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, China
Corresponding author: Xuelei Zhang (e-mail: zhangxuelei@163.com).

This work was supported in part by the Research Innovation Team Fund (award no. 18TD0026) from the Department of Education, Sichuan Province, and in part by the Sichuan Science and Technology Program (project no. 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 [2] [11] [4] [3] 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] [8] [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] [11] [12], and its advantage lies in the ability to capture aspect-specific sentiment information. 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.

With the development of deep learning, a large number of neural network models [15] [16] [28] [18] [29] [17] [19] are
trained language models, standard attention scores for each sentence. One is called Attention Mask Weight (AM-Weight), which sets a weight threshold to filter out irrelevant parts. 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 [23] and GloVe [24], 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 [35], GPT [36], and BERT [25]. 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 [27], 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
Attentional Encoder Network (AEN-BERT) [38] uses BERT achieved significant improvement over it in our experiment.

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) [14] 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) [13] 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) [34] 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) [37] 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) [38] 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 [39] transforms the aspect sentiment classification problem into a special machine reading comprehension problem. BERT-pair-QA [40] 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 [41] 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) [42] 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) [43] 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) [44] utilizes certain prior knowledge to guide the capsule attention process to solve the ABSA task. Knowledge-enabled BERT [45] 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

\[ A = \{A_1, A_2, ..., A_L\} \] (1)

are predefined, and the sentiment polarities are restricted to

\[ P = \{Positive, Negative, Neutral\} \] (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

\[ S = \{w_1, w_2, ..., w_i, ..., w_n\} \] (3)

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

\[ T^S = \{T^S_1, T^S_2, ..., T^S_M\} \] (4)

and each target from the sentence \( S \), is represented as

\[ T^S_i = \{w_i, w_{(i+1)}, ..., w_{(i+m_i-1)}\} \] (5)
with $m_i$ terms. The goal of the ATSC task is to predict the sentiment polarities for M targets, $P^T_M$ represents the sentiment polarity of the $T^*_M$ in the sentence:

$$P^T = \{P^T_1, P^T_2, \ldots, P^T_M\}$$  \hspace{1cm} (6)

Similarly, if a sentence contains N aspect categories, $A^S$ the ACSC task predicts the sentiment polarities for each of them, $P^A_N$ represents the sentiment polarity of the $A^S_N$ in the sentence:

$$P^A = \{P^A_1, P^A_2, \ldots, P^A_N\}$$  \hspace{1cm} (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 aspect and its context, both are encoded in the same way.

2) Attention layer

The attention layer takes the output of the contextualized embedding layer as input. In this layer, we calculate the attention [22][49] 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 vectors, and then apply one of the two mask mechanisms, 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 $\gamma$ 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.

$$\alpha_{ij} = \begin{cases} 0 & \alpha_{ij} < \gamma_{max}w \\ \alpha_{ij} & \alpha_{ij} \geq \gamma_{max}w \end{cases}$$  \hspace{1cm} (12)

where $\gamma$ is the configurable parameter of the AM-Weight model, $\gamma_{max}$ is the maximal attention score in a sentence, and $\alpha_{ij}$ is the original attention score.

4) Attention mask word

When using the other attention mask mechanism in ABSA, a more intuitive method is adopted. $\beta$ 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.

$$sorted_{\alpha_{ij}} = \{\alpha_{ir_1}, \alpha_{ir_2}, \ldots, \alpha_{ir_{\beta n}}, \ldots, \alpha_{ir_n}\}$$  \hspace{1cm} (13)

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

$$\alpha_{ij} = \begin{cases} \alpha_{ij} & \alpha_{ir_1} \geq \alpha_{ij} \geq \alpha_{ir_{\beta n}} \\ 0 & \alpha_{ir_{\beta n}} > \alpha_{ij} \geq \alpha_{ir_n} \end{cases}$$  \hspace{1cm} (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 $\beta n$ words for contextualized embedding using BERT or RoBERTa, and perform aspect sentiment classification with their representation.

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

$$z_i = \sum_{j=1}^{n} \alpha_{ij} (V^j W^V)$$  \hspace{1cm} (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. $W^Q, W^K, W^V \in \mathbb{R}^{d_x \times d_e}$ are attention parameter matrices, separately tuned for each layer and attention head.
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.

6) **Loss**

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

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

**A. TASK DEFINITION**

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

**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

\(^1\)http://alt.qcri.org/semeval2014/task4/
TABLE 1. Statistics of the SemEval-2014 dataset.

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

TABLE 2. Statistics of the MAMS dataset.

| Dataset   | Positive | Negative | Neutral | Total |
|-----------|----------|----------|---------|-------|
| ATSA      |          |          |         |       |
| Train     | 3380     | 2764     | 5042    | 11186 |
| Validation| 403      | 325      | 607     | 1332  |
| Test      | 400      | 329      | 607     | 1336  |
| ACSA      |          |          |         |       |
| Train     | 1929     | 2084     | 1613    | 7090  |
| 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    |

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
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.
where the mask percentage $\beta$ takes the same range. From Figure 3, the optimal value for $\gamma$ 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 $\beta$ 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 "price" are correctly identified. The irrelevant part is removed "staff" and "expensive bottle of wine" for 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-Weight-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 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.

