TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 14/24 pp481–491
DOI: 10.26599/TST.2023.9010021
Volume 29, Number 2, April 2024

Deep Broad Learning for Emotion Classification in Textual Conversations

Sancheng Peng, Rong Zeng, Hongzhan Liu*, Lihong Cao*, Guojun Wang, and Jianguo Xie

Abstract: Emotion classification in textual conversations focuses on classifying the emotion of each utterance from textual conversations. It is becoming one of the most important tasks for natural language processing in recent years. However, it is a challenging task for machines to conduct emotion classification in textual conversations because emotions rely heavily on textual context. To address the challenge, we propose a method to classify emotion in textual conversations, by integrating the advantages of deep learning and broad learning, namely DBL. It aims to provide a more effective solution to capture local contextual information (i.e., utterance-level) in an utterance, as well as global contextual information (i.e., speaker-level) in a conversation, based on Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (Bi-LSTM), and broad learning. Extensive experiments have been conducted on three public textual conversation datasets, which show that the context in both utterance-level and speaker-level is consistently beneficial to the performance of emotion classification. In addition, the results show that our proposed method outperforms the baseline methods on most of the testing datasets in weighted-average F1.

Key words: emotion classification; textual conversation; Convolutional Neural Network (CNN); Bidirectional Long Short-Term Memory (Bi-LSTM); broad learning

1 Introduction

Emotion Classification in Textual Conversations (ECTC) aims to classify the emotion of each utterance from textual conversations. It is becoming one of the most important tasks in Natural Language Processing (NLP) due to its wide range of applications[1, 2], such as public opinion mining, behavior analysis, mental health, recommendation systems, etc. In addition, it can also be utilized in human-machine dialogue systems, question-answering systems, chat robots, and so on.

In recent years, many methods have been proposed for the task of ECTC, which can be divided into two types: sequence-based methods[3–13] and graph-based methods[14–20]. In the sequence-based methods, Bidirectional Long Short-Term Memory (Bi-LSTM)[21], Recurrent Neural Network (RNN)[22], and Gated Recurrent Unit (GRU)[23] are usually adopted to capture contextual information among utterances. However, these methods neglect the relationship between the utterance and the speaker. In the graph-based methods, graph convolutional networks are utilized to capture long-distance dependency among speakers.

• Sancheng Peng and Lihong Cao are with Center for Linguistics and Applied Linguistics, and Laboratory of Language Engineering and Computing, Guangdong University of Foreign Studies, Guangzhou 510006, China. E-mail: psc346@aliyun.com; 201610130@oamail.gdufs.edu.cn.
• Rong Zeng and Hongzhan Liu are with Guangdong Provincial Key Laboratory of Nanophotonic Functional Materials and Devices, South China Normal University, Guangzhou 510006, China. E-mail: zengrong980302@163.com; lhzscnu@163.com.
• Guojun Wang is with School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou 510006, China. E-mail: csgjwang@gzhu.edu.cn.
• Jianguo Xie is with Modern Education Technology Center, Guangdong University of Foreign Studies, Guangzhou 510006, China. E-mail: 200611473@oamail.gdufs.edu.cn.
∗ To whom correspondence should be addressed.
Manuscript received: 2022-11-22; revised: 2023-02-18; accepted: 2023-03-19

However, these methods suffer from graph sparsity and computation complexity.

Despite progress made by the above-mentioned methods, there are still many challenges for ECTC[4]. The first is that the same word may deliver different emotions in different contexts, for instance, the word "yes" may deliver many different emotions, such as "joy", "sadness", and "neutrality". Thus, to identify the emotion of a speaker precisely, it needs to capture the context in textual conversation more effectively. The second is that most recent methods on ECTC capture the contextual information of conversations by using the deep learning models, such as RNN, graph neural network, and transformer. However, there are some disadvantages to these methods, such as a large number of parameters and a large number of computing resources. In addition, RNN is difficult to model the long-distance dependency, and the graph-based neural network is difficult to capture sequential features among utterances.

To address these challenges, we propose a novel method for ECTC tasks, by integrating the advantages of deep learning (Convolutional Neural Network (CNN) and Bi-LSTM) and Broad Learning (BL), namely DBL. In DBL, a popular pre-training language model BERT is utilized to obtain the initial word representations, to capture the context dependency of utterance and the global information of the conversation. Then, two deep broad learning modules (CNN and Bi-LSTM) and BL are leveraged to obtain utterance-level and speaker-level contextual information by capturing local and global contextual features, respectively. The CNN is utilized to extract k-gram features in an utterance, and Bi-LSTM is applied to simultaneously extract the sequential information and long-distance dependencies between two utterances. The BL is utilized to integrate all the features and classify emotions, which can solve the high computational complexity of deep networks.

The contributions of this paper are summarized as follows:

- It is the first effort to combine deep learning and broad learning to classify emotions in textual conversations.
- A more effective framework is designed to capture local contextual information (i.e., utterance-level) in an utterance, and to capture global contextual information (i.e., speaker-level) in a conversation, by using the CNN, Bi-LSTM, and BL.
- Extensive experiments have been conducted on three datasets with baseline methods for comparison. The experimental results demonstrate the superiority of our proposed method DBL and verify the motivation of the combination of deep learning and broad learning for the ECTC task.

2 Related Work

2.1 Emotion classification

Emotion classification aims to predict emotional polarity in a given text, which has been widely studied in recent years. In general, emotion[24] denotes people's attitude experience and corresponding behavioral responses to objective things. Ekman[25] divided emotion into six types: anger, disgust, fear, happiness, sadness, and surprise. Ben-Zeev[26] provided a wheel-shaped emotion classifier on the basis of Ekman's model, which includes four bipolar sets: joy and sadness; anger and fear; trust and disgust; and surprise and anticipation, respectively.

