Computer Science and Information Systems 20(4):1797–1817 https://doi.org/10.2298/CSIS221212031C

Deep Adversarial Neural Network Model Based on Information Fusion for Music Sentiment Analysis

Wenwen Chen

College of Music, Jimei University
21 Yindou Road, Jimei District, Xiamen City, Fujian Province, China
chenwwencww@163.com

Abstract. Natural language processing (NLP) is a computer-based technology used to process natural language information in written and spoken form that is unique to human society. In the process of mining massive text information, a variety of technologies and research directions in the field of NLP have gradually emerged. And sentiment analysis is an important research direction, which has important research value and practical application value for enterprises and social life. Sentiment analysis is basically a single mining of semantic or grammatical information without establishing the correlation between semantic information and grammatical information. In addition, previous models simply embed the relative distance or grammatical distance of words into the model, ignoring the joint influence of relative distance and grammatical distance on the aspect words. In this paper, we propose a new model that combines deep adversarial neural network model based on information fusion for music sentiment analysis. Firstly, the information of music text sequence is captured by the bidirectional short and long time memory network. Then the sequence information is updated according to the tree structure of dependency syntactic tree. Then, the relative distance and syntactic distance position information are embedded into the music text sequence. Thirdly, the adversarial training is used to expand the alignment boundary of the field distribution and effectively alleviate the problem of fuzzy features leading to misclassification. Semantic information and syntactic information are optimized by attention mechanism. Finally, the fused information is input into the Softmax classifier for music sentiment classification. Experimental results on open data sets show that compared with other advanced methods, the recognition accuracy of the proposed method is more than 90%.

Keywords: Natural language processing, deep adversarial neural network, information fusion, music sentiment analysis, attention mechanism.

1. Introduction

Text information on the Internet is growing day by day, and a large number of texts with sentiment information are generated on social media and e-commerce platforms. Sentiment analysis is an important task in natural language processing. Sentiment analysis has attracted wide attention in industry and academia in recent years [1,2].

Aspect level sentiment analysis, also known as target-specific sentiment analysis or fine-grained sentiment analysis, refers to the analysis of the sentiment polarity of the aspects (goals) involved in the text when given a paragraph and specific aspect words [3]. In sentiment expression sentences, different aspects may express different sentiments. Song et al. [4] proposed an attention-based Long Short-Term Memory (LSTM) network, adding aspect words to the sentence. The sentence representation that generates aspect perception predicts the textual sentiment polarity of a given aspect word using an attention-mechanism-based LSTM. Chen et al. [5] used convolutional neural network to extract sentiment features and selectively outputted features related to aspect words through gating mechanism for sentiment classification. Li et al. [6] introduced multiple attention mechanism to obtain more comprehensive attention information, integrate contextual self-attention and specific aspect of word attention, and predict specific aspect of sentiment polarity. A lot of works had been done to obtain better performance of aspect-level emotion classification tasks by constructing different attention networks.

Sentiment expression is highly related to the field. The same music sentiment word may express the same or similar sentiment in different fields, but also may express opposite sentiment. When dealing with such samples, it is relatively easy for sentiment classifiers to carry out domain migration [7]. Without a classifier to carry out domain adaptation to the target domain, it is difficult to correctly distinguish the sentiment polarity of statements containing such words. Traditional music sentiment classification methods require sufficient labeled training data, and require the same distribution of training set and test set [8]. Emerging fields often lack sufficient labeled data. How to use the rich source domain of aspect-level music sentiment category labels to judge the polarity of music sentiment for the specified aspect words in the target domain is a widely concerned issue in the field of sentiment analysis. The cross-domain music sentiment classification task makes full use of the large-scale data labeled in the source domain and migrates it to the new domain with few labeled data by using some algorithms [9], so as to realize the classification of music sentiment in the new domain.

In terms of cross-domain aspect-level sentiment classification task, Zhang et al. [10] proposed Interactive Attention Transfer Network (IATN) to integrate the influence of aspect words into the representation learning of cross-domain sentiment classification model. Tian et al. [11] proposed to extract domain-invariant sentiment features with the assistance of aspect word detection to conduct cross-domain aspect level sentiment classification. Cao et al. [12] jointly extracted aspect words and affective features, carried out unsupervised domain adaptation setting, and aligned automatically captured correlation vectors through selective adversarial learning method to achieve adaptability between domains at the word level.

Recently, with the proposal of Graph Convolutional Network (GCN) [13], GCN has been widely used in sentiment analysis tasks. Huang et al. [14] used GCN for text classification, and then combined it with dependency syntax tree for aspect level sentiment analysis. This method combining dependency tree and graph convolution could shorten the distance between aspect words and sentiment words and used graph convolution network to learn grammatical information. Zhang et al. [15] used the dependency relationships of sentences to construct the adjacency matrix, and used the graph convolutional network to learn the grammatical relations in sentences, proving that the graph convolutional neural network could correctly capture syntactic information and remote word dependencies. Zhao et al. [16] proposed a new network architecture to connect sentiment feature words with aspect words by using multi-layer graph attention network. Xue et al. [17] believed that the common adjacency matrix lacked the dependency and co-occurrence relationship between words, and proposed hierarchical syntactic graph and hierarchical lexical graph to replace the common adjacency matrix for grammar learning. For example, the word ”nothing special” pair in the sentence ”food was okay, nothing special.” appeared five times in the training set, and the word pair indicated negative emotional polarity. In this case, if the word ”food” was to be predicted correctly, other information was needed to counteract the positive appearance of ”okay” in the sentence.

