Applied Intelligence
https://doi.org/10.1007/s10489-022-04198-5

Dual-channel and multi-granularity gated graph attention network for aspect-based sentiment analysis

Yong Wang[1] · **Ningchuang Yang**[1] · **Duoqian Miao**[2] · **Qiuyi Chen**[1]

Accepted: 21 September 2022

Abstract

The Aspect-Based Sentiment Analysis(ABSA) aims to determine the sentiment polarity of a specific aspect. Existing approaches use graph attention networks(GAT) to model syntactic information with dependency trees. However, these methods do not consider the noise of the dependency tree and ignore the sentence-level feature. To this end, we propose the Dual-Channel and Multi-Granularity Gated Graph Attention Network(DMGGAT) to jointly consider semantics and syntactic information of multiple granularity features generated by GAT and BERT, in which BERT alleviates the instability of the dependency tree and enhance the semantic information lost in the graph calculation. First, We propose a two-channel structure composed of BERT and GAT, enabling syntactic and semantic information generated by BERT to assist GAT. Furthermore, an aspect-based attention mechanism is used to generate sentence-level features. Finally, a newly designed gated module is introduced to integrate the aspect(fine-Granularity) and sentence-level (coarse-Granularity) features from the two channels to classify jointly. The experimental results show that our model achieves advanced performance compared to the current model on three extensive datasets.

Keywords Graph attention network · Aspect-based sentiment analysis · Multi-granularity · BERT

1 Introduction

The ABSA [1, 2] is a fine-grained sentiment classification dividing the target aspect polarity in the context into three categories: positive, negative, and neutral. For example, in the following sentence, "The performance of this laptop is excellent, but the screen is terrible", the polarity of "*laptop*" is positive, but the polarity of its "*screen*" is negative.

Unlike traditional machine learning methods [3, 4], deep neural network does not need to design features manually, but automatically learns the semantic information of context and presents it in low latitude. Convolutional neural networks(CNN) [5, 7] has been widely used in ABSA because it can learn the potential semantic representation of the context. However, CNN can only carry out convolution operation for multiple consecutive words and cannot learn the emotion of non-consecutive words. Compared with CNNs, recurrent neural networks(RNN) [10, 11] takes the context dependence into account when modeling the text, so its strengths are more prominent. Although RNN achieved good results in ABSA task, it could not learn accurate information of distant words. Research [3] show that 40% of sentiment classification errors are caused without considering the specific aspect in sentiment classification. For specific aspects, attention mechanisms [6, 13, 14] can effectively capture key information in the context, so many models combine attention mechanisms to generate accurate aspect expression to better improve the accuracy of the model. However, due to the complexity of the syntax and loss of syntax information, the attention mechanism may not be able to provide accurate attention weight. Take the following sentence as an example, when it says "it has a bad

✉ Ningchuang Yang
yangningchuang@163.com

Yong Wang
ywang@cqut.edu.cn

Duoqian Miao
dqmiao@tongji.edu.cn

Qiuyi Chen
chens7cq@163.com

[1] Schoole of Artificial Intelligence, Chongqing University of Technology, Chongqing, 401135, China

[2] Department of Computer Science and Technology, Tongji University, Shanghai, 201804, State, China

Published online: 06 October 2022

memory but a great battery life", the focus of the "*battery life*" is on "*bad*" instead of "*great*", depending on the syntactic information of the task. The gate mechanism [17–19] can filter out important features.

Recently, since the dependency graphs were generated from dependency trees, they have been able to provide more comprehensive and accurate syntactic information. The dependency of the sentence "The meal is delicious although the service is poor" is shown in Fig. 1. Many models use graph neural networks(GNNs) to capture context syntactic information, treating the dependency graph as an adjacency matrix and using graph convolutional networks(GCN) [20–22] and GAT [23, 24] to capture syntactic information. GAT is a variant of GNNs with the ability of extracting spatial features with generalized topology structure. Unlike GCN, it measures the strength of relationships between adjacent words based on attention. However, the dependency tree generated by the parser has some noises and instabilities. As DGEDT [25] and DualGCN [26] share the same idea to solve the dependency tree problem, DGEDT uses transformers to support graph learning, and DualGCN uses SemGCN to mutually enhance syntactic and semantic learning. Multimodal sentiment analysis [27–29] usually also adopts the similar method of learning the information of f two parts and then implementing interactions and iterations. GraphMerge [32] uses learned parameters to fuse multiple trees generated by different parsers to alleviate the noise of the dependency trees. These methods have achieved remarkable success in ABSA. In addition to dependency trees, pre-trained models achieve excellent performance on ABSA tasks, such as BERT [30] based on transformers [31].

However, there are some problems with these methods. First of all, the instability caused by the noise of the dependency tree needs to be further improved., and the semantic information is lost in graph computation. Secondly, most methods only rely on aspect representation to predict emotion and ignore the importance of sentence-level information. To solve the above problems, this paper proposes the Dual-Channel and Multi-Granularity Gated Graph Attention Network, consisting of a dual-channel structure of GAT and BERT. The research [33] shows that the induced tree from fine-tuned PTMs outperforms the parser-provided tree. Therefore, we propose that BERT is used as not only the embedding of the word vector of the model but a separate channel, providing the lost semantic information in the GAT calculation and reducing the noises caused by the dependency tree. Inspired by MGAN [13], single granular may cause the loss of important features. Therefore, we use the aspect-based attention to generate sentence-level features rich in current aspect-based background information. To select the best features for each granularity, a new multi-granularity gated module is used to fuse the information of both channels. The major original contributions of this article in the field include:

1. We propose a two-channel structure composed of BERT and GAT, enabling BERT's syntactic and semantic features to solve the semantic information lost in the GAT computation and the uncertainty of the dependency tree.
2. We propose an attention mechanism based on aspect, using aspect to calculate attention with sentences and then pool it evenly. In the calculation process, the aspect itself is shielded to avoid including aspect information in sentence-level features.
3. We propose a multi-granularity gated module to fuse the futures of both channels dynamically. The fusion of aspect-level and sentence-level features in two channels enable the model to learn two granular features.
4. We conducted extensive experiments on three data to verify the effectiveness of the model in the ABSA task. We also conducted a series of ablation and comparative experiments to verify the effectiveness of each component and the selection of parameters.

In the following sections of the paper, the Section 2 introduces the relevant work in the current field, Section 3 presents the methodology of DMGGAT, Section 4 compares and analyzes the results of the methods along with the ablation experiments conducted to verify the effectiveness of each component, and finally the Section 5 summarizes the paper and future work.

