IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 22, 2020, accepted December 31, 2020, date of publication January 5, 2021, date of current version January 14, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3049294

Multi-Head Self-Attention Transformation Networks for Aspect-Based Sentiment Analysis

YUMING LIN, CHAOQIANG WANG, HAO SONG, AND YOU LI
Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China

Corresponding author: You Li (liyou@guet.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62062027, Grant U1711263, and Grant 61762027; in part by the Guangxi Natural Science Foundation under Grant 2018GXNSFDA281049, Grant 2020GXNSFAA159012, Grant 2018GXNSFAA281326, and Grant 2018GXNSFAA138090; in part by the Science and Technology Major Project of Guangxi Province under Grant AA19046004; in part by the Innovation Project of Guest Graduate Education under Grant 2019YCXS040; and in part by the Guangxi Key Laboratory of Trusted Software under Grant kx201916.

ABSTRACT Aspect-based sentiment analysis (ABSA) aims to analyze the sentiment polarity of an input sentence in a certain aspect. Many existing methods of ABSA employ long short-term memory (LSTM) networks and attention mechanism. However, the attention mechanism only models the local certain dependencies of the input information, which fails to capture the global dependence of the inputs. Simply improving the attention mechanism fails to solve the issue of target-sensitive sentiment expression, which has been proven to degrade the prediction effectiveness. In this work, we propose the multi-head self-attention transformation (MSAT) networks for ABSA tasks, which conducts more effective sentiment analysis with target specific self-attention and dynamic target representation. Given a set of review sentences, MSAT applies multi-head target specific self-attention to better capture the global dependence and introduces target-sensitive transformation to effectively tackle the problem of target-sensitive sentiment at first. Second, the part-of-speech (POS) features are integrated into MSAT to capture the grammatical features of sentences. A series of experiments carried on the SemEval 2014 and Twitter datasets show that the proposed model achieves better effectiveness compared with several state-of-the-art methods.

INDEX TERMS Aspect-based sentiment analysis, self-attention, transformation networks, target-specific transformation.

I. INTRODUCTION

Sentiment analysis for online reviews can provide valuable information for both businesses and consumers. Aspect-based sentiment analysis (ABSA) is a fine-grained sentiment analysis task [1], which aims to analyze the sentiment polarity of users towards specific aspects in a sentence. For example, given a review sentence "*Great food but the service was dreadful!*", ABSA is expected to assign a positive sentiment type to the aspect "*food*" and a negative sentiment type to the aspect "*service*", respectively.

Recently, neural models have been shown to be successful on ABSA. However, the standard recurrent neural network (RNN) encounters the gradient vanishing or exploding problem. Long short-term memory network (LSTM) could handle this problem effectively. Attention mechanism can provide neural networks with the ability to focus on the significant contents related to the given aspect. Thus, many existing models leverage LSTM to distill sentiment information from embedding vectors, and apply attention mechanism to make them focus on the specific scope of a given entity in a sentence. Such models include the attention-based LSTM with aspect embedding (ATAE-LSTM) [2], target-dependent sentiment classification (TD-LSTM) [3], recurrent attention memory network (RAM) [4], and so on. In these work, LSTM can capture sequence information effectively, which solves the problem of gradient vanishing or exploding in some ways, but it rarely captures the interdependence characteristics between words. Moreover, the combination of word-level features based on attention weight may introduce some noise and downgrade the prediction accuracy [5]. This phenomenon can also be observed in machine translation [6] and image captioning [7].

For ABSA, it is essential to capture the interdependent features between the words of a sentence to learn the structural information of the sentence, and then to infer the sentiment polarity on a given aspect. The attention mechanism can obtain the degree of association between the target word and

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

each word in the sentence and generate the target representation depending on each word. However, because natural language expression is delicate and complex, it is not enough to only obtain the dependent information between the target word and other words in the sentence. Such dependent information is called the local dependence in this work. To better represent the complex sentence structures, we need to obtain the dependent information between each word and other words in the sentence. Such dependent information is called the global dependence. Self-attention [8] is an effective mechanism for capturing the global dependency features between words. However, standard self-attention cannot model the word order of a sentence. Thus, we encode the relative position information of words in the sentence and integrate them into our proposed model.

Another important issue is that the sentiment expression of a word is sensitive to its described target (called target-sensitive sentiment in [9]). Then, only relying on attention will degrade the prediction effectiveness. Considering the following two sentences: ''*The price is high.*'' and ''*The screen resolution is high.*''. The word ''*high*'' is used to express the user's sentiment in these two sentences. However, it means a negative sentiment in the first sentence and a positive sentiment in the second one. Therefore, Generating a specific word representation based on the given target is a good solution [5] rather than focusing on improving attention for such situations.

In this paper, we propose an effective model, named multi-head self-attention transformation network (MSAT), to solve the above issues for the ABSA tasks. MSAT firstly encodes the context words and their POS information into word embeddings and generates the contextualized word representations by Bi-LSTM. And then, MSAT introduces the target specific Multi-head Self-Attention mechanism to capture the global dependency features between words effectively. Finally, we integrate target-sensitive Transformation module into the MSAT to tackle the problem of target-sensitive sentiment expression and implement the aspect-level sentiment classification.