Therefore, we propose a new model that combines deep adversarial neural network model based on information fusion for music sentiment analysis. The new model learns semantic information and syntactic information at the same time. Hierarchical vocabulary is used to replace common dependency syntax tree for syntactic information learning, and hierarchical word frequency graph is used to replace common dependency syntax tree for syntactic information learning of graph convolutional network. In addition, relative distance and syntactic distance information are fused and embedded into the model. Contributions of this paper are as follows:

1. The new model integrates relative distance and grammatical distance into the model, that is, the influence of relative distance of words on aspect words is taken into account, as well as the influence of grammatical distance of words on aspect words.
2. The influence of the position of a word in the dependency syntactic tree is proposed, and the importance of the height of a word in the syntactic tree and the degree of the word node in the text sequence is considered.
3. At the same time, the semantic information and grammatical information are studied, and the information contained in the text sequence is fully considered to make it contain richer information, so as to improve the extraction of aspect word information.
4. The hierarchical lexical graph is introduced to replace the common dependency syntax tree, which can not only capture grammatical information, but also pay attention to the co-occurrence relationship between words, so as to improve the efficiency of grammar learning.

The rest of this paper is organized as follows. In the second section, related works are introduced. The third section introduces the proposed work in detail. The fourth part introduces the experiments and results. Finally, the fifth section concludes this paper.

2. Related Works

The neural network model is used as an aspect level sentiment analysis task, and most of them use the word embedding layer [18] as the starting layer of the model. The main reason is that word embedding layer can map text sequence information to low-dimensional vector space one by one, and then it can learn the information mapped to low-dimensional vector space through various deep learning models. Convolutional Neural Network (CNN) is initially used in image processing tasks, mainly by convolutional kernel to extract the local information of the image. Since CNNs have been used in natural language processing tasks [19], researchers have further extracted feature information by combining CNN with other models. Zhang et al. [20] used multiple layers of CNN to model the context in parallel, and then used the attention mechanism to associate the information between context and aspect words. Wang et al. [21] proposed a model of convolutional neural network combined with gating mechanism, which could filter sentiment features unrelated to aspect word information, so as to selectively output relevant sentiment features based on aspect word.

Studies have shown that, in the task of aspect level sentiment analysis, each word in a sentence has a different degree of influence on aspect words. The farther away the word is from aspect words, the smaller the influence on aspect words. Therefore, distance can be divided into relative distance and grammatical distance. Sun et al. [22] used the weighted convolutional network to explore the relative position relation and grammatical distance relation, and the research results proved that the grammatical distance relation was more conducive to the classification of aspect words than the relative distance relation. Zhang et al. [23] integrated the syntactic location information into the model, and the experimental results proved that the syntactic location could help the model better understand the relationship between context and aspect words.

With the continuous in-depth study of sentiment analysis, grammatical information is constantly being mined. For example, by using a grammar parsing tool, The sentence ”The music was excellent as well as service, however, I left the four seasons very disappointed” is resolved into a dependency syntax tree with syntactic information. There are three aspects in the sentence: music, service and seasons. Clauses infer that the aspect words ”music” and ”service” are positive, and ”seasons” is negative. Many models in the past focused on the wrong word in some cases [24]. For the word ”service” shown in Figure 1, many models mistakenly focus on ”well”, but from the whole sentence, the emotional polarity of the word ”service” depends on ”excellent” rather than ”well”. From the dependency syntax tree, ”service” connects with ”excellent” through the dependency relationship ”attr” and ”acomp”, reducing the influence of ”well” on the aspect word.

The advantage of graph convolutional network is that it can capture the structure information of dependency syntactic tree very well, which makes up the defect that other models can not capture syntactic information. Zheng et al. [25] could selectively utilize dependent information through key-value memory network, and finally weighted different dependent information according to memory mechanism to effectively screen useless information. Zhao et al. [26] proposed to use hierarchical syntactic graph and hierarchical lexical graph to learn grammatical information. Hierarchical lexical graph (HLG) uses the frequency of two words in the corpus as connecting factors to construct hierarchical lexical graph for each sentence, as shown in Figure 1.

3. Proposed Music Sentiment Analysis Model

Given a music text sequence $s = w_1, w_2, \cdots, w_n$ and an aspect word sequence $a = w_t, w_{t+1}, \cdots, w_{t+m-1}$. Wherein, the music text sequence contains n words, the aspect word sequence contains m words, and the aspect word sequence a is a subsequence of the text sequence s. The model in this paper is shown in Figure 2. The model consists of seven layers, namely, word deep adversarial layer, hidden layer, location embedding layer, learning layer, attention layer, feature extraction layer and pooling layer. Among them, the learning layer is divided into two modules, namely semantic learning module and grammar learning module. The attention layer is also divided into two parts, namely semantic optimization attention and syntactic optimization attention.

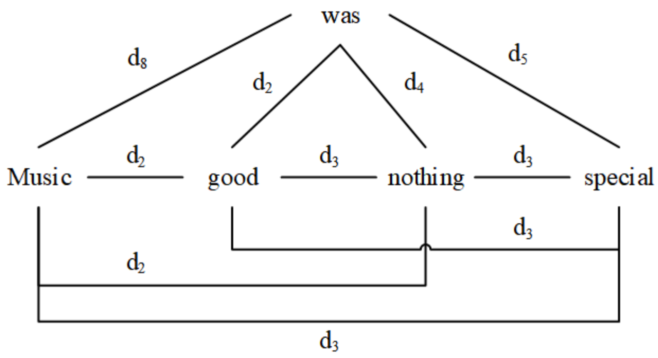

Fig. 1. Hierarchical Lexical Graph

3.1. Deep Adversarial Layer

The expression of music text has various meanings and distributed representation can better express the semantic information of music text. In this paper, the pre-trained Word Vector (Global Vector for Word Representation) [27] is used to generate distributed representations for each word in the text data. The word Embedding search matrix $L \in R^{d_e \times |V|}$ is first generated. Where, d_e represents the dimension size of the word vector. $|V|$ is the vocabulary size. During data preprocessing, each word in the vocabulary is assigned a unique index corresponding to the index number of L, and the required word embedding can be generated for each word in the data set through the mapping relationship. For a given input data:

$$x = [w_1, w_2, \cdots, w_n] \tag{1}$$

Input can be mapped to music context embedding.

$$\hat{E}^c = [e_1, e_2, \cdots, e_n]^T \in R^{n \times d_e} \tag{2}$$

At the same time, a given input aspect word $a = [e_1, e_2, \cdots, e_m]$ is mapped to an aspect word embedded representation.