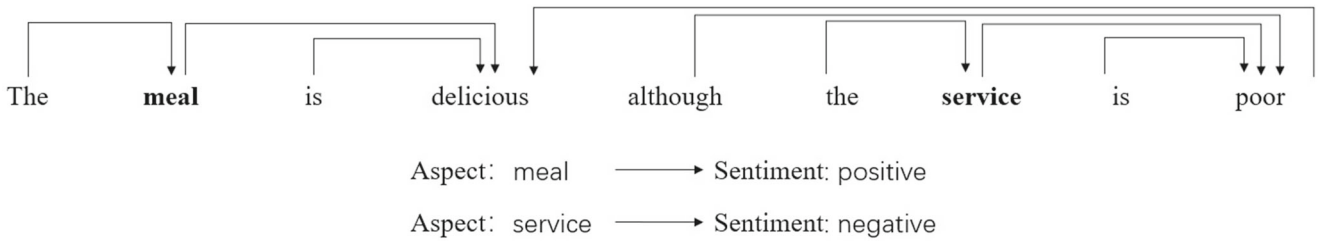

Fig. 1 An restaurant example sentence with its dependency tree. The line represents a dependency between two words. The sentence contains two aspects: meal and service and their corresponding positive and negative emotions

2 Related works

This section mainly introduces three methods to solve aspect-based sentiment analysis: traditional neural network methods, Pre-Training Models, and GNNs.

2.1 ABSA without syntactic dependencies

Early aspect-level sentiment analysis is mainly rule-based methods [34] and statistic-based [3] methods, relying on feature engineering. The deep neural network does not need to manually design features and can automatically learn to use low-dimensional vectors to represent features. The TD-LSTM [11] uses two one-way LSTMs [35] to model the left and right contexts of aspect to capture the opinion words located on both sides of the aspect.

Furthermore, the attention mechanism is used to emphasize contextual words associated with aspect words to obtain more accurate word representations. Wang et al. [12] first used the attention mechanism on LSTM to capture important emotional information related to the aspect and appends the target embedding with each word embedding. Ma et al. [6] uses LSTM to model the aspect and context separately, then the aspect and context representations are generated through interactive attention. Li et al. [7] uses attention mechanisms to incorporate contextual information into each word to model word representation. Zeng et al. [8] incorporates the position-aware vectors to improve adjacent context words. Ma et al. [9] first integrated the position modeling context and then used the position attention Mechanism to solve the multi-faceted word within one opinion. Fan et al. [13] uses fine-grained and coarse-grained attention mechanisms to interactively model aspect and sentence. Song et al. [14] introduce self-supervised attention learning methods to automatically mine useful attention information to model context and aspect. These methods do not use syntactic information to improve their performance further. In addition to attention, Gated mechanism have also been applied to this task. Xue and Li [19] propose gated Tanh-ReLU units to control the path of emotional information flow to the pooling layer based on the given aspect words. Xing et al. [18] proposes a new aspect-aware LSTM, which incorporates aspect information into the LSTM unit through a gated mechanism.

Recently, The pre-training models such as BERT [30] can effectively improve the performance of the model in aspect-level sentiment classification. Song et al. [14] use BERT instead of Glove [41] to encode the context and aspect separately, and model performance is significantly improved. Zhao et al. BERT use BERT to introduce target mentions is effective for the model. Li et al. [42] use BERT for end-to-end target sentiment prediction. BERT adopts a fine-tuning mechanism for different tasks. BERT is generally used in the embedding layer and fine-tuning in the task. First, the sentence is serialized, then Transformer [31] creates three vectors (query,key,value) for each sequence position. And finally applies the attention mechanism for each position. This computation can be presented as:

$$\text{Attention}\ (Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Compared with the performance of a single attention head, the multi-head attention found is more effective. BERT is built on multiple layers of transformers [31], and multi-headed attention is used for each layer. Particularly, we use the $BERT_{BASE}$ and $BERT_{LARGE}$ model with Layer=12, Dim=768, Head=12 and Layer=24, Dim=1024, Head=16 as hyper-parameters respectively.

2.2 ABSA with syntactic dependencies

In the early days, some researchers used the artificial definition of syntactic rules to add syntactic information to improve the ability of model to capture syntax [15, 16]. With the emergence of the dependency tree, Dong et al. [37] proposes that adaptive recursive neural network (AdaRNN) uses syntactic information adaption to propagate emotional information to aspect words. Nguyen et al. [38] uses a simpler method, using a combination of dependency tree and composition tree for further coding. He et al. [39] proposed to use the tree node distance to calculate the attention weight.

Recently, GNNs built on dependency trees have achieved good results in ABAS tasks. Zhao et al. [20] are the first to propose the use of GCN for dependency graphs to model the context by mentioning syntactic dependency information. Zhao et al. [36] further add the word position information and interactively calculates the attention to obtain a new representation, and finally sends it to the graph convolutional network. Huang and Carleyconstructs et al. [23] GAT based on the dependency graph and combine GAT with LSTM to capture information related to aspect. Wang et al. [40] use GAT to encoder dependencies relations and to establish a connection between specific aspect and context. However, these methods do not take into account the instability caused by the dependency tree. Tang et al. [25] proposes to use dual transformers, which contain GCNs and original transformers to interact syntactically and semantically respectively to reduce the instability caused by dependency trees. Li et al. [26] utilizes dual GCN interactions to resolve instability of dependency trees, which contains both syntactic GCN and semantic GCN.

3 Proposed methodology

In this section, we will go into the details of the model in detail. The overall structure of DMGGAT is shown in

Fig. 2. Our methodology is divided into the following parts 1) Vector representations of words are generated by BERT 2) GAT Layer is used to extract syntactic information of dependency graphs 3) Coarse-grained features are produced using aspect-based attention in dual-channels respectively 4) Feature fusion.

3.1 BERT encoder layer

An n-word sentence $w^c = \{w_1^c, w_2^c, \ldots, w_n^c\}$ with the aspect $w^a = \{w_1^a, w_2^a, \ldots, w_m^a\}$ is given, where w^a is sub-sequence of w^c. Each word will be mapped to a low-dimensional vector space via BERT from a lexicon size dimension. In order to comply with the rules of BERT training input, the sentence sequence is conveyed as "[CLS] + Context + [SEP] + aspect + [SEP]". [CLS] and [SEP] are special tokens of BERT. BERT learns each word representation by the transformers. We get the output of the last layer of transformer:

$$h = \{h_{CLS}, h_1, \ldots, h_n, h_{SEP}, h_{n+2}, \ldots, h_{n+m+1}, h_{SEP}\} \quad (2)$$

Where h_{CLS} is BERTPooling, it contains the classification information of BERT. $h_1, .., h_n$ are the embedding vector of the sentence.

