IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 14, 2021, accepted January 5, 2022, date of publication January 11, 2022, date of current version January 24, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3142178

Unrestricted Attention May Not Be All You Need–Masked Attention Mechanism Focuses Better on Relevant Parts in Aspect-Based Sentiment Analysis

AO FENG, XUELEI ZHANG, AND XINYU SONG
School of Computer Science, Chengdu University of Information Technology, Chengdu 610225, China

Corresponding author: Xuelei Zhang (zhangxueleie@163.com)

This work was supported in part by the Research Innovation Team Fund from the Department of Education, Sichuan Province, under Award 18TD0026; and in part by the Sichuan Science and Technology Program under Project 2020YFG0168.

ABSTRACT Aspect-Based Sentiment Analysis (ABSA) is one of the highly challenging tasks in natural language processing. It extracts fine-grained sentiment information in user-generated reviews, as it aims at predicting the polarities towards predefined aspect categories or relevant entities in free text. Previous deep learning approaches usually rely on large-scale pre-trained language models and the attention mechanism, which applies the complete computed attention weights and does not place any restriction on the attention assignment. We argue that the original attention mechanism is not the ideal configuration for ABSA, as for most of the time only a small portion of terms are strongly related to the sentiment polarity of an aspect or entity. In this paper, we propose a masked attention mechanism customized for ABSA, with two different approaches to generate the mask. The first method sets an attention weight threshold that is determined by the maximum of all weights, and keeps only attention scores above the threshold. The second selects the top words with the highest weights. Both remove the lower score parts that are assumed to be less relevant to the aspect of focus. By ignoring part of input that is claimed irrelevant, a large proportion of input noise is removed, keeping the downstream model more focused and reducing calculation cost. Experiments on the Multi-Aspect Multi-Sentiment (MAMS) and SemEval-2014 datasets show significant improvements over state-of-the-art pre-trained language models with full attention, which displays the value of the masked attention mechanism. Recent work shows that simple self-attention in Transformer quickly degenerates to a rank-1 matrix, and masked attention may be another cure for that trend.

INDEX TERMS Sentiment analysis, attention mechanism, pre-trained language model, masked attention.

I. INTRODUCTION

Sentiment analysis [1]–[4] is one of the prevalent tasks in natural language processing (NLP). With the fast development of social media and E-commerce platforms in recent years, sentiment analysis of online reviews aims at mining users' opinions towards social events or products. With effective detection of sentiment polarity expressed in these reviews [5], stakeholders have the opportunity to guide public opinion towards their favorable direction, or improve products for better user experience [6].

Aspect-based sentiment analysis (ABSA) [7]–[9] is a more fine-grained task in comparison to document or sentence-level sentiment analysis. ABSA analyzes the user's emotional state of aspect categories or entities [10]–[12], and its advantage lies in the ability to capture aspect-specific sentiment information [13], [14]. Figure 1 shows two subtasks of ABSA. The first example is for Aspect Term Sentiment Analysis (ATSA), in which two aspect terms ''food'' and ''service'' are associated with positive and negative sentiments respectively. The second review shows the application of Aspect Category Sentiment Analysis (ACSA). The sentence does not mention the two aspect categories ''food'' or ''ambience'' directly, with only description terms that fall within their range.

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

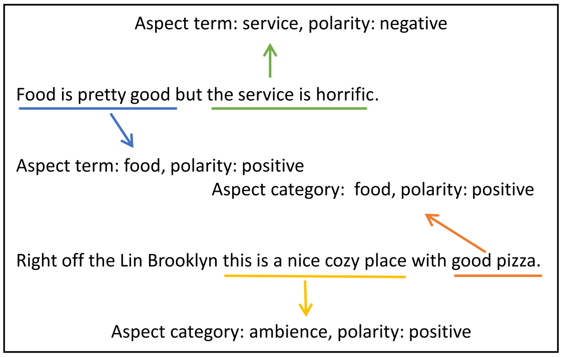

Aspect term: service, polarity: negative
Food is pretty good but the service is horrific.
Aspect term: food, polarity: positive
Aspect category: food, polarity: positive
Right off the Lin Brooklyn this is a nice cozy place with good pizza.
Aspect category: ambience, polarity: positive

FIGURE 1. Example of ATSA and ACSA data. The underlined words are key instances of the aspect polarity. Aspect term is a word or phrase in a sentence, and aspect category is a predefined category that is expressed by specific terms in a sentence.

With the development of deep learning, a large number of neural network models [17]–[21], [30], [31] are present in ABSA tasks. Among those models, Long Short-Term Memory (LSTM) [22] and attention mechanism [54] have achieved good results in many natural language processing tasks. LSTM networks and the attention mechanism are often combined in ABSA, because LSTM can effectively extract most of the information in the sentence, and attention calculates the correlation weights of different words in the comments to determine which part is the most relevant. In order to reduce the high computational complexity of LSTM and attention mechanism, Convolutional Neural Networks (CNN) are also used in ABSA because of their efficient feature extraction capabilities [36], [37]. Traditional deep learning methods are built upon embedding representations, and the commonly used term embedding vectors include Word2Vec [25] and GloVe [26].

More recently, large-scale pre-trained language models gradually become standard baselines in NLP [38], represented by BERT [27], [39], XLNET [28], and RoBERTa [29]. These language models have strong representative capabilities due to their rich text representation learned in large datasets. They have achieved promising results in ABSA [53], in comparison to traditional deep learning models. As sentiment information related to aspect terms is the key to solving the ABSA task, it may be assumed that combining the attention mechanism with pre-trained language models should improve the performance at aspect level. However, we find that a simple addition of aspect attention to BERT hurts its performance on some ABSA datasets. Analysis of the data shows that the rich context of the whole sentence is confusing for a single aspect, especially when a sentence contains multiple aspects.

In this paper, we propose two attention mask mechanisms to limit aspect attention to the most relevant parts of a sentence. One is called Attention Mask Weight (AM-Weight), which sets a weight threshold to filter out irrelevant parts. The other is Attention Mask Word (AM-Word), keeping only the top βn words in weight assignment. Starting from pre-trained language models, standard attention scores for each aspect term or category are calculated. Then the candidates are filtered by one of the mask mechanisms, weighted sum of their embedding representations are calculated, and the output goes through a fully-connected layer and softmax to output the sentiment classification label. Experiments are conducted on SemEval-2014 and the Multi-Aspect Multi-Sentiment (MAMS) datasets, which show the performance improvement with the attention mask.

Main contributions of this article can be summarized as follows:

1) This paper proposes two attention mask mechanisms for the aspect-based sentiment analysis task, and shows their value with detailed experiments.
2) State-of-the-art performance is achieved on three public datasets, with introduction of the attention mask over the RoBERTa baseline.
3) With sensitivity analysis and examination of both success/failure cases, recommended strategy is provided for the two attention mask mechanisms. The model also provides a potential remedy for the deficiency of self-attention mechanism in Transformer networks.