Existing emotion classification methods may be divided as follows[2]: text-oriented monolingual, text-oriented cross-linguistic, and emoji-oriented cross-linguistic. Text-oriented monolingual method[27] focuses on classifying textual emotion in a single language (e.g., Chinese). Text-oriented cross-linguistic method[28] focuses on exploring textual emotion classification from source language text (e.g., English) to target language (e.g., German). Emoji-oriented multi-linguistic method[29] focuses on predicting emoji in multi-linguistic texts (e.g., English and Spanish).

2.2 Emotion classification in conversation

Existing works usually capture contextual features for the ECTC task by using deep learning methods, which can be divided into sequence-based and graph-based methods.

2.2.1 Sequence-based method

MVN[3] utilizes a multi-view network to capture word- and utterance-level dependencies. HiGRU[4] employs a lower-level GRU to model word-level inputs and utilizes an upper-level GRU to capture contexts of utterance-level embeddings. DialogueRNN[5] adopts RNN to classify emotion in a conversation. COSMIC[6] utilizes GRU to model context and commonsense. DialogueCRN[7] adopts LSTM to learn intrinsic logical order and employs attention mechanism to match relevant contextual clues. IDS-ECM[8] utilizes Bi-LSTM for feature extraction. HiTrans[9] utilizes a low-level transformer to generate local utterance representations and employs a high-level transformer to capture global

context information in conversations. Lu et al.[10] proposed an iterative emotion interaction network to model the emotion interaction explicitly. TODKAT[11] utilizes a pointer network and additive attention to integrate commonsense knowledge from multiple sources and dimensions. DialogXL[12] adopts the pre-training language model XLNet for the ECTC task. MDFN[13] utilizes Bi-GRU for the ECTC task by decoupling the utterance-aware and speaker-aware information.

2.2.2 Graph-based method

DialogueGCN[14] utilizes Relational Graph Attention Networks (RGAT) to model both self-dependency and inter-speaker dependency. RGAT[15] employs relational graph attention networks to recognize emotions in conversations. ConGCN[16] regards the utterances and speakers as graph nodes and the dependencies between speakers and utterances as graph edges. KET[17] utilizes hierarchical self-attention to encode contextual utterances and employs a context-aware graph attention mechanism to incorporate commonsense knowledge. DAG-ERC[18] adopts a directed acyclic graph to model conversation context. HGNN[19] utilizes the heterogeneous graph neural networks for the ECTE task.

2.3 Broad learning

BL was proposed by Chen and Liu[30]. Instead of stacking and greatly expand neurons, BL expands neurons with feature nodes (namely F) and enhancement nodes (namely E) in a wide manner. Then, the output weight is calculated by the pseudo inverse. BL is an effective alternative method for deep learning, due to its simple network structure, short training time, and strong generalization ability. Its training process can also be extended to the incremental learning model without retraining when new nodes are added. Chen et al.[31] also proved theoretically the universal approximation ability of BL. This method has achieved better results in many applications, such as cross-domain emotion classification[32, 33] and negative emotion classification[34].

3 Methodology

DBL is a method for classifying emotion in textual conversations. It consists of five components: utterance encoding, utterance-level context encoding, speaker-level encoding, emotion classifier, and prediction. The framework of DBL is shown in Fig. 1.

In utterance encoding, the pre-training model BERT is adopted to obtain the word embedding of each utterance in a conversation. Then, the utterance-level context encoding and speaker-level encoding are utilized to obtain two different types of sequence information. More specifically, CNN is responsible for extracting k-gram features, and Bi-LSTM is responsible for extracting simultaneously sequential information and long-distance dependencies between two utterances. In the emotion classifier, the BL is responsible for integrating all the features and classifying emotions. Finally, the output of BL is input into the softmax to obtain the classification results.

3.1 Problem definition

As to ECTC, we may define a textual conversation as $U = \{u_1, u_2, \ldots, u_K\}$, and define a set of speakers as $S = \{s_1, s_2, \ldots, s_M\}$, where K denotes the number of utterances, and M denotes the number of speakers. Each utterance u_i is uttered by the speaker $s_{\delta(u_i)} \in S$, where δ is employed to map the utterance index into the corresponding speaker.

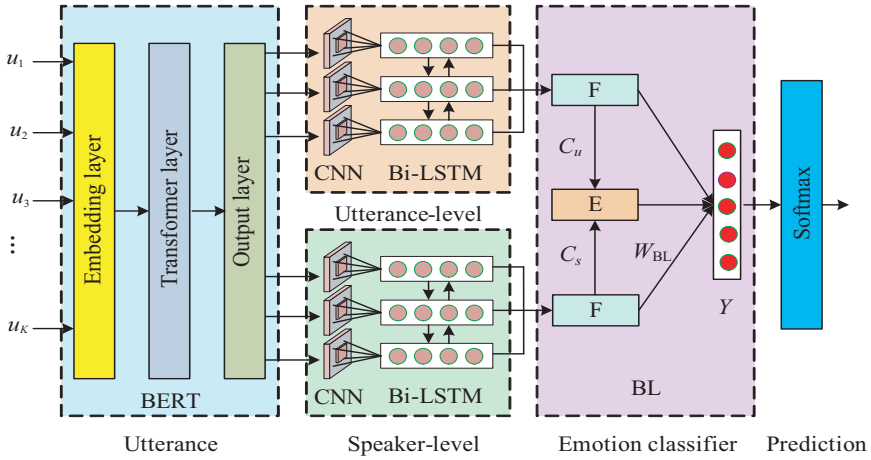


Fig. 1 Framework of the proposed method DBL.

3.2 Utterance encoding

We employ the Chinese BERT[35] to extract the features of utterances and to obtain word vectors, because it not only captures the contextual information but also extracts the key features in the text. Given K utterances in a conversation, we insert a token CLS before each utterance u_i, and then a sequence $\{CLS_1, u_1, CLS_2, u_2, \ldots, CLS_K, u_K\}$ is obtained.

