future internet

MDPI

Article

Feature Fusion Text Classification Model Combining CNN and BiGRU with Multi-Attention Mechanism

Jingren Zhang [1], Fang'ai Liu [1,*], Weizhi Xu [1] and Hui Yu [2]

[1] School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China; Ryan_sdnu@163.com (J.Z.); xuweizhi@sdnu.edu.cn (W.X.)
[2] School of Business, Shandong Normal University, Jinan 250358, China; huiyu0117@163.com
* Correspondence: lfa@sdnu.edu.cn

Received: 3 September 2019; Accepted: 7 November 2019; Published: 12 November 2019

Abstract: Convolutional neural networks (CNN) and long short-term memory (LSTM) have gained wide recognition in the field of natural language processing. However, due to the pre- and post-dependence of natural language structure, relying solely on CNN to implement text categorization will ignore the contextual meaning of words and bidirectional long short-term memory (BiLSTM). The feature fusion model is divided into a multiple attention (MATT) CNN model and a bi-directional gated recurrent unit (BiGRU) model. The CNN model inputs the word vector (word vector attention, part of speech attention, position attention) that has been labeled by the attention mechanism into our multi-attention mechanism CNN model. Obtaining the influence intensity of the target keyword on the sentiment polarity of the sentence, and forming the first dimension of the sentiment classification, the BiGRU model replaces the original BiLSTM and extracts the global semantic features of the sentence level to form the second dimension of sentiment classification. Then, using PCA to reduce the dimension of the two-dimensional fusion vector, we finally obtain a classification result combining two dimensions of keywords and sentences. The experimental results show that the proposed MATT-CNN+BiGRU fusion model has 5.94% and 11.01% higher classification accuracy on the MRD and SemEval2016 datasets, respectively, than the mainstream CNN+BiLSTM method.

Keywords: BiGRU; multi-attention; MATT-CNN+BiGRU; PCA

1. Introduction

Accompanying the continuous development of social networks, the role of Internet users also has changed quietly from the original recipient of information to the creator of information. More and more people are expressing their opinions through the Internet and, gradually, form a short text-based expression. The text information has a large amount of data, and the text content is scattered and disorganized. It is difficult to distinguish. Therefore, how to use natural language processing related techniques to analyze the emotional polarity of short texts in social networks has become one of the hotspots of current research [1,2].

Deep learning is an important branch of machine learning. Deep learning is an algorithm that uses high-order abstraction of multiple nonlinear transformation structures [3]. Recently, more and more researchers have begun to use deep learning technology in the field of image and speech recognition to solve the problem of text sentiment classification.

Kim [4] proposed an English classification model, taking preprocessed word vectors as input and using convolutional neural networks (CNN) to achieve sentence-level classification tasks. Xing et al. [5] used CNN to solve the polarity judgment problem of twitter. Although CNN has made great breakthroughs in the field of text categorization, it pays more attention to local features and ignores the contextual meaning of words, thus affecting the accuracy of classification. Therefore, improvements to

the CNN model have been ongoing. Zhao et al. [6,7] proposed capsule networks and dynamic routing based on convolutional neural networks and achieved better classification results than traditional CNN. Mikolov et al. proposed applying the RNN model to text classification tasks [8]. Since the output value of the current node of the RNN is determined by the current input and output of the previous node, the word structure of the text is considered, but the RNN is prone to problems such as gradient dispersion. Although long short-term memory (LSTM) can solve the gradient dispersion problem existing in RNN, LSTM has a large computational complexity due to its structural complexity and stores a large number of redundant intermediate variables. Therefore, a large amount of training time and memory space is required, and it is overly dependent on historical information and cannot utilize future information. Bidirectional gated recurrent unit (BiGRU) is a variant of bidirectionally long short-term memory (BiLSTM) and GRU. BiGRU structurally streamlines the BiLSTM three-door (forget, input, output) structure into two gates, namely update and reset, so fewer parameters accelerate the convergence of the model. Simultaneously, BiGRU combines the characteristics of short text content with high context, fully considers the meaning of words in context, and overcomes the problem of semantic information after the original GRU cannot consider words.

The attention mechanism was first applied in the field of image processing. Mnih [9], in 2014, proposed to use the attention mechanism in the image classification task and achieved good experimental results. Yin [10], in 2015, (and others) proposed one based on multi-layer attention. The convolutional neural network of the force mechanism is applied to the sentence modeling to better capture the local text features and verify the effectiveness of the combination of the attention mechanism and the convolutional neural network.

Aspect-based sentiment analysis (ABSA), as an important sub-task of sentiment analysis, is a deeper sentiment analysis with the goal of identifying emotional polarity in different aspect contexts (positive, neutral, negative). Given a sentence "The pizza in this restaurant is very delicious, but the quality of the takeaway pasta is very poor.", for example, the emotional polarity of the term "pasta" is negative, while the aspect is of "out of the box". The emotional polarity is opposite to the emotional polarity of "pizza". Therefore, even if it is the same sentence, there may be a completely opposite emotional polarity for different goals. The current research finds that the combination of the CNN network and the attention mechanism can obtain very good target-specific emotional classification results [11]. It can well solve the shortcomings of LSTM which cannot accurately indicate the importance of each word in the sentence. The proposed fusion model combines this advantage and the CNN model with a multiple-attention mechanism that is proposed to obtain the emotional polarity of keywords, which is an important dimension of the emotional classification of fusion models.

The main contributions of this paper are as follows:

1. Aiming at the current problems in the field of short text sentiment classification, this paper proposes a feature fusion text classification model combining CNN and BiGRU with a multi-attention mechanism based on previous research.

2. The model proposed in this paper is divided into two models: CNN and BiGRU. The CNN model combines the specific target sentiment classification method to extract the emotional polarity of the target keyword in the sentence. BiGRU analyzes the sentence-level emotional polarity and, finally, the two features are merged to construct a fusion global feature vector.

3. This paper uses the PCA dimension reduction technology to effectively reduce the feature fusion global vector.

4. The multi-attention convolutional neural network model has a simple structure and does not require external knowledge, such as dependency syntax analysis and semantic dependency analysis and does not require additional vectorization for specific targets.

5. We propose a two-way scanning algorithm that can identify effectively the extent of different words in the sentence so the convolutional neural network can make full use of the location information of the extracted keywords.

The model proposed in this paper combines the advantages of a convolutional neural network to extract local features and uses BiGRU to take into account the global features of the text and consider the characteristics of the contextual semantic information, which overcomes the problem that the original GRU and LSTM cannot consider—the information after the word. Furthermore, the classification accuracy of the feature fusion model is improved. The proposed MATT-CNN+BiGRU fusion model applied to the mainstream datasets of MRD and SemEval2016 results in 5.94% and 11.01% higher accuracy, respectively, than using the mainstream CNN+BiLSTM method.

The rest of this paper is organized as follows: Section 2 mainly discusses the attention mechanism and the contribution of aspect-level sentiment analysis (ASA) to the field; Section 3 mainly is based on the current mainstream models and technologies in the field of scientific research and the step-by-step of our proposed fusion model. It is elaborated, and the structural principles of the three attention mechanisms and the construction method of the fusion model are introduced in detail. Section 4 gives our experimental results from different dimensions. Section 5 summarizes the content of this article and provides some directions for future research.

2. Related Work

Text sentiment classification is one of the important tasks in the field of natural language processing (NLP). So far, many researchers have conducted in-depth research on the field of text sentiment classification. During the early stage of the research, due to the relatively small size of the dataset, the machine learning-based classification algorithm demonstrated its superiority in the text classification task. Among them, Sun et al. [12] proposed an SVM based on unbalanced text classification. The classifier solves this problem and has achieved good research results. However, as the scale of datasets continues to expand, the flourishing development of deep learning algorithms provides new ideas for text classification.

While deep learning has made breakthroughs in various fields, it also has made important developments in data fusion. Deep learning has the ability to fit large amounts of data. Through layer-by-layer extraction, deep neural networks extract different data features at different layers. Jing [13] proposed an adaptive multi-sensor data fusion method based on a deep convolutional neural network (DCNN) for fault diagnosis. The proposed method can learn features from raw data and adaptively optimize different fusions with combination of levels to meet the requirements of any troubleshooting task. Audebert [14] studied the use of a deep full convolutional neural network (DFCNN) in pixel-level scene markers of Earth observation images in the image field and achieved good experimental results. Bakalos et al. [15] used multi-modal data fusion and adaptive deep learning to monitor critical water infrastructure and also gained valuable application results.

Recently, the attention mechanism has been applied widely in the text classification model. The advantage is that the different importance levels of each word of the article can be distinguished in the classification task. Yang et al. [16] proposed to use the attention mechanism on words and sentences while preserving the document structure with a hierarchical structure, so as to distinguish the importance of each sentence and word to the classification category. Wang et al. [17] proposed a joint embedding model, combining words and labels on ACL 2018, which maintains the interpretability of word embedding and has the ability to use other information than the input text sequence. Offering a good experimental effect, Vaswani [18] (and others) proposed a multi-head attention mechanism and a transformer framework in their research, providing a new idea for attention in the field of text classification. Chen et al. [19] used BiLSTM and a positional attention mechanism combined with mixed neural for text classification. Rozental et al. [20] proposed a hybrid model using a BiGRU neural network and convolution maximum pooling to extract text feature information. Kumar et al. [21], using the simple BiLSTM and two-layer attention mechanism to mix and match the model, after extracting the text feature information from the BiLSTM layer, sequentially established the word-level attention mechanism and the sentence-level attention mechanism, and the experimental model was

tested on the SemEval2017Task5. The evaluation results of sub-tracks 1 and 2 were 1.7% and 3.7% higher, respectively, than the current best systems.