II. RELATED WORK

A. TEXT REPRESENTATION

To deal with the issue of dimensional curse and independence assumption of one-hot vectors, low-dimensional embedding vectors were proposed as an alternative representation of individual terms in a large vocabulary. Word2Vec [25] and GloVe [26], as the most popular embedding representations, are widely used in a large number of NLP applications, displaying their clear advantage over one-hot vectors. However, pre-training of a term's representation vector over a large unlabeled dataset learns only the average of its semantics, ignoring its context in each appearance. It explicitly omits the polysemous case of the word, which may introduce large bias when the downstream application is not covered in the source text collection.

Pre-trained language models using large-scale corpora were proposed in 2018 that take more context into consideration, such as ELMo [38], GPT [39], and BERT [27]. These pre-trained models have strong feature extraction capabilities from the data. BERT, with a bidirectional Transformer model for feature extraction, uses a special mask model for context representation and next sentence prediction for semantic coherence. Depending on the downstream task, BERT can be fine-tuned to better fit the task, and its effectiveness has been proved in multiple NLP tasks. Excellent performance of BERT highly relies on its complex Transformer framework, as static vectors from BERT do not show much improvement over Word2Vec or GloVe.

Recent experiments show that the static mask in the original BERT model limits its performance. RoBERTa [29], as an improved version of BERT, replaces its static mask with a dynamic one, expands the pre-training corpus, and removes the next sentence prediction task. In a number of

NLP tasks, RoBERTa shows significant improvements over BERT, making it new state-of-the-art in the BERT family.

B. ASPECT-BASED SENTIMENT ANALYSIS

ABSA aims at predicting sentiment polarity for a specific aspect category or term. It has been widely adopted in many business and social applications, including sentiment analysis of social media [33] and conversational sentiment analysis [34]. Besides, it helps identify the polarity of users' opinions in traffic events [35], which helps determine the accurate condition. It is also applied to the healthcare data to predict drug side effects and abnormal conditions in patients [32].

Most of the recent deep learning work in aspect-based sentiment analysis use static word embeddings as input. LSTM networks and the attention mechanisms are often adopted for feature extraction. Target dependent Long Short-Term Memory (TD-LSTM) [16] model embeds the target word and its context into the vector space separately, and uses two LSTMs to obtain the relationship between the context and the target word to extract the emotional information of the aspect entity. Attention-based LSTM with Aspect Embedding (ATAE-LSTM) [15] embeds the splicing of sentences and aspect information into the same vector space, and applies LSTM to obtain the embedding vector information. The attention mechanism is also applied to obtain the most relevant part within the sentence for aspect sentiment analysis. Aspect-based sentiment analysis with gated convolutional networks (GCAE) [36] uses two convolutional neural networks to extract aspect-related and aspect-independent information, and designs a gate control mechanism to filter the output for aspect-related information. Constrained attention network for multi-aspect sentiment analysis (CAN) [40] restricts the attention weight by introducing sparse regularization, feeding more weight into the higher attention values. It introduces orthogonal regularization to restrict different aspects from focusing on the same parts of a sentence, ensuring more sparse attention for multiple aspects. These methods extract the semantic information of the word embedding through complex network structures, and have achieved competitive results in ABSA.

Recent developments of large-scale pre-trained language models, especially BERT, have made great impact on mainstream NLP tasks. Aspect-based sentiment analysis, as an extension to the coarse-grained sentiment analysis task, also follows the trend. As an extension to the BERT model, BERT-SPC sends "[CLS] + sentence sequence + [SEP] + aspect sequence+ [SEP]" to the hidden layer output of the pre-trained BERT network for aspect sentiment classification. It achieves good performance in ABSA, following the next sentence prediction task in BERT that proves its ability to capture the relationship between aspect information and the whole sentence. As a simple baseline built over BERT, BERT-SPC does not involve any complex network engineering mechanism, but some of the other models has not achieved significant improvement over it in our experiment. Attentional Encoder Network (AEN-BERT) [41] uses BERT to embed the context sequence and aspect information, and then applies the attention mechanism to extract the semantic interaction between an aspect and its context. BERT-PT [42] transforms the aspect sentiment classification problem into a special machine reading comprehension problem. BERT-pair-QA [43] constructs a question answering problem with aspect information, performing aspect sentiment analysis by combining automatic question answering and natural language understanding with BERT. CapsNet-BERT [44] mixes BERT with the capsule network to model the complex relationship between the aspect information and its context. Local context-focus on syntax-ASC (LCFS-ASC) [45] employs local context-focus and the syntactic dependency tree to obtain the sentence components related to the aspect, discarding the rest of the context that are irrelevant. Multi-head self-attention transformation (MSAT) [46] network 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 sentiment analysis. Knowledge guided capsule network (KGCapsAN) [47] utilizes certain prior knowledge to guide the capsule attention process to solve the ABSA task. Knowledge-enabled BERT [48] utilizes the additional information from a sentiment knowledge graph and pre-trained language model to solve ABSA task.

III. METHOD

This section describes the aspect-based sentiment analysis task and how to apply the attention mask mechanism over a BERT-based model. Within the attention mask network, there are two variations of the mask mechanism, named Attention Mask Weight (AM-Weight) and Attention Mask Word (AM-Word), respectively.

A. PROBLEM FORMULATION

The ABSA task includes Aspect Term Sentiment Classification (ATSC) and Aspect Category Sentiment Classification (ACSC). In the ACSC task, the categories

$$A = \{A_1, A_2, \ldots, A_L\} \tag{1}$$

are predefined, and the sentiment polarities are restricted to

$$P = \{Positive, Negative, Neutral\} \tag{2}$$

A sentence or paragraph (In ABSA, we usually process a short piece of text in each instance, i.e., a sentence.) from the text collection is represented as

$$S = \{w_1, w_2, \ldots, w_i, \ldots, w_n\} \tag{3}$$

consisting of n words, w_i represents the i word in sentence S. A sentence may contain M targets

$$T^S = \left\{T_1^S, T_2^S, \ldots, T_M^S\right\} \tag{4}$$

and each target from the sentence S, and is represented as

$$T_i^S = \left\{w_i, w_{(i+1)}, \ldots, w_{(i+m_i-1)}\right\} \tag{5}$$

with m_i terms. The goal of the ATSC task is to predict the sentiment polarities for M targets, P_M^T represents the sentiment polarity of the T_M^S in the sentence:

$$P^T = \left\{P_1^T, P_2^T, \ldots, P_M^T\right\} \tag{6}$$

Similarly, if a sentence contains N aspect categories,

$$A^S = \left\{A_1^S, A_2^S, \ldots, A_N^S\right\} \tag{7}$$

the ACSC task predicts the sentiment polarities for each of them, P_N^A represents the sentiment polarity of the A_N^S in the sentence:

$$P^A = \left\{P_1^A, P_2^A, \ldots, P_N^A\right\} \tag{8}$$

B. ATTENTION MASK NETWORK FOR ABSA

This section introduces our proposed Attention Mask Network for aspect-based sentiment analysis, with its framework shown in Figure 2. The network consists of a contextualized embedding layer, an attention layer, two attention mask layers, and a fully-connected layer.

1) INPUT AND CONTEXTUALIZED EMBEDDING LAYER

Inputs of our model include a sentence and an aspect. When using BERT to embed an aspect and its contextual information, a special classification mark "[CLS]" is added at the beginning of the input sequence, and a separator "[SEP]" is placed at the end. In order to obtain the relationship between the aspect and its context, both are encoded in the same way. When using the RoBERTa pre-trained language model, the inputs are also processed in the same manner with identical input format. The only difference is that the "<s>" separator is used to replace "[CLS]", and "</s>" takes the place of "[SEP]".

