Hindawi
Scientific Programming
Volume 2020, Article ID 4690974, 10 pages
https://doi.org/10.1155/2020/4690974

Research Article

Big Data Management and Analytics in Scientific Programming: A Deep Learning-Based Method for Aspect Category Classification of Question-Answering-Style Reviews

Hanqian Wu,[1,2] **Mumu Liu**,[1,2] **Shangbin Zhang**,[1,2] **Zhike Wang**,[1,2] **and Siliang Cheng**[1,2]

[1]*School of Computer Science and Engineering, Southeast University, Nanjing, China*
[2]*Key Laboratory of Computer Network and Information Integration of Ministry of Education, Southeast University, Nanjing, China*

Correspondence should be addressed to Hanqian Wu; hanqian@seu.edu.cn

Received 23 October 2019; Revised 7 January 2020; Accepted 16 January 2020; Published 8 June 2020

Guest Editor: Zhiang Wu

Online product reviews are exploring on e-commerce platforms, and mining aspect-level product information contained in those reviews has great economic benefit. The aspect category classification task is a basic task for aspect-level sentiment analysis which has become a hot research topic in the natural language processing (NLP) field during the last decades. In various e-commerce platforms, there emerge various user-generated question-answering (QA) reviews which generally contain much aspect-related information of products. Although some researchers have devoted their efforts on the aspect category classification for traditional product reviews, the existing deep learning-based approaches cannot be well applied to represent the QA-style reviews. Thus, we propose a 4-dimension (4D) textual representation model based on QA interaction-level and hyperinteraction-level by modeling with different levels of the text representation, i.e., word-level, sentence-level, QA interaction-level, and hyperinteraction-level. In our experiments, the empirical studies on datasets from three domains demonstrate that our proposals perform better than traditional sentence-level representation approaches, especially in the *Digit* domain.

1. Introduction

Nowadays, Internet finance has evolved rapidly, and e-commerce platform occupies the central position in its development by providing virtual payment means. In the virtual trading environment created by the e-commerce platform, both parties can realize communication on time by online product reviews. To some extent, the content of those reviews can affect the business of the e-commerce platform, thus spreading to the stability of the whole Internet finance. On these grounds, it is very important to mine valuable information contained in those reviews, which can not only help consumers make purchase decisions but also help organizations know customer satisfaction and make adjustment strategies. However, the number of online product reviews is exploding. It is difficult for us to manually collect and manipulate these texts. Thus, sentiment analysis comes into being.

Sentiment analysis, also known as opinion mining, is a task which aims to analyze people's sentimental orientation of a given text. Researchers used to pay attention on document-level or sentence-level sentiment analysis; however, with the rapid development of e-commerce business, consumers tend to learn about products in detail. Thus, aspect-level sentiment analysis has become a research hotspot, which is proposed to have a better understanding of rapid-growing online reviews than traditional opinion mining by extracting fine-grained insights such as aspect terms, aspect categories, and sentiment polarities.

Recently, there appears a new question-answering (QA) style of reviews, namely, "customer questions and answers," in various popular e-commerce platforms at home and abroad, such as Taobao, Jingdong, and Amazon. Figure 1 shows an example of QA-style reviews. In this novel reviewing form, a potential customer can ask questions

Question1: *How long does it last when you are playing games?Is the quality of this phone good?*
Answer1: *The electricity is not very durable, but its appearance is good.*
→ *Battery*

Question2: *I want to buy it for design, how about its running?*
Answer2: *It has lots of running memory, but I sometimes feel a little* choppy
→ *Performance*

Figure 1: Translated examples of QA-style reviews on e-commerce websites. The coloured words are called aspect terms, and words in red are matched in a QA-style review while those in other colours are unmatched.

about a certain product, and others who have purchased the same item can answer these questions. Along with the popularity of this new reviewing form, the relevant research is really worthwhile due to its own characteristics. For one thing, consumers prefer QA-style reviews to traditional reviews. For another, this QA reviewing way can largely reduce fake information which makes product reviews more reliable and convincing. However, there is less existing research on the aspect category classification of QA-style reviews which aims to identify the aspect category of a given QA-style review.

And the existing deep learning-based approaches to aspect category classification on traditional product reviews cannot be directly applied to identify the aspect category of a given QA-style review. On the one hand, for a QA-style review, the aspect category referred in both question and answer texts is the valid aspect. Thus, the aspect-related matching information between the question and answer text is helpful for aspect category classification on QA-style reviews. We argue that the 2D textual representation is difficult to capture the semantic relationship between the question and its corresponding answer due to the possible long distance between them if we simply concatenate them into a sequence as the representation of a QA-style review for classification. On the other hand, the matching information between the question and answer sentence is more or less related to the annotated aspect category, but the 2D textual representation may not explore the correlation degree of the matching information between the question and answer sentence with the annotated aspect category.

Thus, we illustrate our 4-dimension (4D) textual representation model in Figure 2. In our 4D textual representation approach, the word-level representation is leveraged as the first dimension, and the sentence level is leveraged as the second dimension, which is similar to the representation of the traditional textual representation. Furthermore, different from traditional text representation approaches, another two dimensions, namely, QA interaction-level and hyperinteraction-level representations, are proposed as the third and fourth dimensions. Our empirical studies demonstrate the effectiveness of our proposals for the aspect category classification task on QA-style reviews.

