is a phrase consisting of one or more words in a review, where m is the length of the phrase. Aspect-level sentiment classification aims to predict the sentiment polarity corresponding to a specific aspect in review. The sentiment polarity $y_i \in \{-1, 0, 1\}$ corresponds to "positive", "neutral" and "negative" sentiments of a reviewer. The relationships between sentiments and aspects can be represented as follows:

$$y_i \propto f(S, T_i) \tag{1}$$

For a different aspect T_i in the same S, the model must to recognize the corresponding evaluation words in their context.

B. ASPECT-CONTEXT INTERACTIVE ATTENTION

We map each word into a low-dimensional space, and a real-valued vector is used to represent the semantic features of each word. For each word w, the vector is given by $e_w \in R^{d_w \times |V|}$, where d_w is the embedding dimension and $|V|$ is the vocabulary size. This representation is called word embeddings [39]. Word embeddings are used as the parameters to adjust model training, which makes them suitable for our task.

Assume that there is a query vector and a matrix of key-value pairs with corresponding relationships. The output of the attention function is computed as a weighted sum of values. Where the weights are computed using a compatibility function between queries and keys [10]. On aspect-level sentiment classification, the traditional practice is to treat specific aspects as queries, keys and values are mapped from contexts. We propose the Aspect-Context Interactive Attention (ACIA) mechanism, based on interactive attention calculations between aspects and contexts. ACIA can assign different weights to different words in the aspect and context simultaneously. Additionally, the proposed model is an integrated structure that does not require two separate attention mechanisms such as the interactive attention network (IAN) [7]. The architecture of ACIA is shown in Fig.1.

$S = [e_{w_1}, e_{w_2}, \ldots, e_{w_n}]$ and $T_i = [e_{w_j}, \ldots, e_{w_{j+m-1}}]$ are context and aspect matrices respectively. To calculate interactive attention, we first map S and T (For simplicity, we use T to denote T_i) using fully connected layer networks to generate key-value pairs $< K_S, V_S >$ and $< K_T, V_T >$. This process is represented by the following formulas:

$$K_S = ReLU(S \cdot W_{K_S} + b_{K_S}) \tag{2}$$

$$V_S = ReLU(S \cdot W_{V_S} + b_{V_S}) \tag{3}$$

where W_{K_S} and W_{V_S} are the weight matrices of the fully connected layers, b_{K_S} and b_{V_S} are bias vectors, and $ReLU$ is

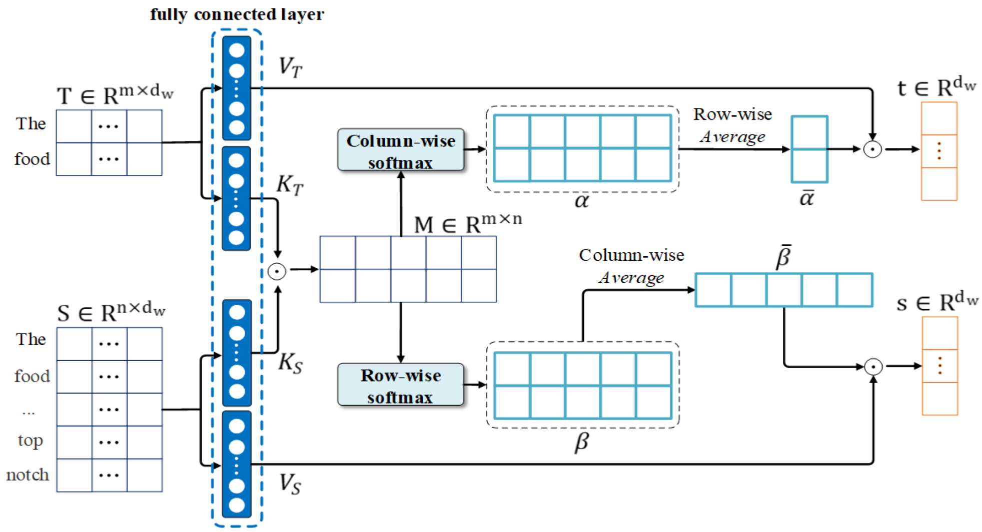


FIGURE 1. The architecture of ACIA.

an activation function. We perform the same mapping on the aspect matrix T to derive K_T and K_T.

After deriving $<K_S, V_S>$ and $<K_T, V_T>$, we perform a dot product operation on K_S and K_T to generate a pair-wise matching matrix $M = K_T \cdot K_S^T$, where the value of each entry in M represents the relevance of words between an aspect and a context. We use the softmax function to convert these values into probabilities between zero to one, where a greater probability indicates a stronger correlation. The softmax function is defined by in the following formulas:

$$\alpha_{ij} = \frac{exp(M_{ij})}{\sum_{i=1}^{m} exp(M_{ij})} \tag{4}$$

$$\beta_{ij} = \frac{exp(M_{ij})}{\sum_{j=1}^{n} exp(M_{ij})} \tag{5}$$

where α denotes context-to-aspect correlations based on column-wise softmax and β denotes aspect-to-context correlations based on row-wise softmax. α contains n columns of weight distributions for aspects, that are based on the correlations of n corresponding words in a context. The final weight of each word in an aspect is determined by averaging α row-wise. Similarly, the final weight of each word in a context is obtained by applying a similar operation to β. These operations can be expressed as follows:

$$\bar{\alpha}_j = \frac{1}{n}\sum_{j=1}^{n} \alpha_{ij} \tag{6}$$

$$\bar{\beta}_i = \frac{1}{m}\sum_{i=1}^{m} \beta_{ij} \tag{7}$$

where $\bar{\alpha} \in R^{m\times 1}$ and $\bar{\beta} \in R^{n}$ are the final weights of the words in an aspect and a context, respectively. Finally weighted sum of V_T and V_S to generate a weighted representation of the aspect $t = (\bar{\alpha})^T \cdot V_T$ and context $s = \bar{\beta} \cdot V_S$. The representation of the aspect is context-based, and the representation of the context is based on a specific aspect.

C. MULTIACIA REPRESENTATION

It is beneficial to generate multiple groups of $<K_S, V_S>$ and $<K_T, V_T>$ pairs for ACIA operations. This concept was inspired by "multi-head" attention [10]. We linearly project aspects and contexts through multiple sets of fully connected layers to generate multiple groups of $<K_S^h, V_S^h>$ and $<K_T^h, V_T^h>$. The ACIA operation is then performed in parallel and multiple ACIA output vectors are concatenated to from the final result. The following formulas describe the "multi-head" concept:

$$s_i, t_i = ACIA(<K_S^i, V_S^i>, <K_T^i, V_T^i>) \tag{8}$$

$$S', T' = [s_1 : s_2 : \ldots : s_h], [t_1 : t_2 : \ldots : t_h] \tag{9}$$

where h is the number of heads in the multi-head and ":" indicates the concatenation of vectors.

We use a multiple ACIA blocks stacked structure to generate sequence-to-sequence representations of aspects and contexts. Between two ACIA blocks, we adopt a residual connection and standardized fully connected layers. Fig.2 presents the connection structure of MultiACIA.

Residual connections do not add any training parameters, but they can solve the problem of "*degradation*" caused by an increasing number of network layers [40]. Such connections can be expressed as follows:

$$S = S + S' \otimes e_n \tag{10}$$

$$T = T + T' \otimes e_m \tag{11}$$

where e_n and e_m are column vectors with n and m 1s respectively, and $S' \otimes e_n$ represents the repeated concatenation of S' for n iterations to form a new matrix with the same dimension as that of S. Additionally, performing layer normalization can

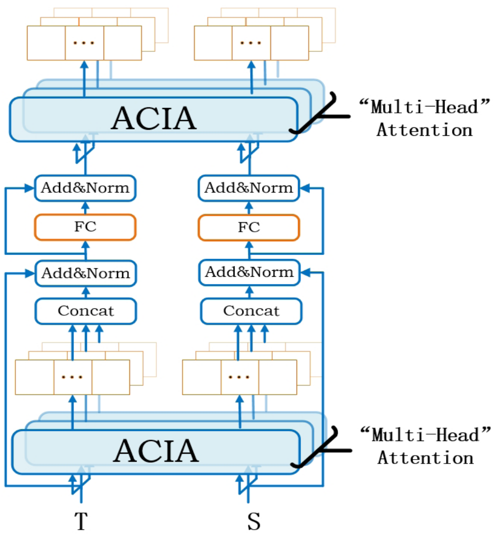


FIGURE 2. MultiACIA connection structure.

improve the training speed of a model and enhance generalization ability [41].

A fully connected feed-forward network is used to map aspect and context representations between two ACIA structures. This network consists of two layers of internal and external linear networks with an activation function between layers. The functionality of the feed-forward network can be expressed as follows:

$$S_{next} = ReLU(S \cdot W^S_{D_1} + b^S_{D_1}) \cdot W^S_{D_2} + b^S_{D_2} \quad (12)$$
$$T_{next} = ReLU(T \cdot W^T_{D_1} + b^T_{D_1}) \cdot W^T_{D_2} + b^T_{D_2} \quad (13)$$

where $W^S_{D_1}$, $W^S_{D_2}$, $W^T_{D_1}$ and $W^T_{D_2}$ are weight matrices; $b^S_{D_1}$, $b^S_{D_2}$, $b^T_{D_1}$ and $b^T_{D_2}$ are bias vectors. $ReLU$ is the activation function. The outputs S_{next} and T_{next} are the inputs for the next ACIA structure after the residual connection and normalization operations.

By using MultiACIA representation learning, the aspect features and sentiment features of a specific aspect can be extracted and integrated into the representation of each word in an aspect or context. As the proposed model is a sequence-to-sequence representation learning structure, it can generate corresponding sequence representations by replacing LSTM to model the aspects and contexts.