In summary, our contributions can be summarized as follows:

1) A MSAT model is proposed for the aspect-based sentiment analysis in an end-to-end fashion, by which the modules in the middle can be trained as a whole. The proposed model is able to capture the information of global interdependence between words in the sentence.
2) A module of dynamic target representation is designed to solve the problem of target-sensitive sentiment expression, and the part of speech features of words are encoded to capture the grammatical features of sentences.

The rest of this paper is organized as follows. Section II gives a brief review of related works on ABSA. Section III describes the proposed MSAT model in detail. In Section IV, we introduce the experiment setup and results to verify the effectiveness of the proposed model. The case study is also provided in this section. Finally, Section V draws the conclusion.

II. RELATED WORK

Aspect-based sentiment analysis is an indispensable task in sentiment analysis. Most of the traditional approaches adopt some classic classification models (such as supported vector machine) based on the extensive hand-coded features [10], such as bag-of-words, sentiment lexicons [11], [12]. However, the results generated by these methods highly depend on the quality of features. In addition, these methods are labor intensive and usually result in high-dimensional, high-sparse text representation. More traditional approaches and their details can be found in [1].

In recent years, many neural networks-based methods have achieved good effectiveness in various sentiment analysis tasks. Zhang *et al.* proposed TD-LSTM and TC-LSTM in [14], they suggested that the context should be split into two parts and the target should be associated with the contextual features separately. A three-way gate control network was proposed in [15], which can better capture the interaction between the target and its surrounding contexts.

With the success of attention mechanism in machine translation, its application in ABSA tasks has received more and more attention. Wang *et al.* proposed an attention-based LSTM with aspect embedding (ATAE-LSTM) model [2], which uses the embedding vectors of aspect words to selectively attend the regions of the representations generated by LSTMs. Tang *et al.* proposed a multi-hop memory network based on attention (MemNet) [15], its improved versions were proposed in [4] and [9]. IAN [16] is an model that can learn attention interactively in context. PBAN [17] is a position-aware bidirectional attention network based on bidirectional gated recurrent unit (GRU). These attention methods only consider the coarse-grained level attention, which only use the simple average embeddings of the words as the target representation. This strategy may be inappropriate in some cases because different words usually do not make equal contribute to the target representation.

Fine-grained attention mechanism is an effective way to capture the word-level interaction between aspect and context. A multi-grained attention network (MGAN) was proposed in [18], which applies coarse-grained and fine-attention mechanisms to capture interactive information between aspect and context. Zhang *et al.* used the multi-head attention (MHA) and point-wise feed-forward networks (PFFN) to interactively obtain the hidden representation of the context and aspect embeddings [19]. Xu *et al.* used a global attention module and a local one to capture different granularity of interactive information between aspect and context in [20]. However, simply improving the attention mechanism cannot solve the issue of target-sensitive sentiment [9] and it cannot be inferred from the context alone, because the sentiment expression depends on the given target.

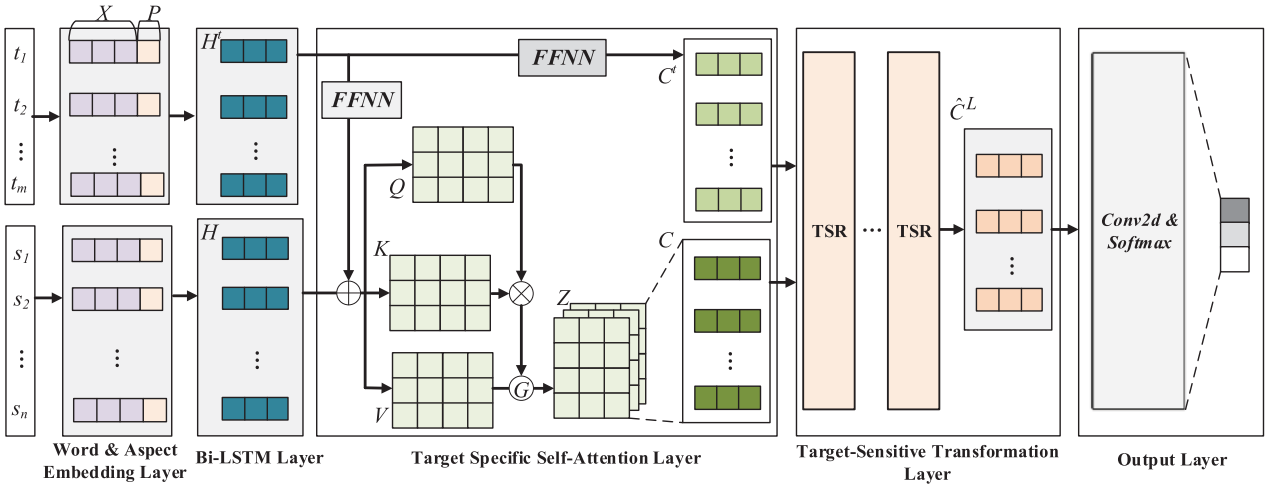

FIGURE 1. Overall framework of the proposed MSAT.

There are many models that use convolutional neural networks (CNN) and graph convolutional networks (GCN) for ABSA tasks. These works show that convolution operations can capture compositional structure of texts with rich semantic information. GCAE [21] is a model based on convolutional neural networks and gating mechanisms, which realizes parallelization during training. Li *et al.* [5] proposed a target-specific transformation networks (TNet) model, which adopts CNN for target-level sentiment classification. RPAEN [22] incorporates the relative position information and aspect attention into the CNN model for aspect-based sentiment analysis. To enhance the effect of model, [23] and [24] used GCN to draw syntactic information and model dependency tree graph for modeling the long-range word dependencies.