The main contributions of our work are summarized as follows:

(i) We introduce the QA interaction-level dimension representation to capture the matching information between the question and answer sentence.

(ii) We propose the hyperinteraction-level dimension representation to explore the correlation degree of the matching information between the question and answer sentence with the annotated aspect category.

(iii) We conduct comparison experiments on our annotated datasets extracted from the domain of *Digit*, *Beauty*, and *Bag* in the famous website Taobao. We find that our proposed classification model, 4-dimension textual representation, tailored for QA-style reviews performs better than the 2D textual representation. Especially in the *Digit* domain, the accuracy and Macro-F1 of our proposed approach are, respectively, 7.5% and 10.7% higher than the 2D representation.

The rest of our paper is organized as follows. Section 2 discusses the related work on the aspect category classification. Section 3 presents data collection and annotation. Section 4 proposes our approach to aspect category classification on QA-style reviews. Section 5 reports the experimental setup, and Section 6 discusses and analyzes the results. Finally, Section 7 gives the conclusion and our future work.

2. Motivation

The existing learning approaches to the aspect category classification on traditional product reviews are deep learning-based approaches which first model each word as a real-valued vector, i.e., word-embedding phase. Then, the whole sentence or document is represented as a word-embedding sequence and trained with a sequence learning model, such as RNN [1] and LSTM [2]. In such approaches, the sentence or document text is represented by two dimensions, i.e., word-embedding dimension and word sequence dimension. In brief, we refer to this kind of representation as 2-dimension (2D) textual representation. Taking an example of traditional product reviews "I like beef very much!," the flow of aspect category classification based on the 2D representation is as shown in Figure 3.

QA-style aspect category classification task aims to identify the aspect category referred in both question and answer texts inside a given QA-style review. One straightforward way to deal with this task is to directly apply existing learning approaches based on the 2D representation to aspect category classification on other kinds of text styles. The part of results based on the 2D representation of QA-style reviews is illustrated in Table 1. We can find that if we adopt methods based on the 2D representation to deal with

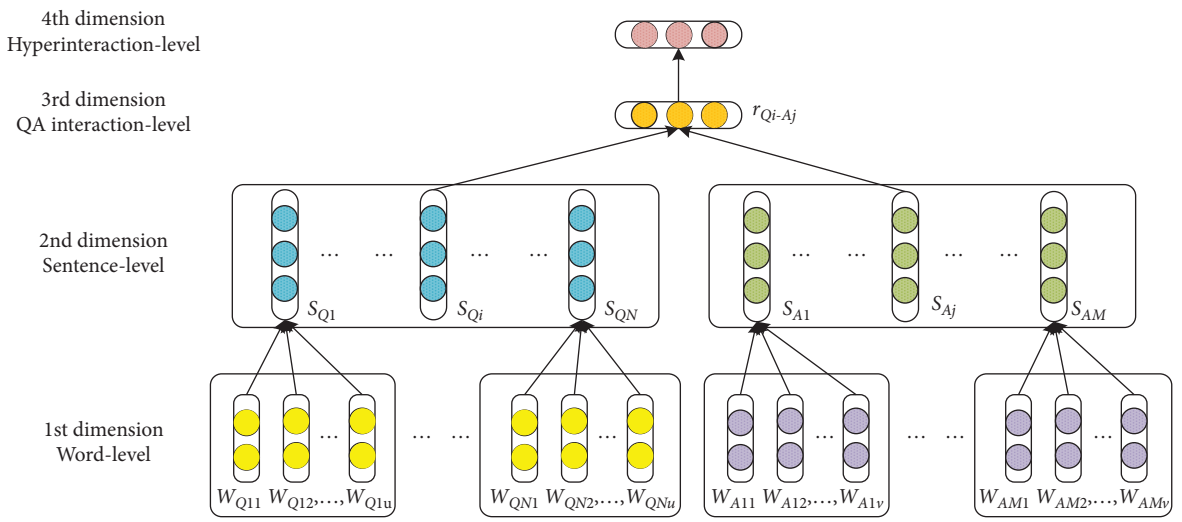

Figure 2: The overall architecture of the aspect category classification based on the 4D textual representation of QA-style reviews.

the aspect category classification task on QA-style reviews, the performance of classification is very bad. Thus, by analysis, the 2D representation is not the best choice for representing QA-style texts due to the following reasons.

First, in the QA-style review, it consists of the question text and answer text. And the question text and answer text are more likely to be two parallel units rather than a sequence form. For instance, in Figure 1, Answer 1 "The electricity is not very durable," is actually not following the last sentence in Question 1 "Is the quality of this phone good?" but is corresponding to the first sentence in Question 1 "How long does it last when you are playing games?" Therefore, when the question text and answer text are presented as two units in a sequence, it is rather difficult to capture the relationship between the question and its corresponding answer due to the possible long distance between them. A better way to handle this issue is to segment the question and answer text into some parallel sentences and capture the matching measurement between the question and answer sentence.

