# Aspect Feature Distillation and Enhancement Network for Aspect-based Sentiment Analysis

Rui Liu
Institute of Information Engineering,
Chinese Academy of Sciences
School of Cyber Security, University of
Chinese Academy of Sciences
Beijing, China
liurui3221@iie.ac.cn

Jiahao Cao
Institute of Information Engineering,
Chinese Academy of Sciences
School of Cyber Security, University of
Chinese Academy of Sciences
Beijing, China
caojiahao@iie.ac.cn

Nannan Sun
Institute of Information Engineering,
Chinese Academy of Sciences
School of Cyber Security, University of
Chinese Academy of Sciences
Beijing, China
sunnannan@iie.ac.cn

Lei Jiang*
Institute of Information Engineering,
Chinese Academy of Sciences
School of Cyber Security, University of
Chinese Academy of Sciences
Beijing, China
jianglei@iie.ac.cn

## ABSTRACT

Aspect-based sentiment analysis (ABSA) is a fine-grained sentiment analysis task designed to identify the polarity of a target aspect. Some works introduce various attention mechanisms to fully mine the relevant context words of different aspects, and use the traditional cross-entropy loss to fine-tune the models for the ABSA task. However, the attention mechanism paying partial attention to aspect-unrelated words inevitably introduces irrelevant noise. Moreover, the cross-entropy loss lacks discriminative learning of features, which makes it difficult to exploit the implicit information of intra-class compactness and inter-class separability. To overcome these challenges, we propose an Aspect Feature Distillation and Enhancement Network (AFDEN) for the ABSA task. We first propose a dual-feature extraction module to extract aspect-related and aspect-unrelated features through the attention mechanisms and graph convolutional networks. Then, to eliminate the interference of aspect-unrelated words, we design a novel aspect-feature distillation module containing a gradient reverse layer that learns aspect-unrelated contextual features through adversarial training, and an aspect-specific orthogonal projection layer to further project aspect-related features into the orthogonal space of aspect-unrelated features. Finally, we propose an aspect-feature enhancement module that leverages supervised contrastive learning to capture the implicit information between the same sentiment labels and between different sentiment labels. Experimental results on three public datasets demonstrate that our AFDEN model achieves state-of-the-art performance and verify the effectiveness and robustness of our model.

*Lei Jiang is the corresponding author.



*SIGIR '22, July 11–15, 2022, Madrid, Spain*

ACM ISBN 978-1-4503-8732-3/22/07.
https://doi.org/10.1145/3477495.3531938

## CCS CONCEPTS

• **Information systems** → **Sentiment analysis**.

## KEYWORDS

Aspect-based sentiment analysis, Orthogonal projection, Adversarial training, Supervised contrastive learning

**ACM Reference Format:**
Rui Liu, Jiahao Cao, Nannan Sun, and Lei Jiang. 2022. Aspect Feature Distillation and Enhancement Network for Aspect-based Sentiment Analysis. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3477495.3531938

## 1 INTRODUCTION

Aspect-based sentiment analysis (ABSA) is a fine-grained sentiment classification task, which aims to infer the sentiment polarity (i.e., positive, neutral or negative) of a given aspect in the whole sentence. It has attracted extensive attention in recent years. Different from traditional text classification or sentence-level sentiment analysis, there may be several aspects in a sentence, and different aspects may have different contexts and sentiment polarities. For example, in a sentence *"The ambience was nice, but service wasn't so great."*, the sentiment for *"ambience"* is positive, while the sentiment for *"service"* is negative. Therefore, it is necessary to mine the context words related to the target aspect to predict its sentiment polarity. However, how to capture the relevant context words of different aspects and make full use of their information is very challenging.

Some works utilize various attention mechanisms [3, 7, 10, 13, 20, 25, 27, 35, 43, 46] to model the semantic relevance of the target aspect and its context words to predict its sentiment polarity. Although these works have achieved good performance, the inherent defects existing in the attention mechanism still introduce a lot of noise for the ABSA task. For example, in the sentence *"The ambience was nice, but service wasn't so great."*, when the aspect is *"ambience"*, the attention mechanism still assigns weights to the unrelated words like *"wasn't so great"*. Although the attention scores of unrelated context words may be small, they inevitably introduce interference. This may lead to some deviation between the aspect feature representation and the standard feature representation, especially when there are multiple aspect words in a sentence.

Besides, most existing works [5, 34, 38, 42, 47] use standard cross-entropy loss to fine-tune their models for the ABSA task. However, the traditional cross-entropy loss lacks the discriminative learning of features [23], ignoring the implicit information of intra-class compactness and inter-class separability. It cares more about the accuracy between the label and the prediction, so it is difficult to capture the potential information within the correct labels and between the correct and incorrect labels. Therefore, in the feature space, the learned aspect features are loose in the same category, while the boundaries between different categories are not clear, which is not good for the ABSA task.

To solve the above problems, we propose a novel architecture from a new perspective named aspect feature distillation and enhancement network (AFDEN), as shown in Figure 1. Specifically, we first design a dual-feature extraction module to extract aspect-related features and aspect-unrelated features respectively. To eliminate the interference of aspect-unrelated words, inspired by [30], we propose an aspect-feature distillation module containing a gradient reverse layer (GRL) that learns aspect-unrelated contextual features through adversarial training, and an aspect-specific orthogonal projection layer (AS-OPL) to further project aspect-related features into the orthogonal space of aspect-unrelated features. In this way, we distill out pure aspect-related features and remove aspect-unrelated features.

Furthermore, to exploit the implicit information of labels ignored by traditional cross-entropy loss, we design an aspect-feature enhancement module that leverages supervised contrastive learning [17] to enhance the representation of aspect-related features after distillation. Supervised contrastive learning can reduce the distances between positive samples and increase the distances between positive and negative samples, which makes it easy to capture the implicit label information. Therefore, through the aspect-feature enhancement module, the same sentiment representations are more centralized in the feature space and the boundaries between the different sentiment representations are more clear, which is more conducive to the ABSA task. We evaluate our method on the benchmark dataset Semeval2014 [29] and the Twitter dataset [6]. Moreover, we verify the effectiveness and robustness of our method on MAMS [15] and ARTS datasets [44].

Our contributions are highlighted as follows:

- We propose an aspect-feature distillation module containing a GRL and an AS-OPL for the ABSA task. The GRL encourages the network to better learn the aspect-unrelated features through adversarial training, whereas the AS-OPL eliminates the interference of aspect-unrelated features through projecting aspect-related features out of the orthogonal space of aspect-unrelated features.
- We design an aspect-feature enhancement module that leverages supervised contrastive learning to learn the implicit information within the same sentiment labels and between the different sentiment labels. This module captures the latent label information missing from the cross-entropy loss, enhancing the sentiment discriminability of aspect features.
- We conduct extensive experiments on the SemEval2014 and Twitter datasets. The experimental results demonstrate the effectiveness of our AFDEN. In addition, our results on the MAMS dataset and ARTS robust dataset also verify the effectiveness and robustness of our model.