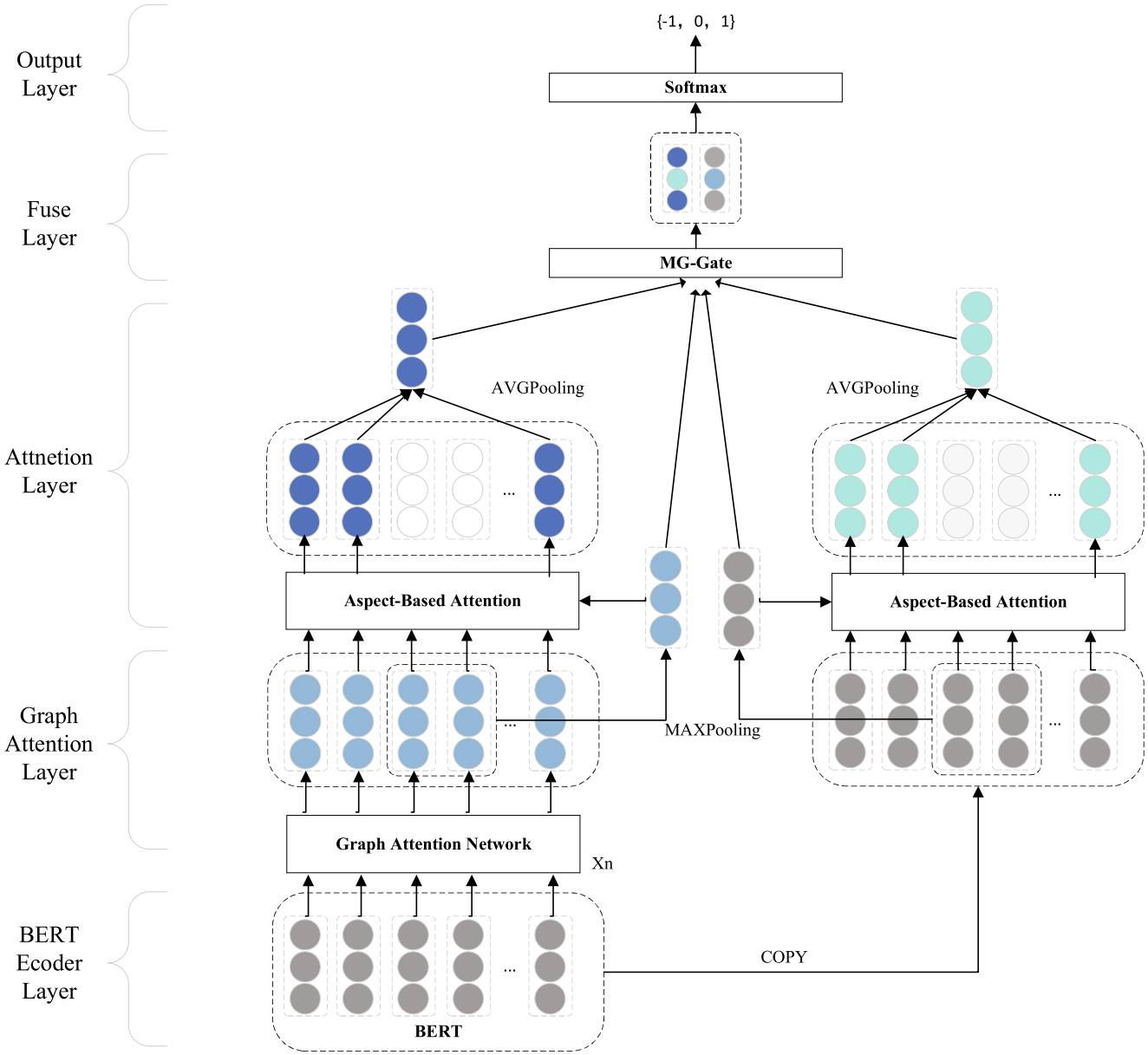

Fig. 2 Framework of dual-channel multi-granularity gated graph attention network

3.2 Graph attention network layer

The graph attention network layer is composed of GAT and point-wise convolution transformation(PCT) as shown in Fig. 3. GAT is a network that counts multi-head attention based on a dependency graph, and PCT is used to transform and collect features extracted by the GAT.

3.2.1 Graph attention network

GAT is a variant of graph neural network, calculating attention weight based on the distance of the syntactic dependence tree. Dependencies can be represented by a syntax graph of N nodes. Each word represents a node in the graph, and the edges in the graph represent the dependencies between words. For example, $G(V, A)$ represents a dependency graph, where V represents all nodes, and A is the adjacency matrix. If the two nodes have a dependency relationship $A_{i,j} = 1$, otherwise $A_{i,j} = 0$. The procedure of generating the adjacency matrix A for each sentence is depicted in Algorithm 1. We use the context representation $h_1, h_2, \ldots, h_n$ obtained from BERT as the input to the first layer of GAT. GAT updates each node by aggregating adjacent nodes information using multi-head attention:

$$h_i^{l+1} = \|_{k=1}^{K} \sum_{j \in N(i)} \alpha_{ij}^{lk} W_V^{lk} h_j^l \tag{3}$$

$$\alpha_{ij}^{lk} = \frac{\exp\left(f\left(h_i^{lk}, h_j^{lk}\right)\right)}{\sum_{j \in N(i)} \exp\left(f\left(h_i^{lk}, h_j^{lk}\right)\right)} \tag{4}$$

Where h_i^{l+1} represents the i-th node of the $l+1$ layer. $\|_{k=1}^{K}$ represents the k attention head. α_{ij}^{lk} represents the attention weight of the k-th attention head of the two nodes

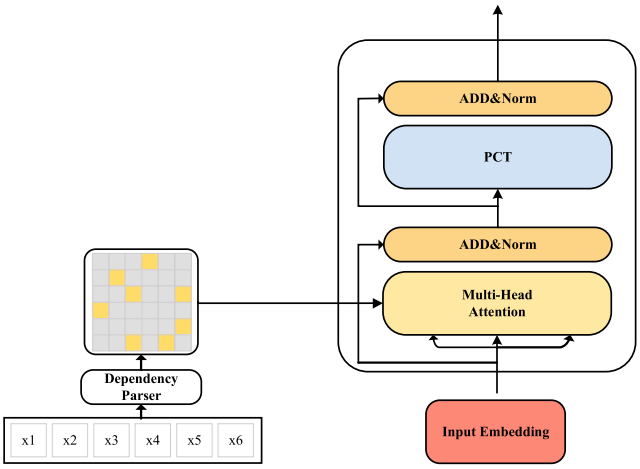

Fig. 3 Detailed structural diagram of the work of the graph attention layers

Require: a sequence of words $w^c = w_1^c, w_2^c, \ldots, w_n^c$;a aspect words $w^a = w_1^a, w_2^a, \ldots, w_m^a$;the dependency tree of the sentence dependency(s);

```
for i = 1 ← n do
    for j = 1 ← n do
        if dependency(w_i, wj)∈ dependency(s) then
            Generated by dependency tree
            A_{i,j} ← 1
        else if i=j then
            A_{i,j} ← 1
        else if w_i ∈ w^a and w_j ∈ w^a then
            A_{i,j} ← 1
        else
            A_{i,j} ← 0
        end if
    end for
end for
```

Algorithm 1 The procedure of deriving the adjacency matrix for each sentence.

i, j at l layer, $\mathrm{W}_{\mathrm{V}}^{\mathrm{lk}} \in \mathrm{R}^{\frac{d}{K} xd}$ an input transformation matrix of the k-th attention head at l layer.

$$f\left(h_i^{lk}, h_j^{lk}\right) = \frac{\left(h_i^l W_Q^{lk}\right)^T h_j^l W_K^{lk}}{\sqrt{d/k}} \tag{5}$$

Where$\mathrm{W}_{\mathrm{Q}}^{\mathrm{lk}} \in \mathrm{R}^{\frac{d}{K} xd}$ and $\mathrm{W}_{\mathrm{K}}^{\mathrm{lk}} \in \mathrm{R}^{\frac{d}{K} xd}$ are the learnable weights of k-th head at l layer.

