





NNDF: A New Neural Detection Network for Aspect-Category Sentiment Analysis

Lijian Li^{1,2} , Yuanpeng He^{1,3} , and Li Li¹

¹ College of Computer and Information Science College of Software,
Southwest University, Chongqing 400715, China

{lljllj, hypppyh010403}@email.swu.edu.cn, lily@swu.edu.cn

² School of Engineering, The Hong Kong University of Science and Technology,
Hong Kong 999077, China

³ School of Computer Science, Peking University, Peking 100871, China

Abstract. Aspect-category sentiment analysis (ACSA) is crucial for capturing and understanding sentiment polarities of aspect categories hidden behind in sentences or documents automatically. Nevertheless, existing methods have not modeled semantic dependencies of aspect terms and specified entity's aspect category in sentences. In this paper, we propose a New Neural Detection Network, named NNDF in short, to enhance the ACSA performance. Specifically, representations of input sentences and aspect categories contained in our method are generated by a CNN-pooling-BiLSTM structure respectively, where sentences are represented based on their contextual words and aspect categories are represented based on word embeddings of entities category-specific. Then, a Transformer-based encoder is used to model implicit dependency of sentence contexts and aspect categories of entities in sentences. Finally, the embedding of aspect-category is learned by the novel bidirectional attention mechanism for the sentiment classification. Besides, experiments conducted on Restaurant and MAMS benchmark datasets for the task demonstrate that NNDF achieves more accurate prediction results as compared to several state-of-the-art baselines.

Keywords: Aspect-category sentiment analysis · Transformer-based encoder · Bidirectional attention mechanism

1 Introduction

Aspect-based sentiment analysis (ABSA) is a fine-grained sentiment analysis task which has attracted increasing attention in industry and academia. Other than traditional sentiment analysis tasks which predict sentiment polarity of a sentence or document, ABSA mainly focuses on identifying emotional polarities of multiple aspects contained in a sentence. Besides, it mainly consists of two tasks, i.e., aspect term sentiment analysis (ATSA) and aspect category sentiment analysis (ACSA). ATSA aims to predict emotional polarities towards aspect terms contained in a sentence. Contrast to the ATSA, ACSA is intended to analyze the sentiment polarity of a set of predefined aspect categories which are possibly not existing in sentences. A typical example of

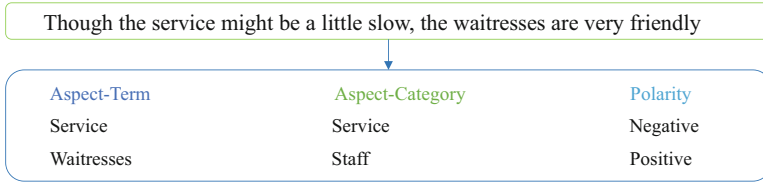


Fig. 1. An example of the sentence manifesting multi-aspect emotions

the comparison ATSA and ACSA, based on the sentence “Though the service might be a little slow, the waitresses are very friendly” is in Fig. 1. In the ATSA, “service” and “waitresses” are aspect terms, which are visible in the sentence denoting the positive and negative emotions, respectively. In the ACSA, two aspect categories are “staff” and “service” which also express the same emotions, but the “staff” category doesn’t appear in the sentence. In this paper, the focus of our research is mainly on ACSA. Therefore, how to accurately identify aspect categories and their contexts in sentences and obtain their relations is the main challenge we face.

Previous sentiment analysis works are almost sentence-based, which only focus on prediction of the emotional orientation of a whole sentence. Therefore, if we still apply the traditional models in ACSA, the outputs will possess some biases with respect to practical conditions. Recently, since neural network models were introduced into ACSA task, the performance of related models has been greatly improved. Based on Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), most early models achieved good performance. By employing convolutional windows-fixed filters, CNN and its derived models can effectively acquire semantic features and dependencies between words of sentences. Nevertheless, complex syntactic information contained in sentences still can not be obtained. e.g., for RNN [1] and its derivations, they are very sufficient for data with sequence characteristics and can mine temporal and semantic information in data. Hence, compared with other models, these sequence models based on RNN achieve better performance in ACSA. However, as the number of aspect categories with different sentiment polarities in sentences increases, they can not accurately obtain semantic features of aspect categories and their dependencies. Then, as the attention mechanism was proposed [2], combining it with RNN or CNN allows these models [3–5] to concentrate on key features for aspect terms which play great roles in sentiment prediction. However, because a sentence may contain various aspect categories, simply employing a single attention module is not enough to adequately obtain semantic features and associations between aspect categories and contexts. In general, these pre-existing models have following problems: (1) Noise data from other aspect categories will interfere with sentiment prediction. (2) These models can not fully acquire aspect-specific features and semantic dependencies of sentences because they contain multiple aspect categories.

To alleviate the aforementioned issues, in this paper, we propose a New Neural Detection Network, named NNDF in short, which applies the Pre-trained Bidirectional Encoder Representations from Transformers (BERT) to encode context words and aspect categories respectively to word embedding and utilizes a global feature

extraction layer to capture both local and long-term feature information of sentences. Then, with the Transformer module [6], NNDF can better acquire semantic dependencies and emotional information and obtain connections between contexts and aspect categories. Finally, we utilize the bidirectional attention mechanism [7] to synchronously learn multi-aspect categories and their relations, which also avoids interference from sentiment information of other aspect categories.