Second, in the question or answer text, there often exist different aspect categories in different sentences. And the matching information between the question and answer sentence is more or less related to the annotated aspect category. For instance, in Figure 1, the matching information between the sentence in Question 1 "How about its running?" and the sentence in Answer 2 "It has lots of running memory" is most related to the annotated aspect category "performance." A better choice to handle this issue is to explore the correlation degree of the matching information between the question and answer sentence with the annotated aspect category.

To summarize, we propose another two dimensions, i.e., QA interaction-level and hyperinteraction-level, to build a novel classification model based on the 4-dimension representation for the aspect category classification task on QA-style reviews. We leverage the QA interaction-level dimension to capture the matching measurement between the question and answer sentence and the hyperinteraction-level dimension to explore the matching information between the question and answer sentence with the annotated aspect category. Then, we will introduce our model in detail in the following section.

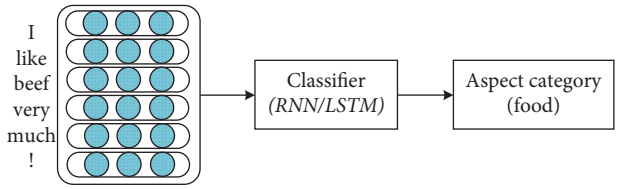

Figure 3: The flow of the aspect category classification on traditional product reviews.

3. Related Work

3.1. Aspect Category Classification. Aspect category classification, a basic task for aspect-level sentiment analysis [3], aims at identifying an aspect category referred in a given review, which is usually treated as a special case of the text classification task. Therefore, text classification approaches can be naturally applied to solve the aspect category classification task, such as SVM [4]. Alvarez-López et al. [5] used SVM to classify aspect categories based on restaurant review corpus in English and Spanish. Machacek [6] used the bigram model to solve the aspect category classification task. Kiritchenko et al. [7] used a set of binary SVMs with different types of n-grams and information from a specially designed lexicon. However, traditional machine learning approaches focus on sparse lexical features such as n-grams, one-hot vector representation, and term frequency-inverse document frequency (TF-IDF) to represent the text; these approaches highly depend on the quality of features, and feature engineering is labor-intensive.

Recently, with the spread of the word2vec model [8], neural network approaches based on CNNs [9, 10] or RNNs have shown promising results. Kim [9] pioneered the use of

Table 1: The part of results based on the 2D representation of QA-style reviews.

Question text	Answer text	Ground aspect	Predicted aspect	Yes/no
How about the quality of this phone? How long is the stand-by time?	I can make few phone calls and then it is off. The sound is not too loud.	Battery	Quality	No
Hello, are the products in this store authentic? How about the quality of products?	The quality is OK.	Quality	Certified product	No
Is it authentic?	Yes, it is. We can take clear pictures and it does not stutter.	Certified product	IO	No
How about its appearance?	It looks very good.	Appearance	Appearance	Yes

convolutional neural networks for text categorization and achieved better results than most traditional methods on related datasets. CNNs have strong feature expression ability and can be easily parallelized. However, it is not easy to determine the optimal size of the convolutional kernel. Tang et al. [11] proposed a method of text classification using RNNs. Compared to CNNs, RNNs are able to process variable-length sentences better and have a strong ability to relate context. Gated recurrent unit (GRU) networks and long short-term memory (LSTM) [2, 12] networks can mitigate the problem of gradient disappearance and gradient explosion in addition to preserving information for a longer time than plain RNNs.

Neural networks with attention mechanisms have achieved very good results in the field of machine reading comprehension, machine translation, etc., causing a lot of research around attention mechanisms [11, 13]. Yang et al. proposed a text classification approach that stratified documents according to their structure and then combined a bidirectional GRU network with an attention mechanism [14].

Considering methods used in the text classification, Toh and Su [15] adopted the sigmoidal feedforward network to train a binary classifier for the aspect category classification. Xue et al. [16] utilized the correlation between the aspect category classification task and the aspect term extraction task to perform joint learning. Wan et al. [17] proposed a representation learning algorithm for the aspect category classification. Those approaches are based on the 2D textual representation for the aspect category classification task. Besides, different from all of the above, our work devotes to the aspect category classification task on QA-style reviews in which little research focus on.

4. Model

In this section, we present our model to identify the aspect category of a given QA-style review. The basic idea of our approach is to learn a 4-dimension textual representation for the question text in a QA-style review. As introduced in Section 1, our model contains four dimensions for representing different levels' texts including word level, sentence level, QA interaction-level, and hyperinteraction level.

The 1st dimension and the 2nd dimension have been widely used in the previous studies in the text classification research area. In this paper, we adopt the word2vec model [18] to obtain the word-level dimension and apply the sequence-to-sequence neural network, i.e., bidirectional gated recurrent unit (Bi-GRU) [19], to obtain the sentence-level representation.

In the following sections, we propose the representation models for the 3rd dimension and the 4th dimension, i.e., QA interaction-level and hyperinteraction level.

4.1. The 3rd Dimension Representation. According to the characteristics of QA-style reviews, we can see the matching measurement between the question and answer text is important for the aspect category classification of QA-style reviews. In this section, we learn the matching representation between a sentence from the question text and a sentence from the answer text, i.e., learning sentence-sentence matching representation.