Despite the effectiveness of those methods, it is hard to capture the interdependence of words effectively. In this work, we employ the target specific self-attention mechanism to capture the global interdependence features between the words of a sentence. Then, the learned word representations are processed by a target-sensitive transformation layer to effectively tackle the problem of target-sensitive sentiment.

III. PROPOSED METHODOLOGY

Given a review sentence $S = \{s_1, s_2, \cdots, s_n\}$ and the aspect set $T = \{t_1, t_2, \cdots, t_m\}$, where s_i stands for the word of the sentence and t_i stands for the aspect term contained in this sentence. The goal of the proposed model is to predict the sentiment type $Y \in \{1, -1, 0\}$ for each aspect target in T, where 1, -1 and 0 denote sentiment types ''positive'', ''negative'' and ''neutral'', respectively.

The overall framework of our proposed MSAT is shown in Figure 1, which is composed of:

- **Word & aspect embedding layer**. Given an input sentence S and the aspect set T included in this sentence, a base encoder is adopted to generate the word representations in this layer.
- **Bi-LSTM layer**. To capture the contextual information of the input sequence, the sentence and aspect word embeddings are fed into the Bi-LSTM layer to generate the context word representations H and aspect representations H^t.
- **Target specific self-attention layer**. The aspect representations and the word representations are treated as the into data of the target specific self-attention layer to better capture the global dependence.
- **Target-sensitive transformation layer**. To tackle the problem of target-sensitive sentiment, we use target-specific sentence representation module to convert the target representations and word representations generated in the previous layer into the sentence representation of the specific target. Then, the convolution is applied to extract emotional features in this layer for sentiment classification in the last layer.
- **Output layer**. This layer predict the sentiment type for each given aspect by a fully connected network with softmax function based on the sentence representation generated in previous layer.

A. WORD & ASPECT EMBEDDING LAYER

Noting that MSAT is a general framework, we can potentially leverage any network as the encoder to learn word-level representations. In this paper, we implement two different encoders for word embedding. One is GloVe [13], which has been widely used in numerous neural-based models for NLP tasks. The other is BERT(Bidirectional Encoder Representations from Transformers) [25], a pre-trained bidirectional transformer encoder which has achieved state-of-the-art performances across a variety of NLP tasks.

For the input sentence S, we obtain word's embedding $X = \{x_1, x_2, \cdots, x_n\} \in \mathbb{R}^{n \times dim_x}$ and its POS embedding $P = \{p_1, p_2, \cdots, p_n\} \in \mathbb{R}^{n \times dim_p}$, where dim_x and dim_p denote the dimension of the word embedding and the POS embedding

respectively. Then, we concatenate X and P to obtain the input representation $W = \{w_1, w_2, \cdots, w_n\} \in \mathbb{R}^{n \times dim_w}$ of Bi-LSTM layer, where $dim_w = dim_x + dim_p$.

B. BI-DIRECTIONAL LSTM LAYER

Combining contextual information with word embeddings is an effective way to represent a word. LSTM has achieved a great success in various sentiment analysis tasks because it can effectively capture input sentence sequence information [2]–[4]. In these works, sentiment analysis needs the context of the keywords to infer its hidden sentiment information. However, LSTM can only process sequence from one direction. To capture the forward and backward sequence information of the input sequence, we encode each word w_i by a forward LSTM to obtain the sequence information representation $\overrightarrow{h}_i$ from left to right, and employ a backward LSTM to obtain the sequence information representation $\overleftarrow{h}_i$ from right to left. Then, we concatenate the vector representations in the two opposite directions to generate the context word representation $h_i \in \mathbb{R}^{n \times 2dim_w}$:

$$\overrightarrow{h}_i = \overrightarrow{\text{LSTM}}(w_i) \tag{1}$$

$$\overleftarrow{h}_i = \overleftarrow{\text{LSTM}}(w_i) \tag{2}$$

$$h_i = \left[\overrightarrow{h}_i : \overleftarrow{h}_i\right], \quad i \in [1, n] \tag{3}$$

C. TARGET SPECIFIC SELF-ATTENTION LAYER

There are two essential issues for accurately predicting the sentiment polarity of a given aspect. The first one is how to distill the global dependence of the inputs accurately, and the second one is the target-dependent problem of sentiment expression. In the proposed model, we employ the target specific self-attention for the former and the target-sensitive transformation for the latter, respectively.

1) SELF-ATTENTION WITH ASPECT REPRESENTATION

To distill the global dependence of the input sentence, the following two essential issues need to be tackled: (1) calculating the attention of each word in the input sentence; (2) capturing sentence sequence information. We solve the first issue by the self-attention and the second one by encoding relative position described in Section III-D2.

To capture the global dependence of words in the sentence, we calculate the attention of each word in this sentence at first. The queries are the set of all words in the sentence. In the field of natural language processing, both keys and values are usually treated as the input sequences, that means queries = keys = values.

Self-attention is a more scalable and parallel attention calculation method proposed in [8]. The self-attention is regarded as a content-based query processing, which computes the attention function on a set of queries and packages them into a matrix Q. At the same time, the keys and values are also packed together into matrices K and V.