All in all, our contributions can be summarized as follows:

- We propose a novel framework called NNDF for the aspect-category sentiment analysis.
- Our method leverages the Transformer-based encoder to capture implicit dependency of the sentence context and aspect categories of entities simultaneously, followed by the novel bidirectional attention mechanism is used to learn the aspect-category embedding.
- We conduct extensive experiments on two benchmark datasets, namely Restaurant and MAMS, and compare our results against several state-of-the-art baselines across on the ACSA task. The experimental results have verified the effectiveness of NNDF.

Organization: The remainder of this paper is organized as follows. We review related research in this area in Sect. 2. In Sect. 3, we formalize the problem and give an overview of the framework of our proposed NNDF model. Section 4 provides the details of the proposed NNDF model. In Sect. 5 and Sect. 6, we conduct extensive experimental evaluations and provide an analysis of the effectiveness of NNDF in terms of the ACSA task. Meanwhile, we also conduct the results of node embeddings for quantitative evaluations. Finally, the conclusion and future work are described in Sect. 7.

2 Related Work

This section briefly reviews related works from different semantic analysis granularity, i.e., sentence-level sentiment analysis methods and aspect-level sentiment analysis methods.

2.1 Sentence-Level Sentiment Analysis Methods

Machine Learning-Based Methods. Bhoi [8] compared the performances of various machine learning methods, including Naive Bayes, Decision Tree, Random Forest, Extra Trees, Extreme Gradient Boosting (XGBoost) and Support Vector Machine (SVM) [9]. Among of them, SVM gets the best classification results. However, these methods need to pre-construct abundant features and do lots of pre-processing for input sentences. The performance of these models greatly depends on the features of artificial construction and sufficient prior knowledge, which will cost more human resources.

Deep Neural Network-Based Methods. Since deep neural networks were applied widely to the field of sentiment classification, some models benefiting from them have obtained great performances. Compared with machine learning based models, deep neural networks are more powerful in capturing complex high-order features with non-linear activation functions, which usually yields better performance. For example, RNNs and CNNs [10–12] are capable of flexibly acquiring features of sentences. Account for positive conditions discussed before, some derived models have been proposed. Tang *et al.* [13] first proposed the idea of considering the semantic relatedness between target aspect term and its context word who also put forward two target-dependent LSTM modules to automatically capture features of aspect terms and contexts. Besides, some researchers combined CNN and LSTM to obtain both local and long-term semantic dependent information. Further, Yoon *et al.* [14] proposed a Multi-Channel Lexicon Integrated CNN-BiLSTM model, which utilized a multi-channel method on lexicon to improve lexicon feature and CNN and BiLSTM to obtain the n-gram features as well as long-term dependent information respectively. Besides, Wang *et al.* [15] put forward a novel CNN-LSTM model, which is composed of a regional CNN (R-CNN) and LSTM. Different from traditional CNN which regards a whole input text as input text, the regional CNN splits sentences into different regions whose useful local features will be effectively extracted. Moreover, by integrating R-CNN and LSTM, both local features and long-term dependent information can be utilized in the prediction process. Inspired by Wang’s work, we designed a global feature extraction layer which is composed of CNN and RNN and inherits main advantages of Wang’s work.

2.2 Aspect-Level Sentiment Analysis Methods

Deep Neural Network-Based Methods. All above models do not successfully take aspect-aware information into consideration and establish correlations between aspect terms and their emotional information in training completely. Then, researchers applied the attention mechanism to address the problem, which achieved good performance so that more and more attention-based models were raised. In view of this, some researchers combined LSTM with the attention mechanism, and also provided some valuable solutions in correlation construction. Wang *et al.* [3] applied the attention mechanism to establish semantic dependencies between contexts and aspect terms by appending aspect term embedding into word vectors as the input vectors. Besides, to generate more comprehensive representations, Ma *et al.* [16] designed more complicated network structures, including two separate attention modules which were used to learn attention weight of aspect terms and sentences. Inspired by Wang, some studies tried to apply the attention mechanism to other network structures. For instance, Tang *et al.* [17] employed the deep memory network and attention mechanism to generate deeper text representation. And Gu *et al.* [18] adopted a bidirectional attention mechanism to mutually establish relationships between sentences and aspect terms. In the last, Xue *et al.* [19] made computations parallel and effectively decreased training time by using CNN and gating units, but accuracy of the proposed model had not been improved significantly.

Transformer-Based Methods. Transformer [6] based methods make great progress in comparison to CNN and RNN in ACSA tasks. Jiang *et al.* [20] combined capsule networks with BERT to compute the deeper representations of sentences and aspect terms. Moreover, Wu *et al.* [21] adopted the pre-trained RoBERTa as backbone network to predict sentiment polarities of multi-aspect terms. And Wu *et al.* [22] put forward the quasi-attention to enable their model to learn both additive and subtractive attention, which effectively calculates the context-aware attention representations. All these works have achieved excellent performances, so we employ BERT as our backbone network to ensure accuracy of downstream classification task of our model.

Graph Neural Network-Based Methods. The prevalence of graph neural network based models, typically GCN [23–25], has led to excellent performance gains on sentiment analysis tasks. Liu *et al.* [7] utilized GCN to obtain sentence structure information and employed the bidirectional attention mechanism to acquire useful interactive information between aspect terms and contexts. By building a dependency tree of sentence, Zhang *et al.* [26] used a GCN module to better acquire semantic dependencies and syntactic information. In addition, Li *et al.* [27] integrated syntactic information and semantic dependencies through the SynGCN and SemGCN module simultaneously and employed two regularizers to model correlations among words.

3 Preliminary

In this section, we first formulate our task. Then, we introduce the framework of our proposed NNDF model.