Thus, we utilize the last hidden layer of the BERT to obtain the vector representation of each word in each utterance u_i. If there are n_i words in an utterance u_i, we can obtain the embedding of a conversation $X = \left\{x_{CLS_1}, x_{u_1}, x_{CLS_2}, x_{u_2}, \ldots, x_{CLS_K}, x_{u_K}\right\}$, where $x_{u_i} \in \mathbf{R}^{n_i \times 768}$ denotes the representations of all words embedding in u_i.

Since there is a length limitation (512 tokens) in the BERT input, if the total length for all utterances in a conversation exceeds 512, it cannot deal with all utterances in a conversation simultaneously. To address this problem, by following the solution proposed in HiTrans[9], the utterances in an overlong conversation are split into chunks whose lengths are less than 512. For example, assuming that there are 4 utterances $\{u_1, u_2, u_3, u_4\}$ in a conversation, the length of u_1 and u_2 together is no more than 512, but it will exceed 512 when u_3 is added. Thus, u_1 and u_2 are taken as a chunk together and input into the BERT, and then whether the length of u_3 and u_4 together exceeds 512 or not is checked. If not, u_3 will be input into the BERT; otherwise, u_3 and u_4 together are input into the BERT.

3.3 Utterance-level context encoding

Utterance-level context encoding aims to consider the emotional change of utterances through the utterance-dependency, which means that it can be affected by the current context and other utterances from other speakers. Since the input length of BERT is limited, it hard to obtain the complete contextual information of entire conversation. Thus, CNN and Bi-LSTM are utilized to capture the k-gram feature and long-distance dependency of the utterances, respectively. First, CNN is employed to extract features of k-gram features and salient features by convolution and max-pooling operation. Then, Bi-LSTM is adopted to learn the correlation among each utterance.

A conversation X embeddings are input into the CNN. The output of the convolutional layer is represented as follows:

$$G_u = \varphi\left(W_u \otimes X + b_u\right) \in \mathbf{R}^{K \times q} \tag{1}$$

where φ denotes the activation function, W_u and b_u denote the weights and bias of convolution kernel, respectively, q denotes the output dimension of the CNN, and $\otimes$ denotes the convolution operation.

After the input data are extracted by the convolutional layer, the max-pooling operation is adopted to extract salient features and reduce the dimension of convolution output. Thus, the output of the max-pooling layer is represented as follows:

$$P_u = \text{MaxPool}\left(G_u\right) \in \mathbf{R}^{K \times r} \tag{2}$$

where MaxPool denotes the max-pooling operation of the CNN and r denotes the output dimension of the max pool layer.

Because the CNN only considers the current input and ignores the previous degenerate information, it is difficult to capture the relevant contextual information among utterances. Thus, the Bi-LSTM is utilized to solve this problem. The CNN features of each utterance $P_u = \left\{p_i^u\right\}_{i=1}^K$ are fed into the Bi-LSTM to learn the correlative information among utterances,

$$\overrightarrow{c_i^u}, \overrightarrow{h_i^u} = \overrightarrow{\text{LSTM}_u}\left(p_i^u, \overrightarrow{h_{i-1}^u}\right),\ i = 1, 2, \ldots, K \tag{3}$$

$$\overleftarrow{c_i^u}, \overleftarrow{h_i^u} = \overleftarrow{\text{LSTM}_u}\left(p_i^u, \overleftarrow{h_{i-1}^u}\right),\ i = 1, 2, \ldots, K \tag{4}$$

where $\overrightarrow{h_i^u}$ and $\overleftarrow{h_i^u}$ denote the working memory of forward LSTM and backward LSTM, respectively, and $\overrightarrow{c_i^u}$ and $\overleftarrow{c_i^u}$ denote the contextual features of two directions of i-th utterance, respectively, $\overrightarrow{\text{LSTM}_u}$ and $\overleftarrow{\text{LSTM}_u}$ denote the forward and backward calculations of LSTM, respectively.

These two features $\overrightarrow{c_i^u}$ and $\overleftarrow{c_i^u}$ are concatenated into $c_i^u = [\overrightarrow{c_i^u}; \overleftarrow{c_i^u}]$ to generate the utterance-level feature for i-th utterance. Thus, the utterance-level feature of a conversation is represented as follows:

$$C_u = \left[c_1^u, c_2^u, \ldots, c_K^u\right] \in \mathbf{R}^{K \times (2l)} \tag{5}$$

where l denotes the dimension of hidden state of the LSTM.

3.4 Speaker-level context encoding

Speaker-level context encoding aims to consider not only the emotional change between two utterances but also the self-dependency utterance between two adjacent utterances of the same speaker.

Let X be the CNN input. The output of the convolutional layer is represented as follows:

$$G_s = \varphi\left(W_s \otimes X + b_s\right) \in \mathbf{R}^{K \times q} \tag{6}$$

where W_s denotes the weight of the convolutional layer and b_s denotes the bias of the current convolutional layer.

Then, the output of the convolutional layer is input into the pooling layer. The specific description is represented as follows:

$$P_s = \text{MaxPool}\,(G_s) \in \mathbf{R}^{K\times r} \tag{7}$$

The CNN features of each utterance $P_s = \{p_i^s\}_{i=1}^K$ are fed into the Bi-LSTM to learn the correlation of each utterance from the same speaker,

$$\overrightarrow{c_i^s}, \overrightarrow{h_{\theta,i}^s} = \overrightarrow{\text{LSTM}_s}(p_i^s, \overrightarrow{h_{\theta,i-1}^s}),\ i = [1, |U_\theta|] \tag{8}$$

$$\overleftarrow{c_i^s}, \overleftarrow{h_{\theta,i}^s} = \overleftarrow{\text{LSTM}_s}(p_i^s, \overleftarrow{h_{\theta,i-1}^s}),\ i = [1, |U_\theta|] \tag{9}$$

where $\theta = \delta\,(u_i)$, $|U_\theta|$ refers to all utterances of the speaker s_θ, $\overrightarrow{h_{\theta,i}^s}$, and $\overleftarrow{h_{\theta,i}^s}$ denote the working memory of the forward LSTM and backward LSTM, respectively, $\overrightarrow{c_i^s}$ and $\overleftarrow{c_i^s}$ denote the contextual features of two directions of i-th speaker; $\overrightarrow{\text{LSTM}_s}$ and $\overleftarrow{\text{LSTM}_s}$ denote the forward and backward calculations, respetively.