D. FINAL REPRESENTATION AND MODEL TRAINING

We tested three methods for generating the final aspect representation S_{final} and context representation T_{final} from MultiACIA outputs. First, we averaged the sequence representations of aspects and contexts to form final representations. Then we concatenated them to form a vector input for the classification layer. Second, we used IAN which proposed by Ma *et al.* [7], that assigns different weights to the words in aspects and contexts when generating a final representation. The calculation process for T_{final} in this method is defined as follows:

$$T_{final} = \sum_{i=1}^{m} \eta_i t_i^{output} \quad (14)$$

where t_i^{output} is the i-th word in the aspect sequence representation of a MultiACIA outputs and η is the target weight vector generated based on contexts. The formula for calculating η is written as follows:

$$\eta_i = softmax(tanh(t_i^{output} \cdot W_t \cdot s_{avg}^{output^T} + b_t)) \quad (15)$$

where $tanh$ is an activation function, s_{avg}^{output} is the average of the context sequence representation, and W_t and b_t are weight matrix and bias, respectively. The calculation method for S_{final} is similar. The two outputs T_{final} and S_{final} are then concatenated to form a vector for classification.

The final method involves the use of the AOA mechanism [8], that focuses on the interactions between word-pairs in contexts and aspects. It also generates a pair-wise matching matrix $I = S^{output} \cdot T^{output^T}$, which is used to generate attention vectors α and β using column-wise softmax and row-wise softmax operations, respectively. However, AOA does not generate a final aspect representation, it utilizes the importance of each word in an aspect when generating a final context representation. The calculation formula for S_{final} is written as follows:

$$\gamma = \alpha \cdot \bar{\beta}^T \quad (16)$$
$$S_{final} = \sum_{j=1}^{n} \gamma_j S_j^{output} \quad (17)$$

where S_j^{output} is the j-th word in the context sequence representation of a MultiACIA outputs and γ is the weight vector for the corresponding contexts.

The final representation is then fed into the classification layer and processed is as follows:

$$y = softmax(r \cdot W_l + b_l) \quad (18)$$

where W_l and b_l are weight matrix and bias, respectively, and r is the final representation, which can be generated using S_{final} and T_{final} connections, or S_{final} alone.

For classification, we use cross entropy with L_2 regularization as the loss function. The loss calculation formula is written as follows:

$$Loss = -\sum_{p \in (a,s)} \sum_{c \in C} \hat{y}_c^p log y_c^p \quad (19)$$

where p is an aspect-context pair and $C = \{1, 0, -1\}$ is a collection of sentiment polarities. The Adam algorithm is used to adjust the parameter sets for our model by minimizing the error between ground truth probability distribution $\hat{y}$ and the predicted probability distribution y [42].

TABLE 1. Distributions of sentiment polarity categories for the three datasets.

Dataset		Positive	Neutral	Negative
Laptop	Train	994	464	870
	Test	341	169	128
Restaurant	Train	2164	637	807
	Test	728	196	196
Twitter	Train	1567	3127	1563
	Test	174	346	174

IV. EXPERIMENTS

A. DATASETS AND HYPERPARAMETER SETTINGS

We conducted comprehensive experiments on three benchmark datasets. The first two are from SemEval-2014 Task4 [43], namely Restaurant and Laptop. The last one is a Twitter dataset collected by Dong *et al.* [44]. Each sentence in these three datasets has been annotated with all aspects (words or phrases) and corresponding sentiment polarities. The statistics for three datasets are shown in Table 1.

In our experiments, we utilized the pre-trained GloVe word vectors developed by Pennington *et al.* [45] to initialize the word embeddings for aspects and contexts. Each word vector contains 300 features. We also employ the same word vectors in existing approaches [7], [8], [12]. In addition, we performed random initialization for approximately 5% of out-of-vocabulary words from the uniform distribution $U \in (-0.01, 0.01)$. The interactive attention representations for aspects and contexts were generated using a stack of four ACIA layers and the tunable parameters for the model were initialized randomly. The learning rate for the Adam optimizer was set to 0.001. The batch size on the Restaurant and Laptop datasets was set to 128 and it was 32 on Twitter dataset.

B. MODEL COMPARISONS

To verify the effectiveness of our method, we compared it with thirteen state-of-the-arts methods, which are summarized as follows:

- **LSTM** models the context information using a single LSTM module without considering aspect information.
- **TD-LSTM** uses two LSTMs to model the left and right context of the aspect, and then feeds the concatenated context representations to the classification layer [46].
- **ATAE-LSTM** is a typical attention-based LSTM network model that generates context representations based on specific aspect via vector concatenation [6].
- **IAN** models both the aspects and contexts using LSTM. It interactively learns the weight for each word using an attention mechanism based on generated aspect and context representations [7].
- **RAM** is a recurrent attention network that uses a multi-attention mechanism to capture the sentiment features of long-distance separation [21].
- **IAD** incorporates Inter-Aspect Dependencies (IAD) by classifying various aspects of a sentence at the same time, and using recurrent networks to time-dependently process their corresponding sentence representations [47].
- **IARM** generates aspect-aware sentence representations for all the aspects by gated recurrent unit and attention mechanism. Then, the specific aspect representation is repeatedly matched with other aspects through the memory network to generate a more accurate representation [48].
- **FANS** first uses unigram, part of speech, and word position features to learn the feature-enhanced word representations. Then interactively models context, target and sentiment words through a multi-view co-attention network, so that the model can learn a better multi-view sentiment-aware and target-specific sentence representation [49].
- **LSTM + SynATT + TarRep** uses two methods to improve attention mechanism performance. One is a novel aspect representation method that maps semantically similar aspects to the same vector. The other is the introduction of external information for constructing a syntax-based attention mechanism using context syntax information [38].
- **AOA-LSTM** applies the AOA mechanism to handle the correlations between aspects and contexts. The words in aspect phrases are also assigned different weights [8].
- **TNet-LF/AS** consolidates context words into aspect representations using target-specific transformation networks, where aspect representations are tailored to context words, and use context maintaining mechanism to prevent context information from being lost during merging [12].
- **TNet w/o transformation** removes the target-specific transformation networks in TNet-LF/AS [12].
- **DMMN-SDCM** proposes a framework of deep mask memory network with semantic dependency and context moment. Firstly, semantic parsing information is used instead of location information to guide attention mechanisms, and to better use information from the nearby aspects when modeling aspects. Secondly, the sentiment distribution of the entire sentence is learned as a background to the sentiment analysis of specific aspect through a auxiliary task [50].
- **MultiACIA** averages the aspect and context sequence representations of MultiACIA outputs, and concatenates them as inputs for the classification layer.
- **IAN-MultiACIA** and **AOA-MultiACIA** replace LSTM in IAN and AOA-LSTM with MultiACIA. The only difference is that a single-LSTM was used in IAN and the bidirectional-LSTM was used in AOA-LSTM.

Table 2 lists the experimental accuracy of each model on two datasets. Based on the random initialization problem, experimental results often fluctuate [51]. For fair comparison, we report the best results for all the methods, following existing research papers [6]–[8], [12], [21], [38].

TABLE 2. Comparisons of experimental accuracy and macro-F1. TD-LSTM's results are retrieved from RAM paper, the results with symbol "†" are our reimplemented, and others baseline results are retrieved from the original papers. The underlined numbers represent the best results on the corresponding datasets. The values marked with "*" represent that MultiACIA is significantly better than LSTM, IAN-MultiACIA is significantly better than IAN, and AOA-MultiACIA is significantly better than AOA-LSTM (p-value < 0.05).

Models	Restaurant		Laptop		Twitter	
	ACC	Macro-F1	ACC	Macro-F1	ACC	Macro-F1
TD-LSTM	0.7800	0.6673	0.7183	0.6843	0.6662	0.6401
ATAE-LSTM	0.7720	-	0.6870	-	-	-
RAM	0.8023	0.7080	0.7449	0.7135	0.6936	0.6730
IAD	0.7900	-	0.7250	-	-	-
IARM	0.8000	-	0.7380	-	-	-
FANS	-	-	-	-	0.7120	0.6880
LSTM+SynATT+TarRep	0.8063	0.7132	0.7194	0.6923	-	-
TNet-LF/AS	0.8079	0.7127	0.7654	0.7175	<u>0.7497</u>	<u>0.7360</u>
TNet w/o transformation	0.7890	0.6586	0.7330	0.6825	0.7210	0.7057
TransCap	0.7955	0.7085	0.7387	0.7010	-	-
DMMN-SDCM	0.8116	0.7150	<u>0.7759</u>	<u>0.7361</u>	-	-
LSTM	0.7430†	0.6413†	0.6650†	0.6398†	0.6814†	0.6556†
Ours:MultiACIA	**0.7991***	**0.6987***	**0.7246***	**0.6800***	**0.7032***	**0.6857***
IAN	0.7860	0.6912†	0.7210	0.6714†	0.7023†	0.6758†
Ours:IAN-MultiACIA	**0.8044***	**0.7141***	**0.7324***	**0.6845***	**0.7124**	**0.6892***
AOA-LSTM	0.8120	0.7105†	0.7405	0.6812†	0.7150†	0.6788†
Ours:AOA-MultiACIA	**<u>0.8259</u>***	**<u>0.7213</u>***	**0.7527***	**0.7024***	**0.7240**	**0.6940***

C. RESULT ANALYSIS

Table 2 reveals that compared with LSTM, MultiACIA obtains average 4.52% accuracy and 4.26% macro-F1 significant improvements on three datasets. Note MultiACIA can be used to derive sequence representations in place of LSTM in some of the baseline models. The accuracy increased by average 1.33% and the macro-F1 increased by average 1.64% after using MultiACIA instead of single-LSTM in IAN model. Similarly, AOA-MultiACIA obtains average 1.17% accuracy and 1.57% macro-F1 improvements after using MultiACIA instead of bidirectional-LSTM in AOA-LSTM model. We analyze the experimental results from two perspectives, namely comparisons with LSTM/Attention-based LSTM and the ability to handle aspects co-occurrence.

1) COMPARISONS WITH LSTM/ATTENTION-BASED LSTM

Because MultiACIA considers aspect information during the process of sequence representation modeling and continuously extracts the features related to a specific aspect in its context using a multi-layer attention stacking structure, its accuracy and macro-F1 is much higher than those of LSTM. Additionally, the performance of MultiACIA is also better than that of IAN. This indicates that a single-layer interactive attention mechanism has limited ability to extract features. Moreover, in addition to the multi-layer structure, the "multi-head" mechanism of each attention layer enables the model to learn useful information regarding different dimensions from different representation subspaces.