## 2 RELATED WORK

Aspect-based sentiment analysis task is an entity-level oriented classification task. Compared with the traditional document level and sentence level sentiment classification, it is a fine-grained sentiment classification task and needs deeper semantic understanding. The early research of aspect sentiment classification mainly used the traditional machine learning algorithm to study this task as a text classification problem. For example, [14, 19, 28, 41] designed bag-of-words, sentiment lexicon and other features to train the support vector machine (SVM) sentiment classifier. However, feature engineering is labor-intensive, and the results are highly dependent on the quality of features, so it is easy to reach the performance bottleneck.

With the development of deep learning, neural networks have greatly promoted the development of aspect sentiment analysis, such as CNN [11, 16, 18, 22], RNNs [1, 36], memory networks [37], because neural networks can automatically learn the low dimensional and continuous features of aspects and their contexts. For example, [36] divided the context into left and right sides of aspect words, modeled the two parts respectively with two LSTMs, and spliced the aspect information with the input word embedding to obtain the sentiment representation of aspect words. Then, the standard cross-entropy loss and softmax layer are used to obtain the prediction results.

However, the model solely based on RNN can not well capture the relationship between aspect words and sentiment polarity words or phrases in sentences, so the attention mechanism is introduced [3, 7, 10, 13, 20, 25, 27, 35, 43, 46] to solve this problem. [43] encoded sentences and given aspect words with LSTM, processed the hidden layer output with attention mechanism, and obtained the sentiment polarity representation of aspect words. [27] calculated not only the attention distribution of sentence hidden layer output but also the attention distribution of aspect words. [7] used a multi granularity attention mechanism to capture word-level interactions between aspects and their contexts. However, the weighting of word-level features by attention mechanism may introduce a lot of noise and reduce the prediction accuracy. Although the attention scores of aspect-unrelated words may be small or even negligible, they will also lead to some deviations between sentiment representation and standard representation.

Recent studies have explored the use of graph neural networks (GNNs) to learn the representation of dependency tree, combined with syntactic aware graph structure [2, 12, 21, 24, 34, 38, 42, 47, 48] to solve this task, and achieved attractive results. [47] introduced aspect-specific graph convolutional networks (ASGCN) and used dependency graph to deal with aspect level sentiment classification tasks. [38] proposed a dependency graph enhanced dual-transformer network (DGEDT), which jointly considers the flat representations learned from Transformer and graph-based representations learned from the dependency graph. [42] effectively encoded grammatical information by reshaping and pruning an ordinary dependency parse tree, and proposed a relational graph attention network (R-GAT) to encode a new tree structure for aspect sentiment prediction. [21] proposed a dual graph convolutional networks (DualGCN) model, which considered the complementarity of syntax structure and semantic correlations simultaneously. Using dependency-based parse tree can provide more comprehensive syntactic information. However, due to the imperfect parsing performance and the randomness of input sentences, it is inevitable to introduce noise information through the dependency tree. Besides, graph convolutional networks have poor ability to model long-distance or incoherent words in dependency trees.

In addition, pre-trained language models such as BERT [5] have achieved good performance in many NLP tasks and have also achieved good results in the field of aspect-based sentiment analysis. [33] transformed ABSA task into sentence pair classification task by constructing auxiliary sentences. [45] proposed a post-training method for BERT to improve the performance during the fine-tuning phase of ABSA tasks. [4] compared induced trees and dependent parsing trees from pre-trained language models on several popular models of ABSA tasks, and found that induced trees from fine-tuned RoBERTa (FT-Roberta) outperformed the parser-provided tree, suggesting that pre-trained language models could learn better implicit task-oriented syntactic information.

# 3 METHODOLOGY

## 3.1 Task Definition

Given an aspect $A$ and the corresponding sentence $S$, aspect-based sentiment classification aims to identify the sentiment polarity $y \in \{positive, negative, neutral\}$ of this aspect, where the sentence $S = [\omega_1, \dots, \omega_{a+1}, \dots, \omega_{a+m}, \dots, \omega_n]$ is a sequence consisting of $n$ words, and $A = [\omega_{a+1}, \dots, \omega_{a+m}]$ stands for the specific aspect with $m$ words.

## 3.2 Overview

Figure 1 provides an overview of our AFDEN model. We first construct a sentence-aspect pair of "[CLS] sentence [SEP] aspect [SEP]" as the input of BERT encoder to obtain aspect-aware hidden representations of the sentence. Then, the representations are input into a dual-feature extraction module to obtain the aspect-related and aspect-unrelated features with rich semantics and syntax information. To eliminate the interference of aspect-unrelated features and separate aspect-related features from the context, a novel aspect-feature distillation module is proposed with a gradient reverse layer (GRL) and an aspect-specific orthogonal projection layer (AS-OPL). The GRL helps the learning of aspect-unrelated features through adversarial training, while the AS-OPL further projects aspect-related features into the orthogonal space of aspect-unrelated features. Furthermore, an aspect-feature enhancement module with supervised contrastive learning is designed to capture the implicit information of different sentiment representations. The training procedure of our AFDEN is depicted in Algorithm 1. Next, we will elaborate on the details of our proposed AFDEN model.

**Algorithm 1** Training procedure of AFDEN

**Input:**
  Batch size $N$ and number of training epochs $t$;
  Sentence-aspect pairs $(x, a)$ from dataset $X$;
  The rate of supervised contrastive learning $\alpha$.
**Output:**
  The predictions $Y_{AR}$ of inputs $(x, a)$.
  Initialize the parameters for the BERT encoder $F$, the ARGCN $G_1$ and the AUGCN $G_2$, the self-attentions $ATT_1$ and $ATT_2$, the classifiers $C_{AR}$ and $C_{AU}$;
  **for** epoch=1 to t **do**
    **for** sampled minibatch $\{(x_k,\ a_k)\}_{k=1}^{N}$ **do**
      $h = F(x_k, a_k)$;
      $h_{AR} = G_1(ATT_1(h), h), \quad h_{AU} = G_2(ATT_2(h), h)$;
      $h_{mask} = h_{AU} \cdot mask$;
      $h_{GRL} = GRL(h_{mask}), \ h_{OPL} = OPL(h_{AR}, h_{AU})$;
      $Y_{AR} = C_{AR}(h_{OPL}), \quad Y_{AU} = C_{AU}(h_{GRL})$;
      Calculate $\mathcal{L}_{AR}$ and $\mathcal{L}_B^{sup}$ as Eq.(12) and (14);
      Define $\mathcal{L}_1 = \alpha\mathcal{L}_B^{sup} + (1-\alpha)\mathcal{L}_{AR}$;
      Calculate $\mathcal{L}_2$ as Eq.(7);
      Update network AFDEN to minimize $\mathcal{L}_1, \mathcal{L}_2$;
    **end for**
  **end for**
  **return** $Y_{AR}$

## 3.3 Dual-feature Extraction

To extract aspect-related and aspect-unrelated features, we design a dual feature extraction module with two self-attention mechanisms, two graph convolutional networks (GCNs) and an aspect-specific masking layer. AR-GCN and AU-GCN have the same structure, but do not share parameters. We feed the attention matrix containing semantic information obtained by self-attention mechanisms and the aspect-aware hidden representations containing rich semantic and syntactic information obtained by BERT encoder into two GCNs. To better obtain aspect-oriented features, we add an aspect-specific masking layer after the AU-GCN to shield the influence of other words.