Aspect-Based Sentiment Analysis

Aspect-based sentiment analysis (ABSA) is an important sub-task of sentiment analysis. Compared with traditional sentiment classification tasks, the language granularity is more precise, the level is deeper, and targets emotional polarity. The judgment depends, not only on the context of the text information but, also, on the feature information of the specific target. Recently, with the rapid development of deep learning technology, many researchers have begun to use deep learning techniques to carry out specific target sentiment classification tasks. Nguyen et al. [22] proposed a feature sentiment analysis model based on RNN and a dependency tree. Ruder et al. [23] proposed a hierarchical two-way LSTM network, which allows text features to be extracted at different levels, which can learn effectively the internal structural relationship and grammatical relationship of sentences so the emotional polarity of a specific target can be discriminated. The method is limited to situations involving a specific target, however. Based on this, Zhou [24] proposed an LSTM network combining an attention mechanism. Focusing on the target word vectorization, the attention mechanism is embedded into the LSTM network, so the network pays more attention to the target word itself during the training process, which can effectively identify the emotional polarity of different targets. Song et al. [25] proposed the AEN model. The AEN model avoids complex recurrent neural networks, uses attention-based encoders to model context and target, and can extract rich introspective and interactive semantic information from word embedding. Additionally, Mohammadi [26–28] also provided some new methods for text data noise processing.

Through the in-depth analysis of the short text language organization of the network, combined with the actual content of the SemEval2016 and MRD datasets, we found that the emotional polarity of key target words often directly reflects the main emotional polarity of sentences in a clause and has a strong indicator meaning. The sentence, "The atmosphere of this movie is magnificent, although it still has room for improvement", for example, the emotional polarity of the target word "atmosphere" is positive, although the latter emotions are retained, but still can be consistent with the emotional polarity of the entire clause. Therefore, the model proposed in this paper adopts the specific target sentiment analysis of the fusion distance attention mechanism in the CNN model to enhance the ability to grasp the fine-grained content. Then, the sentence-level BiGRU model is used to fully consider the context information of the article and construct the text classification fusion model, thereby improving the accuracy of the classification.

3. MATT-CNN+BiGRU Feature Fusion Model

3.1. The Brief Introduction of MATT-CNN+BiGRU

The proposed feature fusion model is divided mainly into two major models: the convolutional neural network model with multiple attention mechanisms, and the BiGRU network model. Subsequent to reading the research results of our predecessors, we found that the CNN model has obvious advantages in the extraction of target keywords [29]. To a certain extent, the target keyword directly determines the emotional polarity of the short text content.

We use the CNN model to analyze the pre-processed target keywords and obtain the emotional polarity of the feature keywords to form a dimension of our proposed fusion model. We then combine BiGRU's content superiority to the short text overall context to get the emotional polarity of the sentence level and form the second dimension. Then, feature vector fusion is performed in the concat layer, and the final classification result is output by using the softmax classifier after the PCA is reduced in dimension. This article describes the models and techniques used in Section 3.1, and the following sections describe the specific structure of the fusion model we propose.

3.1.1. Principal Component Analysis

Principal component analysis, also known as (PCA), is a dimensionality reduction method based on mathematical statistics. PCA uses the idea of dimensionality reduction to first standardize data in high-dimensional data and, then, the covariance matrix of the target matrix and its corresponding eigenvectors are obtained. Finally, the initial data is transformed into a linearly independent representation of arbitrary dimensions by linear transformation, thereby transforming multiple indicators (high dimensions) into a few major feature components. The basic principle of PCA is:

Let the raw data $A_{m\times n}$ set be arranged as a matrix, then each row element of the matrix is zero-averaged and the calculation expression is as shown in Equation (1):

$$A_{ij} = \frac{a_{ij} - \overline{a}_i}{S_i}, \tag{1}$$

where a_{ij} represents the element of the ith row and j-column of the matrix $A_{m\times n}$, and $\overline{a}_i$ represents the ith row of the matrix $A_{m\times n}$. The average value, S_{i} represents the standard deviation of the ith row of the matrix $A_{m\times n}$. Following zero-mean normalization, the covariance matrix S of the matrix X is obtained. Equation (2) is as follows:

$$S = \frac{1}{\mathrm{n}-1}\sum\nolimits_{k=0}^{j} (a_j - \overline{a})(a_j - \overline{a})^T, \tag{2}$$

where n represents the number of samples in the above formula and, after obtaining the covariance matrix S, the eigenvalue $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ and the eigenvector $d_1, d_2 \ldots d_n$; are obtained accordingly.

The feature vector is obtained by arranging the feature values from largest to smallest to obtain the matrix P, and the reduced-dimensional data Y is obtained by Equation (3):

$$Y = PA. \tag{3}$$

Finally, the contribution rate V_{i} of each feature root is calculated according to Equation (4):

$$V_i = \frac{\lambda_i}{\sum_1^n \lambda_i}. \tag{4}$$

3.1.2. Basic Structure of GRU and BiGRU

Along with the rapid development of LSTM in the field of natural language processing, especially text categorization, and the increase in the number of samples, the training time is long, the parameters are many, and the internal computational complexity is high. Based on this, Cho et al. proposed a simpler GRU model in 2014 [30]. The GRU model maintains the original LSTM original effect, with a simpler structure, fewer parameters, a better convergence model, plus a GRU model. It consists of two doors, an update door and a reset door. The update gate determines the extent to which the previous output hidden layer affects the current layer. The larger the value is, the stronger the influence is. The reset gate determines the extent to which the previous hidden layer information is ignored. The smaller the reset gate value is, the more ignored the information is. The specific structure of the GRU is shown in Figure 1:

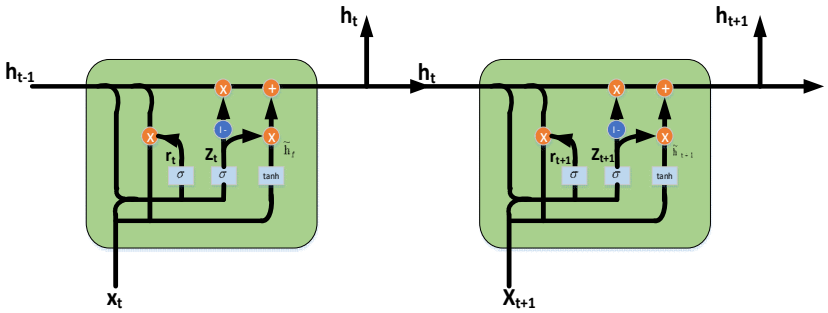

Figure 1. Structure of a gated recurrent unit (GRU).

The GRU model is updated in the following ways:

$$r_t = \sigma(W_r * [h_{t-1}, x_t]) \tag{5}$$

$$z_t = \sigma(W_z * [h_{t-1}, x_t]) \tag{6}$$

$$\widetilde{h}_t = \tanh\left(W_{\widetilde{h}} * [r_t * h_{t-1}, x_t]\right) \tag{7}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \widetilde{h}_t \tag{8}$$

where r_t represents the reset gate at time t, Z_t represents the update gate at time $\widetilde{h}_t$, h_t represents the candidate activation state at time t, h_{t-1} represents the active state at time t, and h_{t-1} represents the hidden layer state at time $(t-1)$. The update gate z is determined by the history information that the current state needs to be forgotten and the new information that is accepted; the reset gate r is determined by the information obtained from the history information of the candidate state.

GRU is a kind of one-way neural network structure, and a one-way neural network is always output from the back. However, in the text sentiment classification, the output of the current moment often has a relationship with the moments before and after. BiGRU is a one-way, opposite direction, and the output is determined by the GRU common state of the two GRUs to construct a neural network model. Occurring at each moment, the input provides two GRUs in opposite directions, and the output is determined by the two unidirectional GRUs. The specific structure of BiGRU is shown in Figure 2.

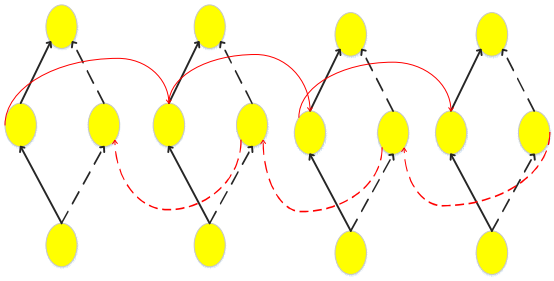

Figure 2. Structure of the bidirectional gated recurrent unit (BiGRU).