The two features $\overrightarrow{c_i^s}$ and $\overleftarrow{c_i^s}$ are concatenated into $c_i^s = [\overrightarrow{c_i^s}; \overleftarrow{c_i^s}]$ to generate the speaker-level feature for i-th speaker. Thus, the speaker-level feature in a conversation is represented as follows:

$$C_s = \left[c_1^s, c_2^s, \ldots, c_K^s\right] \in \mathbf{R}^{K\times(2l)} \tag{10}$$

3.5 Emotion classifier

After the utterance-level and speaker-level encoding are obtained in Sections 3.3 and 3.4, respectively, BL is adopted to calculate the weight of each emotion label and to obtain the emotion label prediction in each utterance.

Since BL is composed of feature node, enhancement node, and an output layer, the feature embeddings are first linearly mapped into n groups of feature nodes, and then feature nodes are nonlinearly mapped into m groups of enhancement nodes. Finally, the feature nodes and enhancement nodes are input into the output layer to obtain the probability distribution of emotions. During the training process of BL, the weights of feature nodes and enhancement nodes are generated randomly and fixed, and the weights of the output layer are optimized by the ridge regression method.

However, in DBL, the deep features of utterances are extracted through the CNN and Bi-LSTM, which need not be linearly transformed into features of BL. Thus, in BL, they are directly treated as the feature nodes, which are nonlinearly transformed into the enhancement nodes. Finally, the feature nodes and enhancement nodes are concatenated to input into the output layer for calculating the weight of each label.

Thus, the utterance-level features C_u and speaker-level features C_s in a conversation are concatenated and nonlinearly mapped into m groups of enhancement nodes. The j-th group of enhancement nodes E_j is represented as follows:

$$E_j = \xi([C_u, C_s]W_{ej} + \beta_{ej}) \in \mathbf{R}^{N\times t},\ j = 1, 2, \ldots, m \tag{11}$$

where t denotes the number of enhancement nodes of each group, W_{ej} and β_{ej} are randomly generated, which denote the weight matrix and bias matrix, respectively, and ξ denotes a nonlinear activate function.

We assume that $E = [E_1, E_2, \ldots, E_m]$ as m groups of enhancement nods. Thus, the output Y can be represented as follows:

$$Y = [C_u, C_s, E]\, W_{\text{BL}} = A \times W_{\text{BL}} \tag{12}$$

where W_{BL} denotes the output weight of BL, and A denotes the actual input of BL.

To shorten the calculation time and to prevent overfitting, the ridge regression is adopted as an objective function in the general BL, which is represented as follows:

$$\arg\min_{W_{\text{BL}}} \|A \times W_{\text{BL}} - Y\|_2^2 + \lambda \|W_{\text{BL}}\|_2^2 \tag{13}$$

where λ denotes the regularization parameters.

Finally, according to the regularized least square method, W_{BL} can be represented as follows:

$$W_{\text{BL}} = (\lambda I + AA^{\text{T}})^{-1} A^{\text{T}} \hat{Y} \tag{14}$$

where I denotes an identity matrix and $\hat{Y}$ denotes the ground truth label of each utterance.

The specific process is shown in Algorithm 1.

3.6 Prediction

After the processing of BL, its output will be input into the prediction component, which consists of two full connection layers and a softmax classifier. In the full connection layer, ReLU function is adopted as the activation function to avoid the problem of gradient explosion. In addition, during the training process of the

Algorithm 1 Learning process of DBL

Input: $U = \{u_1, u_2, \ldots, u_K\}$
Output: Y

```
Generate utterance embedding by BERT;
while stopping criterion is not met do
    Generate utterance-level features by Eqs. (3) and (4);
    Generate speaker-level features by Eqs. (8) and (9);
    Generate enhancement nodes by Eq. (11);
    Compute W_BL by Eq. (14);
end while
Compute Y by Eq. (12).
```

model, the cross-entropy is adopted as the loss function. Finally, the prediction results are output through the softmax.

4 Experiment

In this section, we provide a brief description of the datasets, baseline methods, parameter settings, and evaluation metrics utilized in our experiments, and compare our proposed method with the baseline methods.

4.1 Datasets and evaluation metrics

To verify the DBL effectiveness, the following three benchmark datasets are utilized, i.e., MELD, EmoryNLP, and IEMOCAP.

MELD[36]: It is collected from *Friends* TV series. Its emotion labels include neutral, surprise, fear, sadness, joy, disgust, and anger.

EmoryNLP[37]: It is also collected from *Friends* TV series, which includes neutral, joyful, scared, mad, sad, powerful, and peaceful.

IEMOCAP[38]: It contains video, audio, and text transcriptions, which are annotated with six kinds of emotion labels: neutral, happiness, sadness, anger, frustration, and excitement.

In terms of the evaluation metric in our experiments, we evaluate the overall performance by using a weighted-average F1 (namely W-F1) score. The specific statistics of these datasets are listed in Table 1.

4.2 Implementation detail

DBL is implemented with the BERT, which is composed of 12 transformer blocks and is pre-trained on a large number of English corpus (e.g., Wikipedia, news). The region size of filters for the CNN is set to 2, 3, and 4, and the feature maps of each filter are set to 100. The nodes of the hidden layer for the Bi-LSTM are set to 200. The BL classifier is implemented with 10 groups of enhancement nodes, 50 nodes in each group, and the activation function is tanh. Model optimization is conducted using the AdamW update strategy[39] with the initial learning rate setting to 8×10^{-6} and weights decay setting to 0.01. The hyperparameters corresponding to the best performance on the validation set are obtained by the grid search, and the regularization parameter is set to $\lambda = 0.001$.