3.1 Task Definition

The ACSA task aims to predict emotional polarity of designated aspect categories. Given a sentence $S = \{W_1, W_2, \dots, W_n\}$ and we pre-defined special aspect categories $C = \{C_1, C_2, \dots, C_m\}$ which may be a word or a phrase. The purpose of this paper is to predict the emotional polarities $y \in \{1, \dots, P\}$. P is the number of sentiment categories, and the lengths of sentence S and aspect categories C are n and m . In the following, we take a sentence “The bar area was fairly crowded but service remained friendly and efficient” as an example shown in Fig. 2. Two aspect categories in the sentence are “place” and “service”, and the emotional polarities towards them are negative and positive.

3.2 Solution Framework

The whole framework of NNDF can be illustrated in Fig. 3. The framework is divided into two key components:

- Embedding layer, it is used for forming the unified representations for encoding feature vectors of input nodes with different dimensions.
- Global Feature Extraction layer, Transformer Encoder Layer, Bidirectional Attention Layer, and Classification Layer.

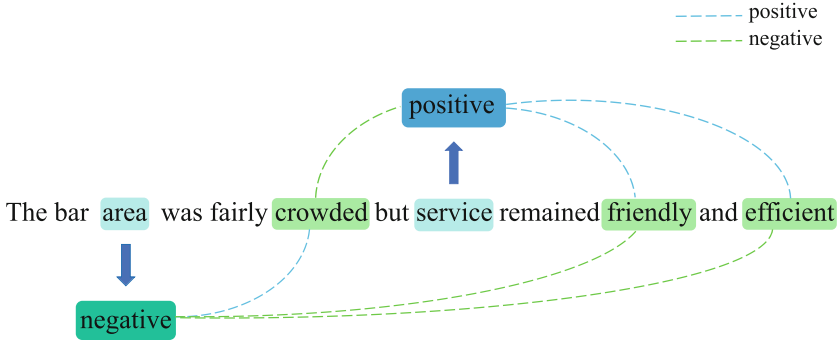


Fig. 2. A sentence from MAMS dataset. The categories “place” and “service” show negative and positive. And “place” is not presented in the sentence.

4 The Proposed Method

In this section, we first formulate our task, and then introduce the framework of our proposed NNDF model.

4.1 Input Representation Layer

We apply the GloVe embedding or BERT embedding in embedding layer to transform the sentence S into word embedding $E = \{E_1, E_2, \dots, E_n\}$. Then, we will briefly introduce the two embedding methods.

GloVe Embedding [28]: The GloVe model has high computational efficiency, and its scale of calculation is proportional to the corpus. When the corpus is small, the GloVe model still works well enough. Therefore, we use the pre-trained GloVe to convert sentence S into word embedding E . The context embedding is presented as $E^s \in R^{d_e \times n}$ and d_e is the dimension of word vector. And the categories embedding is represented as $E^c \in R^{d_e \times m}$. After the sentence passes through the embedding layer, we concatenate categories embedding and sentence embedding into the categories-aware sentence embedding $E^{sc} = [E^s; E^c]$.

BERT Embedding [29]: Compared with traditional embedding methods, BERT has obtained obvious improvement since it was introduced into NLP tasks. The input of BERT consists of a token, segmentation, and position embedding. Therefore, to utilize pre-trained BERT, the sentence and categories are denoted as $\{[CLS], S_1, S_2, \dots, [SEP]\}$ and $\{[CLS], C_1, C_2, \dots, [SEP]\}$. Then, the input will be transformed to the presentation $\{H_{[CLS]}, H_{[S_1]}, \dots, H_{[S_n]}, H_{[SEP]}\}$ and $\{H_{[CLS]}, H_{[C_1]}, \dots, H_{[C_m]}, H_{[SEP]}\}$ respectively.

4.2 Global Feature Extraction

We will introduce the global feature extraction layer (GFE), which is composed of CNN and BiLSTM.

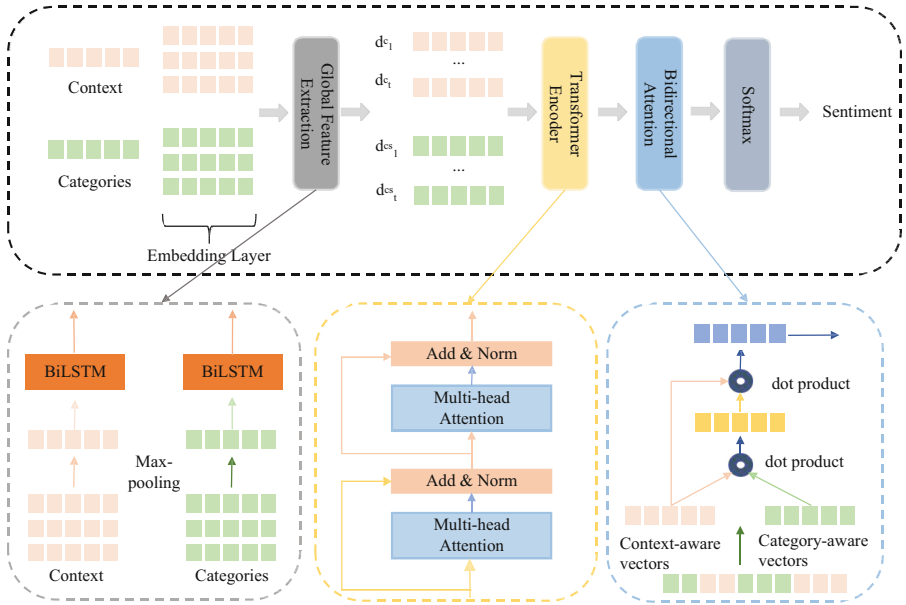


Fig. 3. The overall architecture of NNDF

CNN: We utilize the convolutional layer to extract local features and reduce dimensions of input. Convoluting the window vectors, each filter can generate different features at separate positions. And as filter moves, numerous local features of sentences are captured. In this paper, we set the number and length of filter to be 150 and 3. Besides, we select ReLU which is easy to calculate and speed up convergence of the network as our non-linear transformation function in the convolution process. Then, we applied K-Max pooling to output of convolutional layer, because max-pooling layer not only reduces the amount of computations but also preserves the most significant information. And output of all max-pooling layers will be fused to produce input of BiLSTM.