**V. DATA ANALYSIS**

**A. PARAMETER SENSITIVITY**

Two new parameters $\gamma$ and $\beta$ 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 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 3 shows the macro-F1 value of AM-Weight-RoBERTa, with mask ratio $\gamma$ in the [0.1, 0.9] range. Figure 4 displays the change of performance of AM-Word-RoBERTa, with weight over $\gamma$ times the maximum attention weight in the
Peruvian waiter was very nice, and the food was ok, but forget about ordering an

| words | the | salad | was | good | but | my | friends | hardly | ate | their | pasta |
|-------|-----|-------|-----|------|-----|----|---------|--------|-----|-------|-------|
|       | 0.18| 0.22  | 0.2 | 0.19 | 0   | 0  | 0       | 0      | 0   | 0     | 0     |

![Figure 6](image)

The masked attention mechanism is not always right. In Figure 5 (b) the sentiment polarity toward “manager” should be positive and “waiter” should be negative. However, AM-Weight-RoBERTa assigns neutral sentiments to both “man-

---

**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 $\gamma$ takes 0.6 in all these cases.

**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. $\beta$ is 0.5 for ATSA task and 0.4 for ACSA task.

| 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 us ... 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. |
|------------------|------------------------------------------------------|
| 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. |
|------------------|---------------------------------------------------------------------------------------------------------------|
| 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 shift by waitstaff, who all seem hurried. |
|------------------|-----------------------------------------------------------|
| food             | Cheeses get short shift by waitstaff, who all seem hurried. |
| staff            | Cheeses get short shift by waitstaff, who all seem hurried. |

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 $\beta$. 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” should be positive and “waiter” should be negative. However, AM-Weight-RoBERTa assigns neutral sentiments to both “man-
agger” 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 $\beta$ 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 longer.

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 [52] 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 [53]. 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 Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, pp. 168–177.
[2] R. K. Bakshi, N. Kaur, R. Kaur, and G. Kaur, “Opinion mining and sentiment analysis,” in 2016 3rd international conference on computing for sustainable global development (INDIACom). IEEE, 2016, pp. 452–455.
[3] B. Liu, “Sentiment analysis and opinion mining,” Synthesis lectures on human language technologies, vol. 5, no. 1, pp. 1–167, 2012.
[4] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, “Target-dependent twitter sentiment classification,” in Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, 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. 101 695–101 701, 2021.
[6] M. E. Basiri, S. Nemati, M. Abdar, E. Cambria, and U. R. Acharya, “Abdcm: An attention-based bidirectional cnn-rnn deep model for sentiment analysis,” Future Generation Computer Systems, vol. 115, pp. 279–294, 2021.
[7] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, “SemEval-2014 task 4: Aspect based sentiment analysis,” in Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland.
Association for Computational Linguistics, Aug. 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 Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), 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. De Clercq et al., “Semeval-2016 task 5: Aspect based sentiment analysis,” in International workshop on semantic evaluation, 2016, pp. 19–30.

[10] C. Brum, D. N. Popa, and C. Roux, “Xece: Hybrid classification for aspect-based sentiment analysis,” in SemEval@ COLING. Citeseer, 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 Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), 2014, pp. 437–442.

[12] J. Wagner, P. Aroa, S. C. Vaillo, U. Barman, D. Bogdanova, J. Foster, and L. Tounsi, “Icu: Aspect-based polarity classification for exameval task 4,” in Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014, pp. 223–229.

[13] Y. Wang, M. Huang, X. Zhu, and L. Zhao, “Attention-based lstm for aspect-level sentiment classification,” in Proceedings of the 2016 conference on empirical methods in natural language processing, 2016, pp. 606–615.

[14] D. Tang, B. Qin, X. Feng, and T. Liu, “Effective lstms for target-dependent sentiment classification,” arXiv preprint arXiv:1512.01100, 2015. [Online]. Available: https://arxiv.org/abs/1512.01100

[15] M. Zhang, Y. Zhang, and D.-T. Vo, “Gated neural networks for targeted sentiment analysis,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30, no. 1, 2016.

[16] S. Wang, S. Mazumder, B. Liu, M. Zhou, and Y. Chang, “Target-sensitive memory networks for aspect sentiment classification,” in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 957–967.

[17] Y. Tay, L. A. Tuan, and S. C. Hu, “Learning to attend via word-aspect associative fusion,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, 2018.

[18] P. Chen, Z. Sun, L. Bing, and W. Yang, “Recurrent attention network on memory for aspect sentiment analysis,” in Proceedings of the 2017 conference on empirical methods in natural language processing, 2017, pp. 452–461.

