Incorporating Dynamic Semantics into Pre-Trained Language Model for Aspect-based Sentiment Analysis

Kai Zhang[1], Kun Zhang[2], Mengdi Zhang[3], Hongke Zhao[4], Qi Liu[1,*]
Wei Wu[3], Enhong Chen[1]
[1] School of Data Science, University of Science and Technology of China
[2] School of Computer Science and Information Engineering, Hefei University of Technology
[3] Meituan; [4] College of Management and Economics, Tianjin University
kkzhang0808@mail.ustc.edu.cn; {qiliuql,cheneh}@ustc.edu.cn
{zhang1028kun,wuwei19850318,mdz}@gmail.com; hongke@tju.edu.cn

Abstract

Aspect-based sentiment analysis (ABSA) predicts sentiment polarity towards a specific aspect in the given sentence. While pre-trained language models such as BERT have achieved great success, incorporating dynamic semantic changes into ABSA remains challenging. To this end, in this paper, we propose to address this problem by Dynamic Re-weighting BERT (**DR-BERT**), a novel method designed to learn dynamic aspect-oriented semantics for ABSA. Specifically, we first take the Stack-BERT layers as a primary encoder to grasp the overall semantic of the sentence and then fine-tune it by incorporating a lightweight Dynamic Re-weighting Adapter (DRA). Note that the DRA can pay close attention to a small region of the sentences at each step and re-weigh the vitally important words for better aspect-aware sentiment understanding. Finally, experimental results on three benchmark datasets demonstrate the effectiveness and the rationality of our proposed model and provide good interpretable insights for future semantic modeling.

1 Introduction

Aspect-based sentiment analysis is a branch of sentiment analysis, which aims to identify sentiment polarity of the specific aspect in a sentence (Jiang et al., 2011). For example, given a sentence *"The restaurant has attentive service, but the food is terrible."*, the task aims to predict the sentiment polarities towards *"service"* and *"food"*, which should be positive and negative respectively.

As a fundamental technology, the ABSA task has broad applications, such as recommender system (Chin et al., 2018; Zhang et al., 2021b) and question answering (Wang et al., 2019). Therefore, a great amount of research has been attracted from both academia and industry. Among them, deep neural networks (DNN) (Nguyen and Shirai, 2015;

Tang et al., 2015, 2016; Zheng et al., 2020), attention mechanism (Wang et al., 2016; Ma et al., 2017) and graph neural/attention networks (Huang and Carley, 2019; Zhang et al., 2019a; Wang et al., 2020) have significantly improved the performance through deep feature alignment between the aspect representations and context representations.

Recently, the large-scaled pre-trained language models, such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), realize a breakthrough for improving many language tasks, which further attracts considerable attention to enhance the semantic representations. In ABSA, Xu et al. (2019a) designed BERT-PT, which explores a novel post-training approach on the BERT model. Song et al. (2019) further proposed a text pair classification model BERT-SPC, which prepares the input sequence by appending the aspects into the contextual sentence. Although great success has been achieved by the above studies, some critical problems remain when directly applying attention mechanisms or fine-tuning the pre-trained BERT in the task of ABSA.

Specifically, most of the existing approaches select all the important words from a contextual sentence at one time. However, according to neuroscience studies, the essential words during semantic comprehension are dynamically changing with the reading process and should be repeatedly considered (Kuperberg, 2007; Tononi, 2008; Brouwer et al., 2021). For example, when judging the sentiment polarity of the aspect *"system memory"* in a review sentence *"It could be a perfect laptop if it would have faster system memory and its radeon would have DDR5 instead of DDR3"*, the important words should change from general sentiment words {*"faster", "perfect", "laptop"*} into aspect-aware words {*"would have", "faster", "could", "be", "perfect"*}. Through these dynamic changes, the sentiment polarity will change from positive to the ground truth sentiment label negative.

---

* Corresponding author.

Meanwhile, simply initializing the encoder with a pre-trained BERT does not effectively boost the performance in ABSA as we expected (Huang and Carley, 2019; Xu et al., 2019a; Wang et al., 2020). One possible reason could be that training on two specific tasks, i.e., Next Sentence Prediction and Masked LM, with rich resources leads to better semantic of the overall sentences. However, the ABSA task is conditional, which means the model needs to understand the regional semantics of sentences by fully considering the given aspect. For instance, BERT tends to understand the global sentiment of the above sentence *"It could be a perfect laptop ... of DDR3"* regardless of which aspect is given. But in ABSA, the sentence is more likely to be different sentiment meanings for different aspects (e.g., negative for *"system memory"* while positive for *"DDR5"*). Therefore, the vanilla BERT is hardly to pay closer attention to relevant information for the specific aspect, especially when there are multiple aspects in one sentence.

To equip the pre-trained models with the ability to capture the aspect-aware dynamic semantics, we present a Dynamic Re-weighting BERT (**DR-BERT**) model, which considers the aspect-aware dynamic semantics in a pre-trained learning framework. Specifically, we first take the Stack-BERT layers as primary sentence encoder to learn overall semantics of the whole sentences. Then, we devise a Dynamic Re-weighting Adapter (DRA), which aims to pay most careful attention to a small region of the contextual sentence and dynamically select and re-weight one critical word at each step for better aspect-aware sentiment understanding. Finally, to overcome the limitation of vanilla BERT mentioned above, we incorporate the light-weighted DRA into each BERT encoder layer and fine-tune it to adapt to the ABSA task. We conduct extensive experiments on three widely-used datasets where the results demonstrate the effectiveness, rationality and interpretability of the proposed model.

2 Related Work

2.1 Aspect-based Sentiment Analysis

Aspect-based sentiment analysis identifies specific aspect's sentiment polarity in the sentence. Some approaches (Ding and Liu, 2007; Jiang et al., 2011; Kiritchenko et al., 2014) designed numerous rules-based models for ABSA. For example, Ding and Liu (2007) first performed dependency parsing to determine sentiment polarity about the aspects.

In recent years, most research studies make use of the attention mechanism to learn the word's semantic relation (Tang et al., 2015, 2016; Wang et al., 2016; Ma et al., 2017; Xing et al., 2019; Liang et al., 2019; Zhang et al., 2021a). Among them, Wang et al. (2016) proposed an attention-based LSTM to identify important information relating to the aspect. Ma et al. (2017) developed an interactive attention to model the aspect and sentence interactively. Fan et al. (2018) defined a multi-grained network to link the words from aspect and sentence. Li et al. (2018) designed a target-specific network to integrate aspect information into sentence. Tan et al. (2019) introduced a dual attention to distinguish conflicting opinions.