2) ATTENTION LAYER

The attention layer takes the output of the contextualized embedding layer as input. In this layer, we calculate the attention [24], [52] weight between aspect vectors and sentence vectors, and then apply one of the two mask mechanisms, mask attention score or mask attention word, to produce the final attention weight vector. Ideally, each aspect keeps attention only from the most relevant context after the mask.

$x^s = \left(x_1^s, \ldots, x_n^s\right)$ is the input of sentence contextualized embedding, and $x^a = \left(x_1^a, \ldots, x_n^a\right)$ is the input of aspect contextualized embedding, where $x_i \in \mathbb{R}^{d_s}$. A new sequence $z = (z_1, \ldots, z_n)$ with the same length is calculated as the attention output, where $z_i \in \mathbb{R}^{d_z}$.

e_{ij} is used as a compatibility function that measures the relatedness or similarity of two input elements:

$$e_{ij} = \frac{\left(x_i^a W^Q\right)\left(x_j^s W^K\right)^T}{\sqrt{d_z}} \tag{9}$$

Then each weight coefficient α_{ij} is calculated with the softmax function

$$\alpha_{ij} = \frac{\exp e_{ij}}{\sum_{k=1}^{n} \exp e_{ik}} \tag{10}$$

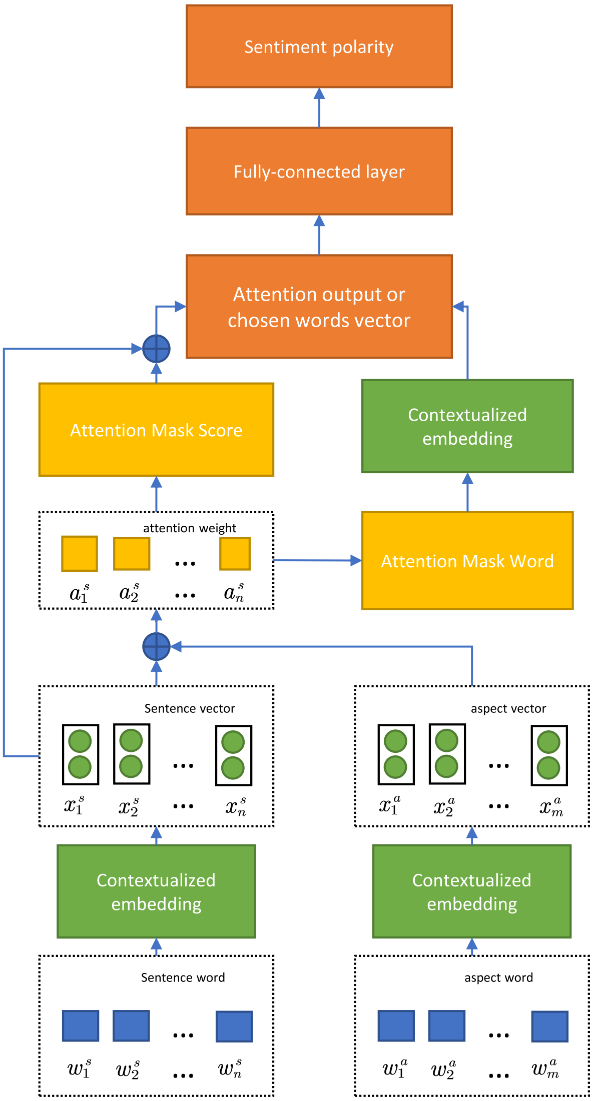


FIGURE 2. Structure of the proposed attention mask network for aspect-based sentiment analysis, displaying both the attention mask weight and attention mask word mechanism.

Each output element, z_i is calculated by linearly transforming the input elements with their weighted sum:

$$z_i = \sum_{j=1}^{n} \alpha_{ij}\left(V_j^s W^V\right) \tag{11}$$

We choose scaled dot product as the compatibility function for effective calculation. Through linear transformation of the input, relevant contextual information for each aspect is extracted from the input. $W^Q, W^K, W^V \in \mathbb{R}^{d_x \times d_z}$ are attention parameter matrices, separately tuned for each layer and attention head.

3) ATTENTION MASK WEIGHT

To retain only the part of sentence that has a significant attention score (highly relevant) for the aspect, we introduce a parameter γ as the threshold ratio of the mask. It is multiplied to the maximum of attention weights in the sentence, and the result is used as the threshold. When the attention value is lower than that threshold, it is masked to zero, other values remain unchanged.

$$\alpha_{ij} = \begin{cases} 0 & \alpha_{ij} < \gamma max_w \\ \alpha_{ij} & \alpha_{ij} >= \gamma max_w \end{cases} \tag{12}$$

where γ is the configurable parameter of the AM-Weight model, max_w is the maximal attention score in a sentence, and α_{ij} is the original attention score.

4) ATTENTION MASK WORD

When using the other attention mask mechanism in ABSA, a more intuitive method is adopted. β is the only parameter for this method, showing the percentage of words to keep in the original sentence. All n attention weights in a sentence are sorted in descending order,

$$sorted\_\alpha_{ij} = \{\alpha_{ir_1}, \alpha_{ir_2}, \ldots, \alpha_{ir_{\beta n}}, \ldots, \alpha_{ir_n}\} \tag{13}$$

in which the r_j subscript shows the rank of each weight in the sorted list. Top βn words with higher scores are kept. Other attention weights are set to zero for the lower values.

$$\alpha_{ij} = \begin{cases} \alpha_{ij} & \alpha_{ir_1} \geq \alpha_{ij} \geq \alpha_{ir_{\beta n}} \\ 0 & \alpha_{ir_{\beta n}} > \alpha_{ij} \geq \alpha_{ir_n} \end{cases} \tag{14}$$

Words with zero attention weights are removed from the sentence, and the rest with non-zero attention are saved in the original order. Then they are fed into the sentiment classification layer.

5) ASPECT SENTIMENT CLASSIFICATION

If the attention mask weight mechanism is used, the weighted sum of masked attention and original term vectors is used as the attention output, which is fed to the final aspect sentiment classification. For the attention mask word mechanism, we only keep the remaining βn words for contextualized embedding using BERT or RoBERTa, and perform aspect sentiment classification with their representation.

6) LOSS

The cross-entropy loss is used to calculate the disagreement between the predicted label and the true label.

$$\text{Loss} = -\sum_{i=1}^{M} y_i \log \hat{y}_i \tag{15}$$

IV. EXPERIMENTS

This section introduces the task definition, datasets, evaluation metrics, baseline models and their performance comparison.

A. TASK DEFINITION

Experiments include two main subtasks of ABSA - Aspect-Term Sentiment Analysis (ATSA) and Aspect-Category Sentiment Analysis (ACSA).

TABLE 1. Statistics of the SemEval-2014 dataset.

Dataset			Positive	Negative	Neutral	Total
ATSA	Laptop	Train	987	866	460	2313
		Test	341	128	169	638
	Restaurant	Train	2164	805	633	3602
		Test	728	196	196	1120
ACSA	Restaurant	Train	2179	839	500	3518
		Test	657	222	94	973

TABLE 2. Statistics of the MAMS dataset.