[19] B. Huang and K. M. Carley, “Parameterized convolutional neural networks for aspect level sentiment classification,” arXiv preprint arXiv:1909.06276, 2019. [Online]. Available: https://arxiv.org/abs/1909.06276

[20] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] 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,” 2017.

[22] F. Ali, S. El-Sappagh, S. 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 Generation Computer Systems, vol. 114, pp. 23–43, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X1931605X

[23] 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, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/2/234

[24] W. Li, W. Shao, S. Ji, and E. Cambria, “Bieru: bidirectional emotional recurrent unit for conversational sentiment analysis,” arXiv preprint arXiv:2006.00492, 2020. [Online]. Available: https://arxiv.org/abs/2006.00492

[25] 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 network data,” Accident Analysis & Prevention, vol. 151, p. 105973, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S000145752100004X

[26] W. Xue and T. Li, “Aspect based sentiment analysis with gated convolutional networks,” arXiv preprint arXiv:1805.07043, 2018. [Online]. Available: https://arxiv.org/abs/1805.07043

[27] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” arXiv preprint arXiv:1802.05365, 2018. [Online]. Available: https://arxiv.org/abs/1802.05365

[28] A. Radford, K. Narasimhan, S. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” 2018.

[29] M. Hu, S. Zhao, L. Zhang, K. Cai, Z. Su, R. Cheng, and X. Shen, “Can: Constrained attention networks for multi-aspect sentiment analysis,” arXiv preprint arXiv:1812.10735, 2018. [Online]. Available: https://arxiv.org/abs/1812.10735

[30] Y. Song, J. Wang, T. Jiang, Z. Liu, and Y. Rao, “Attentional encoder network for targeted sentiment classification,” arXiv preprint arXiv:1902.09314, 2019. [Online]. Available: https://arxiv.org/abs/1902.09314

[31] H. Xu, B. Liu, L. Shu, and P. S. Yu, “Bert post-training for review reading comprehension and aspect-based sentiment analysis,” arXiv preprint arXiv:1904.02232, 2019. [Online]. Available: https://arxiv.org/abs/1904.02232

[32] C. Sun, L. Huang, and X. Qin, “Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence,” arXiv preprint arXiv:1903.09588, 2019. [Online]. Available: https://arxiv.org/abs/1903.09588

[33] P. Jiang, L. Chen, R. Xu, A. Oo, and M. Yang, “A challenge dataset and effective models for aspect-based sentiment analysis,” in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 6281–6286.

[34] M. H. Phan and P. O. Ogunbona, “Guided capsule attention networks for aspect-based sentiment analysis,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3211–3220.

[35] 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.

[36] 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 Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2358–2351, 2020.

[37] A. Zhao and Y. Yu, “Knowledge-enabled bert for aspect-based sentiment analysis,” Knowledge-Based Systems, vol. 227, p. 107220, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705121004822

[38] Z. Gao, A. Feng, X. Song, and X. Wu, “Target-dependent sentiment classification with bert,” IEEE Access, vol. 7, pp. 154 290–154 299, 2019.

[39] Y. Tai, J. Yang, X. Li, and C. Xu, “Memnet: A persistent memory network for review reading comprehension and aspect-based sentiment analysis,” in Proceedings of the 5th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 606–615.

[40] 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, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/2/234

[41] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014. [Online]. Available: https://arxiv.org/abs/1412.6980
Feng et al.: Masked Attention Mechanism Focuses Better on Relevant Parts in Aspect-Based Sentiment Analysis

[49] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical attention networks for document classification,” in Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, 2016, pp. 1480–1489.

[50] X. Li, L. Bing, W. Zhang, and W. Lam, “Exploiting bert for end-to-end aspect-based sentiment analysis,” arXiv preprint arXiv:1910.00883, 2019. [Online]. Available: https://arxiv.org/abs/1910.00883

[51] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” arXiv preprint arXiv:1409.0473, 2014. [Online]. Available: https://arxiv.org/abs/1409.0473

[52] 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,” arXiv preprint arXiv:2107.13586, 2021. [Online]. Available: https://arxiv.org/abs/2107.13586

[53] Y. Dong, J.-B. Cordonnier, and A. Loukas, “Attention is not all you need: Pure attention loses rank doubly exponentially with depth,” arXiv preprint arXiv:2103.03404, 2021. [Online]. Available: https://arxiv.org/abs/2103.03404

AO FENG received the B.E. and M.E. degrees in Automation from Tsinghua University, China in 1999 and 2001, respectively. He received his M.S. and Ph.D. degrees in Computer Science from University of Massachusetts Amherst, U.S. in 2008. He has worked at Amazon.com and Lenovo research, and is currently an associate professor at 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, where he is currently pursuing the M.E. degree in computer technology in 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 received the B.E. degree in computer technology with the Chengdu University of Information Technology, Chengdu, China, in 2021. His research interests include sentiment analysis, information extraction, and deep learning.

***