In addition, another research trend is to leverage syntactic knowledge to learn syntax-aware features of the aspect (Tang et al., 2019; Huang and Carley, 2019; Zhang et al., 2019a; Sun et al., 2019; Wang et al., 2020; Tang et al., 2020; Chen et al., 2020; Li et al., 2021; Tian et al., 2021). For example, Tang et al. (2020) developed dependency graph enhanced dual-transformer network to fuse the flat representations. More recently, pre-trained methods have been proved remarkably successful in the ABSA task. Song et al. (2019) devised an attentional encoder and a BERT-SPC model to learn features between aspect and context. Wang et al. (2020) reshaped the dependency trees and proposed a relational graph attention network to encode the syntax relation feature. Tian et al. (2021) explicitly utilize dependency types with a type-aware graph networks to learn aspect-aware relations.

However, these methods largely ignore the procedure of dynamic semantic comprehension (Kuperberg, 2007; Kuperberg and Jaeger, 2016; Wang et al., 2017; Zhang et al., 2019c; Brouwer et al., 2021) and can not fully reveal dynamic semantic changes of the aspect-related words. Thus, it's hard for ABSA models to achieve the same performance as human-level sentiment understanding.

2.2 Human Semantic Comprehension

Actually, no matter in the early days or now, imitating the procedure of human semantic comprehension has always been one of the original intention of many studies (Bezdek, 1992; Wang et al., 2017; Zheng et al., 2019; Li et al., 2019; Zhang et al., 2019d; Peng et al., 2020; Golan et al., 2020), such as machine reading comprehension (Zhang et al., 2019d; Peng et al., 2020), visual object detecting (Spampinato et al., 2017) and relevance estima-
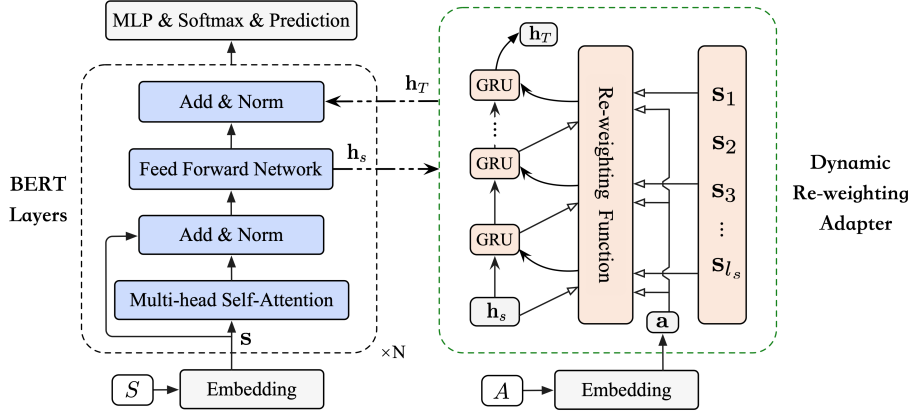
Figure 1: An illustration of the proposed framework. The blue blocks constitute a pre-trained BERT model which are frozen during fine-tuning, and the right block represents the dynamic re-weighting adapter that is inserted after each BERT encoder layer and trained during fine-tuning. Moreover, S and A represent the sentence sequence and the aspect sequence respectively. N indicates the number of layers of the BERT encoder.

tion (Li et al., 2019). For example, attention mechanism (Vaswani et al., 2017) has a widespread influence, which allows the model to focus on important parts of the input as human's attention. Spampinato et al. (2017) aimed to learn human–based features via brain-based visual object. Wang et al. (2017) built a dynamic attention model to model human preferences for article recommendation.

Moreover, some psychologists and psycholinguists have also done many research on the mechanisms of human semantic comprehension (Kuperberg, 2007; Kuperberg and Jaeger, 2016; Brouwer et al., 2021). Specifically, some scholars (Yang and McConkie, 1999; Rayner, 1998) found that most people may focus on 1.5 words. Moreover, Koch and Tsuchiya (2007) and Tononi (2008) assumed that people can only remember the meaning of about 7 to 9 words at each time. The phenomenons indicate that most people only focused on a small region of the sentence at one time and need to repeatedly process important parts for better semantic understanding (Sharmin et al., 2015).

Inspired by the above research and linguistic psychology theories, in this paper, we explore aspect-aware semantic changes of the ABSA task by incorporating the procedure of dynamic semantic comprehension into the pre-trained language model.

3   Dynamic Re-weighting BERT

In this section, we introduce the technical detail of DR-BERT. Specifically, we start with the problem definition, followed by an overall architecture of DR-BERT as illustrated in Figure 1.

Problem Definition   In ABSA, a sentence-aspect pair (S, A) is given. In this paper, the sentence is represented as $S = \{w_1^s, w_2^s, ..., w_{l_s}^s\}$ which consists of a series of l_s words. The specific aspect is denoted as $A = \{w_1^a, w_2^a, ..., w_{l_a}^a\}$ which is a part of S. l_a is the length of aspect words. The goal of ABSA is to learn a sentiment classifier that can precisely predict the sentiment polarity of sentence S for specific aspect A. As the aspect-related information plays a key role in the prediction (Li et al., 2018; Zheng et al., 2020), this paper aims to dynamically select and encode the aspect-aware semantic information through the proposed model.

Overall Architecture   DR-BERT mainly contains two components (i.e., BERT encoder and Dynamic Re-weighting Adapter), together with two modules (i.e., the embedding module and sentiment prediction module). The technical details of each part will be elaborated on as follows.