Dataset		Positive	Negative	Neutral	Total
ATSA	Train	3380	2764	5042	11186
	Validation	403	325	604	1332
	Test	400	329	607	1336
ACSA	Train	1929	2084	1613	5626
	Validation	241	259	388	888
	Test	245	263	393	901

TABLE 3. Hyperparameters used in the experiment.

Parameter	Value
Dropout rate	0.1
Batch size	24
Learning rate	2e-5
Max epoch	10
Max aspect length	8
Max sequence length	80

B. DATASETS

There are two data collections used in our experiment, SemEval-2014 task 4[1] and Multi-Aspect Multi-Sentiment (MAMS) dataset. SemEval-2014 consists of two ATSA datasets - laptop and restaurant, and it also includes an ACSA dataset with only restaurant reviews. MAMS consists of an ATSA dataset and an ACSA dataset, both are from restaurant reviews. Table 1 shows the statistics of the SemEval-2014 dataset, and Table 2 includes the details of the MAMS dataset, respectively. Both tables show the number of training, validation and test samples in each dataset, together with the label breakdown. SemEval-2014 contains only training and test sets, while MAMS also has a validation set in each collection.

C. EXPERIMENT SETTING

All models are implemented with the Pytorch [23] deep learning framework. We use the pre-trained BERT-base-uncased[2] and RoBERTa-base[3] models in English for fine-tuning. The number of transformer layers is fixed at 12, size of the hidden layer is 768, and the number of self-attention heads is 12. The total number of parameters of the pre-trained model is about 110M. We use Adam [51] optimizer for model tuning, and other hyper-parameters in the experiment are shown in Table 3.

D. BASELINE METHODS

In order to verify the performance of our attention mask mechanism, the two variants (AM-Weight and AM-Word) are applied over the base BERT and BoBERTa models and compared to other deep learning methods. Many of these

[1] http://alt.qcri.org/semeval2014/task4/

[2] https://huggingface.co/bert-base-uncased

[3] https://huggingface.co/roberta-base

competitors are optimized versions of BERT with more complex network structures. Evaluation metrics include overall classification accuracy and F1 score, and they are compared on both ATSA and ACSA tasks.

TD-LSTM [16] contains two LSTM networks, which model the context on each side of the target word, respectively. The hidden layer vectors from the two directions are concatenated and the joint vector determines the final aspect term sentiment.

ATAE-LSTM [15] attaches aspect terms to the representation of each word in the sentence. Self-attention mechanism is applied to calculate the attention value of different words in the sentence for aspect sentiment classification.

MemNet [50] designs multiple attention networks to calculate the importance of each word in the context of the aspect information, and uses their calculation result to classify the aspect sentiment.

In **GCAE** [36], two convolutional networks are applied to capture the information relevant and irrelevant to aspect terms, and the Tanh-ReLU unit of the gate control mechanism is used to filter the appropriate information for sentiment classification.

BERT-SPC sends the hidden layer output of the pre-trained BERT network to a fully-connected neural network for aspect sentiment classification. The input sequence of the model is: "[CLS] + sentence sequence + [SEP] + aspect sequence+ [SEP]".

BERT-PT [53] designs a reading comprehension problem based on aspect sentiment classification. BERT is used to form the context embedding of the aspect terms, and a machine reading comprehension task determines the sentiment polarity of each aspect.

In **AEN-BERT** [41], the encoder uses the attention mechanism to model the context and the target word. Then it introduces label smoothing and regularization, and combines the pre-trained BERT model for aspect sentiment classification.

BERT-pair-QA-M [43] constructs an auxiliary question for each aspect. Its BERT model is trained on the question answering problem, and its result can be applied for sentence pair classification.

TD-BERT [49] changes the normal use of [CLS] tag in BERT. Taking the BERT embedding from the target word (instead of [CLS]), its hidden layer output is used for aspect sentiment classification.

CapsNet-BERT [44] inputs the sentence and the aspect word into the embedding layer separately, and the average value of the aspect word embedding is calculated as the input of the next layer. The sentence information is fed into a bidirectional GRU through the residual connection to obtain the contextual representation. Then the final category capsules are computed with the primary capsules, aspect-aware normalized weights, and capsule-guided routing weights.

BERT-Attention binds a single attention layer to the BERT output, and a fully connected layer is added after that for sentiment classification.

RoBERTa uses the hidden layer output from a pre-trained RoBERTa model, then a fully-connected network calculates the aspect sentiment classification. The input sequence of the model is: "<s> + sentence sequence + </s> + aspect sequence + </s>".

RoBERTa-Attention adds an attention layer to RoBERTa, the rest stays the same as the model above.

E. EXPERIMENT RESULTS

Table 4 shows the performance of various models on the ATSA task, and Table 5 contains the result for the ACSA task. Several conclusions can be drawn from the performance comparison in the result tables. First of all, pre-trained language models have a much higher baseline in comparison to the traditional static embedding-based methods. The clear improvement proves that the strong encoding capability of pre-trained language models has successfully learned the rich semantic information in diversified context. Secondly, most of the BERT-based methods do not outperform the simple BERT-SPC model, as they fail to identify the relevant context for each given aspect. Third, AM-Weight-BERT obtains an average of +1.00% accuracy and +1.12% Macro-F1 over BERT-SPC on 5 datasets, indicating that the attention mask mechanism has great potential when combined with a very competitive baseline that is hard to improve upon.

In addition, application of the attention mask mechanism over RoBERTa achieves new state-of-the-art performance on five datasets. AM-Weight-RoBERTa has an average accuracy increase of 0.77% compared to RoBERTa-Attention over five collections, while the average F1 score has increased by 1.26%. Improvement of the AM-Word-RoBERTa model is 0.84% for average accuracy and 1.25% for average F1 score. The weight-based threshold achieves the best result in 2 of the 5 experiments, while limiting to top-N terms wins the other 3. Although it is hard to determine which one might be the best choice for each dataset, their difference is negligible in many cases. RoBERTa is a stronger baseline than BERT, which makes it harder to beat. However, application of the attention mask mechanism, which filters attention weights that are not strongly related to the specified aspect term or category, can still achieve significant performance improvement. That improvement does not rely strongly on the underlying model, which is a desirable attribute for the attention mask mechanism that can be applied to many existing networks.

V. DATA ANALYSIS

A. PARAMETER SENSITIVITY

Two new parameters γ and β have been introduced in the masked attention mechanism, one for AM-Weight and another for AM-Word. Experiments in the previous section show their performance improvement in sentiment analysis, but their sensitivity also needs to be evaluated for different values of the parameter. As MAMS is a larger collection with additional division of validation and test set, performance reported on MAMS tends to be more stable. Considering

TABLE 4. Performance comparison for aspect-term sentiment analysis with classification accuracy and Macro-F1 value of different models. "– –" means that the metric is not reported. For our method or re-implementations from others, we run the program for 5 times with random initialization, and show "mean±std" as its performance. The best performance in each column is bold-typed.