Figure 4 shows the architecture of the representation learning process for the sentence-sentence matching. Formally, we assume that the question text contains N sentences and is denoted as $[S_{Q_1}, S_{Q_2}, \ldots, S_{Q_N}]$, where S_{Q_i} is the sentence representation of the i-th sentence in the question text with words $[w_{Q_{i1}}, w_{Q_{i2}}, \ldots, w_{Q_{iu}}]$. Similarly, S_{A_j} is the sentence representation of the j-th sentence in the answer text with words $[w_{A_{j1}}, w_{A_{j2}}, \ldots, w_{A_{jv}}]$.

Formally, we use bidirectional GRU (Bi-GRU) layers, which can effectively utilize the forward and backward features, to get contextual representations of the question and answer sentences. Through the sequence-to-sequence Bi-GRU layers, the annotation of each word is produced by averaging the forward and backward hidden states. For a question sentence S_{Q_i}, we obtain the hidden state matrix H_{Q_i} by the following formulas. Similarly, we obtain the hidden state matrix of the answer sentence S_{A_j}.

$$\begin{aligned} \overrightarrow{R_{Q_i}} &= \overrightarrow{\mathrm{GRU}}\left(S_{Q_i}\right), \\ \overleftarrow{R_{Q_i}} &= \overleftarrow{\mathrm{GRU}}\left(S_{Q_i}\right), \\ R_Q &= \mathrm{AVE}\left(\overrightarrow{R_{Q_i}}; \overleftarrow{R_{Q_i}}\right). \end{aligned} \tag{1}$$

Then, we calculate the pairwise matching matrix, which represents the matching degree of the question sentence and the answer sentence. Given a question sentence S_{Q_i} and an answer sentence S_{A_j}, we can compute a matching matrix by using the following formula, i.e.,

$$\mathrm{Matching}\left(Q_i, A_j\right) = \tanh\left(W_{ij} \cdot \left(H_{Q_i}^T \cdot H_{A_j}\right) + b_{ij}\right), \tag{2}$$

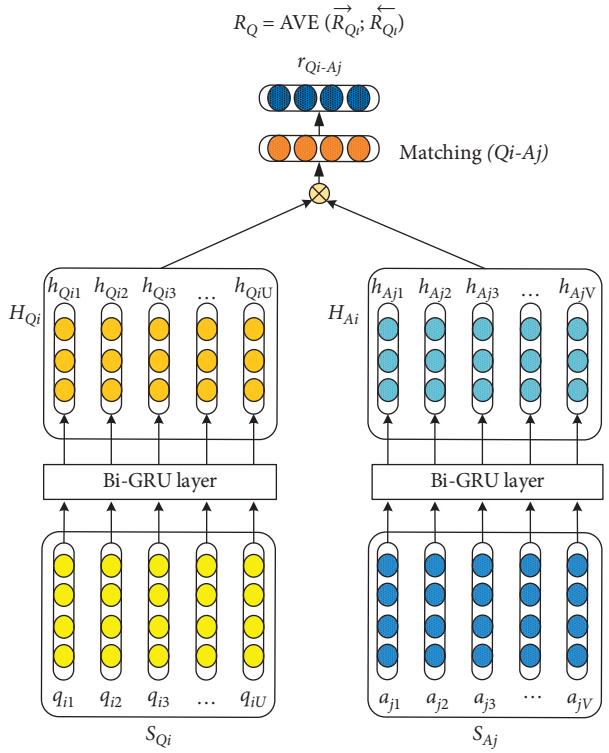

Figure 4: Learning sentence-sentence matching representation.

where Matching(Q_i, A_j) denotes the matching matrix between the two sentences, i.e., S_{Q_i} and S_{A_j}.

On this basis, we obtain the matching representation vector of the question sentence S_{Q_i} by the following formula:

$$r_{Q_i-A_j} = \left(H_{Q_i}\right)^T \cdot \text{softmax}\left(W_r \cdot \text{Matching}\left(Q_i, A_j\right)\right). \quad (3)$$

4.2. The 4th Dimension Representation. Formally, after obtaining all sentence-sentence matching vectors, we concatenate them into a new matrix $M = [\mathbf{r}_{\mathbf{Q}_1-\mathbf{A}_1}, \mathbf{r}_{\mathbf{Q}_1-\mathbf{A}_2}, \ldots, \mathbf{r}_{\mathbf{Q}_N-\mathbf{A}_M}]$. Furthermore, we obtain the correlation degree vector α, in which each value represents the correlation degree of the matching information between the question and answer sentence with the annotated aspect category as follows:

$$\begin{aligned} T &= \tanh\left(W_m \cdot M^T \cdot M + b_m\right), \\ \alpha &= \text{softmax}\left(W_t^T \cdot T\right). \end{aligned} \quad (4)$$

Finally, we can get the 4th dimension representation r, which finally transforms the QA interaction-level representation into the hyperinteraction-level representation, as shown in Figure 5.

$$r = s \cdot \alpha^T. \quad (5)$$

4.3. Classification Model. The hyperinteraction vector $\mathbf{r}$ is a high-level representation for the question text in a QA-style review, and it is concatenated into the sentence-level representation $\mathbf{h_q}$ of the unsplit question text as the final