Seen in Figure 2, the current hidden layer state of BiGRU is determined by the combination of the output $\overleftarrow{h}_{t-1}$ of the hidden layer state forward at the x_t and $(t-1)$ moments and the reverse hidden layer state $\overleftarrow{h}_{t-1}$. BiGRU can be split into two unidirectional GRUs, so the hidden layer state $\overrightarrow{h}_{t-1}$ of BiGRU at time t is obtained by weighting the forward hidden layer state and the reverse hidden layer state $\overleftarrow{h}_{t-1}$:

$$\overrightarrow{h}_t = GRU(x_t, \overrightarrow{h}_{t-1}) \tag{9}$$

$$\overleftarrow{h}_t = GRU(x_t, \overleftarrow{h}_{t-1}) \tag{10}$$

$$h_t = w_t \overrightarrow{h}_t + v_t \overleftarrow{h}_t + b_t \tag{11}$$

The $GRU()$ function represents a nonlinear change to the input word vector, and the word vector is encoded into a corresponding GRU hidden layer state. w_t and v_t represent the weights corresponding to the forward hidden layer state $\overrightarrow{h}_t$ and the reverse hidden state, $\overleftarrow{h}_t$, respectively, corresponding to the bidirectional GRU at the time t; b_t represents the offset value corresponding to the hidden layer state at the time t.

3.2. MATT-CNN Model

The MATT-CNN model combines multiple attention mechanisms to extract features, aiming to obtain more sufficient text emotional feature information, so as to effectively identify the emotional polarity of different labeled keywords, and then positively contribute to the overall model. Its main structure consists of six parts, as shown in Figure 3.

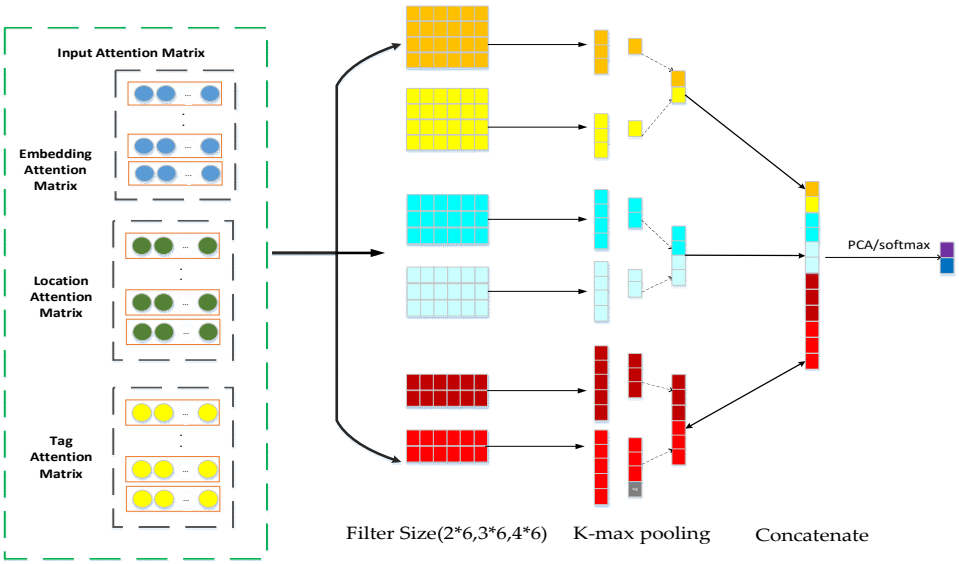

Figure 3. Convolutional neural network (CNN) model.

1. Attention input matrix: This stores attentional feature vector information of different attention mechanisms.

2. Operation layer: This paper uses different arithmetic operations to form a fusion representation of the input text as the input of the fusion model convolutional neural network part, so that the model comes from various aspects in the training process for the two different attention mechanisms listed in this paper. Pay attention to the keywords we have marked.

3. Convolution layer: To extract text information with different semantic depths, our proposed model adopts three window-size sliding windows combined with a multi-attention mechanism to obtain rich local features of input text.

4. Pooling layer: We use the k-max pooling in this model, which considers the influence of the convolution kernel height on the generated graph to perform the downsampling operation. The k-max pooling can express the case where the same type of feature appears multiple times, that is, the strength of a certain type of feature can be expressed and part of the position information can be retained.

5. Merging layers: The most important information of the different attention mechanisms extracted from the pooling layer, in this paper, is operated by the merge layer to form the feature representation of the input text, and the feature representation is reduced by the PCA method, then the convolutional nerve of the fusion model is output by the softmax function. The text classification is the result of the network model.

6. Output layer: This article will output the final classification results through the softmax function.

3.3. Text Preprocessing

3.3.1. Task Definition

Regarding sentence $s = \{w_1, w_2, \ldots t_{vital}, \ldots w_n\}$ of length n, t_{vital} is the keyword that we extracted through the keyword extraction algorithm. A sentence is formed into a word order column in units of words, in this paper, and then each word is mapped into a multi-dimensional continuous value word vector to obtain a word vector matrix $E \in R^{k \times |v|}$.

Here, k is the word vector dimension, that is, each word is mapped to a k-dimensional vector $x_i \in R^k$, and $|v|$ is a dictionary size, that is, the dataset contains the number of all words. Concerning a sentence of length n, it can be represented as a matrix as shown in Equation (12):

$$x_{1:n} = x_1 \oplus x_2 \oplus x_3 \oplus t_i \oplus \ldots \oplus t_j \oplus x_n \tag{12}$$

where $\oplus$ is a splicing operation. The emotional polarity of each target in the target set is judged by the feature information between the word vector set $\{x_1, x_2, \ldots, x_n\}$ and the keyword set $\{t_i, t_j\}$ in the sentence, for the purposes of this paper.

3.3.2. Keyword Extraction Algorithm

Using FastText [31], the mean value of the text word vector is taken as the semantic information of the text. The method is simple and efficient, having achieved good classification results. This paper is inspired by the above method and uses the category feature words to embed the matrix to represent the categories: The correlation between text keywords and categories is evaluated by the combination of tf-idf and cross entropy. Given that the document set for the category is $P = [p^1, p^2, \ldots, p^l]$, we use the top keywords in the relevant category as the category feature words in each category.

$$s = s_1 \oplus s_2 \oplus s_3 \oplus s_4 \oplus s_k \tag{13}$$

$$S = [s_1, s_2, \ldots, s_L] \tag{14}$$

where S is the category matrix and $\oplus$ is the splicing operation. s is the category keyword vector.

Algorithm 1. Keyword extraction algorithm

Input: Document collection P
Output: Keyword and its eigenvector matrix S
Begin:
For $i = 1$ to L
$x_i = segment(p^i)$//The segment function is a word segmentation using NLTK
$u_i = word2vec(list)$//Vectorize the results of the word segmentation
$y_i = desc(Tfidf(x_i))$//Use the tf-idf method for each category of text to get high frequency words, arranged in descending order.
$Z_i = crossEntropy(y_i)$//Use cross entropy to identify whether these high frequency words are used frequently in other categories
$s^i_{1,2\ldots,k} = topk(z_i)$//Get high frequency words representing each category, select the k words with the highest frequency
$s = fulljoin(s^i_1, s^i_2, s^i_3 \ldots s^i_k)$//The k high frequency words of each category are composed into a feature vector matrix of this category using a splicing method.
End For
End

Algorithm 1 first uses NLTK to segment each piece of text and, after word segmentation, it uses word2vec to generate a word vector (third line). Lines 4 and 5 are high frequency words for each category article using the tf-idf method. Then, cross-entropy is used to compare high-frequency words of this category with other high-frequency words to determine whether the high-frequency words of this category appear frequently in other categories. Line 6 gets k keywords for this category and, finally, stitches them into keyword text feature vectors.

To better learn the feature information of different targets and identify the emotional polarity of different targets, this paper uses the word vector attention mechanism and the position distance attention mechanism to focus on learning different information that needs attention in different ways. Considering a sentence containing t keywords, divide it into t clauses and mark its position with the special symbol «». Since the dataset used in the experimental part of this paper is the semEval2014 dataset, the marking work of the target words has completed, which can reduce the workload of keyword extraction. "The atmosphere of this movie is magnificent, after it still have place to improve", for example, the effect after the mark is shown in Figure 4:

```
The 《atmosphere》 of this movie is magnificent,although it still have place to improve
```

Figure 4. Marked example.

Each word is represented as a k-dimensional vector by word vector matrix, and the three characteristics of the target word t_i word vector, part of speech and its position in the sentence are extracted to construct two attention mechanisms of the neural network input layer:

1. Word vector attention mechanism: Via extracting the word vector of the target keyword as its attention matrix, we operate the attention matrix and the word vector matrix. This attention mechanism can correlate well the content of the text.

2. Positional attention mechanism: The attention mechanism used by our model has two representations; it can be added to the word vector attention mechanism in the form of parameters to assist its operation; it also can be input into the network with independent attention mechanism to position the attention. The combination of mechanisms and other attention mechanisms can fully represent the importance of each word in a sentence.

3. The mechanism of the word attention: It is a supplement to the attention mechanism of the word vector. It is the attention mechanism of the part-of-speech in the sentence; by analyzing the part of speech, it learns more hidden information.

3.4. Word Vector Attention Mechanism

The purpose of the attention mechanism is to allow the model to focus on different aspects during the training process and to understand which part of the information is important, so the model pays close attention to this information. The sentence "magnificent" is used to describe the keyword "movie", for example, so the effect of the emotional word "magnificent" on the keyword "movie" in the sentence is it is important.