3.2.2 Point-wise convolution transformation(PCT)

PCT [14] can transform the information collected by GAT. The size of the convolution kernel is 1, and the same conversion is performed on each position. The PCT is defined as:

$$PCT\left(h^l\right) = \sigma\left(h^l W_{pc}^1 + b_{pc}^1\right) W_{pc}^2 + b_{pc}^2 \tag{6}$$

Where σ denoted the RELU activation, $*$ stands for the convolution operation, $\mathrm{W}_{\mathrm{pc}^1} \in \mathrm{R}^{dxd}$ and $\mathrm{W}_{\mathrm{pc}^2} \in \mathrm{R}^{dxd}$ are weights of two convolutional, b_{pc}^1 and b_{pc}^2 are biases of the two convolutional kernels.

3.3 Aspect-based attention(ABA)

Motivated by MGAN, a single granularity of information may lose part of the characteristic information. In addition, by observing the sentence itself, we found that some sentences cannot judge the emotional polarity through aspect but need the information of the whole sentence. However, the coarse-grained information obtained by direct average pooling of sentences is not accurate enough. So we

design ABA to highlight the important parts of the sentence related to the aspect, and then perform average pooling . In addition, to avoid the influence of the aspect itself on the attention calculation, we have added a mask mechanism to shield the influence of the aspect. Let a be the index set of aspect, we get the Mask vector as follows:

$$\text{Mask}_i = \begin{cases} -\inf & \text{if} \quad i \in \text{a}; \\ 0 & \text{if} \quad \text{other}. \end{cases} \tag{7}$$

We get syntactic features $\left\{h_1^{sy}, h_2^{sy}, \ldots, h_n^{sy}\right\}$ and semantic features $\left\{h_1^{se}, h_2^{se}, \ldots, h_n^{se}\right\}$, which are generated by GAT and BERT channels, respectively. The ABA module is used to learn coarse-grained information about a specific aspect.

$$h_{\max}^{se} = \text{maxpooling}\left(h_i^{se}, i \in a\right) \tag{8}$$

$$h_{\max}^{sy} = \text{maxpooling}\left(h_i^{sy}, i \in a\right) \tag{9}$$

$$A^{se} = \text{softmax}\left(h_{\max}^{se} W^{se} h^{se^T} + \text{Mask}\right) \tag{10}$$

$$A^{sy} = \text{softmax}\left(h_{\max}^{sy} W^{sy} h^{sy^T} + \text{Mask}\right) \tag{11}$$

$$h^{tse} = A^{se} h^{se} \tag{12}$$

$$h^{tsy} = A^{sy} h^{sy} \tag{13}$$

Where $\text{W}^{\text{se}} \in \text{R}^{dxd}$ and $\text{W}^{\text{sy}} \in \text{R}^{dxd}$ are the attention weight matrix. We average pool the results of the ABA of the two channels to obtain the accurate coarse-grained features of the context.

$$h_{avg}^{tse} = \text{avgpooling}\left(h_1^{tse}, h_2^{tse}, \ldots, h_n^{tse}\right) \tag{14}$$

$$h_{avg}^{tsy} = \text{avgpooling}\left(h_1^{tsy}, h_2^{tsy}, \ldots, h_n^{tsy}\right) \tag{15}$$

3.4 Multi-granularity gate and feature fusion

We average pool the obtained context features as coarse-grained features From the GAT and BERT channels, we get the fine-grained of the Maxpooling aspect and the coarse-grained features of the context computed by ABA, respectively. In order to better integrate the multi-granularity information of the two channels, we design multi-granularity gate control using two gates to merge separately coarse and fine-grained information. The integrated aspect-context semantic and sufficient syntactic information can effectively cooperate, defined as follows:

$$h_{\text{g}} = \left[G\left(h_{\text{avg}}^{\text{tse}}, h_{\text{avg}}^{\text{tsy}}\right); G\left(h_{\max}^{se}, h_{\max}^{sy}\right)\right] \tag{16}$$

Where $\left[G\left(h_{\text{avg}}^{\text{tse}}, h_{\text{avg}}^{\text{tsy}}\right); G\left(h_{\max}^{se}, h_{\max}^{sy}\right)\right]$ represents multi-granularity information concatenated. G is the gated function, defined as follows:

$$\text{G}(x1, x2) = g \circ x1 + (1 - g) \circ x2 \tag{17}$$

$$g = \sigma\left(W^g[x1; x2] + b^g\right) \tag{18}$$

Where σ and $\circ$ denoted the sigmoid activation and element-wise product operation respectively, $\text{W}^{\text{g}} \in \text{R}^{2dxd}$ and b^g is are model parameters.

It is worth noting that many BERT-based models can achieve better results but only use BERT as the embedding layer. In order to make full use of BERT, we get final representations of the previous outputs by MG Gate and BERTPooling, concatenate them as the final comprehensive representation, and use a fully connected network to project the concatenated vector into the space of the targeted C classes.

$$h^f = W^f\left[h_g; h_{CLS}\right] + b^f \tag{19}$$

$$y = \text{Soft max}\left(h^f\right) \tag{20}$$

Where ; represents concatenation operation, $\text{W}^{\text{f}} \in \text{R}^{2dxd}$ and b^f is learnable parameters.

3.5 Classification

The model uses minimizing the cross entropy with $L2$-regularization term to train, which is defined as:

$$\text{Loss} = -\sum_{i=1}^{N}\sum_{a \in C} \hat{y}^c \log(y^c) + \lambda\|\Theta\|^2 \tag{21}$$

Where $\hat{y}^c$ is the ground truth represented as a one-hot vector, λ is a regularization hyperparameter, and Θ is set of the model parameter.

4 Experiment

For this section, the evaluated datasets and the compared baseline model are introduced first. The results of DMG-GAT are carried out, and finally, the ablation experiments on each module of the model along with the analysis of the results.

4.1 Datasets

The experiment is conducted on three well-known benchmark datasets, including the Lap14, Rest14 of SemEval2014 [43] and Twitter [37] datasets. The ratio of the validation set to the training set is 31%, 28%, and 11% for Restaurant, Laptop, and Twitter, respectively. The datasets are labeled positive, negative, and neutral. The specific information of these datasets are listed in Table 1.