In the field of natural language processing, the self-attention mechanism is used to seek the dependency of each word with the whole sentence S. In order to obtain the correlation of the word s_i ($i \in [1, n]$) and s_j ($j \in [1, n]$) in S, we need to calculate the dot product of s_i and s_j. Then, s_i and s_j are packaged into matrix Q and K, respectively. The matrix obtained by multiplying Q and K^T represents the relationship between every word and the whole sentence. The correlation score of each word is generated by a *softmax* function. Finally, this score is multiplied by the mapping matrix V of S to obtain the words representation for capturing the global dependency information.

The way of using aspect information in Self-Attention is letting aspect embedding play a role in computing the attention weight. Inspired by [2], we append the input aspect information into each word vector:

$$\tilde{h}_i = \left(h_i + FFNN(H^t)\right), \quad i \in [1, n] \tag{4}$$

where $H^t = \{h_1, h_2, \cdots, h_m\} \in \mathbb{R}^{n \times dim_h}$ and $H = \{h_1, h_2, \cdots, h_n\} \in \mathbb{R}^{n \times dim_h}$ are the hidden representations of words and aspects through Bi-LSTM, respectively. *FFNN* is a feed forward neural network.

Then, Q, K and V are the mapped matrixes of the input sentence, which can be initialized by multiplying the input embedding and the corresponding weight matrix:

$$Q = \tilde{H} w^Q \tag{5}$$

$$K = \tilde{H} w^K \tag{6}$$

$$V = \tilde{H} w^V \tag{7}$$

where $Q, K, V \in \mathbb{R}^{n \times 2dim_w}$ are the mappings of the word representations with aspect information $\tilde{H} = \left\{\tilde{h}_1, \tilde{h}_2, \cdots, \tilde{h}_n\right\} \in \mathbb{R}^{n \times dim_h}$, w^Q,w^K and w^V are learnable parameter matrices. We compute their attention scores to obtain the sentence representation of self-attention Z:

$$Z = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{8}$$

where $\sqrt{d_k}$ is the scaling factor and it is set to the dimension of hidden representation.

2) MULTI-HEAD FOR TARGET SPECIFIC SELF-ATTENTION

As pointed out in [8], it is beneficial to linearly project the queries, keys and values h times. On each of their projected versions, we perform the attention function in parallel and concatenate the results as C:

$$C = Concat\left(Z_1, \ldots, Z_h\right) w^o \tag{9}$$

where $w^o \in \mathbb{R}^{2hdim_w \times d_{\text{model}}}$ is the learnable parameter matrix, h is the number of attention heads, d_{model} is the dimension of the self-attention mechanism's output.

Since multi-head self-attention does not contain recurrence and convolution, we inject some position information of the words in a sentence to use the sequence information. Different from the positional encoding method using the sine and cosine functions in [8], we embed the position information by encoding relative position of words in the sentence. We describe the details of this process in Section III-D2.

D. TARGET-SENSITIVE TRANSFORMATION LAYER

After capturing the structural features of the input sentence, we employ a target-sensitive transformation layer shown in the right part of Figure 1 to tackle the problem of target-sensitive sentiment expression.

The target-specific sentence representation (TSR) module is the core component of the target-sensitive transformation module, which consists of the tailor-made dynamic target representation (DTR) component and the relative position encoding, as shown in Figure 2. TSR can be extended to a multi-layer architecture and learn more abstract word-level features by the deep network. Correspondingly, we express the output of the target specific self-attention layer as $C^0 = \{c_1^0, \ldots, c_n^0\} \in \mathbb{R}^{n \times d_{\text{model}}}$ and $C^t = \{c_1^t, \ldots, c_m^t\} \in \mathbb{R}^{m \times d_{\text{model}}}$ which represents the input of target-sensitive transformation Layer.

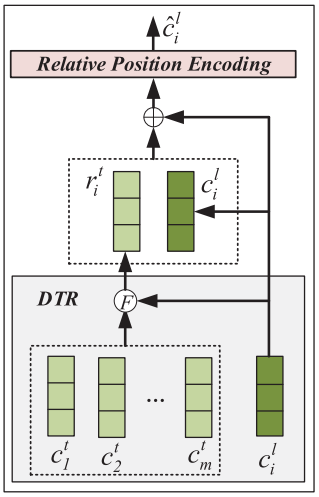

FIGURE 2. Details of the TSR module.

1) DYNAMIC TARGET REPRESENTATION

For the problem of target-sensitive sentiment, it is important to dynamically generate different representation of opinion words for different aspects. For example, in the following two sentences: "*The price is high.*" and "*The screen resolution is high.*", the opinion word *high* expresses two opposite sentiment for different aspect targets. For the aspect *price*, *high* is embedded into the negative space. However, it is embedded into the positive space for the aspect *screen resolution*.

The task of the dynamic target representation component is to generate the representation of the target. Previous methods proposed in [4] and [26] average the embeddings of the target words as the target representation. This strategy may be inappropriate in some cases because different target words usually do not contribute equally [5]. For example, for the opinion target "*amd turin processor*", the word "*processor*" is more important than "*amd*" and "*turin*", because the opinion word is usually conveyed over the phrase head, i.e.,"*processor*", but seldom over modifiers (such as brand name "*amd*"). We dynamically calculate the importance of target words based on each word of the sentence, and generate the target representation r_i^t with the target word and word x_i in the sentence:

$$r_i^t = \sum_{j=1}^{m} c_j^t * F\left(c_i^l, c_j^t\right), \quad j \in [1, m] \tag{10}$$

where F is the *softmax* function and it measures the correlation between the j-th target word representation c_t^j and the i-th word representation c_i^l.