Table 1 Descriptions for training, validation, and test information data in datasets.

Dataset	Conversation (train/ validation/test)	Utterance (train/ validation/test)	Number of classes
MELD	1038/114/280	9989/1109/2610	7
EmoryNLP	659/89/79	7551/954/984	7
IEMOCAP	100/20/31	4810/1000/1523	6

4.3 Compared methods

We compare our proposed method DBL with the following baseline methods in our experiments:

- **TextCNN**[40]**:** It is a convolutional neural network for emotion classification without considering utterance-level and speaker-level contextual information of textual conversation.
- **KET**[17]**:** It is a transformer-based model which adopts external commonsense knowledge to improve contextual utterance representations.
- **DialogueRNN**[5]**:** It is an RNN-based model, which uses three GRUs to track the status of speakers, global contexts, and historical emotions, separately.
- **HiTrans**[9]**:** It is a transformer-based context and speaker-sensitive model for emotion classification in textual conversations.
- **QMNN**[41]**:** It is a quantum-based model which adopts a quantum-inspired neural network for conversational emotion classification.
- **KAITML**[42]**:** It is a graph-based model which introduces a dual-level graph attention mechanism and multi-task learning for conversational emotion classification.
- **DialogXL**[12]**:** It is an XLNet-based model for conversational emotion classification by capturing the useful intra- and inter-speaker dependencies.
- **MVN**[3]**:** It is a multi-view network for the word- and utterance-level dependencies capturing.
- **DBL:** It is our proposed method, which is a deep broad learning method for emotion classification by considering the utterance-level and speaker-level contextual information of textual conversation.

5 Result and Analysis

We evaluate the overall performance by using W-F1 and adopt the results of the baseline methods, like TextCNN, KET, DialogueRNN, and HiTrans, reported in Ref. [9], and the results of QMNN, KAITML, DialogXL, and MVN reported in Refs. [3, 12, 37, 38], respectively.

5.1 Performance comparison

To demonstrate the effectiveness of DBL, we compare it with the baseline methods on the MELD, EmoryNLP,

and IEMOCAP datasets, and the experimental results are shown in Table 2.

Table 2 provides the experimental results on the MELD dataset. It is found that MVN can achieve the best overall performance of 62.47% on W-F1 score among the baseline methods. By comparison, the performance of our proposed DBL outperforms MVN by 1.56%. DBL can improve the overall performance by comparison with the baseline methods.

Table 2 also provides the experimental results on the EmoryNLP dataset. Among all the baseline methods, HiTrans achieves the best overall performance of 36.75% on W-F1. In comparison, the performance of our proposed DBL outperforms HiTrans by 0.24%, which can prove the effectiveness of DBL. It also shows that there is a better capability for DBL to classify emotion in textual conversation.

From Table 2, we can see that MVN has the best overall performance among all the baselines. In particular, MVN and DialogXL perform better than our proposed method on IEMOCAP, which may be attributed to the capturing of useful word- and utterance-level dependencies, and intra- and inter-speaker dependencies.

In conclusion, DBL outperforms other baseline models in most cases. The main reason is that the CNN can effectively capture the k-gram features, Bi-LSTM can effectively capture the long-distance dependencies and sequential information, and BL can effectively integrate these features into a high-dimensional feature space to obtain richer semantic information.

5.2 Model analysis

To verify the effectiveness of our proposed method, we conduct two extensive experiments: one analyzes the impact of the CNN and Bi-LSTM on the emotion classification performance; another analyzes the impact of the CNN, Bi-LSTM, and BL on the emotion classification performance. The experimental results are shown in Table 3.

Table 2 W-F1 for different methods on MELD, EmoryNLP, and IEMOCAP datasets.

(%)

Method	Dataset		
	MELD	EmoryNLP	IEMOCAP
TextCNN	52.55	27.85	43.25
KET	58.18	34.39	59.56
DialogueRNN	57.03	31.70	62.75
HiTrans	61.94	36.75	64.5
QMNN	58.00	–	59.88
KAITML	58.87	35.59	61.43
DialogXL	62.41	34.73	65.94
MVN	62.47	–	**65.99**
DBL(Ours)	**64.03**	**36.99**	62.53

Table 3 provides the experimental results on the MELD, EmoryNLP, and IEMOCAP for different components in DBL. As shown in Table 3, it is found that BERT+CNN+Bi-LSTM can achieve better performance on MELD, EmoryNLP, and IEMOCAP compared with the BERT. As to these three datasets, the improvements of W-F1 are 2.32%, 0.56%, and 7.44%, respectively. The possible reason is that CNN and Bi-LSTM can effectively help BERT+CNN+Bi-LSTM to extract the k-gram features, sequential information, and long-distance dependencies for improving performance.

As shown in Table 3, it is found that DBL (i.e., BERT+CNN+Bi-LSTM+BL) can also achieve better performance on MELD, EmoryNLP, and IEMOCAP compared with BERT. As to these three datasets, the improvements of W-F1 are 3.59%, 2.07%, and 8.38%, respectively. The possible reason is that DBL can effectively model the contextual relations of conversations by capturing the utterance-level and speaker-level features to improve the performance.

In addition, DBL can achieve better performance on MELD, EmoryNLP, and IEMOCAP by comparing with BERT+CNN+Bi-LSTM. As to these three datasets, the improvements of W-F1 are 1.27%, 1.51%, and 0.94%, respectively. The possible reason is that BL can help DBL to integrate the utterance-level and speaker-level features effectively and to transform them into the high-dimensional feature space by introducing the feature and enhancement nodes in BL, thereby improving the performance of DBL.

5.3 Ablation analysis

To comprehensively address the impact of the utterance- and speaker-level encoding on DBL, we conduct an ablation analysis on MELD, EmoryNLP, and IEMOCAP datasets by removing them separately, and investigate their contributions to these datasets. The experimental results are shown in Table 4.