4.2 Parameter settings and evaluation metrics

Our model is built using the PyTorch framework. Two version of BERT [30] encoders are considered: 1) The BERT-large encoder ; 2) The BERT-based encoder. For

Table 1 Detailed statistics of the three data sets used in our experiment

	Positive		Negtive		Neutral	
Dataset	Accuracy	Macro-f1	Accuracy	Macro-f1	Accuracy	Macro-f1
Restaurant	2164	727	807	196	637	196
Laptop	976	337	851	128	455	167
Twitter	1507	172	1528	169	3016	336

the BERT-large encoders, 1024-dimensional as adopted the word representation. the dropout of BERT-large embeddings is 0.3, and regularization term is $\lambda = 2 * 10^{-5}$. For the BERT-base encoders, the dropout rate on BERT-base embeddings is 0.1, and regularization term $\lambda = 2 * 10^{-5}$. The adam [44] optimizer is used as the optimizer of the model, with the learning rate set to 10^{-5} and the $L2$ regularization is set to 10^{-5}. For GAT Encodr, after many experiments, the optimal number of GAT layers and number of attention heads are 2 and 4, respectively, and the dropout is 0.1. The dependency trees are obtained by Deep Biaffine Parser. The batch size is set to 16 and the number of epochs is 10.

ABSA task is a multi-classification task in nature, so F1 and ACC, the two most common measurement standards of the classification task, are compared with other models.

4.3 Baselines

Base Attention or BERT.

- IAN [6]: Context and aspect respectively are modeled through LSTM, and then interactively calculates the attention mechanism to obtain a more accurate feature representation of context and aspect.
- TNet [7]: The aspect is modeled by Bi-LSTM and the relevance is calculated with each word. Then use cnn to extract features.
- MGAN [13]: A Bi-LSTM is used to encode context, then uses a multi granularity attention mechanism to capture the relationship between aspect and context.
- AEN [14]: Attentional encoder network uses attention coding network to encode context and aspect and models the semantic interaction between target and context.
- BERT-SPC [30]: Taking [CLS] + context + [SEP] + aspect + [SEP] as input, use h_{cls} for sentiment classification.

Base GNNs.

- TD-GAT [23]: Huang et al. uses GAT to capture the syntax structure, and uses LSTM to improve the syntax structure to model cross-layer relationships.
- SA-GAT [24]: Uses the GAT on the dependency tree structure and the BERT language model to better model the interaction between context and aspect.
- SDGCN [36]: proposes a bidirectional attention mechanism with location coding to model aspect and context-based word representation and uses GCN to capture emotional dependence between different aspects in a sentence.
- RGAT [40]: Wang et al. use the relational GAT to encode the dependency relations and connections between target aspect and opinions.
- DGEDT [25]: proposes a dual-transformer network to iteratively interact with flat representations learned and graph-based representations learned from dependency graphs to solve the instability and noise of dependency trees.
- DualGCN [26]: proposes to use a dual GCN structure to learn graph dependency representation as well as self-attentive graph representation.

4.4 Overall results

In Table 2, we compare Accuracy and Macro-f1 values of our model DMGGAT with other baseline models on the three datasets of TRIPLE Restaurant, Laptop, and Twitter. We use red, blue and green to represent the top three of each indicator in the table. DMGGAT-Base and DMGGAT-Large represent the performance of the model under BERT's Base and Large versions, respectively. DMGGAT-Large achieved the best results in ACC and F1 at Restaurant and Laptop, with the third best result on Twitter. Even DMGGAT-Base is highly competitive on all three datasets. As shown in the table, we explore comparisons the model of based on attention, dependency graph, and BERT model as follow.

ATAE-LSTM, IAN, TNet, MGAN, and AEN are models that model correlations between specific aspects and contexts through attentional mechanisms only, so they are the least effective in BASELINE. They model the relevant information between specific aspects and contexts through the attention mechanism. This is because in the absence of syntactic information, the attention mechanism may fail

Table 2 Experimental results on three datasets. attention, Graph, and BERT represent whether the model uses the attention mechanism, Graph, and BERT models

	Model	Restaurant		Laptop		Twitter	
		Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
Attention	ATAE-LSTM [12]	77.20	–	68.70	–	–	–
	IAN [6]	78.60	–	72.10	–	–	–
	TNet [7]	80.79	–	76.54	–	73.12	–
	MGAN [13]	81.25	71.94	75.39	72.47	72.54	70.81
	AEN [14]	80.98	72.14	73.15	69.04	72.83	69.81
Graph	TDGAT [23]	81.20	–	74.0	–	–	
	SDGCN [36]	82.95	75.79	75.55	71.35	–	–
	RGAT [40]	83.30	76.08	77.42	73.76	75.57	73.82
	DGEDT [25]	83.90	75.10	76.80	72.30	74.80	73.40
	DualGCN [26]	84.27	78.08	78.48	74.74	75.92	74.29
BERT	BERT-SPC [30]	84.94	78.00	78.47	73.67	74.71	73.13
	SAGAT-BERT [24]	83.12	73.76	79.93	76.31	75.40	74.17
	SDGCN-BERT [36]	83.57	76.47	81.35	78.34	–	–
	RGAT-BERT [40]	**86.60**	**81.35**	78.21	74.07	76.15	74.88
	DGEDT-BERT [25]	86.30	80.00	79.80	75.60	**77.90**	**75.40**
	DualGCN-BERT [26]	**87.13**	81.16	**81.80**	**78.10**	**77.40**	**76.02**
Ours	DMGGAT-Base	**87.13**	**81.19**	**80.78**	**77.57**	75.99	74.56
	DMGGAT-Large	**87.58**	**82.73**	**82.19**	**79.49**	76.63	**74.93**

Red, blue, and green represent the top three models in terms of performance, respectively

due to the complexity of the sentence. It is worth noting that MGAN achieves the best results on Restaurant and Laptop because it proposes that the idea of multi-granularity interaction can help the model learn more accurate aspects of emotions.

Here we analyze the model based on external syntactic dependencies in detail. TD-GAT, SDGGCN, and RGAT use GCN or GAT to extract dependencies from the dependency tree, but they ignore the instability of the dependency tree. DGEDT adds a transformer channel to enhance the learning of graph representation. DGEDT and DualGCN perform better than previous models because they have similar ideas to solve the dependency tree Instability and noise. They add Transformers and semGCN channels, respectively, to enhance the learning of syntactic graph.