$$E^a = [e_1, e_2, \cdots, e_m]^T \in R^{m \times d_e} \tag{3}$$

The task of cross-domain aspect level emotion classification needs to make full use of the given aspect word information and establish the relationship between aspect word and emotion information of music text. In music texts, the feature words that usually express emotion are close to their related aspect words. In this paper, location information is integrated to enrich the embedded representation information, which is helpful to improve the performance of aspect-level sentiment classification. For the embedded sentence representation $\hat{E}^c$, the position information of each word in the input data x is incorporated

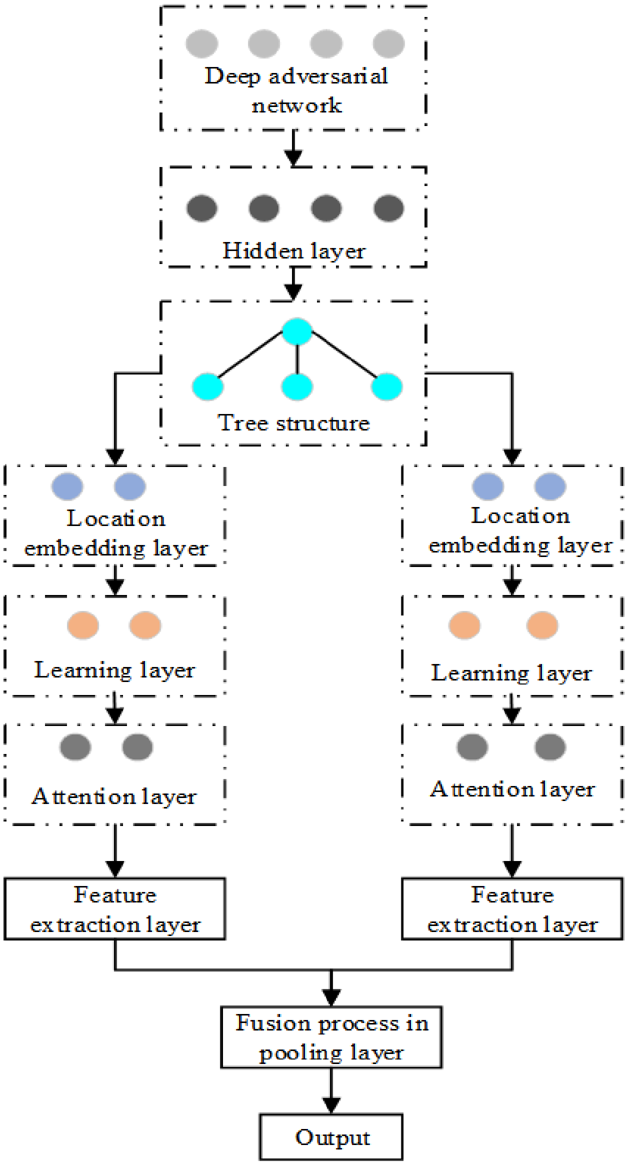

Fig. 2. The overall of proposed model

into $\hat{E}^c$ to obtain the final context-embedded representation E^c. Here, the location encoding method [28] is adopted:

$$PE(pos, 2i) = sin(\frac{pos}{10^{8i/d_e}}) \tag{4}$$

$$PE(pos, 2i+1) = cos(\frac{pos}{10^{8i/d_e}}) \tag{5}$$

$$E^c = \hat{E}^c + PE \tag{6}$$

Where, pos is the position of the word in the sentence, and i stands for the $i - th$ dimension of the corresponding vector at position pos in PE.

Since the aspect words of each domain in a music text may contain more than one word, the characteristics of long-distance dependence need to be learned. For context sentence embedded representation E^c and aspect word embedded representation E^a, feature extraction is carried out using Bidirectional Gated Recurrent Unit (BiGRU) respectively to generate context-rich semantic representations H^c and H^a:

$$H^c = [\overrightarrow{GRU}(E^c); \overleftarrow{GRU}(E^c)] \tag{7}$$

$$H^a = [\overrightarrow{GRU}(E^a); \overleftarrow{GRU}(E^a)] \tag{8}$$

For the context feature representation H^c and aspect word feature representation H^a, the weight of Attention is calculated through the Interaction Attention (IA) layer, and the interaction relationship between context and aspect words is established, so that the context representation can pay more attention to the features with high correlation with aspect words, and the interaction matrix represented by the two features is calculated:

$$I = H^c H^a \tag{9}$$

Through I with column softmax and row softmax, α and β are obtained. Then, the context attention γ, which is closely related to aspect words, is obtained.

$$\gamma = \frac{1}{n} \sum_i \alpha \times \beta^T \tag{10}$$

yγ is applied to the context to represent H^c, focusing on the features closely related to aspect words, and the final feature representation is obtained:

$$H = H^c \gamma \tag{11}$$

3.2. Learning Layer

A music sentence contains semantic and grammatical information, both of which play a role in determining the emotion polarity of aspect words. Two modules are designed to learn semantic information and syntactic information respectively.

A. Grammar learning module

Graph convolution is an improvement on the traditional convolutional neural network. It mainly manipulates graph structure. For each node in the graph structure, the graph convolution should consider the feature information of the node itself as well as the feature information of all neighbors, so that the information between two connected nodes can be obtained effectively. When performing aspect-level sentiment analysis task, it can transfer and obtain information according to the edges in dependency syntax tree. For example, in Figure 3, the aspect word ”service” connects to the ”was” node in the dependency syntax tree, and the ”was” node connects to the ”excellent ”node. Therefore, to obtain information, the ”excellent” node information will pass the feature information to the ”was” node first, and then the feature information to the ”service” node. In the process of information acquisition, the distance of information transmission in the relative position is shortened, and the noise in the process of transmission is effectively reduced.