Whether it is single-LSTM or bidirectional-LSTM, MultiACIA outperforms them on all three datasets when combined with other attention mechanisms. Both IAN and AOA-LSTM are traditional structures in which context and aspect are modeled by LSTMs separately. Unlike these structures, MultiACIA is a completely attention-based integrated structure that uses the correlations between words in aspects and contexts for sequence modeling. Therefore, it utilizes the interactive relationships between contexts and aspects that are ignored in separate LSTM representations. Furthermore, MultiACIA can model dependencies from a global perspective without being limited by the distance between two words in a sequence. Compared with the other baseline models, AOA-MultiACIA exhibits a significant improvement in performance, even though models such as LSTM + SynATT + TarRep has introduced external information.

2) THE ABILITY TO HANDLE ASPECTS CO-OCCURRENCE

MultiACIA achieves the highest accuracy and macro-F1 on Restaurant dataset, as well as very competitive results on Laptop and Twitter datasets. This is due to differences among these datasets. First, we counted the probabilities of co-occurrence aspects in the same review on these datasets. On Restaurant dataset, The probability of co-occurrence aspects is 50.11%, this figure is 37.48% on Laptop dataset and only 0.001% on the Twitter dataset. Second, the Laptop dataset contains more implicit sentiments [52], such as "*lots of preloaded software*". Similarly, Twitter as a social community, the dataset built on it contains more sophisticated speech act [53]. These characteristics make it more difficult to determine the sentiment polarities of aspects.

Based on the above statistics, The problem of co-occurrence aspects is widespread on Restaurant dataset. In contrast, a review often contains only one aspect on Twitter dataset. This important difference on those datasets validates that our method can more accurately find the corresponding evaluation words for different aspects in the same review, because AOA-MultiACIA delivers state-of-the-art performance on Restaurant dataset compared with all baseline models. More specifically, AOA-MultiACIA always extracts the word relationships between the aspect and its context, from the input of the word embeddings to the final

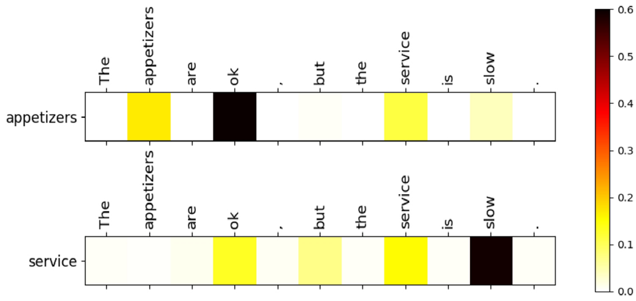


FIGURE 3. Context attention weight distributions.

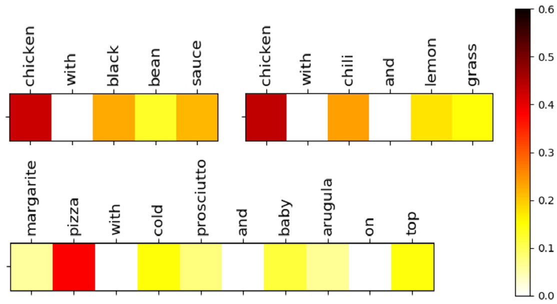


FIGURE 4. Aspect attention weight distributions.

representations, and integrates them into the generated matrix representations in the form of weights. In addition, the attention mechanism allows modeling of dependencies regardless of their distance in the input or output sequences [10]. This ability makes the model more directly focus on whether the current word is related to the specified aspect in its context sequence, ignoring the distance between word and aspect term in the sentence, so that AOA-MultiACIA can more accurately find the evaluation words corresponding to the specified aspect, which is critical for determining the sentiment polarity. Therefore, our method provides the best results for the Restaurant dataset.

DMMN-SDCM integrates the relation information among aspects in the same sentence into deep memory network. Further, it introduces external knowledge and applies semantic dependency information to the attention mechanism. All the above information is ignored by our method, therefore, DMMN-SDCM performs best on the Laptop dataset. With regards to TNet-LF/AS, the performance degradation is noticeable when it is reduced to TNet w/o transformation, this phenomenon reveals that integrating tailor-made aspect information into the word-level representations is crucial for TNet-LF/AS's excellent performance [12]. Both models perform better on Laptop and Twitter datasets where sentiment polarity is more difficult to determine. However, it does not perform as well as our model on Restaurant dataset, which is easier but contains many co-occurrence aspects. This further proves that MultiACIA can handle mixed information of multiple aspects in a review better.

D. ATTENTION VISUALIZATION

To demonstrate the effectiveness of the ACIA structure, we analyze the performance of attention mechanism based on a couple of case studies. Specifically, we visualize the distributions of attention weights in ACIA, which allows users to clearly see which words in an aspect phrase are assigned higher/lower weights, as well as which words in a context determine the sentiment polarities of associated aspects. Finally, we present the differences in attention weights learned in different layers of MultiACIA.