**Self-attention** The self-attention mechanism [40] can fully consider the semantic connections between different words in a sentence by computing the attention score of each pair of elements in parallel. Therefore, we use the self-attention mechanism to calculate the score matrix $A \in R^{n\times n}$ for the representations obtained by BERT. Then the score matrix is fed into graph convolutional networks as the adjacency matrix, it can be expressed as:

$$A = softmax(\frac{QW^Q \times (KW^K)^T}{\sqrt{d}}) \tag{1}$$

where matrices $Q$ and $K$ are both equal to the graph representation of the previous layer and they are initialized to the output representation of BERT in our model. $W^Q$ and $W^K$ are both learnable weight matrices, and $d$ is the dimension of the input node features.

**Graph Convolutional Networks (GCN)** The core idea of graph convolutional networks is to learn a function map. Through the node $v_i$ in the mapping graph, a new representation of the node $v_i$ can be generated by aggregating its own feature $x_i$ and

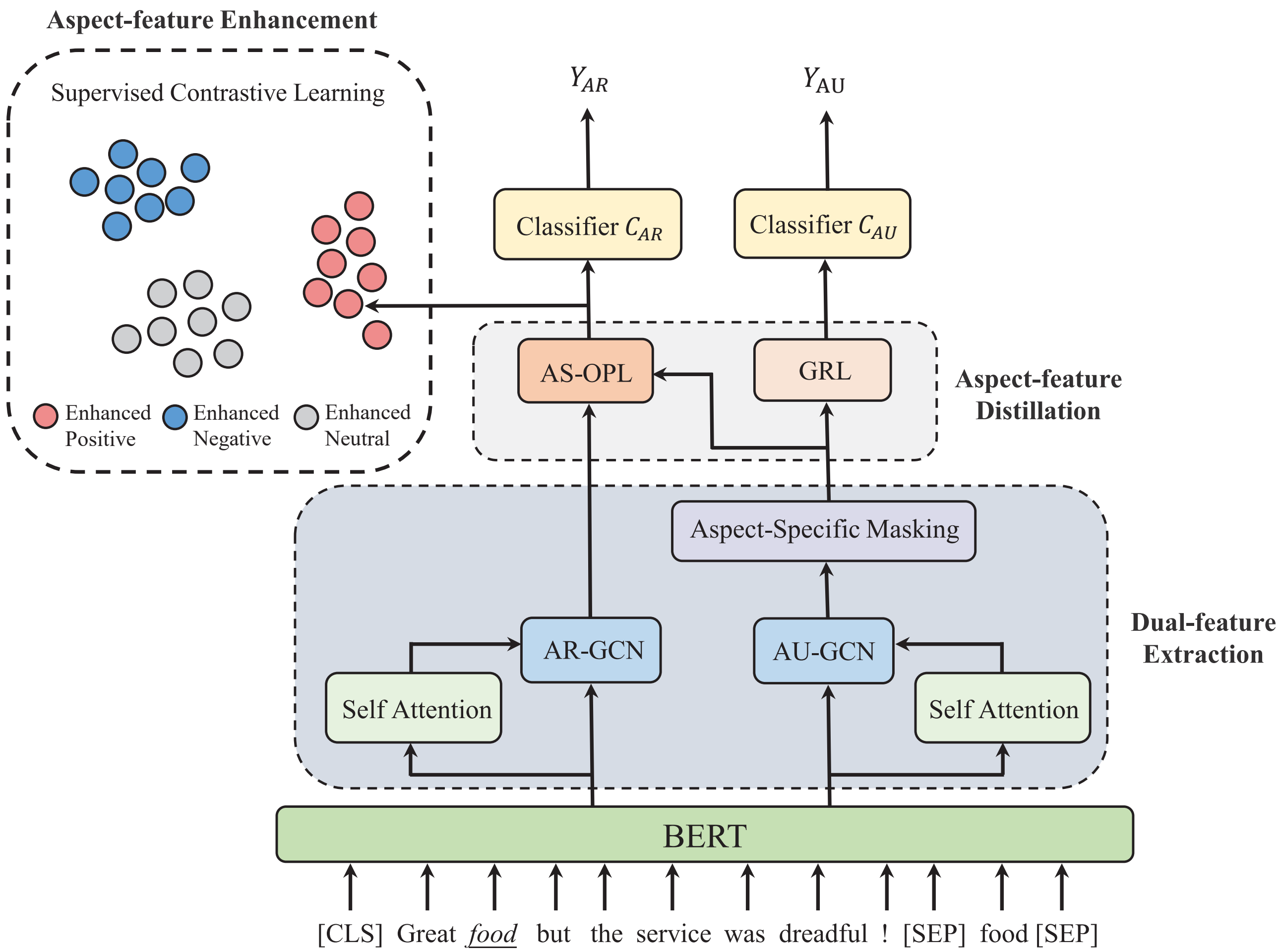



**Figure 1: The overall architecture of AFDEN, which is mainly composed of dual-feature extraction module, aspect-feature distillation module and aspect-feature enhancement module. The details of our model are described in the main text.**

the feature of its neighbor nodes $x_j$ ($j \in N(v_i)$), so each node can learn more contextual representations. For the $L$-layer GCN, $l \in [1, 2, \cdots, L]$. For the $i$-th node of the $l$-th layer of the graph convolutional network, its hidden state representation $h_i^l$ is updated by the following formula:

$$h_i^l = \sigma(\sum_{j=1}^{n} A_{ij} W^l h_j^{l-1} + b^l) \tag{2}$$

where $A_{ij}$ is the adjacency matrix, and $A_{ij}$ is the attention matrix obtained by the self-attention mechanism. $W^l$ is a weight matrix. $b^l$ is a bias term. They are all parametric learnable. $\sigma$ is a ReLU activation function.

**Aspect-specific Masking** In this layer, we mask the hidden state vectors of non-aspect words and keep the state of aspect words unchanged. It can be formulated as:

$$h_t^L = \begin{cases} 0 & 1 \le t < a+1,\ a+m < t \le n \\ h_t^L & a+1 \le t \le a+m \end{cases} \tag{3}$$

Because of the graph convolution in the previous step, the aspect-specific hidden state vector already contains the contextual information related to the aspect. Therefore the output $H_{mask}^L$ obtained by aspect-specific masking is aspect-oriented, and it can be expressed as $H_{mask}^L = \{0, \cdots, h_{a+1}^L, \cdots, h_{a+m}^L, \cdots, 0\}$.