Regarding sentence $s = \{w_1, w_2, \ldots t_i, \ldots, w_n\}$, we extract the word vector of the keyword t_i as the attention matrix. The attention matrix s and the word vector matrix are subjected to the arithmetic operation shown in Equation (15), and the attention feature matrix A^c can be obtained, wherein A^c is a diagonal matrix. The operation process is shown in Figure 5:

$$A^c_{i,i} = \alpha \times \frac{\exp(A_{i,i})}{\sum_{j=1}^{n} \exp\left(A_{j,j}\right)}, \tag{15}$$

where α is an adjustable parameter, similar to the learning rate η in neural networks, which is used to control the influence of different word vectors on target keywords. Initially, α can set the pre-value

manually, or it can be given by the position attention mechanism. The α can indicate the importance of each word. Using the calculated attention feature matrix A^c and the original word vector to calculate the input matrix of the convolutional neural network is shown in Equations (16) and (17):

$$z_i^c = x_i \oplus A_{i,i}^c \tag{16}$$

$$z_i^c = x_i \cdot A_{i,i}^c \tag{17}$$

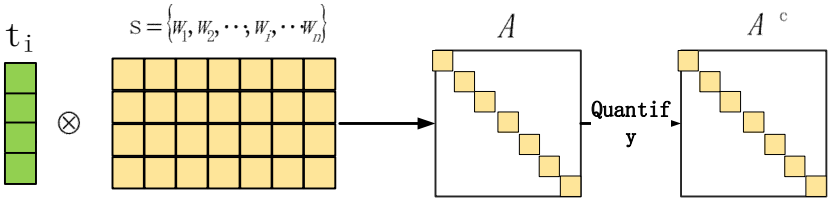

Figure 5. The operation of attention mechanism.

Both of the above methods can be used as the operation of the input matrix, and this paper uses the methods of Equations (16) and (17) vector splicing to operate.

3.5. Position Attention Mechanism

Concerning a specific target sentiment classification, the position between the word and the target keyword often hides key information, such as in the example, see Figure 6, "The 'atmosphere' is magnificent, but the 'performance' of the actors are awful".

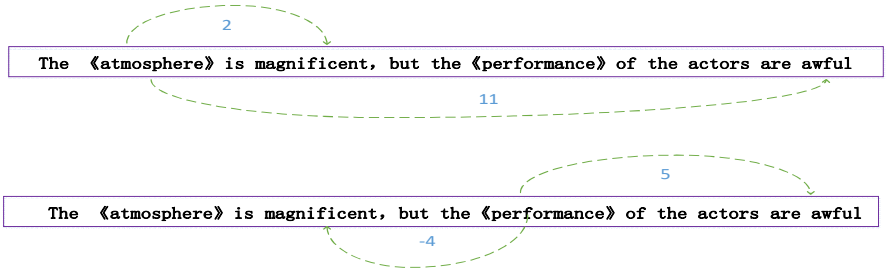

Figure 6. Example of location calculation.

According to past understanding, emotional words that are close to the target keyword often have a greater impact on them. Regarding the target keyword "atmosphere", for example, there is no doubt that the emotional word "magnificent" is closer to it and is also the correct expression of it. Considering the target word "performance", the result is incomplete, because its distance from the emotional word "awful" describing itself is greater than "magnificent", therefore, it is impossible to obtain its true emotional content. To solve this problem, we propose a two-way scanning algorithm to determine the position value between words and targets. The Algorithm 2 is as follows:

Algorithm 2. Two-way scanning algorithm

Input: Sentence s after word segmentation;
output: Positional value set *location* between each word and target word.
Begin:
Step (1) Set the target word and position value to 0 and set the value of the other words to n, which is the length of the sentence.
Step (2) Centering on the target keyword, set two working pointers to scan left and right, respectively (the following steps take one of the pointers as an example, and vice versa)
Step (3) Record the value of the work pointer and the i target relative position $location_i$, if the word at the position is punctuation, perform step (4); else, if the word at the position is the word in the target word set, then perform step (5); else, execute step (6);
Step (4) Use this formula to update the value of $location_i$, add the location $location_i$ to the collection, and continue scanning;

$$location_i = location_i + \min\{5, \frac{n-i}{2}\};$$

Step (5) Update the value of $location_i$ with the formula $location_i = 1$, add the position value i to the set, and continue scanning;
Step (6) Add the position value $location_i$ to the set and continue scanning;
Step (7) When the two pointers respectively reach the start and end positions of the sentence, the position value l_i is added to the set to stop scanning.

This article uses the matrix ψ to store the value of all positions in the dataset, and then to calculate the value of the β:

$$\beta = 1 - \frac{\psi_i + 1}{n+1} \tag{18}$$

Map all the values stored in the matrix ψ to a multidimensional vector, i.e., $\psi_i \in R^k$, and then calculate the input matrix:

$$z_i^{\psi} = \frac{\psi_i + x_i}{2}. \tag{19}$$

3.6. Part of the Attention Mechanism

Regarding some datasets and word segmentation with low emotional word coverage, the accuracy of our emotional classification work based on text content information to classify text emotions will be greatly reduced. To solve the above problems, a part-of-speech mechanism has been introduced in our model. Through re-labeling the part of speech of special words, we can make our fusion model deepen the correlation between target keywords and emotional words. Let us take the sentence, see Figure 7, "The 'atmosphere' is magnificent, but the performance of the actors are awful"as an example.

```
The 【DA】 《 【Target】 Atmosphere 【Tar】 》 【Target】 Is 【LV】
magnificent 【Positive】 , but 【CC】 the 【DA】 performance 【Tar】 of
【Prep】 the 【DA】 Actors 【Noun】 are 【LV】 Awful 【Negative】
```

Figure 7. Sentences re-tagging.

Considering this, we only analyze the key target word "atmosphere". "DA" stands for definite article, "LV" stands for verb, and "Tar" stands for target keyword. "CC" stands for a transition preposition. "Positive" stands for emotional words with positive emotional polarity, and "Negative" stands for negative emotional words.

Like word vectors, this article maps each word into a multidimensional continuous value vector, which we call a part of speech vector $Target_i \in R^l$, where l is the part of speech vector dimension.

Assuming a sentence of length called "m", the part of speech vector can be represented by a vector matrix as shown in Equation (20), where Tar is the part of speech vector of the target keyword.

$$Target_{1:m} = Target_1 \oplus Target_2 \oplus Target_3 \oplus \ldots \oplus Tar \oplus \ldots \oplus Target_m \tag{20}$$

The part-of-speech vector of the target keyword extracted in this paper is used as the part-of-speech attention feature matrix, i.e., $A^t = Tar$, and then the input matrix of the network is calculated:

$$z_i^t = \alpha \times \beta \times \frac{A_i^t + Tar_i}{2}, \tag{21}$$

where β is the weight coefficient value. By adjusting the value of β, you can make full use of the emotional characteristics of the sentence. During this experiment, the value of the emotional word is 1.2, and the other values are 1.0.

3.7. Input Matrix Construction Method

CNN can focus on different grammar and semantic features in the training process through a multi-attention mechanism, thereby capturing deep semantic information to better identify the emotional polarity of the target. This article will introduce the method of constructing the input matrix of MATD-CNN for the multi-attention mechanism.

MATD-CNN: Combine different attention input values into a three-dimensional tensor as input to the neural network. The advantage is that the input matrix can be scrolled through a multi-channel input form. The CNN network can convolute the calculation of multiple words in the text through the convolution kernel and preserve the association between words. Using the sliding window of length h, the operation of the volume on the input matrix is:

$$C_i = f(w \cdot x_{i:i+h-1} + b), \tag{22}$$

where $W \in R^{h \times k}$ is the convolution kernel weight, $b \in R$ is the bias value, f is the activation function, and $x_{i:i+h-1}$ is a text local feature matrix of a convolution window. Concerning a sentence of length n, the feature map shown can be obtained by a convolution kernel operation:

$$c = [c1, c2 \ldots, c_{n-h+1}], \tag{23}$$

where $c \in R^{n-h+1}$. This paper uses the k-max pooling method for downsampling, as shown in Equation (24):

$$\mathrm{m} = \left[\frac{s}{h}\right], \tag{24}$$

where h represents the height of the convolution kernel, i.e., the sliding window, and s represents the length of the short text sentence (controlled within 30 characters). Compared with the maximum pooling strategy, the k-max method can dynamically extract multiple important semantic combination features according to the characteristics of the multi-sliding window convolution layer and preserve the relative order relationship between the features.

The feature vector output from the downsampling layer is used as the input of the fully connected layer. The model sampled by our softmax function outputs the classification result, as shown in Equation (25):

$$\mathrm{y} = soft\max(W_\mathrm{f} X_P + B_\mathrm{f}) \tag{25}$$

where X_p is the downsampling layer output, $W_f \in R^{C \times |x_p|}$ is the fully connected layer weight matrix, and $B_f \in R^C$ is the fully connected bias value. We optimize the loss function by cross entropy, and the cross entropy cost function is given by Equation (26):

$$\text{loss} = -\sum_{i=1}^{D}\sum_{j=1}^{C} \hat{y}_i^j \log y_i^j + \lambda \|\theta\|^2, \tag{26}$$

where, D is the training dataset size, C is the category number, y is the prediction category, $\hat{y}$ is the actual category, and $\lambda\|\theta\|^2$ is a regular term.