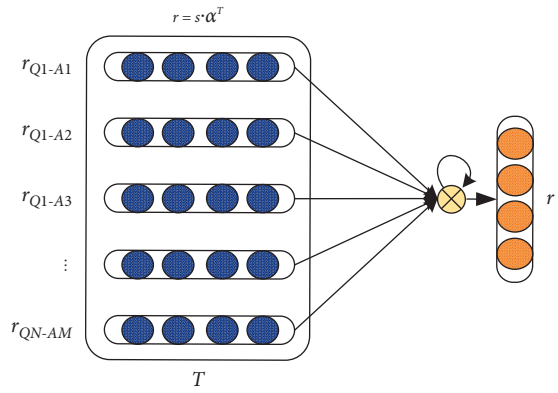

Figure 5: Learning hyperinteraction-level representation.

representation for the classification, where $\mathbf{h_q}$ is the last hidden vector obtained by the bi-GRU mode:

$$h_q^* = \tanh\left(W_p r + W_s h_q\right). \quad (6)$$

Finally, we put the final representation into the softmax layer to compute the conditional probability distribution. Then, the label with the highest probability is the predicted aspect category of a QA-style review:

$$y = \text{softmax}\left(W_l h_q + b_l\right). \quad (7)$$

To learn the whole model, we train our model end-to-end given the training data, and the goal of training is to minimize the cross-entropy loss function:

$$J(\theta) = -\sum_{t=1}^{K}\sum_{k=1}^{C_i} y_t^k \cdot \log\left(S_{Q_t}, S_{A_t}\right) + \frac{l}{2}\|\theta\|_2^2, \quad (8)$$

where K is the number of training data and C is the number of all aspect categories. Besides, $\varnothing(S_{Q_t}, S_{A_t})$ is a black-box function whose output represents a vector representing the probability of aspects, and l is a L_2-regularization term.

5. Experiments

In this section, we first introduce our models and baselines in Section 5.1. Then, we describe experimental settings, i.e., the datasets and evaluation metrics used in our experiment in Section 5.2. Finally, we describe details of the model setup in Section 5.3.

5.1. Model Summary. Since the word-level and sentence-level dimension representation have been widely studied in the natural language processing (NLP) area, we use approaches based on the two dimensions as our baseline methods.

For thorough comparison, we implement several different models for representing the question text in a QA-style review, which are illustrated in the following (the baselines and our proposed models are distinguished with the marker ∘ and ∗, respectively):

(i) A(2D) ∘: this approach employs the word-level and sentence-level dimension representations for answer texts

(ii) Q(2D) ∘: this approach employs the word-level and sentence-level dimension representations for question texts

(iii) Q + A(2D) ∘: this approach employs the word-level and sentence-level dimension representations for the concatenation of question and answer texts

(iv) Q(3D) ∗: this approach employs the word-level, sentence-level, and QA interaction-level dimensions for question texts

(v) Q(4D) ∗: this approach employs the word-level, sentence-level, QA interaction-level, and hyperinteraction-level dimensions for question texts

5.2. Experimental Settings

(i) Datasets: we conduct our experiments on the high-quality annotated corpus composed of QA-style reviews from *Digit*, *Beauty*, and *Bag* domains in Taobao, which conform to our annotation guidelines. Considering the imbalanced distribution of data, we eliminate aspect categories which contain less than 50 reviews. The statistics of the experimental datasets are summarized in Table 2.

(ii) Evaluation metrics: we use the standard accuracy and Macro-F1(F) to evaluate the overall QA aspect classification performance, and Macro-F1 is calculated by the formula $F = 2PR/(P + R)$, where the overall precision P and recall R are the average value of the precision/recall of all aspect categories.

5.3. Model Configuration

(i) Data processing: in order to construct the 4-dimension representation model, we present the sentence segmentation Algorithm 1 based on the Standford CoreNLP toolkit [20] to segment the question and answer text into sentences. Besides, we use word embedding to initialize the words in our datasets and pretrain the word embeddings with our crawled dataset containing 320 thousand QA-style reviews.

(ii) Model setup: in the experiment, we initialize all the out-of-vocabulary words by sampling from the uniform distribution $U(-0.01, 0.01)$. The word-embedding dimension and LSTM hidden state dimension are set to be 100, and the batch size is set to be 32. Considering that QA-style reviews are short texts, the minimum number of words in a sentence $N_{\min}$ is 5. The other parameters are tuned according to the development data. In the training process, we use the gradient descent approach to train our models, and the learning rate is 0.02, and the dropout rate is set to be 0.25 to avoid overfitting. Besides, the optimal number of sentences in the question or answer texts is tuned to be 2.

5.4. Research Question. As described in Section 4, our 4-dimension textual representation model for the QA aspect classification task is based on QA interaction-level and hyperinteraction-level mechanisms, which are built on sentence level. Thus, the impact on performance of the sentence number inside the question and answer texts is the key research question. Accordingly, we will discuss and analyze in the following section.

6. Results and Analysis

In Section 6.1, we compare the classification performance of our proposed approaches and other baselines on the datasets from three domains. Then, Section 6.2 analyzes the influence on the classification performance of the sentence number in the question and answer texts. Besides, the error analysis of misclassified QA-style reviews is illustrated in Section 6.3.

6.1. Performance Comparison. We adopt the holdout method to compare the performance of the approaches described in Section 4.1. In the holdout method, for each dataset from one domain, we set aside 10% from the training data as the development data by stratified sampling to tune learning algorithm parameters. Figures 6 and 7 give the experimental results of all discussed approaches.