### 3.4 Aspect-feature Distillation

To better separate aspect-related and aspect-unrelated information, we propose the aspect-feature distillation module. We design a gradient reverse layer (GRL) to extract aspect-unrelated features, and an aspect-specific orthogonal projection layer (AS-OPL) to distill out pure aspect-related features from the orthogonal projection space of aspect-unrelated features.

**Gradient Reverse Layer (GRL)** To enable the AU-GCN to better learn aspect-unrelated features, we insert a gradient reverse layer (GRL) [8] between aspect-specific masking and the classifier $C_{AU}$ to achieve gradient reverse, and realize the interaction of two-way information through aspect-specific orthogonal projection layer. Then it can create a confrontation with AR-GCN in the

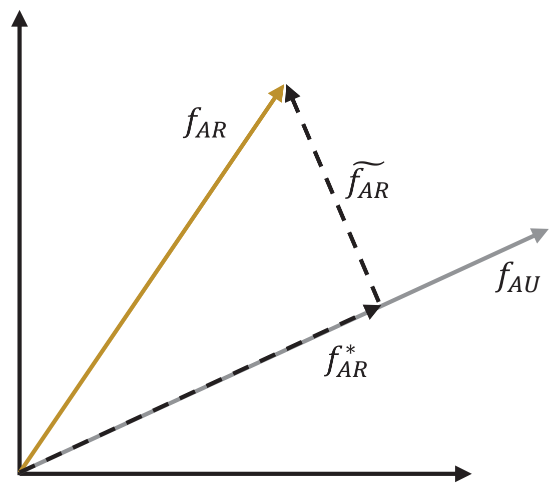



**Figure 2: The principle of AS-OPL.**

dual-feature extraction module, so that the classifier $C_{AU}$ can not predict the sentiment polarity of the embedding correctly while the classifier $C_{AR}$ can predict the sentiment polarity as correctly as possible. Finally, through adversarial training, AU-GCN can learn aspect-unrelated features while AR-GCN can learn aspect-related features as much as possible.

Specifically, GRL acts as an identity transformation during forward propagation, but during back-propagation GRL takes gradients from subsequent layers, changes the sign by multiplying $-\lambda$, and passes it to the previous layer. Mathematically, we can describe the forward and back-propagation behavior of GRL with the following formula:

$$GRL(x) = x \tag{4}$$

$$\frac{\partial GRL(x)}{\partial x} = -\lambda I \tag{5}$$

where $\lambda$ is a hyperparameter and $I$ is an identity matrix. Finally, the output $\widetilde{f_{AU}}$ of GRL is sent to the classifier $C_{AU}$ to obtain the prediction result:

$$Y_{AU} = softmax(\widetilde{f_{AU}} \cdot W_{AU} + b_{AU}) \tag{6}$$

$$\mathcal{L}_{AU} = CrossEntropy(y_{truth}, y_{AU}) \tag{7}$$

where $W_{AU}$ and $b_{AU}$ are the weights and biases of the $C_{AU}$, respectively. They are both parameter-learnable.

**Aspect-specific Orthographic Projection Layer (AS-OPL)** Intuitively, the aspect-related vector representation should be orthogonal to the aspect-unrelated vector representation in a sentence. Therefore, we utilize AS-OPL to project the extracted aspect-related features to the orthogonal direction of the aspect-unrelated features. The aspect-related feature projection space retains the information that is more conducive to the correct sentiment classification of the target aspect, and removes the irrelevant aspect information that does not help or even interferes with the aspect-based sentiment classification. Figure 2 illustrates the principle of AS-OPL from two-dimensional space. Mathematically, we first project the aspect-related feature vector $f_{AR}$ into the direction of the aspect-unrelated feature vector $f_{AU}$:

$$f^*_{AR} = Proj(f_{AR},\ f_{AU}) \tag{8}$$

where $Proj$ is a projection function,

$$Proj(x, y) = \frac{x \cdot y}{|y|}\frac{y}{|y|} \tag{9}$$

where $x$, $y$ are vectors. Then we project in the orthogonal direction of the projected feature $f^*_{AR}$ to get a purer aspect-based classification feature vector:

$$\widetilde{f_{AR}} = Proj(f_{AR},\ (f_{AR} - f^*_{AR})) \tag{10}$$

Following [30], we use Eq.(8) and Eq.(10) to obtain purer aspect-related feature representations after distillation. Specifically, we first extract the aspect-unrelated feature $f^*_{AR}$ in $f_{AR}$ by projecting the original aspect-related feature representation $f_{AR}$ to the direction of the aspect-unrelated feature representation $f_{AU}$ by Eq.(8). Then, we project $f_{AR}$ to the direction of $f_{AR} - f^*_{AR}$ through Eq.(10), that is, the direction perpendicular to $f^*_{AR}$, and obtain the feature representation orthogonal to $f^*_{AR}$. This feature representation is a further purified aspect-related feature obtained by eliminating the interference and redundant information of the aspect-unrelated feature $f^*_{AR}$ in $f_{AR}$. Finally, this purified aspect-related feature vector $\widetilde{f_{AR}}$ is sent to the classifier $C_{AR}$ to obtain the prediction result:

$$Y_{AR} = softmax(\widetilde{f_{AR}} \cdot W_{AR} + b_{AR}) \tag{11}$$

$$\mathcal{L}_{AR} = CrossEntropy(y_{truth}, y_{AR}) \tag{12}$$

where $W_{AR}$ and $b_{AR}$ are the weights and biases of the $C_{AR}$, respectively, and they are both parameter-learnable.

## 3.5 Aspect-feature Enhancement

To further enhance the sentiment representations of the aspect-related embedding, we propose the aspect-feature enhancement module to help the final aspect-based sentiment classification. We introduce supervised contrastive learning in this module. Supervised contrastive learning enables embeddings with the same sentiment label to be close to each other, and embeddings with different sentiment labels to stay away, which is useful for learning high-quality sentiment representations. Specifically, for $((S_i, A_i), y_i)$ in a batch $B$, we first obtain the purified aspect-related sentiment representation $\widetilde{f_{AR}}$ for the sentence-aspect pair through the previous feature extraction and feature distillation modules. We let $z_i = \widetilde{f_{AR}}$, and the supervised contrastive loss in the batch $B$ can be defined as:

$$P^{sup}_B(i, c) = \frac{exp(sim(z_i, z_c)/\tau)}{\sum_{b \in B, b \neq i} exp(sim(z_i, z_b)/\tau)} \tag{13}$$

$$\mathcal{L}^{sup}_B = \sum_{i \in B} -\log \frac{1}{C_i} \sum_{y_i = y_c, c \neq i} P^{sup}_B(i, c) \tag{14}$$

where $P^{sup}_B(i, c)$ indicates the likelihood that $z_c$ is most similar to $z_i$ and $\tau$ is a scalar temperature parameter. We simply use $sim(z_i, z_c) = z_i \cdot z_c$ to measure the similarity. $\mathcal{L}^{sup}_B$ is a supervised contrastive loss computed for each sentiment representation in $B$, where $C_i$ is the number of samples in $B$ in the same sentiment label $y_i$, and $C_i = |\{c | y_c = y_i,\ c \neq i\}|$.