3.1   Embedding Module

To represent semantic information of the aspect words and context words better, we first map each word into a low-dimensional vector. Specifically, the inputs of DR-BERT are the sentence sequence and the corresponding aspect sequence. For the sentence sequence, we construct the BERT input as "[CLS]" + sentence +"[SEP]" and the sentence $S = \{w_1^s, w_2^s, ..., w_{l_s}\}$ can be transformed into the hidden states $\mathbf{s} = \{\mathbf{s}_i \mid i = 1, 2, \ldots, l_s\}$ with BERT embedding. For aspect sequences, we adopt the same method to get the representation vector of each word. Thus, through the embedding module, the aspect sequence $A = \{w_1^a, w_2^a, ..., w_{l_a}^a\}$ is mapped to $\mathbf{a}^s = \{\mathbf{a}_j \mid j = 1, 2, \ldots, l_a\}$. Note that, if the aspect sequence is a single word like *"food"*, the aspect representation is the embedding of the

single word *"food"*. While for the cases where the sequence contains multiple words such as *"system memory"*, the aspect representation is the average of each word embedding (Sun et al., 2015). We can denote the aspect embedding process as:

$$\mathbf{a} = \begin{cases} \mathbf{a}_1, & \text{if } l_a = 1, \\ \\ (\sum_{j=1}^{l_a} \mathbf{a}_j)/l_a, & \text{if } l_a > 1, \end{cases} \quad (1)$$

where $\mathbf{a}_j$ is the embedding of word j in the aspect sequence. $\mathbf{a}$ denotes the embedding of the aspect.

3.2 BERT Encoder

The architecture of BERT (Devlin et al., 2019) is akin to the Transformer (Vaswani et al., 2017). For simplicity, we omit some architecture details such as position encoding, layer normalization (Xu et al., 2019b) and residual connections (He et al., 2016).

1) Multi-head Self-attention Mechanism. In recent years, the multi-head self-attention mechanism (MultiHead) has received a wide range of applications in natural language processing. In the paper, we adopt MultiHead with h heads to obtain the overall semantics of the whole sentence. The product from each self-attention network is then concatenated and finally transformed as:

$$\begin{aligned} \mathbf{m} &= \{\mathbf{m}_i \mid i = 1, 2, \ldots, l_s\} \\ &= \mathbf{MultiHead}(\mathbf{s}\mathbf{W}_h^Q, \mathbf{s}\mathbf{W}_h^K, \mathbf{s}\mathbf{W}_h^V), \end{aligned} \quad (2)$$

where h denotes the h-th attention head, $\mathbf{W}_i^Q$, $\mathbf{W}_i^K$ and $\mathbf{W}_i^V$ are learnable parameters. Finally, the output feature is $\mathbf{m} = \{\mathbf{m}_i \mid i = 1, 2, \ldots, l_s\}$. For detailed implementation of **MultiHead**, please refer to Transformer (Vaswani et al., 2017).

2) Position-wise Feed-Forward Network. Since the multi-head attention is a series of linear transformations, we then apply the position-wise feed-forward network (FFN) to learn the feature's nonlinear transformation. Specifically, the FFN consists of two linear transformations along with a ReLU activation in between. More formally:

$$\begin{aligned} \mathbf{f} &= \{\mathbf{f}_i \mid i = 1, 2, \ldots, l_s\} \\ &= \mathbf{max}(0, \mathbf{m}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2, \end{aligned} \quad (3)$$

where $\mathbf{W}_1$, $\mathbf{b}_1$, $\mathbf{W}_2$ and $\mathbf{b}_2$ are learnable parameters in the linear transformations.

So far, with the input $S = \{w_1^s, w_2^s, ..., w_{l_s}^s\}$, we obtain the hidden states $\mathbf{f} = \{\mathbf{f}_i \mid i = 1, 2, \ldots, l_s\}$ via the BERT encoder. Then, for the words' hidden states of the sentence from FFN, we utilize the max-pooling operation to fairly select crucial features in the sentence (Lai et al., 2015; Zhang et al., 2019b), so as to obtain the original sentence representation $\mathbf{h}_s$ at the beginning of each re-weighting step:

$$\mathbf{h}_s = \text{Max\_Pooling}(\mathbf{f}_i \mid i = 1, 2, \ldots, l_s). \quad (4)$$

3.3 Dynamic Re-weighting Adapter (DRA)

The currently attention mechanism in deep learning is essentially similar to the selective visual attention of human beings (Vaswani et al., 2017; You et al., 2016). However, as for the text semantic understanding, human brain will discover the intentional relationship of words at a sentential level (Taatgen et al., 2007; Sha et al., 2016; Sen et al., 2020) and link the incoming semantic information with pre-existing information stored within memory. Thus, we design a dynamic re-weighting adapter (DRA) which can dynamically emphasize the important aspect-aware words for the ABSA task.

As shown in the right part of Figure 1, based on overall semantics of the whole sentence, DRA further selects the most important word at each step with consideration of the specific aspect representation. Specifically, the inputs of DRA are the final outputs of the BERT encoder (i.e., $\mathbf{h}_s$) and the original aspect embedding (i.e., $\mathbf{a}$). In each step, we first utilize re-weighting attention to choose the word for current input from the input sequence ($\{\mathbf{s}_i \mid i = 1, 2, \ldots, l_s\}$). Then, we utilize Gated Recurrent Unit (GRU)(Cho et al., 2014) to encode the chosen word and update the semantic representation of the review sentence.

Formally, we regard the calculation process as:

$$\begin{aligned} \mathbf{a}_t &= F\left([\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{l_s}], \mathbf{h}_{t-1}, \mathbf{a}\right), \\ \mathbf{h}_t &= GRU\left(\mathbf{a}_t, \mathbf{h}_{t-1}\right), \quad t \in [1, T] \end{aligned} \quad (5)$$

where $\mathbf{a}$ is the original embedding vector of the aspect words. $\mathbf{a}_t$ is the output of re-weighting function F. T denotes the dynamic re-weighting length over the sentences, which represents the cognitive threshold of human beings. $\mathbf{h}_0 = \mathbf{h}_s$ is the initial state and $\mathbf{h}_T$ is the output hidden states of DRA.

1) The Re-weighting Function. More specifically, we utilize the attention mechanism to achieve the re-weighting function F, which aims to select the most important aspect-related word at each step. The calculation can be formulated as:

$$\begin{aligned} \mathbf{S} &= [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{l_s}], \\ \mathbf{M} &= \mathbf{W}_s\mathbf{S} + (\mathbf{W}_d\mathbf{h}_{t-1} + \mathbf{W}_a\mathbf{a}) \otimes \mathbf{w}, \\ \mathbf{m} &= \omega^T \tanh(\mathbf{M}), \end{aligned} \quad (6)$$

where $\mathbf{S}$ denotes the original sentence embedding, $\mathbf{M}$ is the fusion representation of the aspects and the sentences. $\mathbf{W}_s$, $\mathbf{W}_d$, $\mathbf{W}_a$ and ω are trainable parameters. $\mathbf{w} \in \mathbb{R}^{l_s}$ is a row vector of 1 and $\otimes$ denotes the outer product.