BiLSTM: Due to weak ability of CNN to capture long-term features, we put BiLSTM over CNN to acquire long-term semantic information, which deals with the problems of gradient explosion as well as gradient vanishing and utilizes gating units and memory cells to selectively capture useful semantic information in both directions. The update of hidden states and memory cells contents, which includes current input and past state, is determined by gate units, consisting of input, output and forget gate. In this case, we set the dimension of all hidden layers in BiLSTM to be 150. Then, we regard the last hidden state of BiLSTM as the final representation. Therefore, output of GFE will contain local features and long-term semantic information, which will help Transformer module to establish better connections between contexts and aspect categories.

4.3 Transformer Encoder

Transformer parallelly processes all the words and symbols in a sequence without recurrent structure and utilizes the self-attention mechanism to combine context with distant words. It not only trains faster than RNN but also performs better. Hence, we only utilize the encoder of Transformer to obtain associations between aspect categories and contexts.

In the Transformer encoder, multiple scaled dot-product attention constitutes the multi-head attention mechanism. Therefore, MHA can execute attentions simultaneously, which is helpful to obtain connections between contexts and aspect categories. The output of the GFE layer is a matrix $X_{GFE} \in R^{M \times d_e}$. Then, we will randomly initialize three matrixes $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$, and multiple them with X_{GFE} to obtain three same weight matrixes $\mathbf{Q} = (\mathbf{Q}_1, \dots, \mathbf{Q}_M), \mathbf{K} = (\mathbf{K}_1, \dots, \mathbf{K}_M), \mathbf{V} = (\mathbf{V}_1, \dots, \mathbf{V}_M)$, where $q_i, k_i, v_i \in R^{\frac{d_h}{h}}$, and d_h is a hidden dimension. Then, the specific calculation is defined as follows:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Softmax\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (1)$$

Following a series of linear transformations with diverse parameters, three weight matrixes $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ learn different features from contexts and aspect categories severally. Then, through multiple times of transformation, the multi-head attention further captures degrees of associations between aspect categories and its semantic words. The summing of all outputs of scaled dot-product attention will be output of multi-head attention:

$$MHA(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Concat(head_1, \dots, head_h)\mathbf{W}^O \quad (2)$$

$$head_i = Attention(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V) \quad (3)$$

where $\mathbf{W}_i^Q \in R^{d_h \times d_k}, \mathbf{W}_i^K \in R^{d_h \times d_k}, \mathbf{W}_i^V \in R^{d_h \times d_v}, \mathbf{W}^O \in R^{d_h}$, and $d_k = d_v = d_h/h$. In this paper, we set $h = 8$ which is the number of attention layers. And output of GFE layer will be used as input of Transformer encoder and aspect categories representation $MHA^{ca} = [h_1^{ca}, h_2^{ca}, \dots, h_m^{ca}]$ and context representation $MHA^c = [h_1^c, h_2^c, \dots, h_n^c]$ will be calculated, where $\mathbf{h}_i^{ca}, \mathbf{h}_i^c \in R^{d_h}$.

4.4 Bidirectional Attention

We utilize bidirectional attention to further fuse the feature information of contexts and aspect categories. During the process of calculation, attention vectors go into the modeling layer with embedding vectors from the previous layer for each time step, which contributes to decreasing loss of information. And we will fuse semantic dependencies by categories-aware attention and context-aware attention mechanism.

Categories-Aware Attention: Assume the output matrices of contexts and aspect categories are $\mathbf{h}^c = [h_1^c, h_2^c, \dots, h_t^c, \dots, h_n^c]$ and $\mathbf{h}^{ca} = [h_1^{ca}, h_2^{ca}, \dots, h_t^{ca}, \dots, h_m^{ca}]$. The calculation is defined as follows:

$$\alpha = \sum_{i=1}^m \frac{\exp(e\mathbf{h}^{cT} \cdot \mathbf{W}_{ca} \cdot \mathbf{h}_{ca_i}^{ca})}{\sum_{i=1}^m \exp(\mathbf{h}^{cT} \cdot \mathbf{W}_{ca} \cdot \mathbf{h}_{ca_i}^{ca})} \cdot h_{ca_i}^{ca} \quad (4)$$

$$\mathbf{r}^{ca} = \sum_{i=1}^M \alpha \cdot \mathbf{h}_{ca_i}^{ca} \quad (5)$$

where \mathbf{r}^{ca} is representation of categories, which has learned lots of emotional information from contexts. $\bar{\mathbf{h}}^c$ is obtained by average pooling context vectors. $\mathbf{h}_{ca_i}^{ca}$ represents aspect category vectors and \mathbf{W}_{ca} denotes the attention weight matrix.