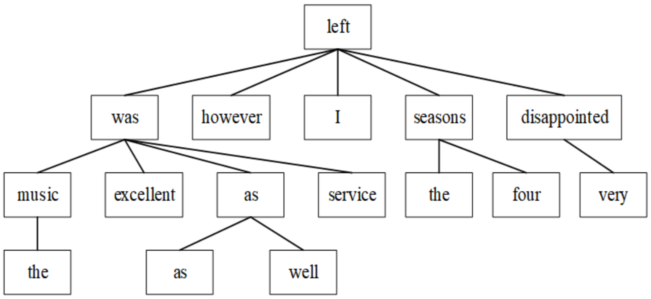

Fig. 3. Simplified dependency syntax tree

Since the common graph convolution operation does not contain graphs with marked edges, this paper uses the hierarchical lexical graph for grammar learning. The graph contains not only the syntactic tree structure, but also the co-occurrence relation between words. Through this graph convolution, words with the same co-occurrence relationship can be merged into virtual nodes, and then all the merged virtual nodes can be updated. Virtual node information update is shown in Formula (12):

$$h_i^l = ReLU(w_l(\oplus_r h_i^{l,r})) \tag{12}$$

Where, $\oplus_r$ represents connections of different co-occurrence relationship types. l indicates the number of layers. w_l is the weight in the $l - th$ layer. $h_i^{l,r}$ represents the representation of each co-occurrence relation r.

B. Semantic learning module

Before graph convolution was proposed, most models used semantic learning to improve the judgment of emotion polarity of aspect words. In this paper, multi-layer convolutional neural networks are used to learn semantic information. Convolutional neural network extracts information between adjacent words by means of sliding window from the matrix corresponding to a music text sequence through convolution kernel. The high pass permanent of the convolution kernel is 3, and the length is the dimension of the word vector. Therefore, the convolution operation is often affected by some noise words. Before semantic learning, the model in this paper weakens the noise words in location encoding. Therefore, this paper reduces some operations on noisy words in semantic learning. We use Formula (13) to learn semantic information:

$$h_i^{c,l} = CNN(h_i^p) \tag{13}$$

Where, CNN represents the convolutional operation, and l represents the layer number of the convolutional neural network.

3.3. Attention Layer

After the semantic and grammatical learning in the learning layer, the two parts of information need to be optimized and integrated. Because there are some differences between semantic information and grammatical information, direct integration of the two will lead to inadequate integration. Therefore, in order to better integrate these two types of information, semantic information and syntactic information are optimized by using tree structured information. The attention layer is divided into two parts: semantic optimization attention and syntactic optimization attention.

A. Semantic optimization attention

Semantic optimization focuses on interactive optimization of information learned by semantic learning modules and structured tree information, as shown in equations (14)-(16).

$$\tilde{\alpha}_i = h_i^t h_i^{c,l} \tag{14}$$

$$\alpha_i = \frac{exp(\tilde{\alpha}_i)}{\sum_{i=1}^{n}(\tilde{\alpha}_i)} \tag{15}$$

$$S_i^{\alpha} = \sum_{i=1}^{n} \alpha_i h_i^{c,l} \tag{16}$$

S_i^{α} represents the output of semantically optimized attention.

B. Syntactic optimization attention

The syntactic optimization attention is to optimize the interaction between the grammar information learned by the grammar learning module and the structured tree information, as shown in equations (17)-(19).

$$\tilde{\beta}_i = h_i^t h_i^l \tag{17}$$

$$\beta_i = \frac{exp(\tilde{\beta}_i)}{\sum_{i=1}^{n}(\tilde{\beta}_i)} \tag{18}$$

$$G_i^\beta = \sum_{i=1}^{n} \beta_i h_i^l \tag{19}$$

Where, G_i^β represents the output of syntactic optimization attention.

3.4. Feature Extraction Layer

The information of aspect words is particularly important for the judgment of the emotion polarity of aspect words. The feature extraction layer is to extract the information of aspect words from semantic information and grammatical information respectively. Aspect word extraction information is shown in Formula (20).

$$a = h_i, t \leq i < t + m \tag{20}$$

If $1 \leq i < t$ or $t + m \leq i \leq n$, then $a = 0$. Where, a represents aspect word information. Formula (20) is used to extract aspect word information in semantic information and aspect word information in grammatical information, denoted as a_s and a_g respectively.

3.5. Pooling Layer

After extracting aspect word information, the extracted aspect word information is integrated, that is, a_s and a_g are spliced. The information integration formula of aspect words is shown as follows:

$$o_a = [a_g, a_s]. \tag{21}$$

The integrated aspect word information is carried out the maximum pooling operation through the pooling layer to further screen the effective information of the aspect word.

$$o_p = max - pooling(o_a) \tag{22}$$

Here $max - pooling$ indicates maximum pooling.

3.6. Classification Module

For the feature representation H obtained by the previous layers of source domain data and target domain data, gradient inversion is performed using GRL_d based on gradient inversion layer:

$$GRL_d(x) = x, \frac{\partial GRL_d(x)}{\partial x} = -\lambda I \tag{23}$$

The domain classifier is trained by reversing the gradient direction so that the representation of source domain and target domain can be mapped to the same distribution space for edge alignment. Here, domain edge refers to the distribution boundary of samples from two different domains when the features of samples from two different domains are mapped to the same vector space. Domain edge alignment or feature alignment means that the feature distributions of samples from different domains are close in the same vector space, so that the classifier can better distinguish the emotional polarity of samples from different domains. Through the domain classifier, it can predict the domain category label:

$$y_d = softmax(W_d GRL_d(H) + b_d) \tag{24}$$

When the domain classifier and feature extractor are trained, the source domain representation and target domain representation are mapped into the same distribution space to narrow the distance between the two domain feature representation. However, there may be fuzzy features near the decision boundary formed by music emotion classifier, which make the samples easy to be wrongly classified. In order to better carry out domain adaptation and solve the problem of misclassification of decision boundary samples, two emotion classifiers C_1 and C_2 are trained to detect the sample points near these decision boundaries that are prone to misclassification. The feature extraction is further trained to generate better features so that the sample points are far away from the decision boundary.

As shown in Figure 4, black represents source domain data, gray represents target domain data, circle represents positive samples, and square represents negative samples.