Subsequently, to better encode aspect-aware semantics, we choose the most important word (i.e., one word) at each step for the specific aspect.

$$\alpha_i = \frac{\exp{(m_i)}}{\sum_{k=1}^{l_s} \exp{(m_k)}} \, , \tag{7}$$
$$\mathbf{a}_t = \mathbf{s}_j, (j = \text{Index}(\max(\alpha_i)))$$

where m_i and α_i are the hidden state and the attention score of i-th word in the sentence. $\mathbf{a}_t$ is the chosen word which is most related to the specific aspect at t-th step. However, Index($\max(\cdot)$) operation has no derivative, which means its gradient could not be calculated. Inspired by softmax function, we modify the Eq.7 and employ the following operation to re-weight the contextual words:

$$\mathbf{a}_t = \sum_{i=1}^{l_s} \frac{\exp{(\lambda m_i)}}{\sum_{k=1}^{l_s} \exp{(\lambda m_k)}} \mathbf{s}_i \, . \tag{8}$$

Note that, we design a hyper-parameter λ to ensure our model achieves the above purpose. Specifically, the softmax function can exponentially increase or decrease the signal, thereby highlighting the information we want to enhance. Thus, when λ is an arbitrarily large value, the attention score of the chosen word is infinitely close to 1, and other words are infinitely close to 0. In this way, the most important word (i.e., one word) will be extract from the context at each re-weighting step.

2) The GRU Function. To better encode semantic of the whole sentence, we also employ GRU to further imitate the procedure of human semantic comprehension under the specific context, which is consistent with the process of people adjusting to a new text based on their understanding behavior. Therefore, given a previous vector embedding, the hidden vectors of GRU are calculated by receiving it as input:

$$\begin{aligned} z_t &= \sigma \left(\mathbf{W}_z \cdot [\mathbf{h}_{t-1}, \mathbf{a}_t] \right) \\ r_t &= \sigma \left(\mathbf{W}_r \cdot [\mathbf{h}_{t-1}, \mathbf{a}_t] \right) \\ \tilde{\mathbf{h}}_t &= \tanh \left(\mathbf{W} \cdot [r_t * \mathbf{h}_{t-1}, \mathbf{a}_t] \right) \\ \mathbf{h}_t &= (1 - z_t) * \mathbf{h}_{t-1} + z_t * \tilde{\mathbf{h}}_t \, , \end{aligned} \tag{9}$$

where σ is the logistic sigmoid function. z_t and r_t denote the update gate and reset gate respectively at the time step t.

Datasets	#Positive		#Negative		#Neural		#L	#M
	Train	Test	Train	Test	Train	Test		
Restaurant	2164	728	807	196	637	196	20	45.5
Laptop	994	341	870	128	464	169	19	36.5
Twitter	1561	173	1560	173	3127	346	16	10.2

Table 1: The statistics of three benchmark datasets. #L is the average length of sentences. #M is the proportion (%) of samples with multiple (i.e., more than 1) aspects.

3.4 Sentiment Predicting

After applying BERT layers and DRA on the input sentence, its root representation (i.e., $\mathbf{s}$) is convert into the feature representation $\mathbf{e}$:

$$\begin{aligned} \mathbf{e} &= \{\mathbf{e}_i \mid i = 1, 2, \ldots, l_s\} \\ &= (\mathbf{W}_e \mathbf{f} + \mathbf{U}_e \mathbf{h}_T + \mathbf{b}_e) \, , \end{aligned} \tag{10}$$

where $\mathbf{W}_e$, $\mathbf{U}_e$ and $\mathbf{b}_e$ are trainable parameters. After N-th stacked BERT layers, we obtain the final representation of the sentence (i.e., $\mathbf{e}_N$). Then, we feed it into a Multilayer Perceptron (MLP) and map it to the probabilities over the different sentiment polarities via a softmax layer:

$$\begin{aligned} \mathbf{R}_l &= \text{Relu}(\mathbf{W}_l \mathbf{R}_{l-1} + \mathbf{b}_l) \, , \\ \hat{\mathbf{y}} &= \text{softmax} \left(\mathbf{W}_o \mathbf{R}_h + \mathbf{b}_o \right) , \end{aligned} \tag{11}$$

where $\mathbf{W}_l$, $\mathbf{W}_o$, $\mathbf{b}_l$ and $\mathbf{b}_o$ are learned parameters. $\mathbf{R}_l$ is the hidden state of l-th layer MLP ($\mathbf{R}_0 = \mathbf{e}_N$, $l \in [1, h]$). $\mathbf{R}_h$ is the state of final layer which is also regard as the output of the MLP. $\hat{\mathbf{y}}$ is the predicted sentiment polarity distribution.

3.5 Model Training

Finally, we applies the cross-entropy loss function for model training:

$$\mathcal{L} = -\sum_{i=1}^{M} \sum_{j=1}^{C} y_i^j \log \left(\hat{y}_i^j \right) + \beta \|\Theta\|_2^2 \, , \tag{12}$$

where y_i^j is the ground truth sentiment polarity. C is the number of labels (i.e, 3 in our task). M is the number of training samples. Θ corresponds to all of the trainable parameters.

4 Experiment

4.1 Datasets

We mainly conduct experiments on three benchmark ABSA datasets, including "Laptop", "Restaurant" (Pontiki et al., 2014) and "Twitter" (Dong et al., 2014). Each data item is labeled with three

Category	Methods / Datasets	Laptop		Restaurant		Twitter	
		Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Attention.	ATAE-LSTM (Wang et al., 2016)	68.57	64.52	76.58	67.39	67.27	66.43
	IAN (Ma et al., 2017)	70.84	65.73	76.88	68.36	68.74	67.61
	MemNet (Tang et al., 2016)	72.32	67.03	78.12	68.99	70.19	68.22
	AOA (Huang et al., 2018)	74.56	68.77	79.42	70.43	71.68	69.25
	MGNet (Fan et al., 2018)	75.37	71.26	81.28	72.07	72.54	70.78
	TNet (Li et al., 2018)	76.54	71.75	80.69	71.27	74.93	73.60
Pre-trained.	BERT (Devlin et al., 2019)	77.29	73.36	82.40	73.17	73.42	72.17
	BERT-PT (Xu et al., 2019a)	78.07	75.08	84.95	76.96	–	–
	BERT-SPC (Song et al., 2019)	78.99	75.03	84.46	76.98	74.13	72.73
	AEN-BERT (Song et al., 2019)	79.93	76.31	83.12	73.76	74.71	73.13
	RGAT-BERT (Wang et al., 2020)	78.21	74.07	<u>86.60</u>	<u>81.35</u>	76.15	74.88
	T-GCN (Tian et al., 2021)	<u>80.88</u>	<u>77.03</u>	86.16	79.95	<u>76.45</u>	<u>75.25</u>
Ours.	**DR-BERT**	**81.45**	**78.16**	**87.72**	**82.31**	**77.24**	**76.10**