Contexts-Aware Attention: Similarly, we will utilize the new categories-aware representation for the same calculation. Then, the new context representation will be calculated as \mathbf{r}^{ca} . The expression of calculation is:

$$\beta = \frac{\exp(\mathbf{r}^{ca\top} \cdot \mathbf{W}_c \cdot \mathbf{h}_i^c)}{\sum_{i=1}^m \exp(\mathbf{r}^{ca\top} \cdot \mathbf{W}_c \cdot \mathbf{h}_i^c)} \quad (6)$$

$$\mathbf{r}^c = \sum_{i=1}^M \beta \cdot \mathbf{h}^c \quad (7)$$

where \mathbf{r}^c , which contains the semantic relevance of aspect categories and contexts, is the final representation for the sentiment prediction.

4.5 Loss Function and Training

We will input the final representation \mathbf{r}^c into the classification layer to predict the emotional polarities towards aspect categories given.

$$p = \text{Softmax}(\mathbf{W}_p \mathbf{r}^c + \mathbf{b}_p), \quad (8)$$

where p is sentiment polarities towards aspect categories, $\mathbf{W}_p, \mathbf{b}_p$ are learnable parameters. And to constrain randomness dropout brings, we apply the Regularized Dropout (R-Drop) [30] to put a regular constraint on prediction, which reduces inconsistency between training and testing. Compared with the traditional training methods, R-Drop only adds a KL-divergence loss function.

$$L_i^{CE} = -\log P_\theta^{(1)}(y_i|x_i) - \log P_\theta^{(2)}(y_i|x_i), \quad (9)$$

$$L_i^{(KL)} = \frac{1}{2} [KL(P_\theta^{(2)}(y|x_i)||P_\theta^{(1)}(y|x_i)) + KL(P_\theta^{(1)}(y|x_i)||P_\theta^{(2)}(y|x_i))], \quad (10)$$

$$L_i = L_i^{CE} + \alpha L_i^{KL}. \quad (11)$$

where x_i, y_i are results of two predictions with the same parameters, L_i^{CE} is the sum of two original cross-entropy functions. L_i^{KL} is KL divergence between two models, α is weight of KL loss. In this paper, α will be set to 3, which is different from the optimal solution proposed in the original paper.

5 Experiments

In this section, we compare the performance of our proposed NNDF model with several state-of-the-art baselines, and a few variants of NNDF itself, using two benchmark datasets.

5.1 Experiment Settings

Datasets: We select two benchmark datasets to conduct a series of experiments to evaluate the performance of NNDF, which includes Restaurant [31]¹ and MAMS datasets [32]². Then, we will briefly introduce the two datasets. Compared with the Restaurant dataset, MAMS adopts five aspect categories from the Restaurant dataset and adds two more aspect categories to deal with some chaotic cases. Different from other datasets, every sentence from MAMS expresses multiple emotional polarities. The release of MAMS pushes forward the development of the ABSA task and prevents it from degenerating to sentence-level sentiment analysis. Table 1 provides specific quantitative information of two datasets.

Table 1. The statistics of both Restaurant and MAMS datasets in the experiments

Dataset		Pos.	Neg.	Neu.	Total.
Restaurant	Train	2164	807	637	3608
	Test	728	196	196	1120
MAMS	Train	1929	2084	3077	7090
	Validation	241	259	388	888
	Test	245	263	393	901

Baselines: In order to evaluate NNDF more comprehensively, we exploit a series of state-of-the-art models as baselines for comparison, including variations of RNN models, CNN with gate units, capsule network, heterogeneous GCN-based models, and Transformer-based models.

- LSTM [1] is a basic RNN network, which utilizes output of the last layer as final sentence representation to conduct emotional categorization.
- TD-LSTM [13] integrates the aspect terms into LSTM to establish correlations between aspects and contexts.
- ATAE-LSTM [3] adds input aspect terms embedding into vector of each word and utilizes the attention mechanism to better establish dependencies between aspect terms and input vector.
- BiLSTM+Attn, based on AT-LSTM, replaces LSTM with BiLSTM to enable the model to take information from both directions of semantic features into accounts.
- IAN [16] uses two same parts which are composed of LSTM and an attention mechanism to learn representations of aspect terms and contexts interactively. Then, concatenate two representations as final representation for emotion prediction.
- MemNet [13] uses multiple computational layers to calculate text representation and representation of the last layer will be used for emotional categorization.
- GCAE [19] utilizes the convolutional layer and gate units to parallelly generate and capture aspect-related sentiment features, which is more efficient than RNN-based models.

¹ The Restaurant dataset of SemEval-2014 Task 4: <https://alt.qcri.org/semeval2014/task4>.

² The MAMS dataset: <http://tcci.ccf.org.cn/conference/2020>.

- PBAN [18] appends positional vectors into input vectors, which can distill aspect-aware information better, and employs bidirectional attention mechanism to establish semantic dependencies between aspect terms and their emotional information.
- BRET [29] utilizes the multi-layer bidirectional transformer encoder to compute more comprehensive representation.
- RoBERTa [33] pre-trains with eight times larger batches and corpora and employs dynamic masking to take place of static masking in BERT, compared with BERT.
- RoBERTa-TMM [34] adopts the pre-trained RoBERTa as backbone network, then fine-tune it on the MAMS dataset.
- CapsNet [20] uses bidirectional gated recurrent unit (BiGRU) to obtain contextualized representation and feeds them into capsule network whose outputs are used to predict emotional polarities.
- CapsNet-BERT combines strength of the capsule network and BERT. The pre-trained BERT is used to compute deep representations of sentences and aspect terms, which will be fed into a capsule network to predict sentiment polarities.
- ASGCN [26] employs Bidirectional LSTM to capture contextual information regarding aspect terms and uses GCN to obtain edge information of syntactical dependencies, which enables the model to capture dependencies among aspect terms.
- QACG-BERT [22] improves the structure of BERT to be context-aware and appends a quasi-attention mechanism. By learning quasi-attention weights which could be negative, the model could learn compositional attention that supports subtractive attention.
- DualGCN [27] obtains syntactic information and semantic dependencies by the SynGCN and SemGCN module. Then, usage of regularizers with semantic constraints is to solve the overlapping problem of semantic information, which makes emotion prediction more accurate.