It is worth noting the power of BERT, because BERT-SPC outperforms all models that do not use BERT without adding any components. As show in the Table 2, the performance of all models has been greatly improved after using BERT as the embedding layer. Previous work [33] has demonstrated that the dependency tree induced by BERT is stronger than the dependency tree generated by the parser. Therefore, We add BERT channels to reduce the instability of dependency trees and solve the lost semantic information in graph computation. DGEDT and DualGCN use the semantic features learned by Transformer and

Table 3 Results of ablation experiments

	Restaurant		Laptop		Twitter	
Model	Accuracy	Macro-f1	Accuracy	Macro-f1	Accuracy	Macro-f1
w/o GAT	85.42	78.29	80.59	77.23	73.92	72.96
w/o BERT	85.87	79.49	79.23	75.22	75.23	74.06
w/o ABA	86.58	80.55	80.62	77.03	75.12	73.83
w/o MG-Gate	85.35	80.36	80.16	76.71	75.88	74.50
w/o Fine	84.27	77.69	78.59	74.61	75.84	74.35
w/o Coarse	84.62	77.78	78.44	74.44	74.24	72.56
DMGGAT-Base	87.13	81.19	80.78	77.57	75.99	74.56

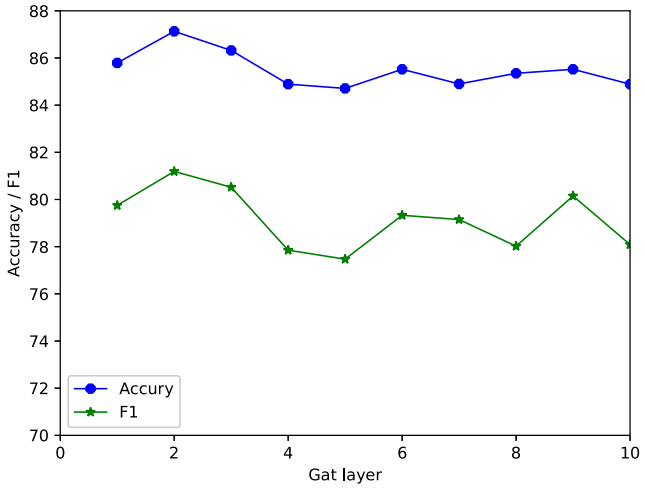

Fig. 4 In the Restaurant dataset, the number of GAT Layer affects the performance of the model

SemGCN to support the learning of graphs respectively. We use BERT to extract syntactic and semantic features, and then use a gate mechanism to obtain more accurate features. In addition, they mostly ignore the importance of sentence-level information, while we use ABA to generate coarse-grained information that is consistent with aspect, and learn sentence-level information.

4.5 Ablation study

To further verify the validity of each component of the model, we performed ablation experiments on each component. The experimental variables are as follows. The experiment is conducted on restaurant, laptop, and Twitter datasets, the results as shown in Table 3.

1. **w/o GAT:** removes the GAT module and uses aspect-specific pooling and sentence-level information generated by BERT to make sentiment judgments.
2. **w/o BERT:** removes the BERT channel and uses only BERT to provide an embedded representation of words for the GAT layer, then uses aspect-specific pooling and sentence-level generated by GAT for sentiment analysis.
3. **w/o ABA:** removes the ABA module to generate aspect-specific sentence-level information and directly averages the pooling of sentences as sentence-level information.
4. **w/o MG-Gate**: removes the MG-Gate module and performs the fusion using a stitching method.
5. **w/o Fine:** We did not consider the information of aspect words and only used the coarse-grained information obtained by ABA for prediction.
6. **w/o Coarse:** Just like w/o Coarse, We do not consider the coarse-grained sentence-level information and only use pooled aspect words generated by two channels for emotion prediction.

Table 3 shows accuracy and F1 values of each DMGGAT component ablation experiment. It can be seen that the performance of the model without any component is inferior to DMGGAT. Compared with the DMGGAT, the accuracy and F1 on the three datasets are reduced by 1 and 2 on average.

Compared with DMGGAT, the performance of w/o GAT is significantly reduced, which is due to the lack of information extraction of dependency tree by GAT, resulting in insufficient syntactic information perception of the model and lack of syntactic information, thus reducing the ability of the model to capture important information. It can be seen from W/O BERT that the performance of the model is significantly reduced, because part of semantic information may be lost during text encoding by GAT, and BERT is absent to supplement semantic information. It can be seen from w/o BERT and o/o GAT that no matter which one is missing, the performance of the model will decline, which indicates that BERT and GAT have mutual achievement effect in modeling, so both are effective.

In Experiment 3, the performance degradation of the model can be seen after the ABA module is removed. It is not accurate to generate sentence-level coarse-grained features by Average direct pooling. Because ABA makes aspect words emphasize the relevant parts of the sentence, the coarse-grained information is more accurately generated by average pooling. In Experiment 4, removing the Mg-gate leads to model performance degradation because the gated mechanism can effectively screen out important features.

Table 4 Effect of pooling method on experimental results

	Restaurant		Laptop		Twitter	
Pooling method	Accuracy	Macro-f1	Accuracy	Macro-f1	Accuracy	Macro-f1
Aspect-avg+sentence-avg	85.06	77.49	78.91	75.64	75.12	73.87
Aspect-avg+sentence-max	85.79	79.42	79.22	75.27	75.26	74.35
Aspect-max+sentence-max	85.79	78.91	79.21	75.95	75.26	73.87
Aspect-max+sentence-avg	87.13	81.19	80.78	77.57	75.99	74.56

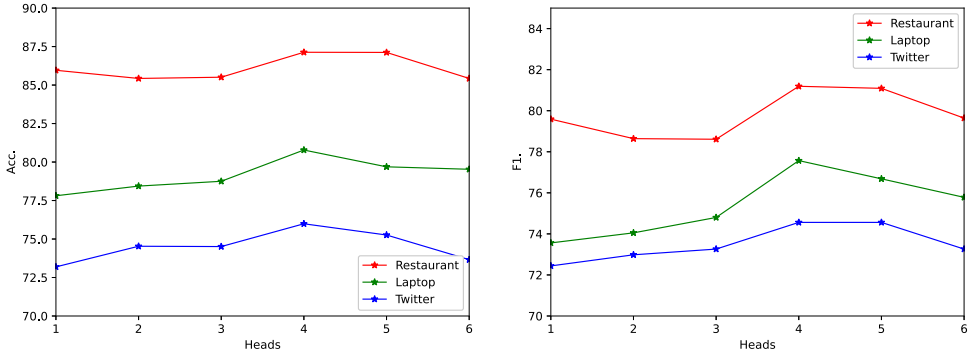

Fig. 5 The effect of the number of attention heads in the graph attention calculation on the acc and f1 values on the three datasets

We use the gated mechanism to screen out important features from two granularity channels to predict emotional results effectively.

In the experiment w/o fine, the fine-grained feature of pooled aspect words is removed, and the information of aspect words is particularly important in ABSA tasks, which leads to a significant decline in performance. Experiment w/o Goarse removed the sentence level coarse-grained feature. Many experiments have shown that context is also essential in judging aspect words. It can be seen that the effect of the model after removing fine grain and coarse grain is not as good as DMGGAT, indicating that both are effective.