Fig.3 presents the attention distributions of context words when different aspects are modeled by ACIA (i.e., $\bar{\beta}$ in (7)). The sentence "*The appetizers are ok, but the service is slow.*" contains two aspects, namely "*appetizers*" and "*service*." When the aspect is "*appetizers*," (the top part of Fig.3) the weight of "*ok*" in the context attention distribution is the largest, indicating our model can capture opinion evaluation words accurately. Similarly, when the aspect is "*service*," (the bottom part of Fig. 3) the weight of "*slow*" is the largest. These results align very well with human judgment. In other words, for different aspects in the same review, ACIA can extract the corresponding evaluation words from a context accurately, leading to better sentimental classification. Additionally, while our attention weights are mainly concentrated on aspects and evaluation words, it also gives relatively high weight to "but" for aspect service, showing that we could leverage it for better identify sentimental polarity as it represents reverse sentiment. On the contrary, prepositions, such as "*the*," "*and*" and "*is*" are assigned extremely small weights by ACIA. As such, ACIA is able to learn syntactic and semantic information to generate high-quality context representations for specific aspects.

Fig.4 presents the attention distributions of longer aspect phrases (i.e., $\bar{\alpha}$ in (6)). First, it presents the two aspect phrases that were discussed in Section 1 (the top part of the figure). We observe that "*chicken*" is assigned the greatest weight in both the phrases, signifying ACIA can extract the commonalities of different aspects when the aspects are in same/similar category. As mentioned previously, this is helpful in determining the sentiment polarity of an aspect. Second, the figure presents the attention distribution of a longer aspect phrase "*margarite pizza with cold prosciutto and baby arugula on top*" (bottom part of the figure). The phrase "*with cold prosciutto and baby arugula on top*" is an attribute of "*pizza*" and serves as a modifier. The importance of these words is well reflected in the distribution of attention weights. Overall, ACIA has strong feature extraction ability when modeling the aspect phrases.

Fig.5 presents the attention distributions of different layers (i.e., $\bar{\beta}$ in (7) of different ACIA layers). Each layer consists of an ACIA attention block. We use a complete review to present what different ACIA blocks have learned. The sentence "*Just got done putting windows 7 ultimate on the laptop ... i love it !!!!*" contains only one aspect "*windows 7*" and its sentiment. It can be observed in attention distribution of the first block that almost every word is noticed, which may be the aspect to

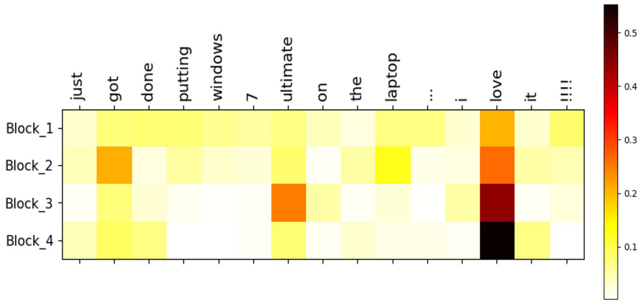


FIGURE 5. Different layer attention weight distributions.

make a rough connection to each word in its context. In the following blocks, the word containing sentiment information is gradually noticed, e.g., "*love*". This indicates that the structure stacked by multiple AICA blocks can enhance the modeling ability, so that the model can more accurately notice the parts of the sentence that are important to the aspect sentiment.

V. CONCLUSION

This paper addressed the biggest challenges on aspect-level sentiment classification that a single review contains multiple aspects and sentiments mixed, and proposed an integrated aspect-context interactive sequence representation model based entirely on attention mechanism. Our proposed MultiACIA can replace the general structure that separately models aspects and contexts with LSTM. The relationships between words in aspects and contexts are considered during the process of sequence modeling. This is helpful when generating representations of both context-based aspects, and aspect-based contexts. Whether compared with LSTM alone or used in combination with other attention mechanisms, our method is more suitable for the aspect-level sentiment classification than LSTM. We verified that the MultiACIA is able to achieve superior performance than other comparison models on Restaurant datasets, as well as very competitive results on Laptop and Twitter datasets.

Finally, although the proposed model achieves very good performances, it ignores the order information of aspect and context sequences. One possibility for future research is to use the semantic dependency information to improve the ability to capture the relationship among words in the same sentence.

ACKNOWLEDGMENT

The authors would like to thank all anonymous reviewers for their valuable comments and suggestions which have significantly improved the quality and presentation of this article.

REFERENCES

[1] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 513–520.

[2] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intell. Syst.*, vol. 31, no. 2, pp. 102–107, Mar. 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7435182/

[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] B. Liu, "Sentiment analysis and opinion mining," *Synth. Lectures Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, May 2012. [Online]. Available: http://www.morganclaypool.com/doi/abs/10.2200

[5] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent twitter sentiment classification," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.* Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 151–160.

[6] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based LSTM for aspect-level sentiment classification," in *Proc. 2016 Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 606–615. [Online]. Available: http://aclweb.org/anthology/D16-1058

[7] D. Ma, S. Li, X. Zhang, and H. Wang, "Interactive attention networks for aspect-level sentiment classification," in *Proc. 26th Int. Joint Conf. Artif. Intell.* Menlo Park, CA, USA: AAAI Press, Aug. 2017, pp. 4068–4074.