Implementation Details: We choose 300-dimension Glove vectors to generate word embedding for non-Transformer-based models. For Transformer-based models, we utilize pre-trained BERT as the backbone network whose embedding dimension and hidden state dimensions are set to 768. For MAMS dataset and Restaurant dataset, we set the size of mini-batch to be 64 and 32 respectively. Then, we employ Adam [35] as our optimization function to update models parameters in iterations. And for non-Transformer-based models and Transformer-based models, we set the initial learning rates to be 0.0003 and 0.00003. The initial dropout rate will be set to 0.5. Finally, we obtain final results by averaging the outputs of 5 round of running.

6 Experimental Results and Analyses

We utilize accuracy and macro-averaged F1-score as our assessment metrics to assess performance of NNDF. The experimental results are shown in Table 2. Obviously, NNDF achieves the best performance in comparison to other baseline methods on two datasets. Then, we will make some discussions and analyses based on the experimental results.

Table 2. The performance results (%) of different methods on the two datasets for the aspect-category Sentiment task. The best and second best results in each column is boldfaced and underlined respectively (the higher, the better). Improvements over the best baseline are shown in the last row

Method	MAMS		Restaurant	
	Acc	F1	Acc	F1
LSTM	47.37	0.432	72.77	0.554
TD-LSTM	<u>62.37</u>	<u>0.497</u>	<u>75.09</u>	<u>0.587</u>
ATAE-LSTM	70.63	0.584	77.58	0.66
BiLSTM+Attn	66.3	0.553	76.36	0.645
IAN	-	-	78.6	0.689
MemNet	63.29	0.541	76.54	0.653
GCAE	72.11	0.613	77.84	0.675
PBAN	-	-	81.16	0.716
CapsNet	73.99	0.629	83.55	0.735
ASGCN	-	-	80.77	0.722
DualGCN	-	-	<u>84.27</u>	<u>0.781</u>
BERT	78.29	0.697	90.44	0.806
RoBERTa	77.44	0.683	-	-
QACG-BERT	-	-	90.67	0.813
RoBERTa-TMM	78.03	0.686	-	-
CapsNet-BERT	<u>79.46</u>	<u>0.698</u>	<u>91.38</u>	<u>0.824</u>
NNDF	76.42	0.672	83.76	0.728
NNDF-BERT	<u>81.53</u>	<u>0.713</u>	<u>92.26</u>	<u>0.835</u>

The performance of all models on the Restaurant dataset is much higher than the one on the MAMS dataset. The result remains consistent with our intuition that the sentences in the MAMS dataset involve more aspect categories, which makes it more difficult for those models to accurately detect emotional information and establish correlations of different aspect categories. The performance of TD-LSTM has made great progress compared with the original LSTM, because TD-LSTM utilizes LSTM to separately capture features of aspect terms and contexts and model their relationships. But one thing needs to be pointed out is that it cannot judge which one of contextual features contributes more to the determination of emotional polarity. To generate a more comprehensive representation, ATAE-LSTM integrates embedding of aspect term into vector of each word. Different from ATAE-LSTM, IAN pays more attention to establishing the relationship between aspect terms and contexts which not only calculates the weight of each word in contexts but also learns weight of each word in corresponding aspect terms. Besides, PBAN achieves outstanding performance by appending positional vectors to input vectors to enable the model to find aspect terms and related information more accurately and utilizing bidirectional attention to model correlations

between aspect terms and contexts. The experimental results proved that adding position information is vital for finding the location of aspect categories in ACSA task.

For GCN-based models, the performance of DualGCN on Restaurant dataset is better than ASGCN. In the process of extracting feature information, problem of overlapping semantic dependencies exists, which means that one category will match some semantic information of another category. For DualGCN model, it uses orthogonal and differential regularizers to learn an semantic attention matrix and features respectively, which solves interference of semantic information of other aspect terms. Previous experimental results have proved that LSTM is not skilled in aspect-specific feature extraction. It is not enough to use a BiLSTM in obtaining aspect-specific features, or it will result in unapparent improvement. For ASGCN model, due to the excellent ability of GCN to employ semantic dependencies and syntactic information despite the usage of LSTM, ASGCN still achieves great performance.

Obviously, compared with non-Transformer-based models, Transformer-based models have made great progress in performance improvement on two datasets, which proves the strong ability of BERT. But at the same time, it also costs generous time. To calculate representations of each layer, the theoretical time complexity of NNDF is $\mathcal{O}(n^2 \cdot d)$, where n is the length of input sequence, d is the dimension of representation. For our model, in comparison with non-Transformer-based models, NNDF achieves the second performance, second only to DualGCN. Compared with Transformer-based methods, NNDF-BERT achieves the best performance on both datasets. The performance of NNDF-BERT on MAMS datasets exceeds CapsNet-BERT by 2.07%. In general, NNDF-BERT accurately extracts feature information and models relationships between aspect categories and contexts, and the relationships enable aspect categories to better match corresponding emotional information.