Text Representation	Method	Laptop		Restaurant		MAMS	
		Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
Embedding	TD-LSTM	62.23±0.92	– –	73.44±1.17	– –	74.60	– –
	ATAE-LSTM	64.38±4.52	– –	73.74±3.01	– –	77.05	– –
	MemNet	70.33	64.09	78.16	65.83	64.57	– –
	GCAE	69.14±0.32	– –	77.28±0.32	– –	77.59	– –
BERT	BERT-SPC	79.07±0.85	75.58±1.12	84.82±0.71	77.84±1.31	82.28±0.29	81.57±0.36
	AEN-BERT	78.35±1.24	73.68±1.19	81.46±0.29	71.73±1.12	– –	– –
	BERT-PT	78.07	75.08	84.95	76.96	– –	– –
	BERT-pair-QA-M	77.93±0.83	73.71±1.72	85.12±0.41	77.31±1.10	– –	– –
	TD-BERT	78.87±1.13	74.38±0.81	85.10±0.20	78.35±1.34	– –	– –
	CapsNet-BERT	– –	– –	**85.93**	– –	**83.39**	– –
	BERT-Attention	79.00±0.62	75.36±0.70	84.92±0.33	77.53±0.61	82.31±0.25	81.61±0.34
	AM-Weight-BERT	79.78±0.77	76.20±0.71	85.66±0.37	78.92±0.97	83.11±0.54	**82.56±0.59**
	AM-Word-BERT	**79.87±0.40**	**76.26±0.87**	85.57±0.71	**79.02±1.07**	83.10±0.31	82.44±0.30
RoBERTa	RoBERTa	81.82±0.85	78.39±1.24	87.50±0.43	80.28±1.07	82.97±0.51	82.62±0.48
	RoBERTa-Attention	81.97±0.71	78.60±0.77	87.25±1.09	80.20±1.64	83.27±1.14	82.62±1.21
	AM-Weight-RoBERTa	82.48±0.54	79.32±0.42	**88.13±0.29**	**82.54±0.68**	83.92±0.92	83.37±0.96
	AM-Word-RoBERTa	**83.04±0.29**	**80.20±0.62**	87.86±0.23	82.05±0.48	**84.15±0.33**	**83.66±0.37**

TABLE 5. Performance comparison for aspect-category sentiment analysis with classification accuracy and Macro-F1 value of different models. "– –" means that the metric is not reported. For our method or re-implementations from others, we run the program for 5 times with random initialization, and show "mean±std" as its performance. The best performance in each column is bold-typed.

Text Representation	Method	Restaurant		MAMS	
		Accuracy	Macro-F1	Accuracy	Macro-F1
Embedding	LSTM	82.00	– –	– –	– –
	ATAE-LSTM	84.00	– –	70.64	– –
	GCAE	79.67±0.35	– –	72.10	– –
	CapsNet	83.56	– –	73.99	– –
BERT	BERT-SPC	90.36±0.24	83.87±0.26	79.02±0.74	78.52±0.69
	BERT-pair-QA-M	89.90	– –	– –	– –
	CapsNet-BERT	**91.38**	– –	79.46	– –
	BERT-Attention	89.96±0.21	83.48±0.51	79.27±1.04	78.77±0.99
	AM-Weight-BERT	91.02±0.53	**84.87±1.22**	80.87±0.55	80.34±0.58
	AM-Word-BERT	90.89±0.32	84.68±0.56	**81.20±1.08**	**80.64±1.14**
RoBERTa	RoBERTa	92.39±0.46	87.09±0.60	80.75±0.84	80.19±0.90
	RoBERTa-Attention	92.10±0.34	86.01±0.73	80.49±0.56	80.04±0.47
	AM-Weight-RoBERTa	**92.99±0.38**	**88.06±0.76**	81.64±1.06	81.20±1.11
	AM-Word-RoBERTa	92.75±0.42	87.23±0.55	**81.71±0.62**	**81.28±1.14**

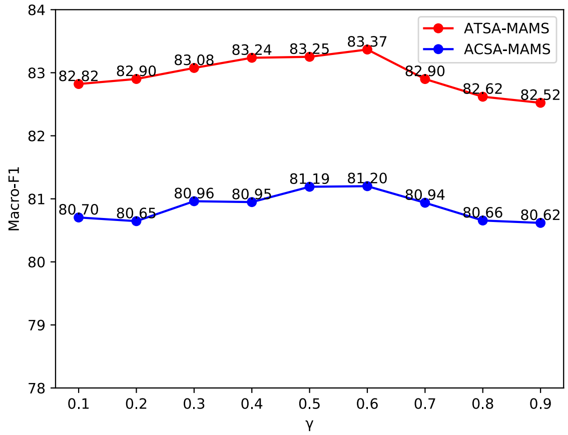


FIGURE 3. Macro-F1 of AM-Weight-RoBERTa on the MAMS dataset with different mask ratio γ.

the imbalanced data distribution in both collections, macro-F1 is a better measure for model performance. Therefore, sensitivity analysis of these parameters is carried out on the MAMS dataset, with macro-F1 as the indicator of overall performance.

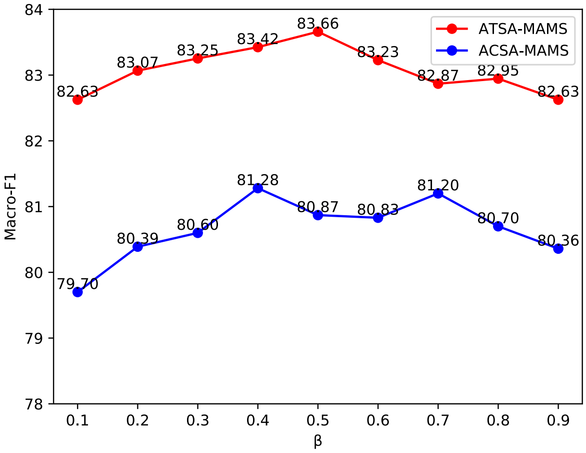
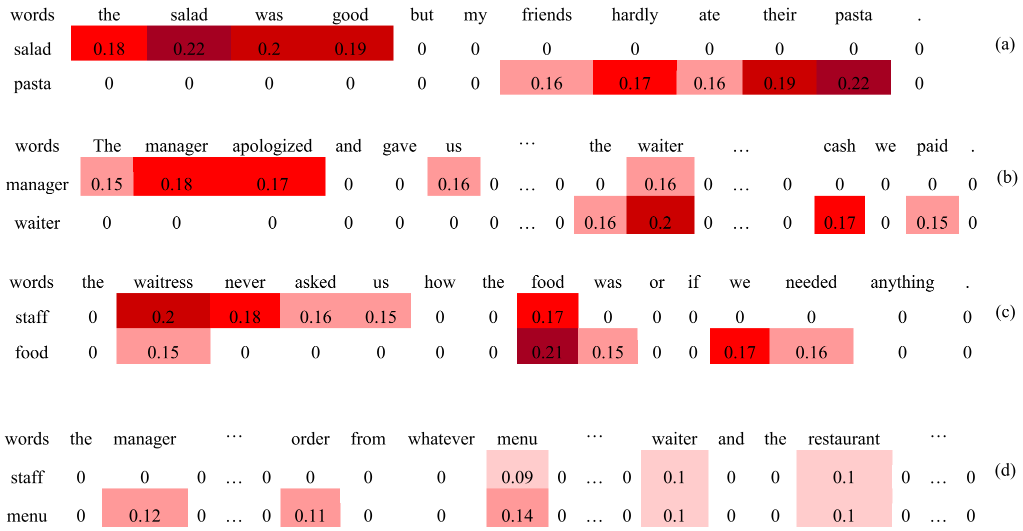


FIGURE 4. Macro-F1 of AM-Word-RoBERTa on the MAMS dataset with different mask percentage β.