From Table 4, it is found that both utterance- and

Table 3 Analysis of the effectiveness with W-F1 for DBL on MELD, EmoryNLP, and IEMOCAP datasets.

(%)

Method	Dataset		
	MELD	EmoryNLP	IEMOCAP
BERT	60.44	34.92	54.15
BERT+CNN+ Bi-LSTM	62.76	35.48	61.59
DBL	**64.03**	**36.99**	**62.53**

Table 4 Ablation analysis with W-F1 for DBL on MELD, EmoryNLP, and IEMOCAP datasets.

(%)

Utterance-level	Speaker-level	MELD	EmoryNLP	IEMOCAP
✓	✓	**64.03**	**36.99**	**62.53**
✓	×	59.84	35.32	62.18
×	✓	59.12	34.75	61.40

speaker-level encodings are essential to the strong performance of DBL on these datasets. There is a relatively greater impact on ECTC from the removal of utterance-level encoding than that of speaker-level encoding, which shows that the contextual information between adjacent utterances is more important for emotion classification.

5.4 Error analysis

Although there is a better performance for DBL, it still fails to classify certain emotions on these three datasets. We provide two cases for the error results in Table 5, where cases A and B are extracted from the EmoryNLP and MELD, respectively.

We find that DBL is difficult to distinguish two pairs of emotional types (i.e., "disgust" vs. "fear", and "peaceful" vs. "powerful"). There are two possible reasons for the model error. One reason is that these two pairs of emotions are very similar. Another reason is that there is only a small number of data available for these emotional texts. Thus, how to precisely distinguish very similar emotions and how to effectively classify emotions under limited data are two challenging works for the ECTC task.

6 Conclusion

In this paper, we explore the importance of the CNN, Bi-LSTM, and BL for the ECTC task. We demonstrate that the combination of DL (i.e., CNN and Bi-LSTM) and BL could also be beneficially utilized in ECTC task by comparison with the existing methods, which usually capture utterance- and speaker-level contextual dependencies based on the deep learning models. Specifically, we propose a novel method to capture utterance-level and speaker-level context dependencies simultaneously in textual conversations on the basis of BERT, CNN, Bi-LSTM, and BL. With these two different representations, the local contextual information (i.e., utterance-level) and global contextual information (i.e., speaker-level) are well captured at the same time. The extensive experiments demonstrates the effectiveness of our proposed method. In our future work, we plan to study how to use our proposed method to solve the multimodal emotion classification task.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 61876205), the National Key Research and Development Program of China (No. 2020YFB1005804), and the MOE Project at Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies.

References

[1] L. Zhou, J. Gao, D. Li, and H. Shum, The design and implementation of XiaoIce, an empathetic social chatbot, *Comput. Linguist.*, vol. 46, no. 1, pp. 53–93, 2020.

[2] S. Peng, L. Cao, Y. Zhou, Z. Ouyang, A. Yang, X. Li, W. Jia, and S. Yu, A survey on deep learning for textual emotion analysis in social networks, *Dig. Commun. Netw.*, vol. 8, no. 5, pp. 745–762, 2022.

[3] H. Ma, J. Wang, H. Lin, X. Pan, Y. Zhang, and Z. Yang,

Table 5 Cases of error analysis for DBL.

Case	Conversation	True/Prediction
A	S1: That's so sweet. I'm gonna get some coffee.	Powerful/Neutral
	S2: Huh? What'd ya say Joe? I'll be right there.	Peaceful/Neutral
	S3: Pheebs...	Powerful/Peaceful
	S4: I just feel so, uhh.....	Powerful/Neutral
	S3: All right!!	Peaceful/Joyful
	S5: I'm coming already!!	Peaceful/Joyful
	S3: Jeez!	Neutral/Neutral
B	S1: I can't believe you would actually say that. I would much rather be Mr. Peanut than Mr. Salty.	Surprise/Neutral
	S2: No way! Mr. Salty is a sailor, all right, he's got to be, like, the toughest snack there is.	Anger/Anger
	S3: I don't know, you don't wanna mess with corn nuts. They're craaazy.	Neutral/Neutral
	S4: Oh my God. You guys! You gotta come see this! There's some creep out there with a telescope!	Fear/Joy
	S3: I can't believe it! He's looking right at us!	Surprise/Surprise
	S5: Oh, that is so sick.	Disgust/Neutral
	S1: I feel violated. And not in a good way.	Disgust/Anger

A multi-view network for real-time emotion recognition in conversations, *Knowl.-Based Syst.*, vol. 236, p. 107751, 2022.

[4] W. Jiao, H. Yang, I. King, and M. R. Lyu, HiGRU: Hierarchical gated recurrent units for utterance-level emotion recognition, in *Proc. 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* (*Long and Short Papers*), Minneapolis, MN, USA, 2019, pp. 397–406.

[5] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, DialogueRNN: An attentive RNN for emotion detection in conversations, in *Proc. Thirty-Third AAAI Conf. on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conf. and Ninth AAAI Symp. on Educational Advances in Artificial Intelligence*, Honolulu, HI, USA, 2019, pp. 6818–6825.

[6] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria, COSMIC: Commonsense knowledge for emotion identification in conversations, in *Proc. Findings of the Association for Computational Linguistics*: *EMNLP 2020*, Virtual Event, 2020, pp. 2470–2481.

[7] D. Hu, L. Wei, and X. Huai, DialogueCRN: Contextual reasoning networks for emotion recognition in conversations, in *Proc. 59th Annu. Meeting of the Association for Computational Linguistics and the 11th Int. Joint Conf. on Natural Language Processing* (*Volume 1*: *Long Papers*), Virtual Event, 2021, pp. 7042–7052.

[8] D. Li, Y. Li, and S. Wang, Interactive double states emotion cell model for textual dialogue emotion prediction, *Knowl.-Based Syst.*, vol. 189, p. 105084, 2020.