6.1 Case Study

To further explore how our model outperforms other ones, we use our model, CapsNet-BERT, ASGCN to predict a sentence from MAMS dataset, which has two aspect categories named “food” and “service”. Figure 4 and Table 3 provides the experimental results, where the underline indicates that the weight of the word is the largest, followed by the wavy underline and the overline is the lowest. Based on the results of experiment, NNDF performs better than the other two models. Through the visual analysis above, when NNDF accurately locates words representing aspect categories in sentences, they are distributed the highest weight. And different with the other two models, it finds out words or phrases that represent emotional information of specific aspect categories and reduces influences brought by other interfering words. Besides, another point which should be pointed out is that NNDF assigns some weight to conjunction word “but”, which sets up a symbol of differentiation of semantic orientation and helps distinguish sentiment information of two aspect categories named “space” and “service” well. Thus, the above analysis proves that NNDF has a better ability to identify emotional information of each category than pre-existing models.

Table 3. The weight visualization of a sentence for CapsNet, DualGCN and NNDF

Model	Category	Attention visualization	Prediction	Label
CapsNet	Place	The <u>bar</u> area was fairly <u>crowded</u> but service remained <u>friendly</u>	Neu.	Neg.
	Service	The bar area was fairly <u>crowded</u> but <u>service</u> remained <u>friendly</u>	Pos.	Pos.
DualGCN	Place	The <u>bar</u> area was fairly <u>crowded</u> but service remained <u>friendly</u>	Neg.	Neg.
	Service	The bar area was fairly <u>crowded</u> but <u>service</u> remained <u>friendly</u>	Pos.	Pos.
NNDF	Place	The <u>bar</u> area was fairly <u>crowded</u> but service remained <u>friendly</u>	Neg.	Neg.
	Service	The bar area was fairly <u>crowded</u> but <u>service</u> remained <u>friendly</u>	Pos.	Pos.

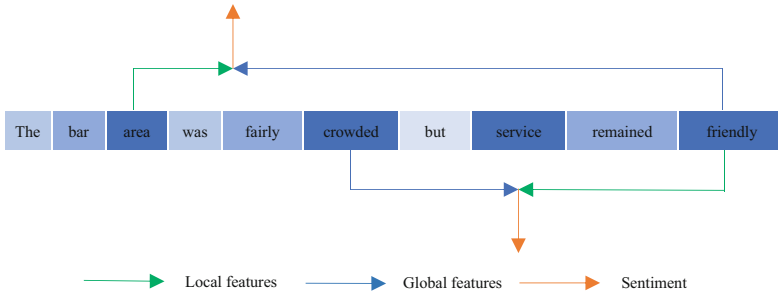


Fig. 4. The weight visualization of a sentence for NNDF

6.2 Ablation Study

We conduct some ablation studies to further explore the function of each part of NNDF in this section. And the experimental results of the ablation study are provided in Table 4.

Table 4. The experimental results of the ablation study

Model	MAMS (ACSA)		Restaurant (ACSA)	
	Acc	F1	Acc	F1
NNDF	76.42	0.672	83.76	0.728
NNDF w/o GFE	75.37	0.665	81.14	0.701
NNDF w/o BA	73.76	0.641	82.37	0.713
NNDF w/o TE	73.45	0.637	82.69	0.716
NNDF w/o R	75.67	0.669	82.57	0.715

where NNDF w/o BA denotes that NNDF removes the bidirectional attention layer and uses the original attention mechanism. NNDF w/o GFE denotes NNDF without the global feature extraction layer. And NNDF w/o TE represents Transformer encoder layer-removed NNDF. Besides, NNDF w/o R symbolizes NNDF without R-Drop and using the original cross-entropy function. According to Table 3, it is interesting to see that the experimental results of NNDF w/o BA are similar to NNDF w/o TE. It implies that both Transformer encoder and bidirectional attention are significant to establish relationships between aspect categories and their emotional information. Obviously, NNDF w/o GFE performs worse on both datasets, which suggests that the GFE layer helps to locate multiple aspect categories and capture their corresponding features. Finally, without R-Drop, the performance of NNDF w/o R receives the same percentage of deterioration on both datasets, which indicates that R-Drop can improve the generalization ability of NNDF.

7 Conclusion

In this paper, we proposed a new Transformer-based model with bidirectional attention mechanism. NNDF mainly focuses on establishing connections between aspect categories and their emotional information and filtering noise information during prediction. And except for the GloVe embeddings, we employ a pre-trained BERT to encode our sentences and categories to further enhance the performance of NNDF. We conduct a series of experiments on Restaurant and MAMS datasets. The experimental results indicate that NNDF-BERT achieves the best performance in comparison to some strong baseline models. Our future work will concentrate on two aspects:

- the time cost of our model is large for the abundant RNNs. Therefore, we will focus on using the GCN to obtain semantic dependencies to decrease the training time.
- we will strengthen associations between semantic dependencies. We tend to remould the structure of the Transformer to further capture the syntactic dependencies and incorporate syntactic and semantic information.

Acknowledgement. We greatly appreciate the valuable suggestions and encouragement from anonymous reviewers and the editor.