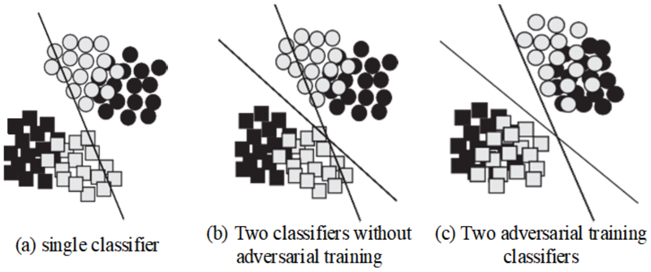
(a) single classifier
(b) Two classifiers without adversarial training
(c) Two adversarial training classifiers

Fig. 4. Decision boundaries formed by different classifiers

It can be observed from Figure 4(a) that part of the target domain samples near the decision boundary are misclassified. When a different emotion classifier is trained to form another different decision plane as shown in Figure 4(b). The effect shown in Figure 4(c) will eventually be achieved after training, alleviating the problem of sample misclassification near the decision plane to a greater extent.

The feature representation H^s is obtained from the source domain label data through the feature extraction module G, which is passed through two emotion classifiers C_1 and C_2 respectively to predict the emotion category label. It minimizes the emotion classification loss to optimize parameters and emotion classification loss is:

$$L_{cls} = -\frac{1}{N_s}\sum_{i=1}^{N_s}\sum_{j=1}^{K} M1 - \frac{1}{N_s}\sum_{i=1}^{N_s}\sum_{j=1}^{K} M2 \tag{25}$$

$$M1 = y_i^s(j) log_2 \tilde{y}_{1i}^s(j) \tag{26}$$

$$M2 = y_i^s(j) log_2 \tilde{y}_{2i}^s(j) \tag{27}$$

$$\tilde{y}_{1i} = C_1(H^s) \tag{28}$$

$$\tilde{y}_{2i} = C_2(H^s) \tag{29}$$

K represents the number of categories of affective polarity categories.

If the decision boundary defined by emotion classifiers C_1 and C_2 is changed, the samples close to the decision boundary will change accordingly. In order to make C_1 and C_2 provide different guidance for these boundary samples and form different decision boundaries for the alignment of feature representations, the difference functions of the two emotion classifiers need to be defined first. Probabilistic output of two emotion classifiers is:

$$p_1 = (y|x), p_2 = (y|x) \tag{30}$$

Defining difference loss:

$$L_{dis} = E_{x \in D_s \cup D_t}[d(p_1 = (y|x), p_2 = (y|x))] \tag{31}$$

The difference function $d(p_1 = (y|x), p_2 = (y|x))$ is realized by calculating the mean value of the absolute value difference of the output probabilities of C_1 and C_2 in each category, i.e.,

$$d(p_1 = (y|x), p_2 = (y|x)) = \frac{1}{K}\sum_{i=1}^{K}|p_{1i} = (y|x) - p_{2i} = (y|x)| \tag{32}$$

Emotion classifiers C_1 and C_2 are trained in the way of adversarial training to align feature distribution representations. In order to make the point close to the decision boundary far away from the decision boundary, it is necessary to maximize the difference loss of the two emotion classifiers. The parameters of other modules are fixed, and only the parameters of two emotion classifiers C_1 and C_2 are trained, and the points near the decision boundary are detected by maximizing the difference function. The training objectives are as follows:

$$max(C_1, C_2)E_{x\in D_s \cup D_t}[\frac{1}{K}\sum_{i=1}^{K}|p_{1i} = (y|x) - p_{2i} = (y|x)|] \tag{33}$$

In order to reclassify the misclassified points near the original decision plane, it is necessary to make them far away from the new decision boundary constructed by two different emotion classifiers. Therefore, the parameters of two emotion classifiers C_1 and C_2 are fixed, and the parameters of part G of feature extractor are trained to minimize the differences. The optimization objectives are as follows:

$$max(G)E_{x\in D_s \cup D_t}[\frac{1}{K}\sum_{i=1}^{K}|p_{1i} = (y|x) - p_{2i} = (y|x)|] \tag{34}$$

Repetitive adversarial training enables the model to locate the misclassified sample points near the original decision plane, further making it far away from the newly formed decision plane for correct classification, rather than simply field edge alignment.

3.7. Training Process

The model uses standard gradient descent algorithm to learn and train parameters. The training algorithm is cross entropy loss function and L2 regularization term.

$$\zeta(\theta) = \sum_{i=1}^{c}\hat{y}_i log y_i + \lambda\sum_{\theta\in\Theta}\theta^2 \tag{35}$$

Where, C represents the number of emotional labels. $\hat{y}_i$ represents the output emotion value of the model. y_i represents the true value in the tag. λ stands for L2 regularization parameter. Θ represents the parameter used in the model.

4. Experiments and Analysis

The new model in this paper is run on the CPU with Intel(R) Xeon(R) W2123 3.60GHz and GPU server with NVIDIA GeForce RTX3060Ti. In this experiment, python language is used for programming. The python version is Python 3.8.8, the programming tool is PyCharm Community Edition.

The model in this paper uses GloVel [14] as the word embedding layer to initialize the music text, and initializes the dimension of the word vector as 300. The hidden layer dimension of the bidirectional short and long time memory network is the same as the word vector dimension, which is set to 300. Adam is selected as the optimizer for the model. The model learning rate is 10^{-4}. The coefficient of L2 regularization term is 10^{-6}. Batch size is 32.

4.1. Experiment Data Set

The data sets used in this experiment are five public English music aspect word sentiment analysis data sets. They are Twitter, RES14, LAP14 in SemEval-2014 Task 4, Res15 in SemEval-2015 Task 12 and Res16 [29] in SemEval-2016 Task 5. The data set label distribution is shown in Table 1.

Table 1. Data sets

Data	Positive	Positive	Neutral	Neutral	Negative	Negative
Class	train	test	train	test	train	test
Twitter	1572	184	3138	357	1571	184
LAP14	995	352	475	181	882	139
RES14	2175	739	648	207	818	207
Res15	923	337	39	38	267	193
Res16	1271	480	72	32	468	128

4.2. Evaluation Index

In order to verify the validity of this experiment, two evaluation indexes are adopted in this paper, one is accuracy (Acc), the other is F1. Acc represents the proportion of samples with correct classification, calculated as shown in Equation (37). F1 is the harmonic average of accuracy rate and recall rate, calculated as shown in equations (38)-(40).