[8] 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*. Washington, DC, USA: Springer, 2018, pp. 197–206.

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

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[11] W. Xue and T. Li, "Aspect based sentiment analysis with gated convolutional networks," in *Proc. 56th Annu. Meeting Assoc. Comput.*, 2018, pp. 2514–2523. [Online]. Available: http://aclweb.org/anthology/P18-1234

[12] X. Li, L. Bing, W. Lam, and B. Shi, "Transformation networks for target-oriented sentiment classification," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 946–956. [Online]. Available: http://aclweb.org/anthology/P18-1087

[13] X. Ding and B. Liu, "The utility of linguistic rules in opinion mining," in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*. New York, NY, USA: ACM, 2007, pp. 811–812.

[14] T. Nasukawa and J. Yi, *Sentiment Analysis, Capturing Favorability Using Natural Language Processing*. New York, NY, USA: ACM, 2003, p. 70. [Online]. Available: http://portal.acm.org/citation.cfm?doid=945645.945658

[15] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*. New York, NY, USA: ACM, 2004, pp. 168–177.

[16] E. Boiy and M.-F. Moens, "A machine learning approach to sentiment analysis in multilingual Web texts," *Inf. Retr.*, vol. 12, no. 5, pp. 526–558, Oct. 2009. [Online]. Available: http://link.springer.com/10.1007/s10791-008-9070-z

[17] S. Kiritchenko, X. Zhu, C. Cherry, and S. Mohammad, "NRC-Canada-2014: Detecting aspects and sentiment in customer reviews," *Assoc. Comput. Linguistics*, 2014, pp. 437–442. [Online]. Available: http://aclweb.org/anthology/S14-2076

[18] J. Wagner, P. Arora, S. Cortes, U. Barman, D. Bogdanova, J. Foster, and L. Tounsi, "DCU: Aspect-based polarity classification for SemEval task 4," in *Proc. 8th Int. Workshop Semantic Eval. (SemEval)*, 2014, pp. 223–229. [Online]. Available: http://aclweb.org/anthology/S14-2036

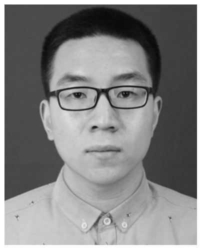
[19] A. Lazaridou, I. Titov, and C. Sporleder, "A Bayesian model for joint unsupervised induction of sentiment, aspect and discourse representations," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics*, 2013, pp. 1630–1639.

[20] T. H. Nguyen and K. Shirai, "PhraseRNN: Phrase recursive neural network for aspect-based sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 2509–2514. [Online]. Available: http://aclweb.org/anthology/D15-1298

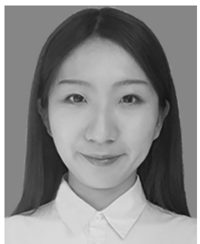
[21] P. Chen, Z. Sun, L. Bing, and W. Yang, "Recurrent attention network on memory for aspect sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 452–461. [Online]. Available: http://aclweb.org/anthology/D17-1047

[22] B. Liang, Q. Liu, J. Xu, Q. Zhou, and P. Zhang, "Aspect-based sentiment analysis based on multi-attention CNN," *J. Comput. Res. Development. Chin.*, vol. 54, no. 8, pp. 1724–1735, 2017.