[9] J. Li, D. Ji, F. Li, M. Zhang, and Y. Liu, HiTrans: A transformer-based context- and speaker-sensitive model for emotion detection in conversations, in *Proc. 28th Int. Conf. on Computational Linguistics*, Barcelona, Spain, 2020, pp. 4190–4200.

[10] X. Lu, Y. Zhao, Y. Wu, Y. Tian, H. Chen, and B. Qin, An iterative emotion interaction network for emotion recognition in conversations, in *Proc. 28th Int. Conf. on Computational Linguistics*, Barcelona, Spain, 2020, pp. 4078–4088.

[11] L. Zhu, G. Pergola, L. Gui, D. Zhou, and Y. He, Topic-driven and knowledge-aware transformer for dialogue emotion detection, in *Proc. 59th Annu. Meeting of the Association for Computational Linguistics and the 11th Int. Joint Conf. on Natural Language Processing* (*Volume 1*: *Long Papers*), Virtual Event, 2021, pp. 1571–1582.

[12] W. Shen, J. Chen, X. Quan, and Z. Xie, DialogXL: All-in- one XLNet for multi-party conversation emotion recognition, *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 15, pp. 13789–13797, 2021.

[13] L. Liu, Z. Zhang, H. Zhao, X. Zhou, and X. Zhou, Filling the gap of utterance-aware and speaker-aware representation for multi-turn dialogue, *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 15, pp. 13406–13414, 2021.

[14] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, DialogueGCN: A graph convolutional neural network for emotion recognition in conversation, in *Proc. Conf. on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing*, Hong Kong, China, 2019, pp. 154–164.

[15] T. Ishiwatari, Y. Yasuda, T. Miyazaki, and J. Goto, Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations, in *Proc. Conf. on Empirical Methods in Natural Language Processing*, Virtual Event, 2020, pp. 7360–7370.

[16] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, and G. Zhou, Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations, in *Proc. Twenty-Eighth Int. Joint Conf. on Artificial Intelligence*, Macao, China, 2019, pp. 5415–5421.

[17] P. Zhong, D. Wang, and C. Miao, Knowledge-enriched transformer for emotion detection in textual conversations, in *Proc. Conf. on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing*, Hong Kong, China, 2019, pp. 165–176.

[18] W. Shen, S. Wu, Y. Yang, and X. Quan, Directed acyclic graph network for conversational emotion recognition, in *Proc. 59th Annu. Meeting of the Association for Computational Linguistics and the 11th Int. Joint Conf. on Natural Language Processing* (*Volume 1*: *Long Papers*), Virtual Event, 2021, pp. 1551–1560.

[19] Y. Liang, F. Meng, Y. Zhang, J. Xu, Y. Chen, and J. Zhou, Infusing multi-source knowledge with heterogeneous graph neural network for emotional conversation generation, in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 15, pp. 13343–13352, 2021.

[20] J. Hu, Y. Liu, J. Zhao, and Q. Jin, MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation, in *Proc. 59th Annu. Meeting of the Association for Computational Linguistics and the 11th Int. Joint Conf. on Natural Language Processing* (*Volume 1*: *Long Papers*), Virtual Event, 2021, pp. 5666–5675.

[21] M. Schuster and K. K. Paliwal, Bidirectional recurrent neural networks, *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, 1997.

[22] C. L. Giles, G. M. Kuhn, and R. J. Williams, Dynamic recurrent neural networks: Theory and applications, *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 153–156, 1994.

[23] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in *Proc. Conf. on Empirical Methods in Natural Language Processing*, Doha, Qatar, 2014, pp. 1724–1734.

[24] X. Huang, *Introduction to Psychology*. Beijing, China: People's Education Press, 1991.

[25] P. Ekman, An argument for basic emotions, *Cogn. Emot.*, vol. 6, nos. 3&4, pp. 169–200, 1992.

[26] A. Ben-Zeev, The nature of emotions, *Philos. Stud.*, vol. 52, no. 3, pp. 393–409, 1987.

[27] S. Peng, R. Zeng, H. Liu, G. Chen, R. Wu, A. Yang, and S. Yu, Emotion classification of text based on BERT and broad learning system, in *Proc. Asia-Pacific Web* (*APWeb*) *and Web-Age Information Management* (*WAIM*) *Joint Int.*

Conf. on Web and Big Data, Guangzhou, China, 2021, pp. 382–396.

[28] Z. Ahmad, R. Jindal, A. Ekbal, and P. Bhattachharyya, Borrow from rich cousin: Transfer learning for emotion detection using cross lingual embedding, *Expert Syst. Appl.*, vol. 139, p. 112851, 2020.

[29] F. Barbieri, J. Camacho-Collados, F. Ronzano, L. Espinosa-Anke, M. Ballesteros, V. Basile, V. Patti, and H. Saggion, SemEval 2018 task 2: Multilingual emoji prediction, in *Proc. 12th Int. Workshop on Semantic Evaluation*, New Orleans, LA, USA, 2018, pp. 24–33.

[30] C. L. P. Chen and Z. Liu, Broad learning system: An effective and efficient incremental learning system without the need for deep architecture, *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 10–24, 2018.

[31] C. L. P. Chen, Z. Liu, and S. Feng, Universal approximation capability of broad learning system and its structural variations, *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 4, pp. 1191–1204, 2019.

[32] R. Zeng, H. Liu, S. Peng, L. Cao, A. Yang, C. Zong, and G. Zhou, CNN-based broad learning for cross-domain emotion classification, *Tsinghua Science and Technology*, vol. 28, no. 2, pp. 360–369, 2023.

[33] S. Peng, R. Zeng, L. Cao, A. Yang, J. Niu, C. Zong, and G. Zhou, Multi-source domain adaptation method for textual emotion classification using deep and broad learning, *Knowl.-Based Syst.*, vol. 260, p. 110173, 2023.