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{36}$$

$$F1 = \frac{2P \times R}{P + R} \tag{37}$$

$$P = \frac{1}{c} \sum_{i=1}^{c} \frac{TP_i}{TP_i + FN_i} \tag{38}$$

$$R = \frac{1}{c}\sum_{i=1}^{c}\frac{TP_i}{TP_i + FP_i} \tag{39}$$

Where TP represents the sample number of correctly classified positive tags. TN represents the sample number of correctly classified negative tags. FP represents the sample number of misclassified positive labels, and FN represents the sample number of misclassified negative labels. C represents the number of label types of emotion categories.

4.3. Comparison Experiments and Analysis

In order to evaluate the effectiveness of the proposed model, we make comparison with some state-of-the-art methods.

SVM: The support vector machine method is used to classify music emotion.

AKSM [30]: This model combines text sequences with aspect words to fuse information, and classifies emotion polarity through LSTM and attention mechanism.

GCNN [31]: This model uses convolutional neural network to extract feature information, then filters emotion features unrelated to aspect word information through gating mechanism, and then selectively outputs relevant emotion features.

LSTM-GAN [32]: This model uses LSTM to model music text sequences and aspect words respectively, uses interactive attention to cross-fuse text sequences and aspect word information, and generates specific music text sequences and aspect word information.

GCNBA [33]: This model uses the attention mechanism to focus on the characteristics of important information, modeling aspect words and sentences in a joint way, and then capturing the interactive information between aspect words and sentences.

HOIE [34] : This model uses dependency syntax tree to build a dependency syntax graph, and then learns syntax information through the graph convolution operation.

DIMM [35]: This model uses the dependency relationship between words to transmit emotional features.

GCNNHM [36] : This model proposes hierarchical syntactic graph and hierarchical lexical graph in dependency syntax tree, and learns syntactic information through double-layer interactive graph convolutional network.

The first five models carry out semantic information learning based on the models in machine learning and deep learning. These models do not learn grammatical information. The last three models learn sentence grammar through dependency syntactic relations. The results of each model on the data set are shown in Tables 2-6.

As can be seen from the above tables, deep learning is generally better than machine learning. This is because deep learning can independently learn the feature information in music sentences and capture richer and more complete information. By comparing the first five models with the last four models, it can be seen from the data in the table that the values of the first five models are significantly less than those of the last four models. This is mainly because the last four models take into account syntactic information, while the first five models do not. It also verifies that the grammatical information of the sentence can effectively help the model identify the emotion polarity of the aspect words.

There is little difference between the HOIE model and the DIMM model in the five data sets, and they all have their own levels. In the Twitter data set, the accuracy rate and

Table 2. Results on Twitter

Model	Acc	F1
SVM	64.51	64.41
AKSM	70.76	68.51
GCNN	72.75	70.99
LSTM-GAN	73.61	71.92
GCNBA	73.41	71.31
HOIE	73.29	71.52
DIMM	73.31	71.56
GCNNHM	75.27	74.46
Proposed	75.39	74.25

Table 3. Results on LAP14

Model	Acc	F1
SVM	71.60	70.39
AKSM	70.25	64.29
GCNN	71.01	66.82
LSTM-GAN	73.16	68.49
GCNBA	73.73	68.63
HOIE	76.66	72.16
DIMM	76.74	71.85
GCNNHM	75.70	72.95
Proposed	78.85	75.38

Table 4. Results on RES14

Model	Acc	F1
SVM	81.27	81.11
AKSM	78.43	66.68
GCNN	79.68	69.17
LSTM-GAN	80.37	71.20
GCNBA	81.08	71.53
HOIE	81.88	73.13
DIMM	82.43	72.83
GCNNHM	83.08	74.59
Proposed	83.61	75.65

Table 5. Results on RES15

Model	Acc	F1
SVM	69.34	56.21
AKSM	76.54	57.45
GCNN	78.96	60.74
LSTM-GAN	79.65	53.76
GCNBA	79.28	58.13
HOIE	81.01	63.02
DIMM	81.49	61.61
GCNNHM	82.27	65.90
Proposed	82.84	67.26

Table 6. Results on RES16

Model	Acc	F1
SVM	73.54	62.37
AKSM	84.36	64.96
GCNN	87.40	66.98
LSTM-GAN	85.85	56.32
GCNBA	88.61	67.32
HOIE	90.10	68.59
DIMM	88.82	68.98
GCNNHM	90.07	71.95
Proposed	90.72	72.68

F1 value of the DIMM model are higher than that of the HOIE model, but the difference is not significant. In Lap14, Res14 and Res15 data sets, the accuracy of the DIMM model is higher than that of the HOIE model, but the accuracy of F1 model is lower than that of the HOIE model. In Res16 data set, although the accuracy of DIMM is lower than that of HOIE, the F1 value is higher than that of HOIE. Therefore, it is difficult to improve the result of learning syntactic information only through dependency trees. Therefore, GCNNHM adds the connection relationship between words in the dependency graph. It can be seen from the data in the table that the value of GCNNHM after adding the connection relationship is significantly higher than that of the previous two models.

Compared with HOIE and DIMM, the accuracy and F1 of the proposed model in this paper are better than those of the two models in the five data sets, which indicates that the new model cannot ignore semantic information when considering grammatical information. This shows that both semantic information and grammatical information are important for the judgment of the affective polarity of aspect words.

Compared with the GCNNHM model, the results of the proposed model in the five data sets are almost improved. All results are higher than GCNNHM except that F1 is lower than GCNNHM on the Twitter dataset. In the other four data sets, F1 values increases by 2.43%, 1.06%, 1.36% and 0.73%, respectively. Among them, the accuracy of five data sets increases by 0.11%, 3.15%, 0.53%, 0.57% and 0.65%, respectively. This also verifies that the new model is meaningful for learning semantic and syntactic information of music text sequences.