Finally, the target representation r_i^t and the current word representation c_i^l are concatenated as the input of a fully connected layer to generate a specific target's representation $\tilde{c}_i^l$ of the i-th word:

$$\tilde{c}_i^l = g\left(\left[c_i^l : r_i^t\right]\right) \tag{11}$$

where g is a non-linear activation function and ":" denotes vector concatenation.

In order to tackle the problem of that the global context information captured by target specific self-attention layer may be lost, we sum the former features c_i^l directly with the current features $\tilde{c}_i^l$ as follows.

$$c_i^{l+1} = c_i^l + \tilde{c}_i^l, \quad i \in [1, n], \ l \in [0, L] \tag{12}$$

where c_i^l is the input of the l-th layer of TSR, $\tilde{c}_i^l$ is output of the l-th layer of DTR.

2) RELATIVE POSITION ENCODING

In a general sentiment analysis model, all words of a sentence possess the same relative position information, which is not enough for predicting the respective sentiments of a specific aspect term [22]. Empirically, the opinion words of a sentence are usually located near the target words. Proximity strategy is observed effective for this goal in [4], in which the position relevance v_i is obtained as follows.

$$v_i = \begin{cases} 1 - \dfrac{(k+m-i)}{a} & i < k+m \\ 1 - \dfrac{i-k}{a} & k+m \le i \le n \\ 0 & i > n \end{cases} \tag{13}$$

where k is the index of the first target word, a is a pre-specified constant, and m is length of the target. Then, v_i is can be integrated into the TSR module by the following equation.

$$\hat{c}_i^l = c_i^l * v_i, \quad i \in [1, n], \ l \in [1, L] \tag{14}$$

E. OUTPUT LAYER

The output of the target-sensitive transformation module is fed into the convolutional layer for extracting sentiment features and generating feature representation $c \in \mathbb{R}^{n-s+1}$:

$$c_i = ReLU\left(w_{conv}^{\top} \hat{c}_{i:i+s-1}^{L} + b_{conv}\right) \tag{15}$$

where $\hat{c}_{i:i+s-1}^{L} \in \mathbb{R}^{s \times d_{\text{model}}}$ is the concatenated vector of $\hat{c}_i^{(L)}, \ldots, \hat{c}_{i+s-1}^{(L)}$, s is the kernel size, w_{conv} and b_{conv} are the

learnable weights of the convolutional kernel. Then, we apply max pooling to obtain the biggest feature of information and get the sentence representation z by employing n_k kernels.

$$z = \left[\max\left(c_1\right), \cdots, \max\left(c_{n_k}\right)\right]^T \tag{16}$$

Finally, a fully connected layer with *softmax* function leverages the vector z to predict the sentiment polarity y:

$$y = softmax\left(w_f z + b_f\right) \tag{17}$$

where w_f and b_f are the parameters of *softmax* function.

IV. EXPERIMENTS

In this section, we evaluate the MSAT on three open datasets (e.g. *Restaurant*, *Laptop* and *Laptop*), where the first two datasets are from Semeval-2014 task [27] and last one is built by Dong *et al.* in [28]. Some statistical information on these datasets is shown in Table 1.

TABLE 1. Statistics of the datasets for ABSA tasks.

Datasets		Number of samples		
		1 (Positive)	-1 (Negative)	0 (Neutral)
Restaurant	Train	2159	800	632
	Test	727	196	196
Laptop	Train	980	858	454
	Test	339	127	169
Twitter	Train	1547	1563	3127
	Test	174	174	346

In the experiments, all word vectors are initialized by the same pre-trained GloVe vector.[1] Then, MSAT-GloVe means GloVe is used to generate the word embeddings for MSAT. To measure the effectiveness of the POS features and the multi-head target specific self-attention mechanism, we remove them from the MSAT respectively and these two simplified models are marked as MSAT-GloVe w/o POS and MSAT-GloVe w/o MSA, separately. The dimension of word vectors dim_x is 300 and that of POS dim_p is 100. Since CNN with single kernel performs better on relatively small datasets [29], we use one convolutional kernel as well as TNet [5]. Adam [30] is used as the optimizer and the initial learning rate (lr) is set to 0.001. Table 2 shows the settings of some MSAT's hyper-parameters. Since BERT[2] has achieved good effectiveness in many NLP tasks, we also implement a BERT-based model marked as MSAT-BERT and compare it with the MSAT-GloVe in our experiments. The experimental workstation houses single 2.90 GHz Intel 8-core CPU, 16 GB of RAM and an NVIDIA GeForce GTX 1660Ti graphics card. The operation system is Linux 4.15.0-107-generic. The POS analysis tool is stanford-corenlp-4.1.0.[3]

To verify the effectiveness of the proposed models, the following baseline methods are compared in our experiments:

[1]https://nlp.stanford.edu/projects/glove/
[2]https://github.com/google-research/bert
[3]https://stanfordnlp.github.io/CoreNLP/

TABLE 2. Hyper-parameters of MSAT.