# 4 EXPERIMENTS

## 4.1 Datasets

We conduct experiments on three public standard datasets. The Restaurant and Laptop datasets are from SemEval2014 task 4 [29], consisting of reviews on the restaurant and laptop domains. The Twitter dataset is originally built by [6] containing twitter posts. Furthermore, we also use a more challenging dataset, Multi-Aspect Multi-Sentiment (MAMS) [15], which shares the same domain to the SemEval2014 Restaurant Review dataset. Each sentence contains at least two different aspects with different sentiment polarities in MAMS. All these datasets have three sentiment polarities: positive, negative and neutral. Each sentence in these datasets is annotated with the aspects and their corresponding polarities. Statistics for the three datasets and the MAMS dataset are shown in Table 1.

## 4.2 Implementation Details

For our AFDEN, we use the bert-base-uncased English version as the encoder. We use AdamW [26] as the optimizer for BERT and set the learning rate to $2 \times 10^{-5}$. To alleviate overfitting, we apply dropout at a rate of 0.1 to BERT. The dropout rates of AR-GCN and AU-GCN are both set to 0.1, and the number of AR-GCN and AU-GCN layers is both set to 2. In GRL, the hyper-parameters $\lambda$ swept [0.05, 0.1, 0.2, 0.4, 0.8, 1.0]. The ratios of cross-entropy loss $\mathcal{L}_{AR}$ and supervised contrastive loss $\mathcal{L}_B^{sup}$ in $\mathcal{L}_1$ are (0.6, 0.4), (0.8, 0.2) and (0.6, 0.4) on the Restaurant, Laptop and Twitter datasets, respectively. The temperature parameter $\tau$ of supervised contrastive learning is 0.14, 0.19 and 0.08 on the three datasets, respectively. The AFDEN model is trained in 20 epochs with a batch size of 32, and the maximum sequence length is set to 80 during the training.

## 4.3 Baseline Methods

To comprehensively evaluate our AFDEN model, we compare it with state-of-the-art baselines. The models are briefly described as follows.
1) **BERT-SPC** [5] constructs the sentence-aspect pair input "[CLS] sentence [SEP] aspect [SEP]" into the basic BERT of sentence pair classification task, and takes the representation of [CLS] for prediction.
2) **AEN+BERT** [32] uses BERT as the encoder and employs an attention encoder network to model between context and aspect words.
3) **BERT-PT** [45] adopts a joint post-training method on BERT to post-train the weights of BERT for multi-task fine-tuning.
4) **TD-BERT** [9] proposes a target-dependent BERT that takes the localization output at the aspect word as the classification input instead of the first [CLS] label.
5) **CapsNet+BERT** [15] combines BERT and capsule network for ABSA task.
6) **SDGCN-BERT** [49] proposes a multi-aspect sentiment classification framework that utilizes GCN to effectively capture sentiment dependencies between different aspects in a sentence.
7) **R-GAT+BERT** [42] obtains an aspect-oriented dependency tree structure by reshaping and pruning, and uses a relational graph attention network to encode a new dependency tree for this task.

**Table 1: Statistics on four datasets of ABSA.**

| Dataset | Division | #Positive | #Negative | #Neutral |
|---|---|---|---|---|
| Restaurant | Train | 2164 | 807 | 637 |
| | Test | 728 | 196 | 196 |
| Laptop | Train | 994 | 870 | 464 |
| | Test | 341 | 128 | 169 |
| Twitter | Train | 1561 | 1560 | 3127 |
| | Test | 173 | 173 | 346 |
| MAMS | Train | 3380 | 2764 | 5042 |
| | Test | 400 | 329 | 607 |

8) **DGEDT+BERT** [38] jointly considers flat representation and graph-based representation through the mutual biaffine module, and proposes a dependency graph enhanced dual-transformer network.
9) **BERT-ADA** [31] first fine-tunes BERT with self-supervised domain-specific data, followed by supervised task-specific fine-tuning.
10) **DualGCN+BERT** [21] proposes a dual graph convolutional network model that considers both syntactic structure and semantic correlation.

## 4.4 Comparison Results

We use the accuracy and macro-averaged F1-score as the main evaluation metrics to evaluate the ABSA models. The main experimental results are shown in Table 2. Our AFDEN model achieves state-of-the-art performance with accuracies of 87.41%, 82.13% and 78.47% on the Restaurant, Laptop, and Twitter datasets, respectively. These results suggest that our model can sufficiently distill out the aspect-related features and enhance them for ABSA tasks. Compared with attention-based methods such as AEN+BERT and R-GAT+BERT, our AFDEN model eliminates the interference of aspect-unrelated contexts, so it can well avoid noises introduced by the attention mechanism. Moreover, compared with DGEDT+BERT, DualGCN+BERT and other syntactic-based methods, our model achieves better performance without introducing additional syntactic knowledge.

## 4.5 Ablation Study

To further investigate the role of different modules in our AFDEN model, we conduct extensive ablation studies on each module separately. The results are shown in Table 3. AFDEN w/o AFE means that we remove the Aspect-feature Enhancement module, so that the model can not enhance the aspect feature representation after distillation by learning the implicit information between the same and different sentiment labels. Therefore, the performance degrades significantly on all three datasets. AFDEN w/o DFE indicates that we have deleted the Dual-feature Extraction module, which means that we directly distill and enhance the aspect feature of the output representation from BERT. Similarly, AFDEN w/o AFD denotes that we remove the Aspect-feature Distillation module and no longer remove the interfering information from the aspect-unrelated context. The experimental results show that our aspect-feature distillation module can remove the interference of aspect-unrelated features well and learn purer aspect sentiment