4.4. Ablation Experiment

In order to verify the importance of each component in the proposed model in this paper, a series of ablation experiments are set up for verification. The experimental results are shown in Table 7 and table 8. A represents the ablation of position information. B indicates that tree structured information is dissolved. C represents the ablation of the attention layer. D represents the ablation of grammar learning module. E represents the ablation of semantic learning module. F represents the ablation of the feature extraction layer.

As can be seen from the ablation experimental data in Table 7 and table 8, after the ablation of position information and tree structure information, experimental results of A and B declines sharply, indicating that the position information and tree structure information of each word would have an impact on the judgment of related words. Compared

Table 7. Ablation Experiment Results on Twitter, LAP14 and RES14

Model	Twitter	Twitter	LAP14	LAP14	RES14	RES14
Value	Acc	F1	Acc	F1	Acc	F1
A	74.66	73.01	77.59	73.23	82.09	73.76
B	74.37	72.60	76.65	72.99	83.34	75.29
C	75.24	73.67	79.95	76.18	81.73	72.10
D	74.66	72.68	76.81	73.23	82.53	74.91
E	74.23	72.79	75.71	70.46	81.11	71.91
F	74.80	73.03	75.56	72.06	80.75	71.52
Proposed	75.38	74.25	78.85	75.38	83.61	75.65

Table 8. Ablation Experiment Results on RES15 and RES16

Model	RES15	RES15	RES16	RES16
Value	Acc	F1	Acc	F1
A	81.55	67.14	90.07	71.12
B	81.36	66.78	89.25	71.82
C	81.18	66.10	89.90	71.06
D	81.36	65.78	90.23	74.45
E	80.81	67.17	89.58	71.70
F	80.81	63.52	90.39	70.97
Proposed	82.84	67.26	90.72	72.68

with the experimental results of the five data sets, the decline amplitude of A in Res14 data set is greater than that of B, and the decline amplitude of B in the other four data sets is greater, indicating that the influence of tree structure information on aspect word is greater than that of location information. After the ablation of the attention layer, the experimental results increases in Lap14 data set, while decreases in other data sets, indicating that the semantic and grammatical information learned in Lap14 data set can express the emotion polarity of the aspect words. However, the excessive focus of attention mechanism on information, on the contrary, inhibits the fusion of semantic information and grammatical information. In the other four data sets, the attention mechanism effectively fuses semantic and syntactic information. After the ablation of the syntactic learning module and the semantic learning module, the results of both C and D show a decrease trend, but the decrease of the results of E is a little larger than that of E, indicating that the semantic learning module and the grammatical learning module are equally important in the new model, but the new model has a higher dependence on the learning of semantic information. Finally, after the ablation of the feature extraction layer, the result of F experiment decreases, indicating that it is necessary to extract the feature of aspect words at the end. From the whole data, the experimental results of the model decreases significantly after the ablation of some components in the model, indicating that every part of the model is very important and cannot be absent. The magnitude of the decline in the results after the ablation of each component also is varied, indicating that each component is sensitive to different data sets to varying degrees, which is limited by the completeness and length of the sentences in each data set.

5. Conclusions

The proposed model in this paper combines semantics and syntax for simultaneous learning. Before the model learning semantic and syntactic information, the music text sequence is updated through the position information of the music text sequence in the dependency syntax tree as well as the relative position information and grammatical position information of the text sequence, which not only effectively reduces the influence of some noise words in the text sequence, but also makes the original text sequence contain grammatically related information before information learning. The experimental results show that the proposed model can effectively learn semantic and grammatical information, and each part of the model is very important. In the task of music-level emotion analysis, the learning of semantic and grammatical information is very important for the judgment of emotion polarity of aspect words, neither of which can be ignored. In the future works, we will research more advanced algorithms and apply them in real engineering application.

Acknowledgments. This work was supported by "2022 Fujian Education System Philosophy and Social Science Research Project "Young and middle-aged Teachers Education Research Project" Research results of Fujian and Taiwan Intangible Cultural Heritage in Ethnic instrumental Music Education in Colleges and Universities, project number: JAS22059".

References

1. Birjali, M., Kasri, M., Beni-Hssane, A.: "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowledge-Based Systems*, Vol. 226, 107134. (2021)
2. Yadav, A., Vishwakarma, D.: "Sentiment analysis using deep learning architectures: a review," *Artificial Intelligence Review*, Vol. 53, No. 6, 4335-4385. (2020)
3. Jiang, D., Li, H., and Yin, S.: "Speech Emotion Recognition Method Based on Improved Long Short-term Memory Networks," *International Journal of Electronics and Information Engineering*, Vol. 12, No. 4, 147-154. (2020)
4. Song, M., Park, H., Shin, K.: "Attention-based long short-term memory network using sentiment lexicon embedding for aspect-level sentiment analysis in Korean," *Information Processing & Management*, Vol. 56, No. 3, 637-653. (2019)
5. Chen, Y., Zhuang, T., Guo, K.: "Memory network with hierarchical multi-head attention for aspect-based sentiment analysis," *Applied Intelligence*, Vol. 51, 4287-4304. (2021)
6. Li, W., Qi, F., Tang, M., et al.: "Bidirectional LSTM with self-attention mechanism and multichannel features for sentiment classification," *Neurocomputing*, Vol. 387, 63-77. (2020)
7. Xu, X., Zhu, G., Wu, H., Zhang, S., Li, K.: "SEE-3D: Sentiment-driven Emotion-Cause Pair Extraction Based on 3D-CNN," *Computer Science and Information Systems*, Vol. 20, No. 1, 77C93. (2023), https://doi.org/10.2298/CSIS220303047X
8. Zhao, Y., Li, H., Yin, S.: "A Multi-channel Character Relationship Classification Model Based on Attention Mechanism," *International Journal of Mathematical Sciences and Computing(IJMSC)*, vol. 8, no. 1, 28-36. (2022)
9. Yang, M., Tjuawinata, I., Lam, K., Zhu, T., and Zhao, J.: "Differentially Private Distributed Frequency Estimation," *IEEE Transactions on Dependable and Secure Computing*, doi: 10.1109/TDSC.2022.3227654.
10. Zhang, K., Zhang, H., Liu, Q., et al.: "Interactive attention transfer network for cross-domain sentiment classification," *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, No. 01, 5773-5780. (2019)