Hyper-params	Value	Descriptions
dim_x	300	the dimension of the word emdedding
$batch\ size$	64	the number of data samples in one training
dim_p	100	the dimension of POS emdedding
dim_w	400	the dimension model input
dim_h	800	the dimension of the output of the Bi-LSTM
d_{model}	800	the dimension of the output of the self-attention mechanism
L	2	the depth of the CPT
a	40	a pre-specified constant
s	3	the size of convolution kernel
$dropout$	0.1	the percentage of randomly selected neurons that are ignored during training
lr	0.001	learning rate
n_k	50	the number of kernels
h	6	the number of self-attention heads
$\sqrt{d_k}$	8	scaling factor

1) **LSTM**: This is a classical model for processing text sequences, and it has achieved a great success in various NLP tasks. LSTM uses the last hidden state vector to predict sentiment polarity.
2) **ATAE-LSTM** (Wang *et al.* 2016 [2]): This is an attention-based model, which extends the input by taking the aspect word embedding as a part of the input data and training them in the RNN networks. This model combines aspect embedding with the embedding of each word to make full use of aspect term information.
3) **TD-LSTM** (Tang *et al.* 2016 [3]): This model encodes the left contexts and right contexts of the target by two LSTMs, and represents the target by averaging the hidden outputs. Finally, it concatenates the two target-specific representations for predicting the sentiment polarity of the aspect.
4) **MemNet** (Tang *et al.* 2016 [15]): It combines the multiple-hop attention, by which the model can only focus on the most informative context area, to infer the sentiment polarity towards the target word.
5) **RAM** (Chen *et al.* 2017 [4]): This model strengthens MemNet by representing memory with Bi-LSTM and using a GRU network to combine the multiple attention outputs for sentence representation.
6) **BILSTM-ATT-G** (Liu and Zhang 2017 [26]): This model adopts attention-based LSTMs and introduces gates to measure the importance of the left context, the right context and the entire sentence information for the prediction.
7) **TNet-LF**(Li *et al.* 2018 [5]): This model employs a Bi-LSTM to accumulate the context information for each word of the input sentence, and proposes target specific transformation to dynamically generate word representations.
8) **RPAEN** (Wu *et al.* 2020 [22]): This model introduces the relative position encoding to encode the relative position of the aspect term in the text, and

TABLE 3. Comparisons of accuracy and F1 on different datasets (%). The results with symbol '*' are retrieved from the original papers, Top 2 scores are in bold.

	Models	Restaurant		Laptop		Twitter	
		ACC	F1	ACC	F1	ACC	F1
Baselines	LSTM	78.21	68.33	70.85	65.09	70.81	69.08
	ATAE-LSTM	77.50	66.03	69.28	62.68	68.79	66.37
	TD-LSTM	79.29	70.25	71.64	66.49	72.69	70.48
	MenNet	79.64	69.08	71.63	66.73	71.97	70.06
	RAM	80.18	69.74	73.67	69.27	72.40	70.03
	BILSTM-ATT-G	80.38*	70.78*	74.37*	69.90*	72.70*	70.84*
	TNet-LF	80.63	70.02	75.54	71.63	73.27	71.24
	RPAEN	81.20*	72.50*	74.10*	70.05*	73.00*	70.10*
The proposed models	MSAT-GloVe w/o POS	80.09	71.28	76.48	73.63	73.55	71.46
	MSAT-GloVe w/o MSA	80.89	72.19	76.17	71.96	73.12	71.29
	MSAT-GloVe	**81.43**	**72.63**	**78.21**	**74.31**	**74.63**	**73.22**
	MSAT-BERT	**83.39**	**76.10**	**80.25**	**76.79**	**75.87**	**74.36**

utilizes attention mechanisms to model the relationship between aspect terms and all the words in the sentence.

To make the experiment more equitable, we follow the operations applying in [5] on these datasets: all words are lowercased without deleting stop words, symbols and digits, and sentences are filled with zero until the longest sentence length in the dataset. Accuracy (ACC) and Macro-Average F1 are used as kthe evaluation indicators in our experiments.

A. RESULTS AND DISCUSSION

Table 3 shows the comparisons of accuracy and F1 for different methods on above three datasets. Compared with the baseline models, we can find that the proposed model MSAT-GloVe achieves better effectiveness. Moreover, MSAT-GloVe model has been well generalized on different kinds of reviews, such as the reviews using relatively formal sentences in datasets *Restaurant* and *Laptop*, and the twitters consisting of more non-grammatical sentences. The main reason is that CNN-based feature extraction has the ability to extract accurately features from the ungrammatical sentences [5]. When we replace GloVe with Bert to produce the word embeddings in our model, the effect of MSAT has been further improved because of Bert's strong encoding ability. It is worth noting that we follow the designs in original papers for all the baseline models, which means GloVe is used for these models. Therefore, we focus on the comparisons between MSAT-GloVe and baseline models in our experiments.

Another observation is that LSTM-based methods can be effectively implemented for relatively formal sentences, such as those in dataset *Restaurant*. Consequently, the LSTM-based models such as TNet-LF, BILSTM-ATT-G and LSTM have similar effectiveness in our aspect-based sentiment tasks. These models only rely on the word sequence information captured from the LSTM, which is not enough for improving the models' effect. RPAEN is based on the CNN model, which achieves competitive results on the *Restaurant* dataset by introducing the relative position information. However, it performs poorly on *Laptop* and *Twitter*. Therefore, LSTM-based methods can handle language sequence tasks more effectively than CNN-based methods. We integrate the POS information of words into our model by inputting the POS feature vector to the Bi-LSTM layer, which can make our model more effective for the aspect-based sentiment analysis tasks.