**Table 2: Experimental results comparison on three publicly available datasets.**

| **Models** | **Rest14** | | **Lap14** | | **Twitter** | |
|---|---|---|---|---|---|---|
| | **Accuracy** | **Macro-F1** | **Accuracy** | **Macro-F1** | **Accuracy** | **Macro-F1** |
| BERT-SPC | 84.46 | 76.98 | 78.99 | 75.03 | 73.55 | 72.14 |
| AEN+BERT | 83.12 | 73.76 | 79.93 | 76.31 | 74.71 | 73.13 |
| BERT-PT | 84.95 | 76.96 | 78.07 | 75.08 | - | - |
| TD-BERT | 85.10 | 78.40 | 78.90 | 74.40 | 76.70 | 74.30 |
| CapsNet+BERT | 85.09 | 77.75 | 78.21 | 73.34 | - | - |
| SDGCN-BERT | 83.57 | 76.47 | 81.35 | 78.34 | - | - |
| R-GAT+BERT | 86.60 | 81.35 | 78.21 | 74.07 | 76.15 | 74.88 |
| DGEDT+BERT | 86.30 | 80.00 | 79.80 | 75.60 | 77.90 | 75.40 |
| BERT-ADA | 87.14 | 80.05 | 79.19 | 74.18 | - | - |
| DualGCN+BERT | 87.13 | 81.16 | 81.80 | 78.10 | 77.40 | 76.02 |
| Our AFDEN | **87.41** | **82.21** | **82.13** | **78.81** | **78.47** | **77.27** |

**Table 3: Experimental results of ablation study**

| Models | Rest14 | | Lap14 | | Twitter | |
|---|---|---|---|---|---|---|
| | Accuracy | Macro-F1 | Accuracy | Macro-F1 | Accuracy | Macro-F1 |
| AFDEN w/o AFE | 86.16 | 80.14 | 79.62 | 76.19 | 75.72 | 73.75 |
| AFDEN w/o DFE | 86.16 | 80.25 | 78.68 | 74.71 | 75.87 | 74.82 |
| AFDEN w/o AFD | 86.07 | 79.67 | 79.93 | 76.34 | 76.30 | 74.99 |
| AFDEN | **87.41** | **82.21** | **82.13** | **78.81** | **78.47** | **77.27** |

correlation representation. Overall, our AFDEN with all modules achieves the best performance.

## 4.6 Case Study

Table 4 shows some cases for aspect sentiment prediction using different models. The symbols $P$, $N$ and $O$ represent positive, negative and neutral sentiments respectively. The red and blue colors in the table represent the aspect words that need to be predicted. For the first sentence, when the aspect is *"focacchia bread"*, all three baselines predict it as positive, and only our AFDEN predicts it as neutral. For the attention-based model AEN+BERT, it tends to focus on the noisy words *"to die for"*, which means excellent. Besides, for the DualGCN+BERT model, although the syntactic dependency provides some direct connections between the target aspect and some words, the complexity of sentences and the instability of dependency parsing performance may lead to the deviation of aspect and its expressions. Compared with other models, our AFDEN can directly eliminate the interference of aspect-unrelated words and obtain aspect-related opinion expressions more accurately. The following cases also fully demonstrate that our AFDEN model can capture the relevant features of the target aspect more effectively, and obtain more accurate prediction results.

## 4.7 Visualization for Aspect-related Features

To more intuitively verify the effectiveness of our model, we visualize the embedding distribution of aspect-related features with T-distributed Stochastic Neighbor Embedding (t-SNE) [39], which is a nonlinear dimensionality reduction algorithm. It is very suitable for reducing high-dimensional data to two or three dimensions for visualization. We take the final high-dimensional aspect-related feature representations for visualization. Figure 3 and Figure 4 show the visualization results on the Laptop and Restaurant datasets, respectively. The red, green and blue dots represent positive, neutral and negative aspect-related feature representations, respectively. Figure 3(a) shows that the aspect-related feature enhancement module is removed and only the standard cross-entropy loss is used to fine-tune our model. It can be seen that the distributions of the same sentiment embeddings are relatively loose, and the distance between the three different sentiments is comparatively small. Figure 3(b) shows the embedding distributions of our AFDEN model. The Embeddings within the same sentiment are more aggregated, and the boundaries between different sentiments are more distinct, which is more conducive to the ABSA task. In addition, the temperature coefficient $\tau$ in supervised contrastive learning can moderate the degree of attention to difficult samples. To investigate the effect of temperature coefficients, we conduct experiments on three datasets, as shown in Figure 5. The model achieves the best performance when the temperature coefficients are 0.14, 0.19 and 0.08 on Restaurant, Laptop and Twitter, respectively.

## 4.8 Aspect Robustness Study

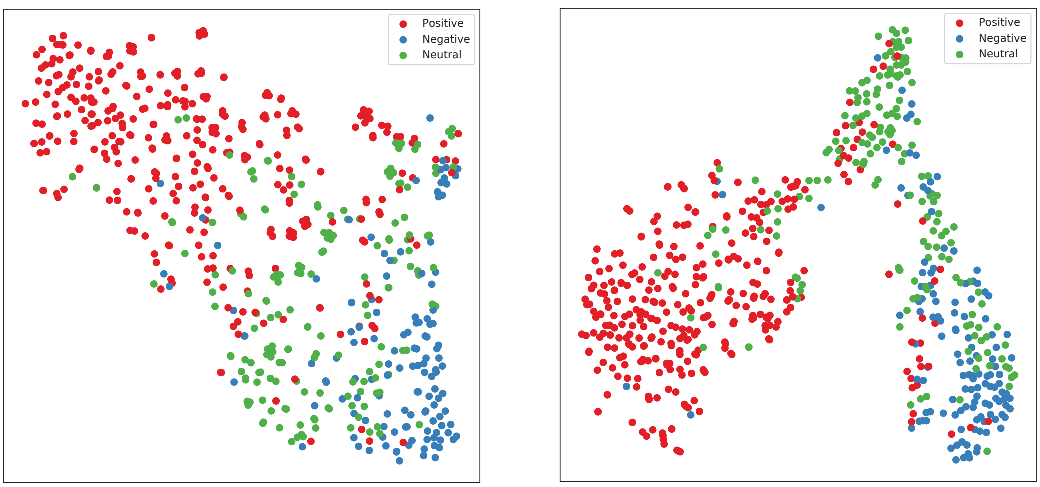
To analyze the performance of our AFDEN in aspect robustness, we use Aspect Robustness Test Set (ARTS) [44] for testing. The testsets apply several perturbations to the reviews from Restaurants and Laptops. The perturbations include reversing the original sentiment of the target aspect (REVTGT), perturbing the sentiments of the non-target aspects (REVNON) and generating more non-target aspect terms that have opposite sentiment polarities to the target (ADDDIFF). The testsets are designed to probe whether the models

**Table 4: Case studies of our AFDEN model compared with other baselines.**

| # | Review | AEN+BERT | BERT-SPC | DualGCN+BERT | AFDEN |
|---|---|---|---|---|---|
| 1 | They are served on focacchia bread and are to die for . | ($P_{\times}$, $P_{\times}$) | ($P_{\times}$, $P_{\times}$) | ($P_{\times}$, $P_{\times}$) | ($O_{\checkmark}$, $O_{\checkmark}$) |
| 2 | Great beer selection too , something like 50 beers . | $P_{\times}$ | $P_{\times}$ | $P_{\times}$ | $O_{\checkmark}$ |
| 3 | I do not like too much windows 8 . | $P_{\times}$ | $P_{\times}$ | $P_{\times}$ | $N_{\checkmark}$ |
| 4 | A beautiful atmosphere , perfect for drinks and / or appetizers . | ($P_{\times}$, $P_{\times}$) | ($P_{\times}$, $P_{\times}$) | ($P_{\times}$, $P_{\times}$) | ($P_{\times}$, $O_{\checkmark}$) |
| 5 | It's good to go there for drinks if you don't want to get drunk because you'll be lucky if you can get one drink an hour . | ($N_{\times}$, $N_{\times}$) | ($P_{\times}$, $P_{\times}$) | ($P_{\times}$, $O_{\checkmark}$) | ($O_{\checkmark}$, $O_{\checkmark}$) |