Figure 3 shows the macro-F1 value of AM-Weight-RoBERTa, with mask ratio γ in the [0.1, 0.9] range. Figure 4 displays the change of performance of AM-Word-RoBERTa, where the mask percentage β takes the same range. From



words	the	salad	was	good	but	my	friends	hardly	ate	their	pasta	.
salad	0.18	0.22	0.2	0.19	0	0	0	0	0	0	0	0
pasta	0	0	0	0	0	0	0.16	0.17	0.16	0.19	0.22	0

(a)

words	The	manager	apologized	and	gave	us		…		the	waiter		…		cash	we	paid	.
manager	0.15	0.18	0.17	0	0	0.16	0	…	0	0	0.16	0	…	0	0	0	0	0
waiter	0	0	0	0	0	0	0	…	0	0.16	0.2	0	…	0	0.17	0	0.15	0

(b)

words	the	waitress	never	asked	us	how	the	food	was	or	if	we	needed	anything	.
staff	0	0.2	0.18	0.16	0.15	0	0	0.17	0	0	0	0	0	0	0
food	0	0.15	0	0	0	0	0	0.21	0.15	0	0	0.17	0.16	0	0

(c)

words	the	manager		…		order	from	whatever	menu		…		waiter	and	the	restaurant		…	
staff	0	0	0	…	0	0	0	0	0.09	0	…	0	0.1	0	0	0.1	0	…	0
menu	0	0.12	0	…	0	0.11	0	0	0.14	0	…	0	0.1	0	0	0.1	0	…	0

(d)

FIGURE 5. Visualization of the AM-Weight-RoBERTa model attention weights, while the darker color means a higher weight in the attention vector. For each sub-figure, the attention weights are taken from the attention mask weight layer, and are separately calculated for each aspect. Sub-figure (a) and (b) are instances from the ATSA task, and (c) and (d) are taken from the ACSA task. Parameter γ takes 0.6 in all these cases.

Figure 3, the optimal value for γ is between 0.5 and 0.6 for both the ATSA and ACSA tasks. The result shows that relevant parts for a given aspect tend to gain similar attention scores. While weight assignments over half of the maximum ($\gamma = 0.5$) are meaningful, scores lower than that seems to bring in mostly noise. From Figure 4, 0.4-0.5 is the ideal range for β in the AM-Word model. Since each piece of review in the MAMS dataset contains at least two aspects, it is safe to assume that at most half of the content is directly related to an aspect. Although the peak values are very close for AM-Weight and AM-Word mechanisms, the curve is more stable in Figure 3. Overall, AM-Weight becomes a better choice when the optimal parameter is unknown, as it does not make any prior assumption what percentage of the context is relevant.

B. QUALITY ANALYSIS

Experiments in the previous section and sensitivity analysis show the average performance on the given datasets. Next we will look into individual cases how the masked attention mechanism works.

1) CASE STUDY

Figure 5 visualizes the output weights of the attention mask score layer in different aspects of four sentences with the AM-Weight-RoBERTa model. Figure 6 shows the selected attention words of the attention mask word layer in different aspects of the same sentences under the AM-Word-RoBERTa model.

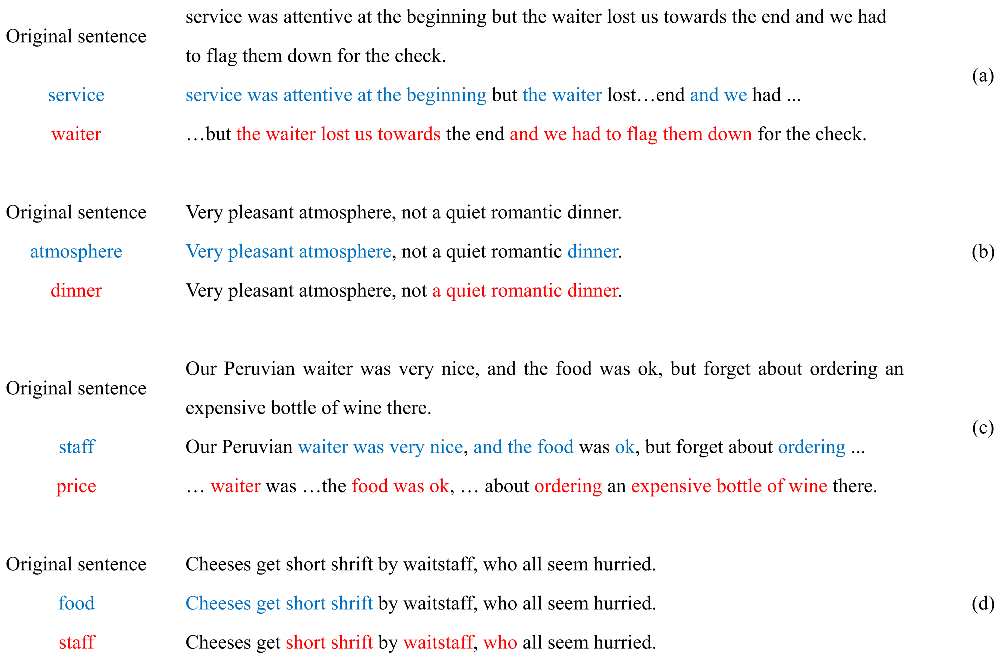
In Figure 5 (a) and (c) show that the AM-Weight-RoBERTa model accurately identifies relevant information and filters out irrelevant parts, with its attention mask mechanism on the ATSA and ACSA tasks. In (a), our model correctly finds the key descriptor "good" for the aspect term "salad" and "hardly ate" for "pasta". In (c), the key context of "staff" is identified as "waitress never asked us", and "needed" correctly matches aspect category "food". After passing the attention mask score layer, irrelevant information with low attention score is set to 0. It removes contextual noise to correctly predict the sentiment polarity of a given aspect.

In Figure 6 (a) and (c) show the same effect of the AM-Word-RoBERTa model, by accurately finding the relevant information and correctly predict the sentiments of the aspects on the ATSA and ACSA tasks. In (a), our model finds the key instances "service was attentive" for aspect term "service" and "the waiter lost us" for aspect term "waiter". In (c), the key instances "waiter was very nice" for aspect category "staff" and "expensive bottle of wine" for aspect category "price" are correctly identified. The irrelevant part is removed after passing the attention mask word layer, so that part will not enter the contextualized embedding layer of RoBERTa.

The AM-Weight mechanism filters by the attention weight. Figure 5 shows that effect by marking relevant parts with weight over γ times the maximum attention weight in the sentence. When the maximum attention weight is large, the attention weights are more focused, and AM-Weight has a better chance of limiting to the aspect-related words. AM-Word reserves a certain number of terms in the context, which is determined by the length of the sentence and β. That phenomenon can be observed from Figure 6, when the number of words in a sentence is large, more terms are kept, including highly relevant parts and remotely related context of the aspect.