Based on above observations and analysis, we can draw some inspirations for the task of aspect-based sentiment analysis:

1) The LSTM-based models can achieve good effectiveness by capturing more useful contextual features based on the sequence information,and the CNN-based models have some advantages for the ungrammatical texts.
2) Adding auxiliary information, such as POS features of words, can make the LSTM-based methods more effective for processing relatively formal sentences.

To investigate whether the words' POS features and multi-head target specific self-attention mechanism are effective for our ABSA tasks, we compare the MSAT model with its simplified versions MSAT-GloVe w/o POS and MSAT-GloVe w/o MSA. When the POS features are not integrated into our model, both accuracy and F1 of MSAT-GloVe w/o POS are reduced. This means that integrating of POS information into the word-level representations is helpful to improve the model's effectiveness. When the multi-head target specific self-attention mechanism is removed from MSAT, a similar situation occurs for MSAT-GloVe w/o MSA. Thus, both POS features and multi-head target specific self-attention mechanism can make a positive impact for the ABSA tasks in the proposed MSAT model.

B. CASE STUDY

Table 4 lists eight sample sentences from the testing set of the datasets. The targets in the input sentence are enclosed in square brackets and highlighted with different colors. The actual sentiment labels are given as subscripts, where 1, −1, and 0 indicate the sentiment types "positive", "negative" and "neutral", respectively. The symbol '×' indicates that the corresponding prediction is incorrect.

TABLE 4. The predictions of some samples.

Sentences	LSTM	ATAE-LSTM	TD-LSTM	MenNet	RAM	TNet-LF	MSAT-GloVe
1. *Great* [food]$_1$ *but the* [service]$_{-1}$ *is dreadful.*	1, 1 ×	1, -1	1, 1 ×	1, -1	1, -1	1, -1	1,-1
2. *The* [staff]$_{-1}$ *should be a bit more friendly.*	1 ×	1 ×	1 ×	1 ×	1 ×	1 ×	-1
3. *The* [food]$_1$ *is so good and so popular that* [waiting]$_{-1}$ *can really be a nightmare.*	1,1 ×	1,1 ×	1,1 ×	1,1 ×	1,1 ×	1, 1 ×	1,-1
4. [Startup times]$_{-1}$ *are incredibly long : over two minutes .*	1 ×	1 ×	1 ×	1 ×	1 ×	-1	-1
5. *I am pleased with the fast* [log on]$_1$ *, speedy* [WiFi connection]$_1$ *and the long* [battery life]$_1$ *-LRB- > 6 hrs -RRB- .*	1,1,1	1,1,1	1,1,1	1,1,1	1,1,1	1,1,1	1,1,1
6. *I stopped by for some* [brunch]$_0$ *today and had the vegan cranberry pancakes and some rice milk.*	1 ×	1 ×	1 ×	1 ×	1 ×	1 ×	0
7. *After dinner I heard music playing and discovered that there is a* [lounge]$_0$ *downstairs.*	1 ×	1 ×	1 ×	1 ×	1 ×	1 ×	0
8. *I know real* [Indian food]$_{-1}$ *and this was n't it.*	1 ×	1 ×	1 ×	1 ×	1 ×	1 ×	0 ×
9. [Host]$_{-1}$ *and* [hostess]$_{-1}$ *were quite rude.*	1 ×, -1	-1, -1	-1, -1	-1, -1	-1, -1	-1, -1	-1, -1

1) TARGET SPECIFIC SELF-ATTENTION MECHANISM CAN CAPTURE THE GLOBAL INTERDEPENDENCE FEATURES

In the first sentence in Table 4, two clauses containing different targets *food* and *service* are connected by the word *but*. This sentence contains obvious opinion words and they are close to the corresponding aspect targets. Then, this issue can be resolved by the attention-based model (ATAE-LSTM, MenNet, RAM) and the TNet-LF model with a convolution kernel. However, when the aspect target in a sentence is far away from the opinion word or closer to the non-target opinion word (such as the second sentence and the third sentence), none of the baseline models can make a correct prediction. On the other hand, all models except LSTM can deal with the situation correctly that multiple aspects share one opinion word in a sentence, i.e., the case of the ninth sentence, in which the word *rude* is used to modify the aspects *host* and *hostess*.

2) DYNAMIC TARGET REPRESENTATION IS HELP FOR SOLVING THE TARGET-SENSITIVE SENTIMENT PROBLEM

The opinion word *long*, which describes the aspects *startup times* and *battery life* in the fourth sentence and the fifth sentence, indicates two opposite sentiment polarities. Thus, this means the dynamic target representation of MSAT-GloVe is capable of handling such cases.

3) INTEGRATING THE POS FEATURES CAN CAPTURE THE GRAMMATICAL FEATURES OF SENTENCE

For the prediction of neutral sentiment, such as the sixth sentence and the seventh sentence, there are no specific opinion words in these sentences. All baseline models predict those as the type of positive. The reason may be that the sentences expressing neutral sentiment occupy a relatively small percentage in the datasets. On the other hand, the sentiment polarity of the target word is often determined by the opinion words with specific POS, this may be another reason. MSAT-GloVe model integrates POS features of words into the input sentence, which can improve effectively the F1 value of prediction results.

At last, we find that all above models cannot make a correct prediction for the eighth sentence. We think that is because there are no explicit opinion words in this sentence. In such cases, we need to combine the second half of this sentence to make an implicit semantic reasoning. For a sentence where an explicit opinion word modifies multiple targets, such as the last sentence, using LSTM alone fails to predict the polarity of all targets.