Table 2: Experimental results (%) in three benchmark datasets. We underline the best performed baseline.

sentiment polarities (i.e., positive, negative and neutral). The statistics of the datasets are presented in Table 1. Moreover, we follow the dataset configurations of previous studies strictly. For all datasets, we randomly sample 10% items from the training set and regard them as the development set.

4.2 Hyperparameters Settings

In the implementation, we build our framework based on the official bert-base models (n_{layers}=12, n_{heads}=12, n_{hidden}=768). The hidden size of GRUs and re-weighting length of DRA are set to 256 and 7. The learning rate is tuned amongst [2e-5, 5e-5 and 1e-3] and the batch size is manually tested in [16, 32, 64, 128]. The dropout rate is set to 0.2. The hyper-parameter l, β and λ have been carefully adjusted, and final values are set to 3, 0.8 and 100 respectively. The model is trained using the Adam optimizer and evaluated by two widely used metrics. The parameters of baseline models are in accordance with the default configuration of the original paper. We run our model three times with different seeds and report the average performance.

4.3 Baselines

- **Attention-based Models:** MemNet (Tang et al., 2016), ATAE-LSTM (Wang et al., 2016), IAN (Ma et al., 2017), AOA (Huang et al., 2018), MGNet (Fan et al., 2018), TNet (Li et al., 2018).
- **Pre-trained Models:** Fine-tune BERT (Devlin et al., 2019), BERT-PT (Xu et al., 2019a), BERT-SPC, AEN-BERT (Song et al., 2019), RGAT-BERT (Wang et al., 2020), T-GCN (Tian et al., 2021).

The baseline methods have comprehensive coverage of the recent related SOTA models recently. Most of them are detailed in Section 2.1. For space-saving, we do not detail them in this section.

4.4 Experimental Results

From the results in Table 2, we have the following observations. First, BERT-based methods beat most of the attention-based methods (e.g., IAN and TNet) in both metrics. The phenomenon indicates the powerful ability of the pre-trained language models. That is also why we adopt BERT as base encoder to learn the overall semantic representation of the whole sentences.

Second, by comparing non-specific BERT models (i.e., BERT and BERT-PT) with task-specific models (e.g., RGAT-BERT) for ABSA, we find that the task-specific BERT models perform better than the non-specific models. Specifically, we can also observe the performance trend that T-GCN&RGAT-BERT >AEN-BERT>BERT-PT>BERT, which is consistent with the previous assumption that aspect-related information is the crucial influence factor for the performance of the ABSA model.

Finally, despite the outstanding performance of previous models, our DR-BERT still outperforms the most advanced baseline (i.e., T-GCN or RGAT-BERT) no matter in terms of Accuracy or F1-score. The results demonstrate the effectiveness of the dynamic modeling strategy based on the procedure of semantic comprehension. Meantime, it also indicates that our proposed DRA can better grasp the aspect-aware semantics of the sentence than other BERT plus-in components in previous methods.

Model Variants	Laptop	
	Accuracy	F1-score
BERT-Base	77.29	73.36
(1): + MLP	77.94	74.42
(2): + DRA	80.66	77.13
(3): + DRA on top 3 layers	78.64	75.16
(4): + DRA on top 6 layers	79.17	75.93
(5): + DRA on top 9 layers	80.22	76.49
(6): DR-BERT	**81.45**	**78.16**

Table 3: The ablation study on different components which conducted on the test set of the Laptop dataset. "BERT-Base" indicates the vanilla BERT. "+" indicates the setting with plus-in components.

(a) The performance (Accuracy) in three dataset.

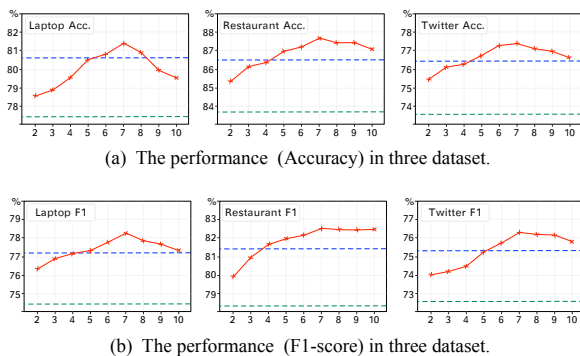(b) The performance (F1-score) in three dataset.

Figure 2: The ablation study on the re-weighting length of the adapter. Red lines indicate Accuracy/ F1 scores while blue and green lines indicate the performance of the best baseline and BERT-base model respectively.

4.5 Ablation Study

Ablations on the Proposed Components. In Table 3, we study the influence of different components in our framework, including the DRA and MLPs. We can find that without utilizing adapters and MLPs, DR-BERT degenerates into the BERT model, which gains the worst performance among all the variants. The phenomenon indicates the effective of the DRA and MLP modules. Moreover, through comparing (1) and (2), we can easily conclude that DRA plays a more crucial role in the final sentiment prediction than MLPs.

Since BERT models are usually quite deep (e.g., 12 layers), we only insert the dynamic re-weighting adapter into top layers (i.e., 3-th, 6-th, and 9-th layers) to further verify the effectiveness of the DRA module. The results are shown in Table 3 (3), (4), and (5). We observe that when introducing adapters to the top layers of DR-BERT, our framework still outperforms the BERT model, showing that the DRA is efficient in encoding the aspect-aware semantics over the whole sentence. In addition, we can also find that the more adapter incorporated
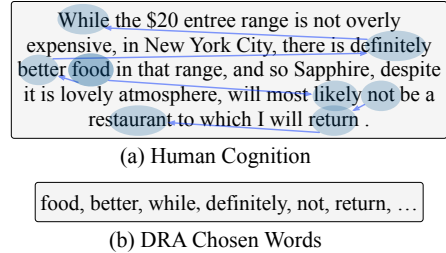

(a) Human Cognition

food, better, while, definitely, not, return, …

(b) DRA Chosen Words

Figure 3: Comparison of the semantic understanding process between human reading and DRA when judging the sentiment polarity of aspect *"food"*. (a) is the visualization of the human understanding process from the eye tracker[†]. (b) denotes aspect-aware words from re-weighting function.

in BERT layers the higher performance gained, illustrating the importance of modeling the deep dynamic semantics over the sentence.