3.8. Text Feature Fusion Model

Regarding the above unit, we introduce the multi-attention convolutional layer model of the fusion model. During this model, we will introduce the proposed BiGRU and multi-attention CNN fusion model; the overall architecture is shown in Figure 8.

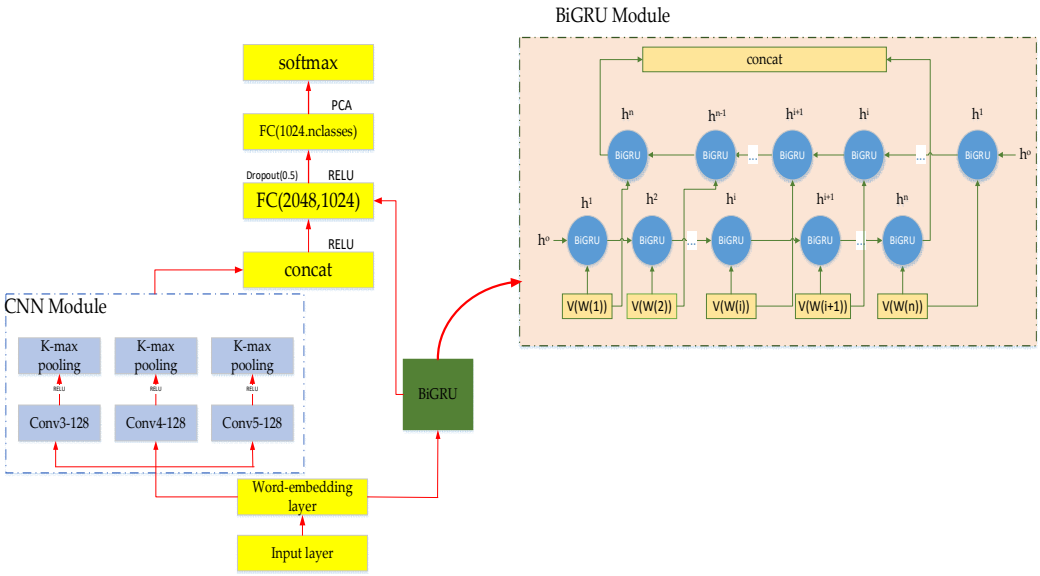

Figure 8. MATT-CNN+BiGRU model diagram.

The first layer of the convolutional neural network model is the word embedding layer. We construct the input matrix by inputting different attention inputs into a three-dimensional tensor as the input of the network (the method has been described above). The second layer is the convolution operation layer where we extract its local semantic information. Following the example of previous literature [32], when the word vector is controlled at 100 dimensions, the filters are 3×100, 4×100, and 5×100, which will achieve better classification results; therefore, we directly use the currently accepted hyperparameters for experiments. We use 128 filters, the stride size value is set to 1, and the padding is VALID. The third layer is downsampled by the k-max pooling method. The purpose is to discard the redundant features and extract as many key features as possible while reducing the dimension. The three pooled operational features are then stitched together as part of the first layer of fully connected layer input features.

The first layer of the BiGRU model is the word embedding layer. The sentence matrix of the embedding layer is taken as input. The word vector dimension we set is 100-dimensional, consistent with the CNN model, while for other hyperparameter settings, such as the number of hidden layer neurons, the corresponding value is 128. The current input information of the model is related to the sequence before and after. Therefore, we give full play to the advantages of BiGRU bidirectional operation, input the fusion model from two directions, save the historical information and future

information in two directions through the hidden layer and, finally, regarding the two hidden layers the output part is spliced to get the output of the last BiGRU. The code is as follows:

output_bgru=rnn.static_bidirectional_rnn(fw,bw,inputs).

The global features of the words in the text are extracted by extracting the contextual semantic information of the words by using the BiGRU model. During the first FC layer, the concat() method in the TensorFlow framework is used to fuse the two models of CNN and BiGRU. The tf code is as follows:

output=tf.concat({[output_cnn,ouput_bgru],axis=1).

We save the merged features in the output. As the first fully connected layer input, we introduce the dropout mechanism in the two FC layers to prevent overfitting. Each iteration will give up some of the trained parameters, making the weight updates no longer rely on some of the inherent features.

Prior to using the feature fusion vector to output the result through the softmax classifier, we use the PCA method introduced earlier to reduce the dimension and retain the useful main information, thus improving the convergence speed of the model. The probability of classifying x as a softmax regression in this model is:

$$p(y^{(i)} = j|x^{(i)}; \theta) = \frac{\exp(\theta_j^T x^{(i)})}{\sum_{l=1}^{k} \exp(\theta_l^T x^{(i)})}, \quad (27)$$

4. Experimental Results and Analysis

4.1. Datasets

To verify the validity of the model proposed in this paper, the experiment used movie review data (MRD) created by Cornell University's film evaluation data and adopted SemEval2016 datasets. Among them, the MRD consists of movie review data, with a positive attitude review of about 1000 articles, a negative attitude review of 1000 articles, a label of five-character sentences of 5331 sentences, and a sentence with 5000 subject sentences. During the experiment, 1000 sentences were randomly selected as the training set and 400 were used as the test set. SemEval2016 was the dataset of the semantic evaluation game task 4, which contains user reviews in the fields of laptop and restaurant, and the emotional polarity of the data samples was divided into positive, negative and neutral. Table 1 gives the statistics of the experimental use data in this paper.

Table 1. Statistics of experimental data.

Dataset	Positive	Negative	Neutral
Laptop-train	872	718	490
Laptop-test	361	152	122
Restaurant-train	2158	804	644
Restaurant-test	748	189	210
MRD-train	483	272	245
MRD-test	189	79	132

4.2. Model Parameter Setting

During this experiment, a variety of window convolution kernels were used to convolve the input matrix. The convolution kernel function was rectified linear units. The training procedure uses the Adadelta update rule proposed by Zeiler [33]. The other super parameters are as follows, see Table 2.

Table 2. Experimental parameters.

Parameter	Parameter Description	Value
h	Windows Size	3, 4, 5
n	Features Map	128
p	Dropout Rate	0.5
s	Constrain L	3
b	Mini-Batch Size	64
m	Pooling Method	k-max

4.3. Experimental Environment

The experimental machine selected had Intel's eighth-generation I5-4590 CPU, 16 GB memory, a GTX750Ti graphics card with 2 GB memory, and a 256 GB solid state drive (SSD). The experimental system used the Linux operating system, after repeated experiments to obtain the following experimental results.

4.4. Model Comparison

The experiment used the accuracy index (*Accuracy*) to measure the performance of the evaluation sentiment classification algorithm. The calculation method is as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}, \tag{28}$$

where TP indicates the number of samples that are positive and the prediction is positive; FN indicates the number of samples that are positive and the prediction is negative; TN indicates the number of samples that are actually negative and the prediction is negative; FP indicates that the actual sample is negative and the prediction is positive.

To verify the validity of the model proposed in this paper, the experiment sets up six baseline methods:

(1) SVM: The traditional feature-based SVM classification model proposed by Kiritchenko et al. [34]. It uses a series of manually annotated data to train the model, which yields better classification results than previous studies.

(2) CNN: The convolutional neural network model based on Kim [5] is the most basic convolutional neural network model.

(3) AM-CNN: Discussed by Wang et al. [35], it is based on the attention mechanism convolutional neural network. The algorithm uses the cyclic neural network to capture the context information of the text and introduces the single-level attention mechanism to obtain the text category matrix, thereby improving the classification accuracy of the model.

(4) CNN+BiLSTM: Li et al. [36] proposed a fusion model based on CNN and BiLSTM. The model organically integrates CNN in the BiLSTM model, and the experiment proves that it has a significant improvement over the single model classification effect.

(5) ATT-LSTM: An attention-based LSTM network proposed by Wang et al. [37], the model incorporates the attention information of a specific target and uses the pre-trained word2vec word vector as input to train the model. This model achieves a better classification effect in the field of sentiment classification than the traditional LSTM network.

(6) BiLSTM-ATT-G [38]: This uses two attention-based LSTMs to model the context information on the left and right sides of the target to extract a distributed representation of the words on the sentence and then applies attention to the hidden nodes. It estimates the importance of each word and introduces a gate function to calculate the left and right sides of the context and the sentimental tendencies of the entire sentence.

(7) IAN [39]: IAN has designed a model for interactive computing of aspect terms and sentences, which leverages attention in context and aspect terms to generate representations of aspect terms and context, respectively. Finally, the sentiment polarity of an aspect term in context is predicted by combining the aspect term and context representation.

4.5. Experimental Results and Analysis

We conducted the MATT-CNN+BiGRU fusion model and the above five models on the SemEval2016 and MRD datasets. The experimental results are shown in Table 3.

Table 3. Experimental results.

Model	Laptop	Restaurant	MRD
SVM	65.17	74.18	70.13
CNN	65.23	69.90	68.43
AM-CNN	65.42	77.67	74.32
CNN+BiLSTM	63.20	**79.54**	73.28
ATT-LSTM	68.22	75.30	67.23
BiLSTM-ATT-G	73.34	79.12	69.89
IAN	**73.24**	77.40	78.19
MATT-CNN+BiGRU	**74.21**	78.47	**79.22**

Following comparison, we found that the proposed MATT-CNN+BiGRU model had a dominant position in the ratios of Laptop and MRD datasets. The Restaurant dataset lagged behind slightly in the CNN+BiLSTM model and achieved relatively good experimental results.