V. CONCLUSION

In ABSA tasks, the attention mechanism often fails to capture the global dependence of input information, which provides an opportunity to improve the effect of the models for such tasks. In this paper, we propose a MSAT model based on multi-head self-attention transformation networks, which conducts more effective sentiment analysis with target specific self-attention and target-sensitive transformation. In this model, target-sensitive transformation component is used to better transform target information into word representations, which can tackle the problem of the target-sensitive sentiment expression. Moreover, to further improve the model's effectiveness, we integrate the POS features of input words into the model to enrich the feature information. Finally, we carry out a series of experiments on three open datasets to demonstrate the effectiveness of the proposed models for aspect-based sentiment analysis. The experimental results show that our models outperform the state-of-the-art methods mentioned in Section IV significantly.

Since ABSA is a fine-grained and complex task, there are still many open problems in this field. For example, how can we make the semantic inference by combining different parts of a sentence? This difficulty does not come from the sentiment analysis of explicit opinion words but from the implicit semantic reasoning, which will be explored in our future work.

REFERENCES

[1] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining Text Data*, C. C. Aggarwal and C. Zhai, Eds. New York, NY, USA: Springer, 2012, pp. 415–463.

[2] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based LSTM for aspect-level sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 606–615.

[3] D. Tang, B. Qin, X. Feng, and T. Liu, "Effective LSTMs for target-dependent sentiment classification," in *Proc. COLING*, 2016, pp. 3298–3307.

[4] 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.
[5] X. Li, L. Bing, W. Lam, and B. Shi, "Transformation networks for target-oriented sentiment classification," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 946–956.
[6] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1412–1421.
[7] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 2048–2057.
[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[9] S. Wang, S. Mazumder, B. Liu, M. Zhou, and Y. Chang, "Target-sensitive memory networks for aspect sentiment classification," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 957–967.
[10] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent Twitter sentiment classification," in *Proc. ACL*, 2011, pp. 151–160.
[11] D. Rao and D. Ravichandran, "Semi-supervised polarity lexicon induction," in *Proc. 12th Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL)*, 2009, pp. 675–0682.
[12] V. Perez-Rosas, C. Banea, and R. Mihalcea, "Learning sentiment lexicons in Spanish," *LREC*, vol. 12, p. 73, May 2012.
[13] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[14] M. Zhang, Y. Zhang, and D. T. Vo, "Gated neural networks for targeted sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 3087–3093.
[15] D. Tang, B. Qin, and T. Liu, "Aspect level sentiment classification with deep memory network," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 214–224.
[16] 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.*, Aug. 2017, pp. 4068–4074.
[17] S. Gu, L. Zhang, Y. Hou, and Y. Song, "A position-aware bidirectional attention network for aspect-level sentiment analysis," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 774–784.
[18] F. Fan, Y. Feng, and D. Zhao, "Multi-grained attention network for aspect-level sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 3422–3433.
[19] Q. Zhang and R. Lu, "A multi-attention network for aspect-level sentiment analysis," *Future Internet*, vol. 11, no. 7, p. 157, Jul. 2019.
[20] Q. Xu, L. Zhu, T. Dai, and C. Yan, "Aspect-based sentiment classification with multi-attention network," *Neurocomputing*, vol. 388, pp. 135–143, May 2020.
[21] W. Xue and T. Li, "Aspect based sentiment analysis with gated convolutional networks," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 2514–2523.
[22] C. Wu *et al.*, "A relative position attention network for aspect-based sentiment analysis," *Knowl. Inf. Syst.*, 2020, doi: 10.1007/s10115-020-01512-w.
[23] B. Zhang, X. Li, X. Xu, K.-C. Leung, Z. Chen, and Y. Ye, "Knowledge guided capsule attention network for aspect-based sentiment analysis," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2538–2551, 2020.
[24] C. Zhang, Q. Li, and D. Song, "Aspect-based sentiment classification with aspect-specific graph convolutional networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 4560–4570.
[25] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. ACL*, 2019, pp. 4171–4186.
[26] J. Liu and Y. Zhang, "Attention modeling for targeted sentiment," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, vol. 2, 2017, pp. 572–577.
[27] 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.
[28] 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*, vol. 2, 2014, pp. 49–54.
[29] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," in *Proc. IJCNLP*, 2015, pp. 253–263.
[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014, p. 41.

YUMING LIN received the B.S. and M.S. degrees in computer science and technology from the Guilin University of Electronic Technology, Guilin, China, and the Ph.D. degree in computer application from East China Normal University, Shanghai, China. He is currently a Professor of Computer Science with the Guilin University of Electronic Technology and a Visiting Scholar with East China Normal University. His current research interests include opinion mining, knowledge graph, and massive data management.

CHAOQIANG WANG is currently pursuing the master's degree with the School of Computer and Information Security, Guilin University of Electronic Technology. His research interest includes opinion mining.

HAO SONG is currently pursuing the master's degree with the School of Computer and Information Security, Guilin University of Electronic Technology. His research interest includes massive data management.

YOU LI received the B.S. degree in computer software from the Guilin University of Electronic Technology, China, and the M.S. degree in computer science from the Dalian University of Technology, Dalian, China. She is currently an Associate Professor of Computer Science with the Guilin University of Electronic Technology. Her current research interests include natural language processing and machine learning.