2) ERROR ANALYSIS

The masked attention mechanism is not always right. In Figure 5 (b) the sentiment polarity toward "manager"



Original sentence	service was attentive at the beginning but the waiter lost us towards the end and we had to flag them down for the check.	(a)
service	service was attentive at the beginning but the waiter lost…end and we had ...	
waiter	…but the waiter lost us towards the end and we had to flag them down for the check.	
Original sentence	Very pleasant atmosphere, not a quiet romantic dinner.	(b)
atmosphere	Very pleasant atmosphere, not a quiet romantic dinner.	
dinner	Very pleasant atmosphere, not a quiet romantic dinner.	
Original sentence	Our Peruvian waiter was very nice, and the food was ok, but forget about ordering an expensive bottle of wine there.	(c)
staff	Our Peruvian waiter was very nice, and the food was ok, but forget about ordering ...	
price	… waiter was …the food was ok, … about ordering an expensive bottle of wine there.	
Original sentence	Cheeses get short shrift by waitstaff, who all seem hurried.	(d)
food	Cheeses get short shrift by waitstaff, who all seem hurried.	
staff	Cheeses get short shrift by waitstaff, who all seem hurried.	

FIGURE 6. **Visualization of the selected words in the AM-Word-RoBERTa model. In each sub-figure, different sentence components are selected for an aspect by the AM-Word mechanism. the selected part of each aspect marked with the same color as the aspect term or aspect category. Sub-figure (a) and (b) are instances of the ATSA task, and (c) and (d) are instances of from the ACSA task. β is 0.5 for ATSA task and o.4 for ACSA task.**

should be positive and "waiter" should be negative. However, AM-Weight-RoBERTa assigns neutral sentiments to both "manager" and "waiter". In Figure 5 (d), the correct sentiment toward "staff" is negative and "menu" should be neutral, but AM-Weight-RoBERTa also assigns a neutral label to "staff". Here the context is quite long for both aspect terms, and the mask erroneously filters out the right parts that are relevant for the sentiment judgment of the aspect. With only the remaining context terms, it is hard for the downstream classifier to identify the correct sentiment polarity.

In Figure 6 (b), AM-Word-RoBERTa assigns positive sentiment to "dinner" while the correct label should be negative, mainly because the keyword "not" is filtered out by the attention mask. Also in Figure 6 (d), the sentiments toward "food" should be neutral, but AM-Word-RoBERTa provides a positive judgment. These sentences are from the MAMS dataset, with multiple aspects in each sentence. Given the small β parameter in the AM-Word model, the number of candidate terms in relatively small, so the model is likely to ignore aspect-related context terms in the highly competitive selection. The AM-Weight model is more tolerant to a certain extent, as it keeps all contextual terms as long as they have sufficient attention weights.

When it comes to the choice between the weight-based and number-based masked mechanisms, the length of context is the main attribute for consideration. The AM-Weight model is usually a better option than AM-Word, as it tends to keep more candidates that have attention weights close to the maximum value. When the dataset mainly contains long sentences, the AM-Word is better with its fixed ratio of candidates. Masked attention has exhibited its power in filtering out irrelevant information, but it also runs the risk of losing relevant parts.

VI. CONCLUSION AND FUTURE WORK

Pre-trained language models show clear advantages over static embedding-based representations, as they have the ability of representing more complex context. They have displayed excellent performance in multiple tasks of NLP, including aspect-based sentiment analysis. However, early applications of BERT in ABSA tasks, often with complex custom-designed network structures, do not achieve significant improvements in comparison to a simple expansion of the base BERT model. That proves the representative power of pre-trained language models, which also sets a competitive baseline for further research.

In this paper, we propose two masked attention mechanisms, namely AM-Weight and AM-Word, and combine them with BERT-like language models for a better solution to the ABSA task. AM-Weight sets an attention weight threshold based on the maximum of attention scores, and AM-Word decides how many terms to keep based on a preset percentage of input. Both filter out the lower score parts that are assumed to be less relevant to the aspect of focus. By ignoring part of input that is claimed irrelevant, a large proportion of input noise is removed, keeping the downstream model more focused and reducing calculation cost. Experiments with multiple ABSA datasets prove the effectiveness of the masked attention mechanism, as both show significant improvements over the baseline BERT and RoBERTa model on ATSA and ACSA. We also analyze the sensitivity of their parameters with recommended optimal region. Results are examined with both success and failure cases. Depending on the length of input, preference is provided for the two options. AM-Weight usually works better for shorter sentences, and AM-Word is safer when the input is long.

From failure analysis, the masked attention mechanism does not work well when the input contains vague or semantically overlapping data, which is also one of the main challenges in sentiment analysis. Adaptive mask strategy selection and parameter setting will be what we focus on next, and incorporation of the word order information will be another key issue in improving the semantic expressiveness of attention-based language models. Besides, as a recent concept, the prompt framework [55] for sentiment analysis is proposed, which converts sentiment classification into a cloze-like task, making full use of the masked language model's representation power. It is an innovative paradigm in the NLP domain, and we will try modeling the ABSA task with that framework for richer in-depth semantic representation.

Recent research notes that a Transformer network with only self attention mechanism quickly degenerates to a rank-1 matrix, and the main remedy is the introduction of shortcut connections. Without the diversity of single-head networks brought in by the shortcuts, the expressiveness of an attention network is quite limited [56]. In this paper, we combine the Transformer-based language models with the mask attention mechanism, which also limits the complexity of attention distribution and introduces numerous context term combinations. It may be another cure for the rank collapse problem of Transformer networks, which requires further analysis and verification.

REFERENCES

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

[2] R. K. Bakshi, N. Kaur, R. Kaur, and G. Kaur, "Opinion mining and sentiment analysis," in *Proc. 3rd Int. Conf. Comput. Sustain. Global Develop. (INDIACom)*, Mar. 2016, pp. 452–455.

[3] B. Liu, "Sentiment analysis and opinion mining," *Synth. Lectures Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, May 2012.

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

[5] X. Yan, F. Jian, and B. Sun, "SAKG-BERT: Enabling language representation with knowledge graphs for Chinese sentiment analysis," *IEEE Access*, vol. 9, pp. 101695–101701, 2021.

[6] M. E. Basiri, S. Nemati, M. Abdar, E. Cambria, and U. R. Acharya, "ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis," *Future Gener. Comput. Syst.*, vol. 115, pp. 279–294, Feb. 2021.

[7] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "SemEval-2014 task 4: Aspect based sentiment analysis," in *Proc. 8th Int. Workshop Semantic Eval. (SemEval)*, 2014, pp. 27–35. [Online]. Available: https://www.aclweb.org/anthology/S14-2004

[8] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos, "SemEval-2015 task 12: Aspect based sentiment analysis," in *Proc. 9th Int. Workshop Semantic Eval. (SemEval)*, 2015, pp. 486–495.

[9] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. Al-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. D. Clercq, and V. Hoste, "Semeval-2016 task 5: Aspect based sentiment analysis," in *Proc. Int. Workshop Semantic Eval.*, 2016, pp. 19–30.