Ablations on the Scale of Adapter. In this subsection, we investigate the influence of the scale of adapters on different datasets. As shown in Figure 2, we tune the adapter's dynamic re-weighting length (T) in a wide range (i.e., 2 to 10). Specifically, the performance of DR-BERT first becomes better with the increasing of re-weighting length and achieving the best result at around 7. Then, as the length continues to increase, the performance continues to decline. This phenomenon is consistent with the psychological findings that human memory focuses on nearly seven words (Tononi, 2008; Koch and Tsuchiya, 2007), which further indicates the effectiveness of DRA in modeling human-like (dynamic) semantic comprehension.

Besides, compared with the best-performed baseline (blue lines), our model can achieve better performance with only 4 or 5 times of re-weighting at most test sets, illustrating the efficiency of the re-weighting adapter. On the other hand, we can also find that DR-BERT always gives superior performance compared to the BERT-based model (green lines), even with the lowest re-weighting length. All those results show that DR-BERT could better comprehend aspect-aware dynamic semantics in aspect-based sentiment analysis.

4.6 Interpretability Verification

Comparison of Semantic Comprehension. To evaluate model rationality and interpretability, we conduct an study for dynamic semantic comprehension by eye tracker. As shown in Figure 3 (a),

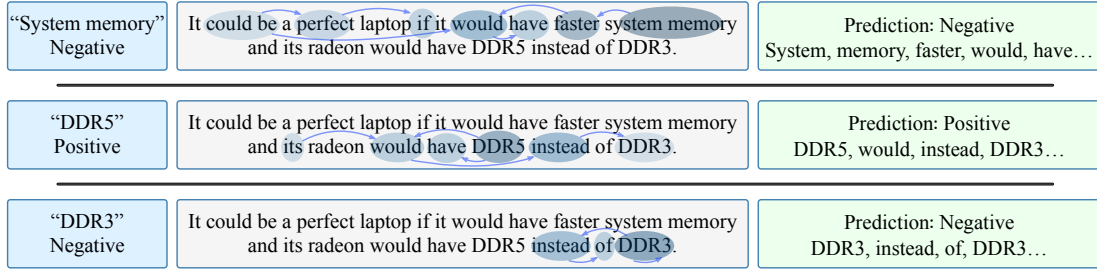[†]The procedure of the human semantic comprehension is generated by the eye tracker: https://www.tobiipro.com/product-listing/nano/

"System memory" Negative	It could be a perfect laptop if it would have faster system memory and its radeon would have DDR5 instead of DDR3.	Prediction: Negative System, memory, faster, would, have…
"DDR5" Positive	It could be a perfect laptop if it would have faster system memory and its radeon would have DDR5 instead of DDR3.	Prediction: Positive DDR5, would, instead, DDR3…
"DDR3" Negative	It could be a perfect laptop if it would have faster system memory and its radeon would have DDR5 instead of DDR3.	Prediction: Negative DDR3, instead, of, DDR3…

Figure 4: Visualization results of multiple aspects in the same sentence. The blue part indicates the aspect and its ground truth. The middle subfigures represent the procedure of human's semantic comprehension which is targeted at one specific aspect. The green subfigures are the predicted labels and the chosen word sequences from DRA.

Case Examples. The label in brackets represents ground truth.	BERT-base	RGAT-BERT	DR-BERT
Aspects: "system memory"(Neg.), "DDR5"(Pos.), "DDR3"(Neg.) **Sentence**: It could be a perfect laptop if it would have faster system memory and its radeon would have DDR5 instead of DDR3.	Pos/Neg/Neg ✗ / ✗ / ✗	Neg/Pos/Pos ✓ / ✓ / ✗	Neg/Pos/Neg ✓ / ✓ / ✓
Aspects: "Supplied software" (Neu.), "software" (Pos.), "Windows" (Neg.) **Sentence**: Supplied software: The software that comes with this machine is greatly welcomed compared to what Windows comes with.	Pos/ Pos/ Pos ✗ / ✓ / ✗	Pos/Pos/Neu ✗ / ✓ / ✗	Pos/Pos/Neg ✗ / ✓ / ✓
Aspects: "waiter" (Neg.), "served" (Neg.), "specials" (Pos.) **Sentence**: First, the waiter who served us neglected to fill us in on the specials, which I would have chosen had I known about them.	Neg/Neg/Neg ✓ / ✓ / ✗	Neg/Neg/Neu ✓ / ✓ / ✗	Neg/Neg/Pos ✓ / ✓ / ✓

Table 4: Error analysis of two review items from laptop and restaurant. The colored words in brackets represents ground truth sentiment label of the corresponding aspect. The symbol ✓ means the predicting sentiment is correct, and the other symbol means the predicting sentiment is wrong.

when a person tries to understand a relatively long sentence, he/she first read the entire sentence. Subsequently, after giving a specific aspect, he/she will dynamically select related words based on the previous memory state until he/she fully understands the sentiment polarity of the given aspect.

Interestingly, the above phenomenon is consistent with our dynamic re-weighting adapter's chosen result. Specifically, as Figure 3 (b) shows, with the re-weighting function F (i.e., Equation 5 and 6), our model dynamically choose the words *"food, better, while, definitely, not, ..."*, which have proven to be very important for predicting the sentiment of aspect *"food"* in Figure 3 (a). Those experimental results again fully indicate the effectiveness and interpretability of our proposed model in dynamic learning aspect-aware information.

The Influence of multiple Aspects. As aspect-related information plays a key role in ABSA and at least 10.2% of reviews contain multiple aspects as shown in Table 1, we are curious about the model's performance in the complex scenarios, e.g., a review sentence contains multiple aspects. Therefore, we randomly choose an example to explore how the selection of the context words will correspondingly change with different inputs. The visualization re-

sults are shown in Figure 4. Specifically, the chosen sentence has three different aspects with their sentiment polarity, i.e., *"System memory"*-negative, *"DDR5"*-positive and *"DDR3"*-negative. Take the aspect *"DDR5"* as example, it is positive which is contrary to *"DDR3"*. After receiving the overall semantic of the whole sentence, readers tend to associate *"DDR5"* with the context words {*"would", "have"*} to predict the correct sentiment "positive". For other two aspects, the observations are consistent with *"DDR5"*. In summary, all those results show that DR-BERT could dynamically extract the vital information to achieve aspect-aware semantic understanding even in a more complex scenario.