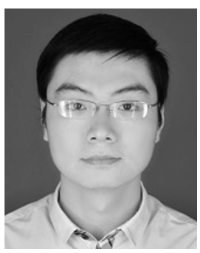
[23] M. Zhang, Y. Zhang, and D.-T. Vo, "Gated neural networks for targeted sentiment analysis," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 3087–3093.
[24] Z. Lei, Y. Yang, M. Yang, and Y. Liu, "A Multi-sentiment-resource Enhanced Attention Network for Sentiment Classification," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 758–763. [Online]. Available: http://aclweb.org/anthology/P18-2120
[25] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-20, no. 11, pp. 1254–1259, Nov. 1998.
[26] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," Sep. 2014, *arXiv:1409.0473*. [Online]. Available: https://arxiv.org/abs/1409.0473
[27] D. Tang, B. Qin, and T. Liu, "Aspect level sentiment classification with deep memory network," in *Proc. 2016 Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 214–224. [Online]. Available: http://aclweb.org/anthology/D16-1021
[28] Y. Cui, Z. Chen, S. Wei, S. Wang, T. Liu, and G. Hu, "Attention-over-attention neural networks for reading comprehension," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, Vancouver, BC, Canada, 2017, pp. 593–602.
[29] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier, "Exploiting document knowledge for aspect-level sentiment classification," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 579–585. [Online]. Available: http://aclweb.org/anthology/P18-2092
[30] Z. Chen and T. Qian, "Transfer capsule network for aspect level sentiment classification," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 547–556. [Online]. Available: https://www.aclweb.org/anthology/P19-1052
[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186.
[32] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. 2018 Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2018, pp. 2227–2237.
[33] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. (2018). *Improving Language Understanding by Generative Pre-Training*. [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf
[34] H. Xu, B. Liu, L. Shu, and S. Y. Philip, "Bert post-training for review reading comprehension and aspect-based sentiment analysis," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2019, pp. 2324–2335.
[35] A. Rietzler, S. Stabinger, P. Opitz, and S. Engl, "Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification," Nov. 2019, *arXiv:1908.11860*. [Online]. Available: https://arxiv.org/abs/1908.11860
[36] B. Huang and K. Carley, "Syntax-aware aspect level sentiment classification with graph attention networks," in *Proc. 2019 Conf. Empirical Methods Natural Lang.Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 5472–5480.
[37] C. Sun, L. Huang, and X. Qiu, "Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2019, pp. 380–385.
[38] H. T. Nguyen and M. Le Nguyen, "Effective attention networks for aspect-level sentiment classification," in *Proc. 10th Int. Conf. Knowl. Syst. Eng. (KSE)*, Nov. 2018, pp. 1121–1131.
[39] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, Feb. 2003.
[40] 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.
[41] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," Jul. 2016, *arXiv:1607.06450*. [Online]. Available: https://arxiv.org/abs/1607.06450
[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980
[43] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "SemEval-2014 task 4: Aspect based sentiment analysis," in *Proc. 8th Int. Workshop Semantic Eval. (SemEval)*, 2014, pp. 27–35. [Online]. Available: http://aclweb.org/anthology/S14-2004
[44] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, and K. Xu, "Adaptive recursive neural network for target-dependent twitter sentiment classification," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 49–54. [Online]. Available: http://aclweb.org/anthology/P14-2009
[45] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. 2014 Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[46] D. Tang, B. Qin, X. Feng, and T. Liu, "Effective LSTMs for target-dependent sentiment classification," in *Proc. COLING 26th Int. Conf. Comput. Linguistics, Tech. Papers*, 2016, pp. 3298–3307.
[47] D. Hazarika, S. Poria, P. Vij, G. Krishnamurthy, E. Cambria, and R. Zimmermann, "Modeling inter-aspect dependencies for aspect-based sentiment analysis," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 2, 2018, pp. 266–270.
[48] N. Majumder, S. Poria, A. Gelbukh, M. S. Akhtar, E. Cambria, and A. Ekbal, "IARM: Inter-aspect relation modeling with memory networks in aspect-based sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 3402–3411. [Online]. Available: http://aclweb.org/anthology/D18-1377
[49] M. Yang, Q. Qu, X. Chen, C. Guo, Y. Shen, and K. Lei, "Feature-enhanced attention network for target-dependent sentiment classification," *Neurocomputing*, vol. 307, pp. 91–97, Sep. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0925231218304764
[50] P. Lin, M. Yang, and J. Lai, "Deep mask memory network with semantic dependency and context moment for aspect level sentiment classification," in *Proc. 28th Int. Joint Conf. Artif. Intell.* Menlo Park, CA, USA: AAAI Press, Aug. 2019, pp. 5088–5094.
[51] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1139–1147.
[52] J. Liao, S. Wang, and D. Li, "Identification of fact-implied implicit sentiment based on multi-level semantic fused representation," *Knowl.-Based Syst.*, vol. 165, pp. 197–207, Feb. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0950705118305756
[53] Y. Tay, A. T. Luu, S. C. Hui, and J. Su, "Reasoning with sarcasm by reading in-between," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 1010–1020. [Online]. Available: http://aclweb.org/anthology/P18-1093



ZHUOJIA WU received the B.S. degree from the School of Computer and Information Technology, Shanxi University, China, in 2016, where he is currently pursuing the M.S. degree in software engineering. His research interests include natural language processing (NLP) and aspect-level sentiment analysis.



YANG LI is currently pursuing the Ph.D. degree with the School of Computer and Information Technology, Shanxi University, China. Her research interest focuses on sentiment analysis.



JIAN LIAO received the Ph.D. degree from the School of Computer and Information Technology, Shanxi University, China. His research interests include implicit sentiment analysis and representation learning.

DEYU LI received the Ph.D. degree from Xi'an Jiaotong University. He is currently a Professor with the Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University. He has published more than 60 articles in international journals. His research interests include artificial intelligence, granular computing, data mining, and machine learning.

XIAOLI LI is currently the Department Head and the Principal Scientist with the Institute for Infocomm Research, A*STAR, Singapore. He also holds an Adjunct Professor position at Nanyang Technological University. His research interests include data mining, machine learning, AI, and bioinformatics. He has been serving as a (Senior) PC Member/Workshop Chair/Session Chair in leading data mining and AI related conferences, including KDD, ICDM, SDM, PKDD/ECML, WWW, IJCAI, AAAI, ACL, and CIKM. He has published more than 180 high quality articles and won numerous best paper/benchmark competition awards.

SUGE WANG received the Ph.D. degree from Shanghai University. She is currently a Professor with the Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University. She has published more than 30 articles in international journals. Her research interests include natural language processing, text sentiment analysis, and machine learning.