References

1. Sepp, H., Jürgen, S.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate (2015)
3. Wang, Y., Huang, M., Zhu, X., Zhao, L.: Attention-based LSTM for aspect-level sentiment classification. In: EMNLP, pp. 606–615 (2016)
4. Chen, P., Sun, Z., Bing, L., Yang, W.: Recurrent attention network on memory for aspect sentiment analysis. In: EMNLP, pp. 452–461 (2017)
5. Huang, B., Ou, Y., Carley, K.M.: Aspect level sentiment classification with attention-over-attention neural networks. In: SBP-BRiMS, pp. 197–206 (2018)
6. Vaswani, A., et al.: Attention is all you need. In: NIPS, pp. 5998–6008 (2017)

7. Liu, J., Liu, P., Zhu, Z., Li, X., Xu, G.: Graph convolutional networks with bidirectional attention for aspect-based sentiment classification. *Appl. Sci.* **11**(4), 1528 (2021)
8. Bhoi, A., Joshi, S.: Various approaches to aspect-based sentiment analysis. *CoRR*, abs/1805.01984 (2018)
9. Vicente, I.S., Saralegi, X., Agerri, R.: EliXa: a modular and flexible ABSA platform. *CoRR*, abs/1702.01944 (2017)
10. Castellucci, G., Filice, S., Croce, D., Basili, R.: UNITOR: aspect based sentiment analysis with structured learning. In: *SemEval*, pp. 761–767 (2014)
11. Johnson, R., Zhang, T.: Semi-supervised convolutional neural networks for text categorization via region embedding. In: *NIPS*, pp. 919–927 (2015)
12. Kim, Y.: Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882 (2014)
13. Tang, D., Qin, B., Feng, X., Liu, T.: Effective LSTMS for target-dependent sentiment classification. In: *COLING*, pp. 3298–3307 (2016)
14. Yoon, J., Kim, H.: Multi-channel lexicon integrated CNN-BILSTM models for sentiment analysis. In: *ROCLING*, pp. 244–253. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP) (2017)
15. Wang, J., Yu, L., Lai, K.R., Zhang, X.: Dimensional sentiment analysis using a regional CNN-LSTM model. In: *ACL* (2016)
16. Ma, D., Li, S., Zhang, X., Wang, H.: Interactive attention networks for aspect-level sentiment classification. In: *IJCAI*, pp. 4068–4074 (2017)
17. Tang, D., Qin, B., Liu, T.: Aspect level sentiment classification with deep memory network. In: *EMNLP*, pp. 214–224 (2016)
18. Gu, S., Zhang, L., Hou, Y., Song, Y.: A position-aware bidirectional attention network for aspect-level sentiment analysis. In: *COLING*, pp. 774–784 (2018)
19. Xue, W., Li, T.: Aspect based sentiment analysis with gated convolutional networks. In: *ACL*, pp. 2514–2523 (2018)
20. Jiang, Q., Chen, L., Xu, R., Ao, X., Yang, M.: A challenge dataset and effective models for aspect-based sentiment analysis. In: *EMNLP-IJCNLP*, pp. 6279–6284 (2019)
21. Wu, Z., Ying, C., Dai, X., Huang, S., Chen, J.: Transformer-based multi-aspect modeling for multi-aspect multi-sentiment analysis. In: *NLPCC*, pp. 546–557 (2020)
22. Wu, Z., Ong, D.C.: Context-guided BERT for targeted aspect-based sentiment analysis. In: *AAAI/EAAI*, pp. 14094–14102 (2021)
23. Msahli, M., Qiu, H., Zheng, Q., Memmi, G., Lu, J.: Topological graph convolutional network-based urban traffic flow and density prediction. *IEEE TITS* **22**(7), 4560–4569 (2020)
24. Li, Y., Song, Y., Jia, L., Gao, S., Qiu, M.: Intelligent fault diagnosis by fusing domain adversarial training and maximum mean discrepancy via ensemble learning. *IEEE TII* **17**(4), 2833–2841 (2020)
25. Fei, H., Lakdawala, S., Qi, H., Qiu, M.: Low-power, intelligent sensor hardware interface for medical data preprocessing. *IEEE Trans. Inf. Technol. Biomed.* **13**(4), 656–663 (2009)
26. Zhang, C., Li, Q., Song, D.: Aspect-based sentiment classification with aspect-specific graph convolutional networks. In: *EMNLP-IJCNLP*, pp. 4567–4577 (2019)
27. Li, R., Chen, H., Feng, F., Ma, Z., Wang, X., Hovy, E.H.: Dual graph convolutional networks for aspect-based sentiment analysis. In: *ACL/IJCNLP*, pp. 6319–6329 (2021)
28. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: *EMNLP*, pp. 1532–1543 (2014)
29. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: *NAACL-HLT*, pp. 4171–4186 (2019)
30. Liang, X., et al.: R-drop: regularized dropout for neural networks. *CoRR*, abs/2106.14448 (2021)

31. Kirange, D.K., Deshmukh, R.R., Kirange, M.D.K.: Aspect based sentiment analysis SemEval-2014 task 4. *AJCSIT* **4**, 72–75 (2014)
32. Chen, L., Xu, R., Yang, M.: Overview of the NLPCC 2020 shared task: multi-aspect-based multi-sentiment analysis (MAMS). In: Zhu, X., Zhang, M., Hong, Yu., He, R. (eds.) *NLPCC 2020. LNCS (LNAI)*, vol. 12431, pp. 579–585. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60457-8_48
33. Liu, Y., et al.: Roberta: a robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692 (2019)
34. Wu, Z., Ying, C., Dai, X., Huang, S., Chen, J.: Transformer-based multi-aspect modeling for multi-aspect multi-sentiment analysis. In: Zhu, X., Zhang, M., Hong, Yu., He, R. (eds.) *NLPCC 2020. LNCS (LNAI)*, vol. 12431, pp. 546–557. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60457-8_45
35. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: *ICLR* (2015)