**Table 5: Model performance on Aspect Robustness Test Set (ARTS). We compare the model accuracy on the original and new testsets, and calculate the accuracy decline of prediction between them.**

| Models | Restaurant-ARTS | | Laptop-ARTS | |
|---|---|---|---|---|
| | Ori→New | Decline | Ori→New | Decline |
| AEN+BERT | 83.12→25.45 | -57.67 | 79.93→30.09 | -49.84 |
| BERT-SPC | 83.04→54.82 | -29.22 | 77.59→50.94 | -26.65 |
| CapsNet+BERT | 83.48→55.36 | -28.12 | 77.12→25.86 | -51.46 |
| BERT-PT | 86.70→59.29 | -27.41 | 78.53→53.29 | -25.24 |
| DualGCN+BERT | 87.13→63.57 | -23.56 | 81.80→57.99 | -23.81 |
| AFDEN | **87.41→65.18** | **-22.23** | **82.13→59.87** | **-22.26** |




(a) AFDEN w/o AFE (b) Our AFDEN

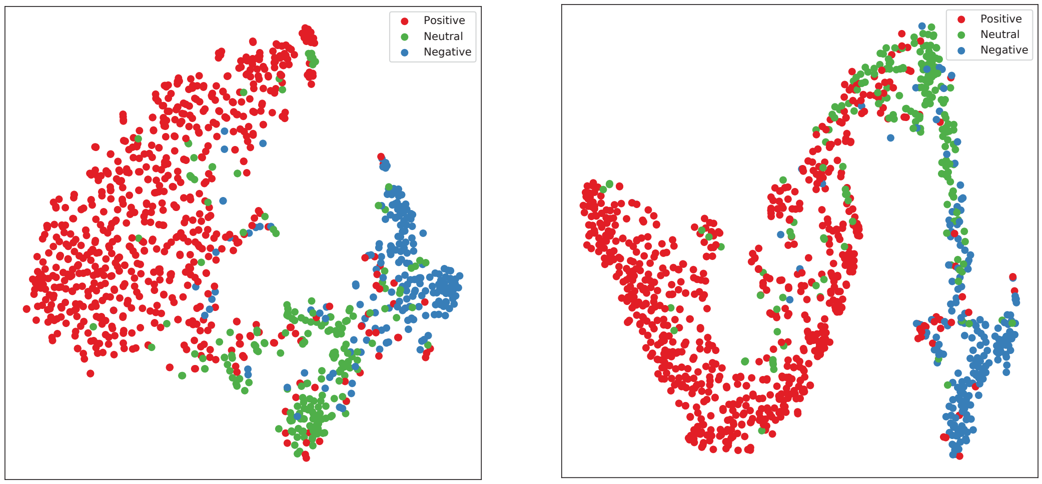
**Figure 3: The visualization of aspect-related embeddings on Laptop dataset.**

can distinguish the sentiment of the target aspect from the non-target aspects and aspect-unrelated information.

Table 5 lists the performance of the tested models, among which our AFDEN model achieves the optimal results, which fully verifies the aspect robustness of our model. Compared to the obvious performance degradation of the baseline models, AFDEN experiences a 22.23% and 22.26% decrease on Restaurant and Laptop. The results show that the perturbation of aspect words can be more robust by using AFDEN. This is mainly because our model can fully explore the relationship between aspect words and context information, and remove the interference of aspect-unrelated context through aspect feature distillation, so the negative effects of perturbation can be avoided to some extent. Moreover, the aspect-feature enhancement module can learn the implicit information between the same label and between the different labels, so it is more robust to the noise generated by perturbation.

## 4.9 Multi-aspect Effectiveness Study

The performance of baselines and our AFDEN in the MAMS dataset is shown in Table 6. Compared with the three datasets of Restaurant, Laptop and Twitter, where most sentences contain only one aspect or multiple aspects with the same sentiment, each sentence in the MAMS dataset contains at least two aspect words and at least two aspects in the same sentence have different sentiment




(a) AFDEN w/o AFE (b) Our AFDEN

**Figure 4: The visualization of aspect-related embeddings on Restaurant dataset.**

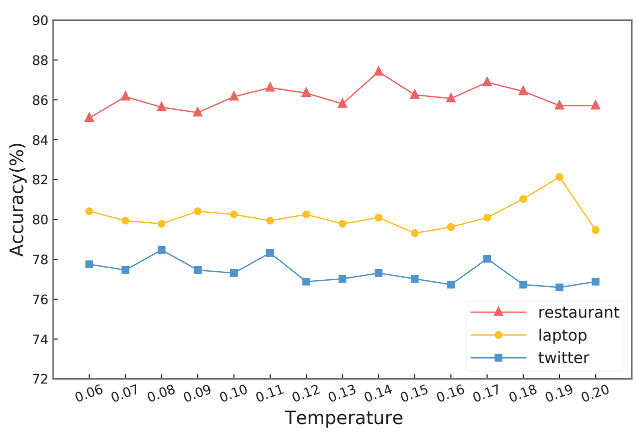



**Figure 5: Effect of different temperatures on three datasets.**

polarities. This is extremely challenging for the ABSA task. The results show that our AFDEN achieves state-of-the-art performance. The efficiency of our model can be attributed to the distillation and enhancement of aspect features, because they remove the interference of aspect-unrelated features and make it easier to distinguish the relevant context of different aspects.

## 5 CONCLUSION

In this paper, we propose an AFDEN architecture to address the disadvantages of the attention mechanism and the traditional cross-entropy loss for the ABSA task. To eliminate the interference of aspect-unrelated features, our AFDEN model first extracts the aspect-related and aspect-unrelated features through the dual-feature extraction module, and then distills out the aspect-related features through the aspect-feature distillation module. The aspect-feature distillation module contains the GRL that learns aspect-unrelated features through adversarial training, and the AS-OPL to further project aspect-related features into the orthogonal space of aspect-unrelated features. Moreover, to effectively capture the implicit label information, we design the aspect-feature enhancement module that leverages supervised contrastive learning to further enhance the representations of the pure aspect-related features. Extensive experiments on the benchmark datasets, the MAMS dataset and the ARTS dataset show that our AFDEN model has better performance and robustness than the baseline models.