Through in-depth analysis of traditional CNN and AM-CNN, it was found that the CNN model, without any attention mechanism, discriminated different target keywords in a large number of sentences into the same emotional polarity, and there was no way to combine the contribution of the first dimension of emotion generated by target keywords to the classification of sentence emotion. Therefore, the traditional CNN was completely behind the AM-CNN model, based on the single attention mechanism.

Comparing the AM-CNN and CNN+BiLSTM models it can be seen that, under the Restaurant dataset, the CNN+BiLSTM model ranked first in all cross-matching models, demonstrating that the BiLSTM+CNN fusion model combined BiLSTM with the advantages of the global features of text sequences. It also compensated for the problem that CNN itself ignored, the contextual meaning of words in text categorization, and improved the accuracy of the feature fusion model in text categorization. However, the datasets of Laptop and MRD were 2.22% and 1.04% behind the AM-CNN model, respectively, which indicates that the attention mechanism made the model pay attention to the feature information of the target keywords in the training process, so as to better recognize polarity of the emotions in the sentences. This proves that the attention mechanism has an obvious lifting value for improving the classification accuracy of the model.

Comparing the three models of SVM, ATT-LSTM and BiLSTM-ATT-G, we found that the LSTM model also achieved better results on the Restaurant and Laptop datasets than the traditional SVM model, based on sequence information. The LSTM model handled relatively formal sentences well by capturing more useful contextual features. BiLSTM-ATT-G, which used two LSTMs to model the left and right sides of the target keyword, achieved better experimental results than the traditional single attention ATT-LSTM, probably because it could be from both directions. The last hidden state joined the emotional features used for sentiment classification and then predicted based on the contextual representation of the cascade. Additionally, on the MRD with strong semantic randomness, the classification effect was not as good as the above two datasets. The reason was that the model based on

LSTM relied on sequence information, and the text with irregular syntax limited the context of such model's ability.

Comparing CNN-based models with LSTM-based models on the MRD dataset, we found that CNN-based models had certain advantages for illegal texts, because CNN's advantage lies in extracting the most critical and rich n-grams information. Features were, therefore, less sensitive to formal text, and instead had a more dazzling performance in the highly random MRD dataset. Comparing the MATT-CNN+BiGRU model and the single attention mechanism AM-CNN model, we saw that the classification accuracy of the proposed model in the three datasets was significantly higher than that of the single attention AM-CNN. When comparing the most obvious Laptop datasets, our proposed model had an accuracy improvement of 8.79% over AM-CNN. The results showed that, compared with the single attention mechanism AM-CNN model, MATT-CNN+BiGRU combined with multiple attention mechanisms made the network focus on and learn different focuses in the training process through different attention mechanisms. The emotional information of the target keywords, as well as the extraction of more hidden information through the interconnection of different attention mechanisms, effectively compensated for the lack of a single attention mechanism.

Comparing the proposed MATT-CNN+BiGRU model with the CNN+BiLSTM model, we found that the fusion model of the two significantly could improve the classification accuracy compared to the variants of the single LSTM and CNN models. The accuracy of MATT-CNN+BiGRU in the Laptop and MRD datasets was higher than that of the CNN+BiLSTM model. It demonstrated the positive effect of our proposed fusion model on the CNN model to strengthen the target keywords for sentence sentiment classification. It directly reflected the important influence of keywords on the classification results in the field of text sentiment classification. However, we analyze two reasons why MATT-CNN+BiGRU lost in the Restaurant dataset:

(1) The Restaurant dataset was significantly larger than the remaining two datasets in our experiment, so the characteristics learned by the training set were limited. The internal parameters of the model need to be improved, which is the focus of our continued research in the future.

(2) Although the BiLSTM three-door structure was cumbersome, it still had a place where BiGRU could not be replaced perfectly in the internal structure, especially in the forgetting door and parameter optimization methods.

Compared with IAN, MATT-CNN+BiGRU increased in accuracy by 0.97% and 1.03% in MRD and SemEval2016, respectively, and the accuracy of the two models was not much different. The main reason for the analysis was that IAN reinforced the interaction between aspect term and context, which used connected attention networks, so the attention mechanism was used to model the target and some results were achieved.

4.5.1. Comparison of Loss Functions

To verify the validity of the MATT-CNN+BiGRU model, we compared the MATT-CNN+BiGRU with the traditional single CNN and BiLSTM in the MRD dataset and ensured that the other parameters of the three models were the same; the learning rate was set to 0.01. The Figure 9 below shows the loss function change graph.

Following comparison, we found that the traditional CNN model loss value was lower than the BiLSTM and the MATT-CNN+BiGRU loss function iteration stability values, but the MATT-CNN+BiGRU model was faster than BiLSTM. Both dropped to a very low value and the convergence effect was good.

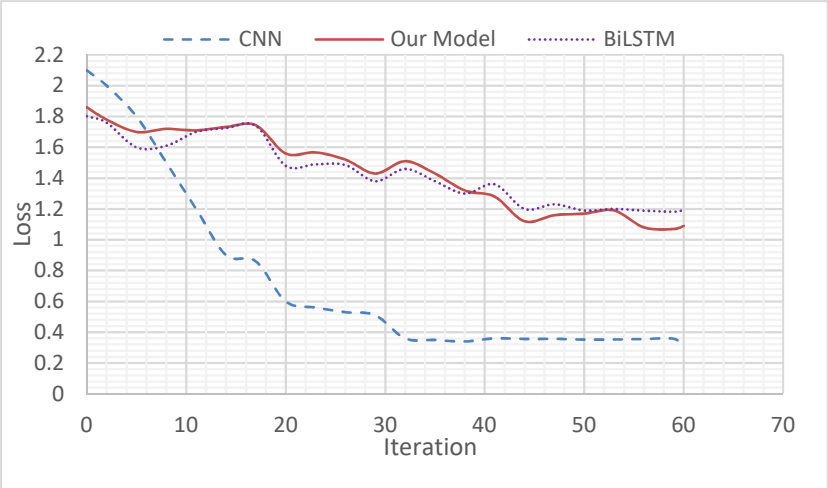

Figure 9. Loss function comparison.

4.5.2. Effect of Learning Rate on MATT-CNN+BiGRU

Figure 10 shows the effect of different learning rates on the performance of MATT-CNN+BiGRU under the MRD dataset. Looking at the gradient descent algorithm, if the initial value of the learning rate setting was too small, the number of iterations would increase, or fall into local minima. The optimal solution formed an infinite loop; if the learning rate was set too large, the cost function would be unstable and could not reach the actual minimum, making the algorithm slow. Figure 8 shows that when the learning rate was 0.01, the accuracy of the MATT-CNN+BiGRU model reached its peak value; when the learning rate increased again, the correct rate decreased, so 0.01 was the resultant empirical learning rate.

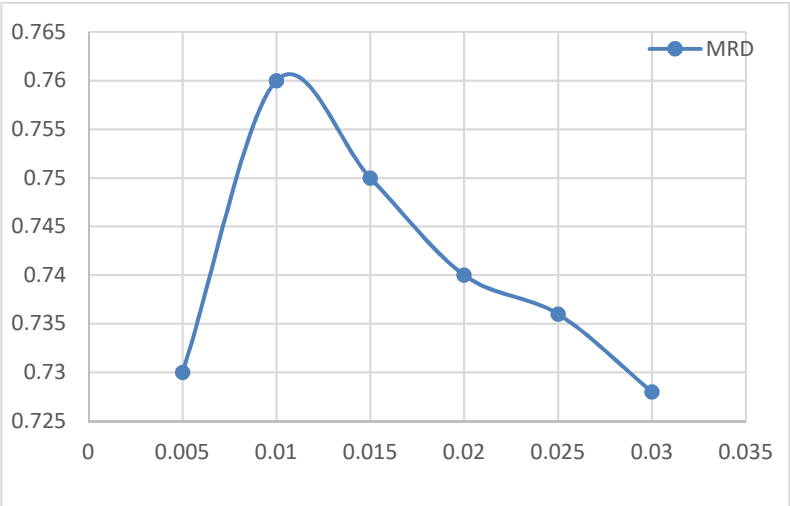

Figure 10. The impact of learning rate on the model.

4.5.3. Analysis of the Impact of PCA on Training Time

To verify the positive impact of PCA on feature fusion vector training after dimension reduction, we analyzed different network models to complete all experiments under the same CPU, GPU and framework. Simultaneously, other super parameters, such as a word vector construction method, also were consistent. Figure 11 shows the comparison of training time for different iterations of different network models on the Laptop dataset.

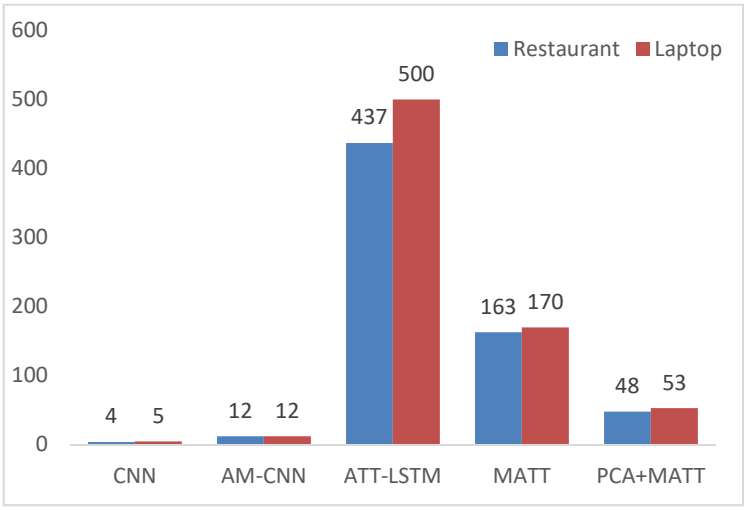

Figure 11. Analysis of the influence of PCA on training time.