4.7 Error Analysis

Table 4 displays three review examples and their prediction results by BERT, RGAT-BERT, and our DR-BERT. As we can see from the "BERT-base" column, when there are multiple aspects, the vanilla BERT often makes the wrong classification since it tends to learn the overall sentiment polarity of the sentences instead of the aspect-aware semantic. While RGAT-BERT can alleviate the problem to a certain extent, it is also hard to predict the accurate sentiment label with few dependency relations. For

Methods	Laptop			Restaurant			Twitter		
	S	**E**	**T**	**S**	**E**	**T**	**S**	**E**	**T**
(1) DR-BERT	157s	10	26.1m	183s	10	30.5m	379s	10	63.2m
(2) T-GCN-BERT	168s	10	28.0m	188s	10	31.3m	411s	10	68.5m
(3) BERT-base	133s	10	22.2m	158s	10	26.3m	242s	10	40.3m
(4) ATAE-LSTM	3s	30	1.50m	4s	30	2.00m	5s	30	2.50m

Table 5: Runtime comparison between DR-BERT, T-GCN-BERT, BERT-base and ATAE-LSTM. Specifically, "S" represents the training time (seconds) for a single epoch, "E" denotes the number of training epochs, and "T" is the total training time (minutes).

example, in the first sentence, *"DDR3"* has few helpful syntactic dependency relations. Therefore, RGAT-BERT makes a wrong sentiment prediction. However, our DR-BERT model, succeeding in predicting most sentiment labels by considering the dynamic changing of the aspect-aware semantic. For other two case examples, the observations are consistent. Note that, for aspect *"Supplied software"* in second sentence, two overlap aspects appear in the same sentence makes it more difficult to distinguish the different sentiment between them. Thus, precisely determine its sentiment polarity is a big challenge for human, let alone deep learning models. This also leaves space for future exploration.

5 Computation Time Comparison

We also compared the computation runtime of three baseline methods. All of the models are performed on a Linux server with 64 Intel(R) CPUs and 4 Tesla V100 32GB GPUs. From the results shown in Table 5, we can first observe that the training time of a single epoch in DR-BERT performs better than T-GCN, which is based on GCN. Meanwhile, the training time of all these BERT-based models is similar (i.e., there is no significant difference). The possible reason is that the official datasets are small, and it is hard to influence the overall runtime of PLMs with such a small amount of data. Second, compared with other models, the training time of the ATAE-LSTM model is less (always an order of magnitude lower). For example, the ATAE-LSTM only needs about two minutes to achieve optimal performance in the restaurant dataset, while BERT-based models require more than 26 minutes. Therefore, though DR-BERT contains a Dynamic Re-weighting adapter based on GRU, the computation time is much lower than the BERT-based framework. In summary, the observations above show that the computation time of our DR-BERT model is within an acceptable range.

6 Conclusion and Future Works

This paper introduced a new approach named Dynamic Re-weighting BERT (DR-BERT) for aspect-based sentiment analysis. Specifically, we first employed the BERT layers as a base encoder to learn the overall semantic features of the whole sentence. Then, inspired by human semantic comprehension, we devised a new Dynamic Re-weighting Adapter (DRA) to enhance aspect-aware semantic features in the sentiment learning process. In addition, we inserted the DRA into the BERT layers to address the limitations of the vanilla pre-trained model in ABSA task. Extensive experiments on three benchmark datasets demonstrated the effectiveness and interpretability of the proposed model, with good semantic comprehension insights for future nature language modeling. Moreover, the error analysis was performed on incorrectly predicted examples, leading to some insights into the ABSA task.

We hope our research can help boost excellent work for aspect-based sentiment analysis from different perspectives. In the future, we plan to extend our method to other tasks like Sentence Semantic Matching, Relation Extraction, etc., which can also benefit from utilizing the dynamic semantics. Besides, we will explore whether DR-BERT can make any positive changes based on previous mistakes during the dynamic semantic understanding.

7 Acknowledgments

References

James C Bezdek. 1992. On the relationship between neural networks, pattern recognition and intelligence. *International journal of approximate reasoning*, 6(2):85–107.

Harm Brouwer, Francesca Delogu, Noortje J Venhuizen, and Matthew W Crocker. 2021. Neurobehavioral correlates of surprisal in language comprehension: A neurocomputational model. *Frontiers in Psychology*, 12:110.

Chenhua Chen, Zhiyang Teng, and Yue Zhang. 2020. Inducing target-specific latent structures for aspect sentiment classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5596–5607.

Jin Yao Chin, Kaiqi Zhao, Shafiq Joty, and Gao Cong. 2018. Anr: Aspect-based neural recommender. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 147–156.

Kyunghyun Cho, Bart van, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*, pages 4171–4186.

Xiaowen Ding and Bing Liu. 2007. The utility of linguistic rules in opinion mining. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 811–812.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 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)*, pages 49–54.

Feifan Fan, Yansong Feng, and Dongyan Zhao. 2018. Multi-grained attention network for aspect-level sentiment classification. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3433–3442.

Tal Golan, Prashant C Raju, and Nikolaus Kriegeskorte. 2020. Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, 117(47):29330–29337.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778.

Binxuan Huang and Kathleen M Carley. 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)*, pages 5469–5477.

Binxuan Huang, Yanglan Ou, and Kathleen M Carley. 2018. Aspect level sentiment classification with attention-over-attention neural networks. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 197–206. Springer.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proceedings of the 49th annual meeting of the association for computational linguistics*, pages 151–160.

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 437–442.

Christof Koch and Naotsugu Tsuchiya. 2007. Attention and consciousness: two distinct brain processes. *Trends in cognitive sciences*, 11(1):16–22.

Gina R Kuperberg. 2007. Neural mechanisms of language comprehension: Challenges to syntax. *Brain research*, 1146:23–49.