From the results, we can see that

(1) In three domains, all Q(2D) approaches are fairly superior to A(2D) approaches, which demonstrates that question texts contain more aspect-related information than answer texts and contribute more in identifying the aspect category of a given QA-style review.

(2) Clearly, among all 2D approaches, when the question and answer texts are both employed in our task, Q + A(2D) approaches perform best. This means that we can utilize auxiliary information contained in answer texts to further improve the performance of the aspect category classification task on QA-style reviews.

(3) In the *Digit* domain, our proposed approach Q(3D) based on the QA interaction-level dimension achieves a definite improvement of 7.5% (accuracy) and 9.5% (Macro-F1) compared with Q(2D) approach, which indicates the importance of capturing the matching information between the sentence inside the question text and the sentence inside the answer text. Furthermore, in *Beauty* and *Bag* domains, the accuracy and Macro-F1 are both increased, but the increase rate is not significant.

(4) Note that our 4-dimension textual representation approach Q(4D) using both QA interaction-level and hyperinteraction-level dimensions achieves the best performance among all approaches. In the *Digit* domain, the accuracy and Macro-F1 of Q(4D) approach are 1.3% and 1.2% higher than Q(3D) approach which only employ the QA interaction-level dimension, respectively. In the *Beauty* domain,

Table 2: Statistics of experimental datasets (the QA-style reviews not in accordance with our annotation guidelines are excluded).

Domains	Digit	Beauty	Bag
Aspect categories	7	10	11
Total QA-style reviews	2,566	3,065	3,077
Maximal number of question sentences	6	5	4
Proportion of one question sentence and one answer sentence (%)	26.4	38.0	28.0
Maximal length of question texts	71	48	42
Aspect containing most QA-style reviews	IO	Effect	Quality
Number of most QA-style reviews in one aspect	1,044	911	868

Input: Question or Answer text $S = \{w_i \mid w_i \text{is a word}\}$;
$N_{\min}$: the minimum number of words in a sentence;
$N_{\max}$: the maximum number of sentences in the answer text
Output: All sentences (Stored in $C = \{c_i\}$) mined from S that satisfy $N_{\min}$ and $N_{\max}$

(1) $C = \varnothing$
(2) C_{temp} = null // the candidate sentence
(3) Segment S into n sentences $\{\sigma_1, \sigma_2, \ldots, \sigma_v\}$ with Stanford CoreNLP toolkit;
(4) **for** $i = 1; i \le |S| - 1; i+ = 1$ **do**
(5) **if** $|C| \ge N_{\max}$ **then**
(6) **break**;
(7) **end**
(8) **if** s_i.length $> N_{\min}$ **then**
(9) $C_{\text{temp}} = s_i$;
(10) $C = C \cup C_{\text{temp}}$;
(11) **else**
(12) $j = i + 1$;
(13) **while** $j \le |S_A| - 1$ **do**
(14) $C_{\text{temp}} = s_i + s_j$;
(15) **if** C_{temp}.length $\ge s_j$ **then**
(16) $C = C \cup C_{\text{temp}}$;
(17) **else**
(18) $j += l$;
(19) **end**
(20) **end**
(21) $i = i + j$;
(22) **end**
(23) **End**

Algorithm 1: Sentence segmentation algorithm.

Figure 6: The comparison of accuracy in three domains.

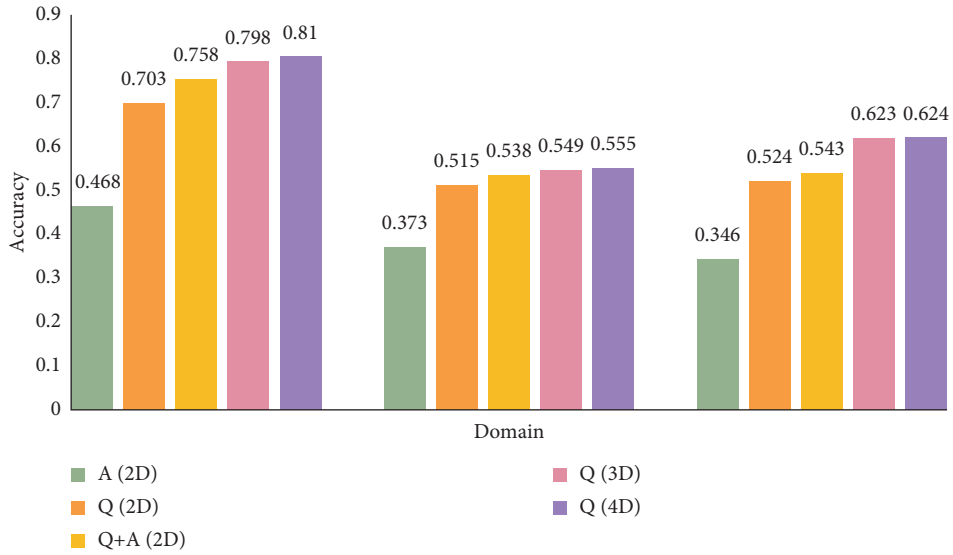

Figure 7: The comparison of F1 value in three domains.