Seen from the results in Figure 11, the training time cost of the LSTM network was very high, mainly because the LSTM network trained serial data, and was 437 and 500 s in the training times of Restaurant and Laptop, respectively. Additionally, CNN, without any attention mechanism, was the fastest, taking only 4 and 5 s for Restaurant and Laptop, respectively. This was in line with our understanding of CNN's inertia. Our proposed MATT-CNN+BiGRU changed the original cumbersome LSTM model and had a significant improvement over LSTM on both datasets. Subsequent to adding the PCA dimension reduction, the training times of MATT-CNN+BiGRU on the Restaurant and Laptop datasets were reduced to 48 and 53 s, respectively. Therefore, the positive impact of PCA on the training cost of our proposed model is verified.

4.5.4. Effect of Cumulative Contribution Rate of Eigenvalues on Classification Accuracy

To study the influence of the cumulative contribution rate of principal component eigenvalues on the classification performance of MATT-CNN+BiGRU model, this paper analyzed the classification accuracy of different cumulative contribution rates for the MATT-CNN+BiGRU model under an MRD dataset. The relationship between classification accuracy and eigenvalue cumulative contribution rate is shown in Figure 12.

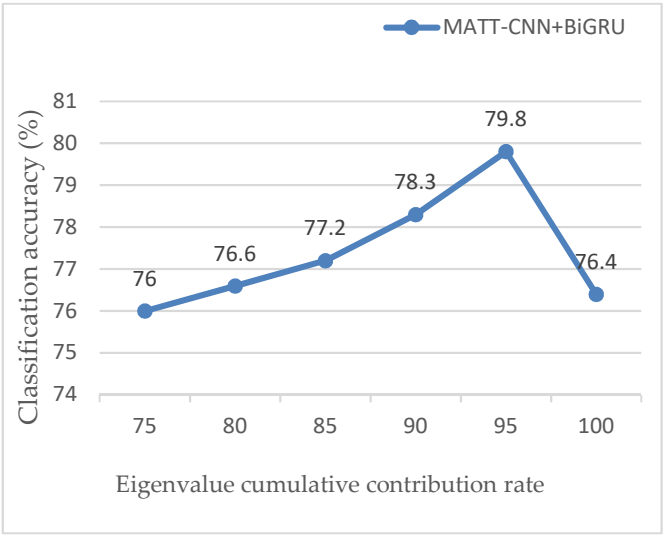

Figure 12. Cumulative contribution rate.

Figure 12 shows the cumulative contribution rate of eigenvalues (CCRE), from 100% to 95%; the text features extracted by the MATT-CNN model were processed by PCA, and the redundancy in the text features was eliminated gradually. Accuracy increased when CCRE = 95%, the redundant information in the text feature was eliminated more completely, and the classification accuracy was also the highest; when CCRE gradually was reduced from 95%, the partially used text also was eliminated, resulting in a decrease in classification accuracy. It follows that the dimension of the principal component is critical to the accuracy of the classification.

4.5.5. Multi-Attention Mechanism Effectiveness Analysis

Based on the previous single attention, this paper combined the three attentional annotation methods of part of speech, word vector, and position to form a multi-attention mechanism. To verify the effectiveness of our proposed part-of-speech mechanism, we worked selectively in Laptop. Two thousand samples were extracted from Restaurant for cross-examination. To control the variables, the remaining hyperparameters were consistent with the previous experiments. The experimental results are shown in Table 4.

Table 4. Multi-attention effectiveness analysis result.

Model	Laptop	Restaurant
MATT+wvatt	70.42	75.89
MATT+patt	71.37	74.46
MATT+latt	72.20	76.51
MATT+latt+DDS	73.98	77.15
MATT+allatt	75.21	79.22

Here, MATT is the abbreviation for our MATT-CNN+BiGRU model. MATT+wvatt represents only the attention mechanism of the word vector. MATT+patt represents only the part-of-sex attention mechanism. MATT+latt indicates that only positional attention has been added. The mechanism, MATT+latt+DDS, represents a single positional attention mechanism after the fusion of the two-way scanning algorithm. MATT+allatt represents the classification model that incorporates the three attention mechanisms.

It can be seen from the results in Table 4 that the positional attention mechanism worked best in the two datasets under the vertical comparison of the three attention mechanisms. Especially after adding the DDS two-way scanning algorithm, the accuracy was improved slightly before the DDS was added. The MATT, which combined the three attention mechanisms, had the highest performance in the Laptop and Restaurant datasets, with accuracies of 75.21% and 79.22%, respectively.

4.5.6. The Influence of Word Vector Dimension

We have added part of the attention mechanism. To verify the effectiveness of our proposed attention mechanism, 2500 and 2000 data points were extracted from the MRD dataset and the Restaurant dataset, respectively. Among them, the part of speech vector dimension was 0, meaning that the part-of-speech attention mechanism was not applicable. The result is shown in Figure 13.

Viewing Figure 13, after adding the part-of-sex attention mechanism, the classification effect on the MATT-CNN+BiGRU model under both datasets was improved significantly. When we calculated by peak, MATT, on the MRD dataset CNN+BiGRU, increased by 8.9% and increased by 9.9% in the Restaurant dataset. This proves from one aspect that the addition of a part-of-speech attention mechanism can make the model more fully learn the emotional information of the text and achieve a better emotional classification effect. When the word vector was greater than 100 dimensions, the model fluctuated in both datasets. Following analysis, the reason was that the input matrix of the MATT-CNN+BiGRU model was the input of the attention mechanism. As the main feature of

model training, word vector attention affected the parameter adjustment when the dimension of the part-of-speech vector exceeded a certain threshold, which reduced the learning effect of feature information at the content level. Additionally, as the vector dimension increased, the training cost of the model increased accordingly. Therefore, in the above experiment, we used the empirical parameter value 100 as the dimension of the part-of-speech attention vector.

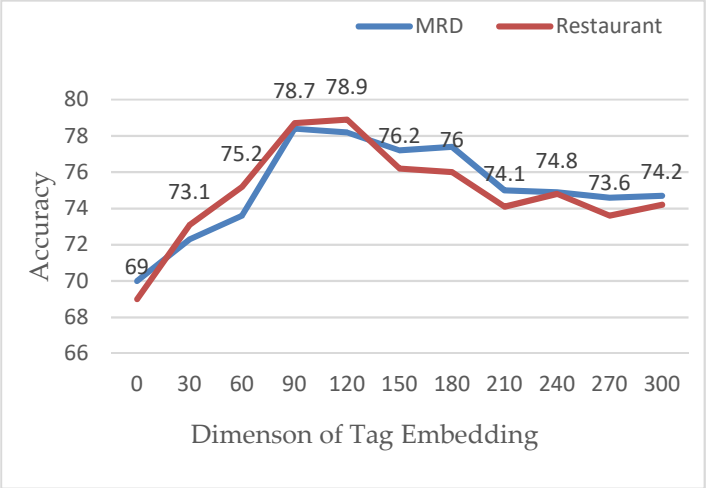

Figure 13. Dimension of tag embedding.

4.6. Case Study

To verify the attention effectiveness of our proposed MATT-CNN+BiGRU model, we visually show the sentence polarity classification results of MATT-CNN+BiGRU under two datasets, see Table 5.

Table 5. Attention weights.

Aspect	Sentence	Polarity	Sen-Polarity
priced	Boot time is super fast, around anywhere from 35 seconds to 1 min, But quite unreasonable priced.	positive	positive
Boot time	Boot time is super fast, around anywhere from 35 seconds to 1 min, But quite unreasonable priced.	negative	neutral
Atmosphere	The atmosphere of this movie is magnificent, although it still have place to improve.	positive	positive

The red box is the target keyword we marked. We visualize the focus of the attention mechanism by the depth of the color. The darker the color, the higher the attention. We clearly can see in the above table that, in the two clauses, the target keyword of the annotation is the same as the sentiment orientation of the sentence. The second sentence, because the keyword selection is negative, combines the influence of BiGRU to define the emotional tendency of the sentence as neutral, which moderates the negative emotion of the sentence to a certain extent.

5. Conclusions and Future Work

This paper proposed a feature fusion model based on multi-attention CNN and BiGRU networks for text classification research. The model can not only take advantage of the multi-attention CNN for n-gram feature extraction and target keyword local feature extraction, but also combines the BiGRU model structure with relatively simple structure and can consider the global characteristics of the text to fully consider the contextual semantic information of the word. The multi-attention CNN model uses three kinds of attention (word vector attention, part of speech attention, position

attention) to extract the semantic weighting of keywords in sentences, which constitutes the first dimension of sentiment classification. The BiGRU model, combined with the BiLSTM model structure, is relatively simple and can take into account the global characteristics of the text to fully consider the advantages of the contextual semantic information of the word, and constitute the second dimension of the semantic vector, which is used to obtain the classification result of the sentence level. Then, we use PCA to reduce the dimension of the two-dimensional fusion vector and, finally, obtain a classification result combining two dimensions of keywords and sentences. Under the two datasets of MRD and SemEval2016, each experiment proved that the MATT-CNN+BiGRU fusion model can improve the accuracy and reduce the training time overhead of the model compared with the widely used attention-aware classification model.