11. Tian, Y., Yang, L., Sun, Y., et al.: ”Cross-domain end-to-end aspect-based sentiment analysis with domain-dependent embeddings,” *Complexity*, Vol. 2021, 1-11. (2021)
12. Cao, Z., Zhou, Y., Yang, A., et al.: ”Deep transfer learning mechanism for fine-grained cross-domain sentiment classification,” *Connection Science*, Vol. 33, No. 4, 911-928. (2021)
13. Liang, B., Su, H., Gui, L., et al.: ”Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks,” *Knowledge-Based Systems*, Vol. 235, 107643. (2022)
14. Huang, B., Zhang, J., Ju, J., et al.: ”CRF-GCN: An effective syntactic dependency model for aspect-level sentiment analysis,” *Knowledge-Based Systems*, Vol. 260, 110125. (2023)
15. Zhang, Z., Hu, C., Pan, H., et al.: ”Aspect-Dependent Heterogeneous Graph Convolutional Network for Aspect-Level Sentiment Analysis,” *2022 International Joint Conference on Neural Networks (IJCNN). IEEE*, 1-8. (2022)
16. Zhao, A., Yu, Y.: ”Knowledge-enabled BERT for aspect-based sentiment analysis,” *Knowledge-Based Systems*, Vol. 227, 107220. (2021)
17. Xue, B., Zhu, C., Wang, X., et al.: ”The Study on the Text Classification Based on Graph Convolutional Network and BiLSTM,” *Proceedings of the 8th International Conference on Computing and Artificial Intelligence*, 323-331. (2022)
18. Wang, L., Yin, S., Hashem, A., et al.: ”A novel deep learning-based single shot multibox detector model for object detection in optical remote sensing images,” *Geoscience Data Journal*, (2022). https://doi.org/10.1002/gdj3.162
19. Chen, H., Nemni, E., Vallecorsa, S., Li, X., Wu, C., Bromley, L.: ”Dual-Tasks Siamese Transformer Framework for Building Damage Assessment,” *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia*, 1600-1603, (2022). doi: 10.1109/IGARSS46834.2022.9883139.
20. Wu, Y., Li, W.: ”Aspect-level sentiment classification based on location and hybrid multi attention mechanism,” *Applied Intelligence*, Vol. 52, No. 10, 11539-11554. (2022)
21. Wang, Y., Chen, Q., Ahmed, M., et al.: ”Joint inference for aspect-level sentiment analysis by deep neural networks and linguistic hints,” *IEEE transactions on knowledge and data engineering*, Vol. 33, No. 5, 2002-2014. (2019)
22. Sun, K., Zhang, R., Mao, Y., et al.: ”Relation extraction with convolutional network over learnable syntax-transport graph,” *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, No. 05, 8928-8935. (2020)
23. Zhang, R., Chen, Q., Zheng, Y., et al.: ”Aspect-Level Sentiment Analysis via a Syntax-Based Neural Network,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 30, 2568-2583. (2022)
24. Yang, L., Li, Y., Wang, J., et al., ”Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning,” *IEEE access*, Vol. 8, 23522-23530. (2020)
25. Zheng, Y., Gao, Z., Shen, J., et al., ”Optimising Automatic Text Classification Approach in Adaptive Online Collaborative Discussion-A perspective of Attention Mechanism-Based Bi-LSTM,” *IEEE Transactions on Learning Technologies*, 1-14. (2022)
26. Zhao, G., Yang, P.: ”Pretrained embeddings for stance detection with hierarchical capsule network on social media,” *ACM Transactions on Information Systems (TOIS)*, Vol. 39, No. 1, 1-32. (2020)
27. Pennington, J., Socher, R., Manning, C.: ”Glove: Global vectors for word representation,” *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532-1543. (2014)
28. Xue, Y., Li, Y., Liu, S., et al.: ”Oriented localization of surgical tools by location encoding,” *IEEE Transactions on Biomedical Engineering*, Vol. 69, No. 4, 1469-1480. (2021)
29. Pontiki, M., Galanis, D., Papageorgiou, H., et al.: ”Semeval-2015 task 12: Aspect based sentiment analysis,” *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, 486-495. (2015)
30. Sindhu, C., Vadivu, G.: ”Fine grained sentiment polarity classification using augmented knowledge sequence-attention mechanism,” *Microprocessors and Microsystems*, 81, 103365. (2021)

31. Zhang, J., Liu, F., Xu, W., et al.: ”Feature fusion text classification model combining CNN and BiGRU with multi-attention mechanism,” *Future Internet*, Vol. 11, No. 11, 237. (2019)
32. Yu, Y., Srivastava, A., Canales, S.: ”Conditional lstm-gan for melody generation from lyrics,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, Vol. 17, No. 1, 1-20. (2021)
33. Liu, J., Liu, P., Zhu, Z., et al.: ”Graph convolutional networks with bidirectional attention for aspect-based sentiment classification,” *Applied Sciences*, Vol. 11, No. 4, 1528. (2021)
34. Li, B., Fan, Y., Sataer, Y., et al.: ”Improving semantic dependency parsing with higher-order information encoded by graph neural networks,” *Applied Sciences*, Vol. 12, No. 8, 4089. (2022)
35. Wen, J., Jiang, D., Tu, G., et al.: ”Dynamic interactive multiview memory network for emotion recognition in conversation,” *Information Fusion*, Vol. 91, 123-133. (2023)
36. Li, X., Lu, R., Liu, P., et al.: ”Graph convolutional networks with hierarchical multi-head attention for aspect-level sentiment classification,” *The Journal of Supercomputing*, Vol. 78, No. 13, 14846-14865. (2022)

Wenwen Chen is with the Jimei University School of Music. She has published several papers and won several awards. Her research direction: Emotion recognition, artistic feature extraction.

Received: December 12, 2022; Accepted: May 05, 2023.