Gina R and T Florian Jaeger. 2016. What do we mean by prediction in language comprehension? *Language, cognition and neuroscience*, 31(1):32–59.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Ruifan Li, Hao Chen, Fangxiang Feng, Zhanyu Ma, Xiaojie Wang, and Eduard Hovy. 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 (Volume 1: Long Papers)*, pages 6319–6329.

Xiangsheng Li, Jiaxin Mao, Chao Wang, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Teach machine how to read: reading behavior inspired relevance estimation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 795–804.

Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018. Transformation networks for target-oriented sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 946–956.

Yunlong Liang, Fandong Meng, Jinchao Zhang, Jinan Xu, Yufeng Chen, and Jie Zhou. 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)*, pages 5569–5580.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4068–4074.

Thien Hai Nguyen and Kiyoaki Shirai. 2015. Phrasernn: Phrase recursive neural network for aspect-based sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2509–2514.

Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Jing Yu, Yajing Sun, and Xiangpeng Wei. 2020. Bidirectional cognitivethinking network for machine reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2613–2623.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372.

Cansu Sen, Thomas Hartvigsen, Biao Yin, Xiangnan Kong, and Elke Runden. 2020. Human attention maps for text classification: Do humans and neural networks focus on the same words? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4596–4608.

Lei Sha, Baobao Chang, Zhifang Sui, and Sujian Li. 2016. Reading and thinking: Re-read lstm unit for textual entailment recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2870–2879.

Selina Sharmin, Oleg Špakov, and Kari-Jouko Räihä. 2015. Dynamic text presentation in print interpreting–an eye movement study of reading behaviour. *International Journal of Human-Computer Studies*, 78:17–30.

Youwei Song, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao. 2019. Attentional encoder network for targeted sentiment classification. *arXiv preprint arXiv:1902.09314*.

Concetto Spampinato, Simone Palazzo, Isaak Kavasidis, Daniela Giordano, Nasim Souly, and Mubarak Shah. 2017. Deep learning human mind for automated visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6809–6817.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–385.

Yaming Sun, Lei Lin, Duyu Tang, Nan Yang, Zhenzhou Ji, and Xiaolong Wang. 2015. Modeling mention, context and entity with neural networks for entity disambiguation. In *Twenty-fourth international joint conference on artificial intelligence*.

Niels A Taatgen, Hedderik Van Rijn, and John Anderson. 2007. An integrated theory of prospective time interval estimation: The role of cognition, attention, and learning. *Psychological review*, 114(3):577.

Xingwei Tan, Yi Cai, and Changxi Zhu. 2019. Recognizing conflict opinions in aspect-level sentiment classification with dual attention networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pages 3426–3431.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2015. Effective lstms for target-dependent sentiment classification. *arXiv preprint arXiv:1512.01100*.

Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 214–224.

Hao Tang, Donghong Ji, Chenliang Li, and Qiji Zhou. 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*, pages 6578–6588.

Jialong Tang, Ziyao Lu, Jinsong Su, Yubin Ge, Linfeng Song, Le Sun, and Jiebo Luo. 2019. Progressive self-supervised attention learning for aspect-level sentiment analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 557–566.

Yuanhe Tian, Guimin Chen, and Yan Song. 2021. Aspect-based sentiment analysis with type-aware graph convolutional networks and layer ensemble. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2910–2922.

Giulio Tononi. 2008. Consciousness as integrated information: a provisional manifesto. *The Biological Bulletin*, 215(3):216–242.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Jingjing Wang, Changlong Sun, Shoushan Li, Xiaozhong Liu, Luo Si, Min Zhang, and Guodong Zhou. 2019. Aspect sentiment classification towards question-answering with reinforced bidirectional attention network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3548–3557.

Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. 2020. Relational graph attention network for aspect-based sentiment analysis. In *Proceedings of 58th Annual Meeting of the Association for Computational Linguistics*, pages 3229–3238.

Xuejian Wang, Lantao Yu, Kan Ren, Guanyu Tao, Weinan Zhang, and et al. 2017. Dynamic attention deep model for article recommendation by learning human editors' demonstration. In *Proceedings of the 23rd international conference on knowledge discovery and data mining*, pages 2051–2059.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

Bowen Xing, Lejian Liao, Dandan Song, and et al. 2019. Earlier attention? aspect-aware lstm for aspect-based sentiment analysis. In *IJCAI*.

Hu Xu, Bing Liu, Lei Shu, and S Yu Philip. 2019a. Bert post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2324–2335.

Hu Xu, Lei Shu, S Yu Philip, and Bing Liu. 2020. Understanding pre-trained bert for aspect-based sentiment analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 244–250.

Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. 2019b. Understanding and improving layer normalization. *Advances in Neural Information Processing Systems*, 32.

Hsien-Ming Yang and George W McConkie. 1999. Reading chinese: Some basic eye-movement characteristics. *Reading Chinese script: A cognitive analysis*, pages 207–222.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659.

Chen Zhang, Qiuchi Li, and Dawei Song. 2019a. 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)*, pages 4568–4578.

Kai Zhang, Qi Liu, Hao Qian, Biao Xiang, Qing Cui, Jun Zhou, and Enhong Chen. 2021a. Eatn: An efficient adaptive transfer network for aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*.

Kai Zhang, Hao Qian, Qi Liu, Zhiqiang Zhang, Jun Zhou, and et al. 2021b. Sifn: A sentiment-aware interactive fusion network for review-based item recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3627–3631.

Kai Zhang, Hefu Zhang, Qi Liu, Hongke Zhao, Hengshu Zhu, and Enhong Chen. 2019b. Interactive attention transfer network for cross-domain sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5773–5780.

Kai Zhang, Hongke Zhao, Qi Liu, Zhen Pan, and Enhong Chen. 2019c. A dynamic and cooperative tracking system for crowdfunding. *arXiv preprint arXiv:2002.00847*.

Kun Zhang, Guangyi Lv, Linyuan Wang, Le Wu, Enhong Chen, Fangzhao Wu, and Xing Xie. 2019d. Drr-net: Dynamic re-read network for sentence semantic matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7442–7449.

Yaowei Zheng, Richong Zhang, Samuel Mensah, and Yongyi Mao. 2020. Replicate, walk, and stop on syntax: An effective neural network model for aspect-level sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9685–9692.

Yukun Zheng, Jiaxin Mao, Yiqun Liu, Zixin Ye, Min Zhang, and Shaoping Ma. 2019. Human behavior inspired machine reading comprehension. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 425–434.