4.6 Parameter analysis

4.6.1 Impact of GAT layer number

The number of GAT layers is a significant parameter impacting the result of the model. We conduct different GAT layer experiments to judge the influence on the model. The Fig. 4 shows the performance effects of GAT layers 1-10 on Restaurant datasets. When the number of GAT layers is 2, the effect of the model is better. As expected, when the layer number is greater than 2, the performance of the model decreases as the number of layers increases. The performance degradation may be influenced by the excessive number of layers leading to the overfitting of the training datasets.

4.6.2 Impact of pooling method

In this section, we have discussed the way the model is pooled. As shown in Table 4, it is divided into four main discussion cases. The experimental results show that the best experimental results are obtained by using maximum pooling for aspect words and average pooling for sentences.

4.6.3 Impact of attention head

The number of attention heads in a graph attention network is an essential parameter. To demonstrate the optimal number of attention heads, we conduct experiments with different numbers of heads from 1 to 6. The experimental results are shown in Fig. 5. The results show that the best results are achieved when the number of heads is about

Fig. 6 Several case shows are selected from Restaurant and Laptops. [Neg, Pos, Neu] denotes predicted sentiment distribution: [NEGATIVE, POSITIVE, NEUTRAL]

#	Sentence	[Neg, Pos, Neu]
1.	Not to Sound too negative but be wary of the **delivary**	M0[0.74✔, 0.2, 0.06]
		M1[0.12, 0.87✖, 0.01]
2.	Great beer selection too , something like 50 **beers**	M1[0.02, 0.25, 0.73✔]
		M0[0, 0.99✖, 0.01]
3.	Air has higher resolution but the **fonts** are small	M0[0.87✔, 0.07, 0.06]
		M1[0.35, 0.56✖, 0.09]
4.	The apple engineers have not yet discovered the **delete key**	M0[0.77✔, 0.22, 0.01]
		M1[0.41, 0.01, 058✖]
5.	Go with the **specials** , and stay away from the salmon	M0[0.04, 0.51✔, 0.46]
		M1[0.01, 0.24, 0.75✖]

4. When set to 1-3 attention heads, the model does not learn enough information leading to poor performance of the model. On the contrary, more attention heads lead to poor performance of the model due to overfitting.

4.7 Case analysis

To better show the effectiveness of the model. We conduct a case study between M0(DMGGAT) and M1(GAT). As shown in the Fig. 6, several cases are selected to be discussed from the results of M0 and M1. We observe that M0 can accurately predict the correct label in all cases, while M1 fails in all cases.

M1 without BERT channel, the model cannot capture sufficient semantics and there are syntactic errors. In case 1, M1 has the ability to capture that the aspect "$delivery$" is related to "$wary$", but the misunderstanding of "$wary$" leads to classification failure. While M0 can correctly understand "$wary$" is negative because BERT provides semantic information. In case 2, M1 captures "$beers$" is related to "$like$" but thinks like is positive. M0 can understand that "$like$" here is neutral because it can learn sentence-level information. In case3 "$small$" does not have a clear emotional polarity. It is necessary to infer that small is negative based on the meaning of the first half of the sentence. M0 adds coarse-grained information of the sentence to help the model understand the meaning of the whole sentence. In case4 and case5, we can find no apparent emotional word in the sentence. Even if the relevant word is captured, it will fail. Thanks to the BERT channel and coarse-grained information added M0 captures the semantic information of sentences and aspect words to predict the results jointly.

5 Conclusion and future works

In this paper, we propose a novel Dual-Channel and Multi-Granularity Gated Graph Attention Network to solve ABSA tasks. Previous work has shown that PTMs can also induce syntactic dependency trees that outperform parser-generated dependency trees. In contrast, one of the major original contributions of our research is that BERT is used not only as an embedding layer of the model, but also to reduce the noise introduced by the parser-generated dependency trees. Also, BERT can provide semantic information lost in the graph attention computation. Most of the work only considers the information of aspect words and ignores the sentence-level information, so we design ABA to generate aspect-specific sentence-level information. To make the two granularity information of the two channels cooperate better to predict the results, we designed a new multi-granularity gated mechanism to fuse the same granularity features of the two channels, respectively. This method takes full account of syntactic and semantic features as well as word level and sentence level features and can generate more accurate and comprehensive representations. We conducted ablation experiments to verify the effectiveness of the model components. We have also conducted extensive experiments on three benchmark datasets, and the experimental results show that our model has achieved advanced results

In the future we will introduce external knowledge embedding for the model to understand the meaning of the words. I am sure this will yield better results.

Acknowledgements The authors would like to thank Editor-in-Chief, editor, and anonymous reviewers for their valuable comments and helpful suggestions. This work is supported by the National Natural Science Foundation of China under Grant 61976158 and Grant 61673301.

Declarations

The datasets analysed during the current study are available from the corresponding author on reasonable request.

References

1. Bakshi RK, Kaur N, Kaur R, Kaur G (2016) Opinion mining and sentiment analysis. In: 2016 3rd international conference on computing for sustainable global development (INDIACom). IEEE, pp 452–455
2. Liu B (2012) Sentiment analysis and opinion mining. Synthesis Lect Human Lang Technol 5(1):1–167
3. Jiang L, Yu M, Zhou M, Liu X, Zhao T (2011) Target-dependent twitter sentiment classification. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, pp 151–160
4. Wagner J, Arora P, Cortes S, Barman U, Bogdanova D, Foster J, Tounsi L (2014) Dcu: aspect-based polarity classification for semeval task 4
5. Huang B, Carley KM (2018) Parameterized convolutional neural networks for aspect level sentiment classification. In: Proceedings of the 2018 conference on empirical methods in natural language processing, pp 1091–1096
6. Ma D, Li S, Zhang X, Wang H (2017) Interactive attention networks for aspect-level sentiment classification. In: Proceedings of the twenty-sixth international joint conference on artificial intelligence, IJCAI 2017, pp 4068–4074
7. Li X, Bing L, Lam W, Shi B (2018) Transformation networks for target oriented sentiment classification. In: Proceedings of the 56th annual meeting of the association for computational linguistics, ACL 2018, pp 946–956
8. Zeng J, Ma X, Zhou K (2019) Enhancing attention-based lstm with position context for aspect-level sentiment classification. IEEE Access 7:20462–20471
9. Ma X, Zeng J, Peng L, Fortino G, Zhang Y (2019) Modeling multi-aspects within one opinionated sentence simultaneously for aspect-level sentiment analysis. Futur Gener Comput Syst 93:304–311