Q(4D) approach achieves the improvement of 1.7% (accuracy) and 0.5% (Macro-F1), and in the *Bag* domain, the improvement of accuracy and Macro-F1 is 3.4% and 0.1%, respectively. This demonstrates that our proposed approach Q(4D) using the hyperinteraction-level dimension can effectively capture the importance degree of the question sentence and its aspect-related matching answer sentence for our task and can further improve the performance of the classification task. Furthermore, significance test, t-test, shows that this improvement is significant (p value <0.05).

(5) In our previous paper, the accuracy of the aspect category classification is 0.865 in the *Digit* domain. We can see that the performance of the multi-attention model is worse than our 4-dimension textual representation model.

On the basis of the above analysis, our proposals can be applied to QA-style reviews in three domains and improve the classification performance compared with the traditional sentence-level textual representation method, especially in the *Digit* domain.

6.2. Impact of Sentence Numbers on the Question or Answer Text. To answer the research question in Section 5.4, we examine the performance of our 4-dimension textual representation model with various sentence numbers of question and answer texts ranging from one to four according to the statistics of experimental data in each domain. We present results on classification performance in Tables 3–5 on QA-style reviews from *Digit*, *Beauty*, and *Bag* domains, in which the columns represent the number of question sentences, and the rows represent the number of answer sentences.

Clearly, for three datasets, we can find that, under the circumstances that the number of question sentences is not equal to the number of answer sentences, when the number of question sentences is fixed, the accuracy and Macro-F1 are both improved with the increase in the number of answer sentences. However, when the number of answer sentences is fixed, the classification performance becomes worse with the increase in the number of answer sentences, even worse than the traditional Q(2D) representation approaches which employs word-level and sentence-level representations. This would be that our classification task mainly depends on question texts, while the answer texts can assist question texts to further improve the performance of the aspect category classification task on QA-style reviews.

When they are equal in number, our model can achieve a tradeoff performance improvement. In particular, when the number of sentences in the question and answer texts is equal to 2, our proposals achieve the best performance in *Digit*, *Beauty*, and *Bag* domains, especially in the *Digit* domain. Besides, the number of QA-style reviews confirming to our annotation guidelines is not very large, and in *Beauty* and *Bag* domains, there are 10 and 11 aspect categories, respectively, which are both more than those in the *Digit* domain. Thus, the performance improvement in the *Beauty* or *Bag* domain is less significant than that in the *Digit* domain.

In addition, because QA-style reviews from the Taobao website are short texts, the number of sentences inside question or answer texts used in our 4-dimension textual representation model is not the more, the better. Particularly for question texts, as shown in Table 1, the maximal length of question texts in three domains is 71. If we segment them into more sentences, it could be counterproductive.

According to the above analysis, considering that QA-style reviews from other e-commerce platforms are similar to our corpus in format and expression, our proposals could be applied to the aspect category classification task on QA-style reviews on these QA-style reviews.

Table 3: Accuracy and Macro-F1 on the aspect category classification task on QA-style reviews with various sentence numbers in the *Digit* domain.

	Question							
Answer	1		2		3		4	
	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
1	—	—	0.783	0.680	0.752	0.641	0.742	0.611
2	0.792	0.727	**0.871**	**0.810**	0.781	0.676	0.744	0.623
3	0.816	0.753	0.785	0.698	**0.806**	**0.682**	0.767	0.655
4	0.818	0.755	0.808	0.704	0.800	0.692	**0.755**	**0.677**

Table 4: Accuracy and Macro-F1 on the aspect category classification task on QA-style reviews with various sentence numbers in the *Beauty* domain.

	Question							
Answer	1		2		3		4	
	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
1	—	—	0.636	0.504	0.634	0.498	0.612	0.492
2	0.677	0.535	**0.702**	**0.555**	0.650	0.513	0.629	0.511
3	0.681	0.544	0.645	0.515	**0.664**	**0.538**	0.648	0.516
4	0.683	0.542	0.662	0.553	0.652	0.517	**0.654**	**0.533**

Table 5: Accuracy and Macro-F1 on the aspect category classification task on QA-style reviews with various sentence numbers in the *Bag* domain.

	Question							
Answer	1		2		3		4	
	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
1	—	—	0.654	0.562	0.648	0.558	0.631	0.551
2	0.664	0.603	**0.712**	**0.624**	0.659	0.560	0.641	0.558
3	0.689	0.606	0.668	0.591	**0.666**	**0.598**	0.659	0.577
4	0.701	0.619	0.671	0.594	0.664	0.574	**0.661**	**0.583**

6.3. Error Analysis. According to the analysis of misclassified QA-style reviews in *Digit*, *Beauty* and *Bag* domains, we find some main reasons for misclassification as follows.

First, the imbalanced distribution of experiment data may lead to that these QA-style reviews tend to be predicted as the aspects which contain more QA-style reviews. For instance, in the *Digit* domain, the predicted aspect of 22.95% of misclassification QA-style reviews is IO. Similarly, in the *Beauty* domain is Effect and in *Bag* is Quality.

Second, the QA-style reviews are colloquial, so the existing word segmentation toolkits may not segment words well, which can influence the pretraining of word embeddings and the word-level textual representation.