[10] C. Brun, D. N. Popa, and C. Roux, "XRCE: Hybrid classification for aspect-based sentiment analysis," in *Proc. 8th Int. Workshop Semantic Eval. (SemEval)*, 2014, pp. 838–842.

[11] S. Kiritchenko, X. Zhu, C. Cherry, and S. Mohammad, "NRC-Canada-2014: Detecting aspects and sentiment in customer reviews," in *Proc. 8th Int. Workshop Semantic Eval. (SemEval)*, 2014, pp. 437–442.

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

[13] Z. Toh and W. Wang, "DLIREC: Aspect term extraction and term polarity classification system," in *Proc. 8th Int. Workshop Semantic Eval. (SemEval)*, 2014, pp. 235–240.

[14] R. Piryani, V. Gupta, and V. Kumar Singh, "Generating aspect-based extractive opinion summary: Drawing inferences from social media texts," *Computación Y Sistemas*, vol. 22, no. 1, pp. 83–91, Mar. 2018.

[15] 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.

[16] D. Tang, B. Qin, X. Feng, and T. Liu, "Effective LSTMs for target-dependent sentiment classification," 2015, *arXiv:1512.01100*.

[17] M. Zhang, Y. Zhang, and D.-T. Vo, "Gated neural networks for targeted sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, vol. 30, no. 1, pp. 3087–3093, 2016.

[18] 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.

[19] Y. Tay, L. A. Tuan, and S. C. Hui, "Learning to attend via word-aspect associative fusion for aspect-based sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, pp. 1–9, 2018.

[20] 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.

[21] B. Huang and K. M. Carley, "Parameterized convolutional neural networks for aspect level sentiment classification," 2019, *arXiv:1909.06276*.

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

[23] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch. NIPS-W," in *Proc. 31st Conf. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, 2017, pp. 4–9.

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

[25] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.

[26] 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.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[28] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 5753–5763.

[29] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[30] D. Tang, B. Qin, and T. Liu, "Aspect level sentiment classification with deep memory network," 2016, *arXiv:1605.08900*.

[31] D. Ma, S. Li, X. Zhang, and H. Wang, "Interactive attention networks for aspect-level sentiment classification," 2017, *arXiv:1709.00893*.

[32] F. Ali, S. El-Sappagh, S. M. R. Islam, A. Ali, M. Attique, M. Imran, and K.-S. Kwak, "An intelligent healthcare monitoring framework using wearable sensors and social networking data," *Future Gener. Comput. Syst.*, vol. 114, pp. 23–43, Jan. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X1931605X

[33] F. Ali, S. El-Sappagh, and D. Kwak, "Fuzzy ontology and LSTM-based text mining: A transportation network monitoring system for assisting travel," *Sensors*, vol. 19, no. 2, p. 234, Jan. 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/2/234

[34] W. Li, W. Shao, S. Ji, and E. Cambria, "BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis," 2020, *arXiv:2006.00492*.

[35] F. Ali, A. Ali, M. Imran, R. A. Naqvi, M. H. Siddiqi, and K.-S. Kwak, "Traffic accident detection and condition analysis based on social networking data," *Accident Anal. Prevention*, vol. 151, Mar. 2021, Art. no. 105973. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S000145752100004X

[36] W. Xue and T. Li, "Aspect based sentiment analysis with gated convolutional networks," 2018, *arXiv:1805.07043*.

[37] S. Poria, E. Cambria, and A. Gelbukh, "Aspect extraction for opinion mining with a deep convolutional neural network," *Knowl.-Based Syst.*, vol. 108, pp. 42–49, Sep. 2016.

[38] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018, *arXiv:1802.05365*.

[39] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," Tech. Rep., 2018. [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

[40] M. Hu, S. Zhao, L. Zhang, K. Cai, Z. Su, R. Cheng, and X. Shen, "CAN: Constrained attention networks for multi-aspect sentiment analysis," 2018, *arXiv:1812.10735*.

[41] Y. Song, J. Wang, T. Jiang, Z. Liu, and Y. Rao, "Attentional encoder network for targeted sentiment classification," 2019, *arXiv:1902.09314*.

[42] H. Xu, B. Liu, L. Shu, and P. S. Yu, "BERT post-training for review reading comprehension and aspect-based sentiment analysis," 2019, *arXiv:1904.02232*.

[43] C. Sun, L. Huang, and X. Qiu, "Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence," 2019, *arXiv:1903.09588*.

[44] Q. Jiang, L. Chen, R. Xu, X. Ao, and M. Yang, "A challenge dataset and effective models for aspect-based sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 6281–6286.

[45] M. H. Phan and P. O. Ogunbona, "Modelling context and syntactical features for aspect-based sentiment analysis," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3211–3220.

[46] Y. Lin, C. Wang, H. Song, and Y. Li, "Multi-head self-attention transformation networks for aspect-based sentiment analysis," *IEEE Access*, vol. 9, pp. 8762–8770, 2021.

[47] 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, Lang. Process.*, vol. 28, pp. 2538–2551, 2020.

[48] A. Zhao and Y. Yu, "Knowledge-enabled BERT for aspect-based sentiment analysis," *Knowl.-Based Syst.*, vol. 227, Sep. 2021, Art. no. 107220. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705121004822

[49] Z. Gao, A. Feng, X. Song, and X. Wu, "Target-dependent sentiment classification with BERT," *IEEE Access*, vol. 7, pp. 154290–154299, 2019.

[50] Y. Tai, J. Yang, X. Liu, and C. Xu, "MemNet: A persistent memory network for image restoration," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4539–4547.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[52] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 1480–1489.

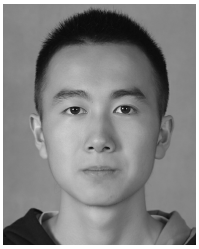
[53] X. Li, L. Bing, W. Zhang, and W. Lam, "Exploiting BERT for end-to-end aspect-based sentiment analysis," 2019, *arXiv:1910.00883*.

[54] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[55] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," 2021, *arXiv:2107.13586*.

[56] Y. Dong, J.-B. Cordonnier, and A. Loukas, "Attention is not all you need: Pure attention loses rank doubly exponentially with depth," 2021, *arXiv:2103.03404*.

AO FENG received the B.E. and M.E. degrees in automation from Tsinghua University, China, in 1999 and 2001, respectively, and the M.S. and Ph.D. degrees in computer science from the University of Massachusetts Amherst, USA, in 2008. He has worked at Amazon.com and Lenovo Research. He is currently an Associate Professor with the Department of Computer Science, Chengdu University of Information Technology. His research interests include information retrieval, data mining, natural language processing, and machine learning.



XUELEI ZHANG received the B.E. degree from Yibin University, Yibin, China, in 2019. He is currently pursuing the M.E. degree in computer technology with the Chengdu University of Information Technology. His research interests include sentiment analysis, text classification, and deep learning.

XINYU SONG received the B.E. degree from Heilongjiang Bayi Agricultural University, Daqing, China, in 2017, and the B.E. degree in computer technology from the Chengdu University of Information Technology, Chengdu, China, in 2021. His research interests include sentiment analysis, information extraction, and deep learning.