10. Chen P, Sun Z, Bing L, Yang W (2017) Recurrent attention network on memory for aspect sentiment analysis. In: Proceedings of the 2017 conference on empirical methods in natural language processing, pp 452–461
11. Tang D, Qin B, Feng X, Liu T (2016) Effective lstms for target-dependent sentiment classification. In: COLING 2016, 26th international conference on computational linguistics, proceedings of the conference: technical papers, December 11-16, 2016, pp 3298–3307
12. Wang Y, Huang M, Zhu X, Zhao L (2016) Attention-based lstm for aspect-level sentiment classification. In: Proceedings of the 2016 conference on empirical methods in natural language processing, pp 606–615
13. Fan F, Feng Y, Zhao D (2018) Multi-grained attention network for aspect-level sentiment classification. In: Proceedings of the 2018 conference on empirical methods in natural language processing, pp 3433–3442
14. Song Y, Wang J, Jiang T, Liu Z, Rao Y (2019) Targeted sentiment classification with attentional encoder network. In: International conference on artificial neural networks, pp 93–103
15. Qiu G, Liu B, Bu J, Chen C (2011) Opinion word expansion and target extraction through double propagation. Comput Linguist 37(1):9–27
16. Liu K, Xu HL, Liu Y, Zhao J (2013) Opinion target extraction using partially-supervised word alignment model. In: IJCAI, vol 13, pp 2134–2140
17. Xing B, Liao L, Song D, Wang J, Zhang F, Wang Z, Huang H (2019) Earlier attention? aspect-aware LSTM for aspect-based sentiment analysis. In: Proceedings of the twenty-eighth international joint conference on artificial intelligence, IJCAI 2019, pp 5313–5319
18. Liang Y, Meng F, Zhang J, Xu J, Chen Y, Zhou J (2019) A novel aspect-guided deep transition model for aspect based sentiment analysis. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, EMNLP-IJCNLP 2019, pp 5568–5579
19. Xue W, Li T (2018) Aspect based sentiment analysis with gated convolutional networks. In: Proceedings of the 56th annual meeting of the association for computational linguistics, ACL 2018, pp 2514–2523
20. Zhang C, Li Q, Song D (2022) Aspect-based sentiment classification with aspect-specific graph convolutional networks. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, EMNLP-IJCNLP 2019, pp 4567–4577
21. Xu K, Zhao H, Liu T (2020) Aspect-specific heterogeneous graph convolutional network for aspect-based sentiment classification. IEEE Access 8:139346–139355
22. Liang Y, Meng F, Zhang J, Chen Y, Xu J, Zhou J (2021) A dependency syntactic knowledge augmented interactive architecture for end-to-end aspect-based sentiment analysis. Neurocomputing 454:291–302
23. Huang B, Carley KM (2019) Syntax-aware aspect level sentiment classification with graph attention networks. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, EMNLP-IJCNLP 2019, pp 5468–5476
24. Huang L, Sun X, Li S, Zhang L, Wang H (2020) Syntax-aware graph attention network for aspect-level sentiment classification. In: Proceedings of the 28th international conference on computational linguistics, pp 799–810
25. Tang H, Ji D, Li C, Zhou Q (2020) Dependency graph enhanced dual-transformer structure for aspect-based sentiment classification. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 6578–6588
26. Li R, Chen H, Feng F, Ma Z, Wang X, Hovy E (2021) Dual graph convolutional networks for aspect-based sentiment analysis. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing, pp 6319–6329
27. Zhou K, Zeng J, Liu Y, Zou F (2018) Deep sentiment hashing for text retrieval in social ciot. Futur Gener Comput Syst 86:362–371
28. Truong Q-T, Lauw HW (2019) Vistanet: visual aspect attention network for multimodal sentiment analysis. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 305–312
29. Dumpala SH, Sheikh I, Chakraborty R, Kopparapu SK (2019) Audio-visual fusion for sentiment classification using cross-modal autoencoder. In: 32nd conference on neural information processing systems (NIPS 2018), pp 1–4
30. Devlin J, Chang M, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, NAACL-HLT 2019, pp 4171–4186
31. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł., Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
32. Hou X, Qi P, Wang G, Ying R, Huang J, He X, Zhou B Graph ensemble learning over multiple dependency trees for aspect-level sentiment classification. In: Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies, NAACL-HLT 2021, pp 2884–2894
33. Dai J, Yan H, Sun T, Liu P, Qiu X (2021) Does syntax matter? A strong baseline for aspect-based sentiment analysis with roberta. In: Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies, NAACL-HLT 2021, pp 1816–1829
34. Ding X, Liu B, Yu PS (2008) A holistic lexicon-based approach to opinion mining. In: Proceedings of the 2008 international conference on web search and data mining, pp 231–240
35. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
36. Zhao P, Hou L, Wu O (2020) Modeling sentiment dependencies with graph convolutional networks for aspect-level sentiment classification. Knowl-Based Syst 193:105443
37. Dong L, Wei F, Tan C, Tang D, Zhou M, Xu K (2014) Adaptive recursive neural network for target-dependent twitter sentiment classification. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers), pp 49–54
38. Nguyen TH, Shirai K (2015) Phrasernn: Phrase recursive neural network for aspect-based sentiment analysis. In: Proceedings of the 2015 conference on empirical methods in natural language processing, pp 2509–2514
39. He R, Lee WS, Ng HT, Dahlmeier D (2018) Effective attention modeling for aspect-level sentiment classification. In: Proceedings of the 27th international conference on computational linguistics, pp 1121–1131
40. Wang K, Shen W, Yang Y, Quan X, Wang R (2020) Relational graph attention network for aspect-based sentiment analysis. In: Proceedings of the 58th annual meeting of the association for computational linguistics, ACL 2020, pp 3229–3238
41. Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: Proceedings of the 2014

conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
42. Li X, Bing L, Zhang W, Lam W (2019) Exploiting BERT for end-to-end aspect-based sentiment analysis. In: Proceedings of the 5th workshop on noisy user-generated Text, W-NUT@EMNLP 2019, pp 34–41
43. Pontiki M, Galanis D, Pavlopoulos J, Papageorgiou H, Androutsopoulos I, Manandhar S (2014) Semeval-2014 task 4: aspect based sentiment analysis. In: Proceedings of the 8th international workshop on semantic evaluation, SemEval@COLING 2014, pp 27–35
44. Kingma DP, Ba J (2015) Adam: a method for stochastic optimization. In: 3rd International conference on learning representations, ICLR 2015

Yong Wang received the Ph.D. degree from East China Normal University, in 2007. He is currently an Associate Professor with the School of Artificial Intelligence, Chongqing University of Technology. His research interests include deep learning, natural language processing, multimedia, and big data technology.

Ningchuang Yang received the bachlor's degree from the College of Mathematics and Big Data, Chongqing University of Arts and Sciences in 2020. He is currently pursuing the master's degree with the Liangjiang School of Artificial Intelligence, Chongqing University of Technology. His research interests include deep learning, natural language processing, and sentiment analysis.

Duoqian Miao received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, in 1997. He is currently a Professor with the Department of Computer Science and Technology, Tongji University. His research interests include artificial intelligence, machine learning, big data analysis, and granular computing.

Qiuyi Chen received the bachlor's degree from the Liangjiang International College, Chongqing University of Technology in 2020. She is currently pursuing the master's degree with the School of Artificial Intelligence, Chongqing University of Technology. Her research interests include deep learning, natural language processing, and machine reading comprehension.