[34] G. Chen, S. Peng, R. Zeng, Z. Hu, L. Cao, Y. Zhou, Z. Ouyang, and X. Nie, p-norm broad learning for negative emotion classification in social networks, *Big Data Mining and Analytics*, vol. 5, no. 3, pp. 245–256, 2022.

[35] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, Pre-training with whole word masking for Chinese BERT, *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3504–3514, 2021.

[36] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, MELD: A multimodal multi-party dataset for emotion recognition in conversations, in *Proc. 57th Annu. Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 527–536.

[37] S. M. Zahiri and J. D. Choi, Emotion detection on TV show transcripts with sequence-based convolutional neural networks, in *Proc. Workshops of the Thirty-Second AAAI Conf. on Artificial Intelligence*, New Orleans, LA, USA, 2018, pp. 44–52.

[38] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, IEMOCAP: Interactive emotional dyadic motion capture database, *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

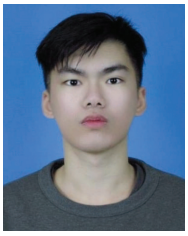
[39] I. Loshchilov and F. Hutter, Fixing weight decay regularization in Adam, in *Proc. of Int. Conf. on Learning Representations*, https://openreview.net/forum?id=rk6qdGgCZ, 2018.

[40] Y. Kim, Convolutional neural networks for sentence classification, in *Proc. Conf. on Empirical Methods in Natural Language Processing*, Doha, Qatar, 2014, pp. 1746–1751.

[41] Q. Li, D. Gkoumas, A. Sordoni, J. Y. Nie, and M. Melucci, Quantum-inspired neural network for conversational emotion recognition, in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 15, pp. 13270–13278, 2021.

[42] D. Zhang, X. Chen, S. Xu, and B. Xu, Knowledge aware emotion recognition in textual conversations via multi-task incremental transformer, in *Proc. 28th Int. Conf. on Computational Linguistics*, Barcelona, Spain, 2020, pp. 4429–4440.

Sancheng Peng received the PhD degree in computer science from Central South University, China in 2010. He is currently a professor at Guangdong University of Foreign Studies, China. He was a research associate at City University of Hong Kong, China from 2008 to 2009. He has authored or co-authored over 70 technical papers in both journals and conferences, such as *IEEE Communications Surveys and Tutorials*, *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Dependable and Secure Computing*, *IEEE Wireless Communications*, *IEEE Network*, *IEEE Internet of Things Journal*, *Journal of Network and Computer Applications*, *Computer Networks*, *Computer and Security*, *Knowledge-Based Systems*, *Information Sciences*, *Future Generation Computer Sciences*, *Journal of Computer and System Sciences*, *Journal of Computer Science and Technology*, IEEE TrustCom, IEEE CBD, ICA3PP, SpaCCS, and EUC. His current research interests include natural language processing, social networks, trusted computing, and mobile computing. He has served as the guest editor of *Future Generation Computer Systems* and as a PC member for various prestige international conferences. He is a senior member of the CCF and a member of ACM.



Rong Zeng received the BEng degree in electronic and information engineering from Hefei University of Technology, China in 2020. He is currently a master student at South China Normal University, China. He has authored or co-authored over 5 technical papers in both journals and conferences, such as *Knowledge-Based Systems*, *Tsinghua Science and Technology*, *Big Data Mining and Analytics*, and APWeb-WAIM. His research interests include natural language processing and machine learning.

Hongzhan Liu received the PhD degree from Graduate School of Chinese Academy of Sciences, China in 2003. Now he is a professor at South China Normal University, and he is a member of Guangdong Provincial Key Laboratory of Nanophotonic Functional Materials and Devices. His research interests include satellite-ground laser communication, precision measurement, and sensing and artificial intelligence technology.

Lihong Cao received the MEng degree in applied linguistics from Hunan Normal University, China in 2010. She is currently a lecturer at Guangdong University of Foreign Studies. She has authored or co-authored over 10 technical papers in both journals and conferences, such as *Journal of Network and Computer Applications*, *Knowledge-Based Systems, Information Sciences*, *Tsinghua Science and Technology*, *Big Data Mining and Analytics*, APWeb-WAIM, and iSCI. Her research interests include applied linguistics, natural language processing, and intelligent computing.

Guojun Wang received the BS degree in geophysics, the MEng degree in computer science, and the PhD degree in computer science from Central South University, China in 1992, 1996, 2002, respectively. He is a Pearl River Scholarship Distinguished Professor of Higher Education in Guangdong Province, a doctoral supervisor at School of Computer Science and Cyber Engineering, Guangzhou University, China, and the director at Institute of Computer Networks, Guangzhou University. He has been listed in Chinese most cited researchers (cyberspace security) by Elsevier in the past nine consecutive years (2014−2022). His h-index is 63 (as of the date March, 2023). His research interests include artificial intelligence, big data, cloud computing, Internet of Things (IoT), blockchain, risk evaluation, trustworthy/dependable computing, network security, privacy preserving, recommendation systems, smart cities, and medical information systems. He has published more than 400 technical papers and books/chapters in the above areas, including top international journals like *ACM TAAS/TOSN/CSUR* and *IEEE TC/TPDS/TDSC*, and top international conferences like CCS/WWW/INFOCOM/CIKM/DSN/ESORICS. He is an associate editor or on editorial board of some international journals, including *IEEE Transactions on Parallel and Distributed Systems* (*TPDS*), *Security and Communication Networks* (*SCN*), *International Journal of Parallel, Emergent and Distributed Systems* (*IJPEDS*), and *International Journal of Computational Science and Engineering* (*IJCSE*).

Jianguo Xie received the BS degree in physics from Fudan University, China in 1986, the MEng degree in computer science and the PhD degree in computer science from Central South University, China, in 1996 and 2002, respectively. He is currently a professor at Guangdong University of Foreign Studies, China. He has published more than 30 technical papers. His research interests include intelligent computing and computer networks.