Author Contributions: Conceptualization: J.Z. and F.L.; methodology: J.Z.; software: W.X.; validation: H.Y.; resource: J.Z.; writing—original draft preparation: J.Z.; writing—review and editing: J.Z. and W.X.; supervision: J.Z; visualization: W.X. and J.Z.

Funding: This work was supported by the following grants: National Natural Science Foundation of China 61772321,61602284,61602285, Natural Science Foundation of Shandong Province ZR2016FP07, the Innovation Foundation of Science and Technology Development Center of Ministry of Education and New H3C Group 2017A15047 and CERNET Innovation Project NGII20170508.

Acknowledgments: Thanks to all commenters for their valuable and constructive comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nasukawa, T.; Yi, J. Sentiment analysis: Capturing favorability using natural language processing. In Proceedings of the International Conference on Knowledge Capture, New York, NY, USA, 23–25 October 2003.
2. Chen, K.; Liang, B.; Ke, W.; Xu, B.; Zeng, G.C. Chinese Micro—Blog Sentiment Analysis Based on Multi-Channels Convolutional Neural Networks. *J. Comput. Res. Dev.* **2018**, *55*, 945–957.
3. Liu, J.Z.; Chang, W.C.; Wu, Y.X.; Yang, Y.M. Deep Learning for Extreme Multi-label Text Classification. In Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval, Tokyo, Japan, 7–11 August 2017.
4. Kim, Y. Convolutional Neural Networks for Sentence Classification. *arXiv* **2014**, arXiv:1408.5882.
5. Xing, S.; Wang, Q.; Zhao, X.; Li, T. A hierarchical attention model for rating prediction by leveraging user and product reviews. *Neurocomputing* **2019**, *332*, 417–427. [CrossRef]
6. Zhao, W.; Ye, J.B.; Yang, M.; Lei, Z.Y.; Zhang, S.F.; Zhao, Z. Investigating Capsule Networks with Dynamic Routing for Text Classification. *arXiv* **2018**, arXiv:1804.00538.
7. Ntalianis, K.; Doulamis, A.D.; Tsapatsoulis, N.; Mastorakis, N.E. Social relevance feedback based on multimedia content power. *IEEE Trans. Comput. Soc. Syst.* **2017**, *5*, 109–117. [CrossRef]
8. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 3111–3119.
9. Mnih, V.; Heess, N.; Graves, A. Recurrent models of visual attention. *Adv. Neural Inf. Process. Syst.* **2014**, *3*, 2204–2212.
10. Yin, W.; Schütze, H.; Xiang, B.; Zhou, B. ABCNN: Attention-Based Convolutional Neural Network for Modeling Sentence Pairs. *Comput. Sci.* **2016**, *4*, 259–272.
11. He, R.; Lee, W.S.; Ng, H.T.; Dahlmeier, D. An Unsupervised Neural Attention Model for Aspect Extraction. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 388–397.
12. Sun, A.; Lim, E.P.; Liu, Y. On strategies for imbalanced text classification using SVM: A comparative study. *Decis. Support Syst.* **2010**, *481*, 191–201. [CrossRef]
13. Jing, L.; Wang, T.; Zhao, M.; Wang, P. An adaptive multi-sensor data fusion method based on deep convolutional neural networks for fault diagnosis of planetary gearbox. *Sensors* **2017**, *17*, 414. [CrossRef] [PubMed]

14. Audebert, N.; Le Saux, B.; Lefèvre, S. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In Proceedings of the Asian Conference on Computer Vision, Taipei, Taiwan, 22 September 2016; Springer: Cham, Switzerland, 2016; pp. 180–196.
15. Bakalos, N.; Voulodimos, A.; Doulamis, N.; Doulamis, A.; Ostfeld, A.; Salomons, E.; Caubet, J.; Jimenez, V.; Li, P. Protecting Water Infrastructure from Cyber and Physical Threats: Using Multimodal Data Fusion and Adaptive Deep Learning to Monitor Critical Systems. *IEEE Signal Process. Mag.* **2019**, *36*, 36–48. [CrossRef]
16. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. 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, San Diego, CA, USA, 12–17 June 2016.
17. Wang, G.; Li, C.; Wang, W.; Zhang, Y.; Shen, D.; Zhang, X.; Henao, R.; Carin, L. Joint Embedding of Words and Labels for Text Classification. *arXiv* **2018**, arXiv:1805.04174.
18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. Available online: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf (accessed on 8 November 2019).
19. Chen, P.; Sun, Z.; Bing, L.; Yang, W. Recurrent Attention Network on Memory for Aspect Sentiment Analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017.
20. Rozental, A.; Fleischer, D. Amobee at SemEval-2018 Task 1: GRU Neural Network with a CNN Attention Mechanism for Sentiment Classification. In Proceedings of the 12th International Workshop on Semantic Evaluation, New Orleans, LA, USA, 5–6 June 2018.
21. Kumar, A.; Kawahara, D.; Kurohashi, S. Knowledge-enriched Two-layered Attention Network for Sentiment Analysis. *arXiv* **2018**, arXiv:1805.07819.
22. Nguyen, T.H.; Shirai, K.; Velcin, J. Sentiment analysis on social media for stock movement prediction. *Expert Syst. Appl.* **2015**, *42*, 9603–9611. [CrossRef]
23. Ruder, S.; Ghaffari, P.; Breslin, J.G. A Hierarchical Model of Reviews for Aspect-based Sentiment Analysis. *arXiv* **2016**, arXiv:1609.02745.
24. Zhou, X.; Wan, X.; Xiao, J. Attention-based LSTM Network for Cross-Lingual Sentiment Classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016.
25. Song, Y.; Wang, J.; Jiang, T.; Liu, Z.; Rao, Y. Attentional Encoder Network for Targeted Sentiment Classification. *arXiv* **2019**, arXiv:1902.09314.
26. Mohammadi, F.; Zheng, C.; Su, R. Fault Diagnosis in Smart Grid Based on Data-Driven Computational Methods. In Proceedings of the 5th International Conference on Applied Research in Electrical, Mechanical, and Mechatronics Engineering, Tehran, Iran, 24 January 2019.
27. Mohammadi, F.; Zheng, C. A Precise SVM Classification Model for Predictions with Missing Data. In Proceedings of the 4th National Conference on Applied Research in Electrical, Mechanical Computer and IT Engineering, Tehran, Iran, 4 Octorber 2018.
28. Mohammadi, F.; Nazri, G.-A.; Saif, M. A Fast Fault Detection and Identification Approach in Power Distribution Systems. In Proceedings of the 5th International Conference on Power Generation Systems and Renewable Energy Technologies (PGSRET 2019), Istanbul, Turkey, 26–27 August 2019.
29. Liang, B.; Liu, Q.; Xu, J.; Zhou, Q.; Zhang, P. Aspect-Based Sentiment Analysis Based on Multi-Attention CNN. *J. Comput. Res. Dev.* **2017**, *54*, 1724–1735.
30. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv* **2014**, arXiv:1412.3555.
31. Joulin, A.; Grave, E.; Bojanowski, P.; Mikolov, T. Bag of Tricks for Efficient Text Classification. *arXiv* **2016**, arXiv:1607.01759.
32. Lei, T.; Barzilay, R.; Jaakkola, T. Molding CNNs for text: Non-linear, non-consecutive convolutions. *Indiana Univ. Math. J.* **2015**, *58*, 1151–1186.
33. Zeiler, M.D. ADADELTA: An Adaptive Learning Rate Method. Available online: https://arxiv.org/pdf/1212.5701.pdf (accessed on 8 November 2019).
34. Kiritchenko, S.; Zhu, X.; Cherry, C.; Mohammad, S. NRC–Canada–2014: Detecting Aspects and Sentiment in Customer Reviews. In Proceedings of the International Workshop on Semantic Evaluation, Dublin, Ireland, 23–24 August 2014.

35. Wang, J.Z.; Peng, D.L.; Chen, Z.; Liu, C. AM–CNN: A Concentration-Based Convolutional Neural Network Text Classification Model. *Mini-Micro Syst.* **2019**, *40*, 710–714.
36. LI, Y.; Dong, H.B. Text sentiment analysis based on feature fusion of convolution neural network and bidirectional long short-term memory network. *J. Comput. Appl.* **2018**, *38*, 3075–3080.
37. Wang, Y.; Huang, M.; Zhao, L. Attention-based LSTM for Aspect-level Sentiment Classification. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–4 November 2017; pp. 606–615.
38. Tang, D.; Qin, B.; Feng, X.; Liu, T. Effective LSTMs for Target-Dependent Sentiment Classification. Available online: https://arxiv.org/pdf/1512.01100.pdf (accessed on 8 November 2019).
39. Ma, D.; Li, S.; Zhang, X.; Wang, H. Interactive attention networks for aspect-level sentiment classification. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017.