**Table 6: Model performance on MAMS**

| Models | MAMS | |
|---|---|---|
| | Accuracy | Macro-F1 |
| AEN | 66.72 | - |
| CapsNet | 79.78 | - |
| AEN+BERT | 72.08 | 71.46 |
| BERT-SPC | 82.22 | - |
| CapsNet+BERT | 83.39 | - |
| AFDEN | **85.33** | **84.73** |

## ACKNOWLEDGMENTS

This paper is supported by Pilot Projects of Chinese Academy of Sciences (No.Y9W0013401).

## REFERENCES

[1] Giuseppe Castellucci, Simone Filice, Danilo Croce, and Roberto Basili. 2014. Unitor: Aspect based sentiment analysis with structured learning. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*. 761–767.

[2] Chenhua Chen, Zhiyang Teng, and Yue Zhang. 2020. Inducing Target-Specific Latent Structures for Aspect Sentiment Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 5596–5607.

[3] Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*. 452–461.

[4] Junqi Dai, Hang Yan, Tianxiang Sun, Pengfei Liu, and Xipeng Qiu. 2021. Does syntax matter? A strong baseline for Aspect-based Sentiment Analysis with RoBERTa. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1816–1829.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[6] Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)*. 49–54.

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

[8] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research* 17, 1 (2016), 2096–2030.

[9] Zhengjie Gao, Ao Feng, Xinyu Song, and Xi Wu. 2019. Target-dependent sentiment classification with BERT. *IEEE Access* 7 (2019), 154290–154299.

[10] Shuqin Gu, Lipeng Zhang, Yuexian Hou, and Yin Song. 2018. A position-aware bidirectional attention network for aspect-level sentiment analysis. In *Proceedings of the 27th international conference on computational linguistics*. 774–784.

[11] Binxuan Huang and Kathleen Carley. 2018. Parameterized Convolutional Neural Networks for Aspect Level Sentiment Classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 1091–1096. https://doi.org/10.18653/v1/D18-1136

[12] Binxuan Huang and Kathleen M Carley. 2019. Syntax-Aware Aspect Level Sentiment Classification with Graph Attention Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5469–5477.

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

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

[15] Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. 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)*. 6280–6285.

[16] Rie Johnson and Tong Zhang. 2015. Semi-supervised convolutional neural networks for text categorization via region embedding. *Advances in neural information processing systems* 28 (2015), 919.

[17] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. *Advances in Neural Information Processing Systems* 33 (2020).

[18] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1746–1751. https://doi.org/10.3115/v1/D14-1181

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

[20] Lishuang Li, Yang Liu, and AnQiao Zhou. 2018. Hierarchical attention based position-aware network for aspect-level sentiment analysis. In *Proceedings of the 22nd conference on computational natural language learning*. 181–189.

[21] Ruifan Li, Hao Chen, Fangxiang Feng, Zhanyu Ma, Xiaojie Wang, and Eduard Hovy. 2021. Dual graph convolutional networks for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 6319–6329.

[22] Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018. Transformation Networks for Target-Oriented Sentiment Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 946–956. https://doi.org/10.18653/v1/P18-1087

[23] Zheng Lian, Ya Li, Jianhua Tao, and Jian Huang. 2018. Speech emotion recognition via contrastive loss under siamese networks. In *Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data*. 21–26.

[24] Bin Liang, Rongdi Yin, Lin Gui, Jiachen Du, and Ruifeng Xu. 2020. Jointly Learning Aspect-Focused and Inter-Aspect Relations with Graph Convolutional Networks for Aspect Sentiment Analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*. 150–161.

[25] Jiangming Liu and Yue Zhang. 2017. Attention modeling for targeted sentiment. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. 572–577.

[26] Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

[27] Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive Attention Networks for Aspect-Level Sentiment Classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (Melbourne, Australia) *(IJCAI'17)*. AAAI Press, 4068–4074.

[28] Veronica Perez-Rosas, Carmen Banea, and Rada Mihalcea. 2012. Learning Sentiment Lexicons in Spanish.. In *LREC*, Vol. 12. Citeseer, 73.

[29] Maria Pontiki, Dimitrios Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. 27–35.

[30] Qi Qin, Wenpeng Hu, and Bing Liu. 2020. Feature projection for improved text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 8161–8171.

[31] Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2020. Adapt or Get Left Behind: Domain Adaptation through BERT Language Model Finetuning for Aspect-Target Sentiment Classification. In *Proceedings of the 12th Language Resources and Evaluation Conference*. 4933–4941.

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

[33] Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 380–385.

[34] Kai Sun, Richong Zhang, Samuel Mensah, Yongyi Mao, and Xudong Liu. 2019. Aspect-level sentiment analysis via convolution over dependency tree. 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)*. 5679–5688.

[35] Xingwei Tan, Yi Cai, and Changxi Zhu. 2019. Recognizing conflict opinions in aspect-level sentiment classification with dual attention networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3426–3431.

[36] Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016. Effective LSTMs for Target-Dependent Sentiment Classification. In *Proceedings of COLING 2016, the 26th International ConferenceComputational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, 3298–3307. https://aclanthology.org/C16-1311

[37] Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect Level Sentiment Classification with Deep Memory Network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 214–224. https://doi.org/10.18653/v1/D16-1021

[38] Hao Tang, Donghong Ji, Chenliang Li, and Qiji Zhou. 2020. Dependency graph enhanced dual-transformer structure for aspect-based sentiment classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 6578–6588.

[39] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[41] Duy-Tin Vo and Yue Zhang. 2015. Target-dependent twitter sentiment classification with rich automatic features. In *Twenty-fourth international joint conference on artificial intelligence*.

[42] Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. 2020. Relational Graph Attention Network for Aspect-based Sentiment Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 3229–3238.

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

[44] Xiaoyu Xing, Zhijing Jin, Di Jin, Bingning Wang, Qi Zhang, and Xuan-Jing Huang. 2020. Tasty Burgers, Soggy Fries: Probing Aspect Robustness in Aspect-Based Sentiment Analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 3594–3605.

[45] Hu Xu, Bing Liu, Lei Shu, and S Yu Philip. 2019. BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2324–2335.

[46] Min Yang, Wenting Tu, Jingxuan Wang, Fei Xu, and Xiaojun Chen. 2017. Attention based LSTM for target dependent sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[47] Chen Zhang, Qiuchi Li, and Dawei Song. 2019. Aspect-based Sentiment Classification with Aspect-specific Graph Convolutional Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 4568–4578. https://doi.org/10.18653/v1/D19-1464

[48] Mi Zhang and Tieyun Qian. 2020. Convolution over Hierarchical Syntactic and Lexical Graphs for Aspect Level Sentiment Analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 3540–3549.

[49] Pinlong Zhao, Linlin Hou, and Ou Wu. 2020. Modeling sentiment dependencies with graph convolutional networks for aspect-level sentiment classification. *Knowledge-Based Systems* 193 (2020), 105443.