Third, some annotated aspect terms are ambiguous in different contexts. We can consciously determine the meaning of these words according to the context and categorize them into the correct aspect category, which is still difficult for a well-trained machine at present.

Fourth, QA-style reviews used in our experiments are short texts, and our classification task mainly depends on question texts. However, as shown in Table 1, the maximal length of question texts is 71 in the *Digit* domain, while in the other two domains, its value is 42. After sentence segmentation, the inputs for the neural network are much shorter. This may lead to that our model could not be trained very well.

Last but not the least, sentence segmentation algorithm based on the Stanford CoreNLP toolkit is not good enough for QA-style reviews because it may ignore the syntax between the sentences in the question or answer text.

7. Conclusions

In this paper, we address a novel aspect category classification task against QA-style reviews, which aims at automatically classifying the aspect category of a given QA-style review and builds a high-quality annotated corpus. To solve this task, we propose a 4-dimension textual representation model based on QA interaction-level and hyperinteraction-level dimensions to capture the aspect-related matching information between the question and answer texts as much as possible.

Our experiment results on three manually annotated datasets, i.e., *Digit*, *Beauty*, and *Bag* datasets, demonstrate that our proposed approaches significantly outperform the baseline approaches, i.e., the textual representation based on sentence-level and word-level dimensions. For our proposed approaches Q(3D) and Q(4D), Q(4D) clearly performs better than Q(3D). In detail, Q(4D) presents an improvement ranging from 3.7% to 5.8% in terms of accuracy against the best baseline, i.e., Q + A(2D).

In our future work, we would like to solve some challenges in the aspect category classification task on QA-style reviews according to the error analysis, such as short question texts, imbalanced data distribution, and syntax parsing, to further improve the performance of this task. Furthermore, we would like to combine char embeddings with word embeddings to better represent the colloquial reviews.

Data Availability

Training data used in our experiment are mainly from "Asking All" in Taobao. The data used to support the findings of this study are available from the corresponding author upon request.

Disclosure

The funding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported in part by Industrial Prospective Project of Jiangsu Technology Department under Grant no. BE2017081 and the National Natural Science Foundation of China under Grant no. 61572129.

References

[1] T. Mikolov, M. Karafiat, L. Burget, J. Cemocky, and S. Khudanpur, "Recurrent neural network based language model," in *Proceedings of the Eleventh Annual Conference of the International Speech Communication Association*, Chiba, Japan, 2010.

[2] Y. Wang, M. Huang, L. Zhao et al., "Attention-based Istm for aspect-level sentiment classification," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 606–615, Austin, TX, USA, 2016.

[3] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[4] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent twitter sentiment classification," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics*, pp. 151–160, Portland, OR, USA, 2011.

[5] T. Alvarez-López, J. Juncal-Martinez, M. Fernández-Gavilanes et al., "Gti at semeval-2016 task 5: svm and crf for aspect detection and unsupervised aspect based sentiment analysis," in *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pp. 306–311, San Diego, CA, USA, 2016.

[6] J. Machacek, "BUTknot at SemEval-2016 Task 5: supervised machine learning with term substitution approach in aspect category detection," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 301–305, San Diego, CA, USA, 2016.

[7] S. Kiritchenko, X. Zhu, C. Cherry, and S. M. Mohammad, "NRCCananda-2014: detecting aspects and sentiment in customer reviews," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 437–442, Dublin, Ireland, 2014.

[8] Z. Su, H. Xu, D. Zhang, and Y. Xu, "Chinese sentiment classification using a neural network tool—word2vec," in *Proceedings of the 2014 International Conference on Multisensor Fusion and Information Integration for Intelligent Systems (MFI)*, IEEE, Beijing, China, pp. 1–6, 2014.

[9] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, Doha, Qatar, 2014.

[10] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," 2014, https://arxiv.org/abs/1404.2188.

[11] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in *Proceedings of the EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, pp. 1422–1432, Lisbon, Portugal, September 2015.

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

[13] N. Vaswani, N. Shazeer, J. Parmar et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[14] Z. Yang, D. Yang, C. Dyer, X. He et al., "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1480–1489, 2016.

[15] Z. Toh and J. Su, "NLANGP: supervised machine learning system for aspect category classification and opinion target extraction," in *Proceedings of the International Workshop on Semantic Evaluation*, pp. 496–501, Denver, CO, USA, 2015.

[16] W. Xue, W. Zhou, T. Li, and Q. Wang, "MTNA: a neural multi-task model for aspect category classification and aspect term extraction on restaurant reviews," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, Asian Federation of Natural Language Processing, Taipei, Taiwan, pp. 151–156, 2017.

[17] X. Zhou, X. Wan, and J. Xiao, "Representation learning for aspect category detection in online reviews," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, Austin, TX, USA, 2015.

[18] T. Mikolov, I. Sutskever, K. Chen et al., "Distributed representations of words andphrases and their compositionality," *Advances in Neural Information Processing Systems*, pp. 3111–3119, 2013.

[19] B. Dhingra, H. Liu, Z. Yang, W. Cohen, and R. Salakhutdinov, "Gated-attention readers for text comprehension," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1832–1846, Vancouver, Canada, 2017.

[20] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60, Baltimore, MD, USA, 2014.