Modeling both Context- and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations

Dong Zhang\textsuperscript{1}, Liangqing Wu\textsuperscript{1}, Changlong Sun\textsuperscript{2}, Shoushan Li\textsuperscript{1}, Qiaoming Zhu\textsuperscript{1\*} and Guodong Zhou\textsuperscript{1}

\textsuperscript{1}School of Computer Science and Technology, Soochow University, China
\textsuperscript{2}Alibaba Group, China
1\{dzhang17, lqwu\}@stu.suda.edu.cn, \{lishoushan, qmzhu, gdzhou\}@suda.edu.cn, 2changlong.scl@taobao.com

Abstract

Recently, emotion detection in conversations becomes a hot research topic in the Natural Language Processing community. In this paper, we focus on emotion detection in multi-speaker conversations instead of traditional two-speaker conversations in existing studies. Different from non-conversation text, emotion detection in conversation text has one specific challenge in modeling the context-sensitive dependence. Besides, emotion detection in multi-speaker conversations endorses another specific challenge in modeling the speaker-sensitive dependence. To address above two challenges, we propose a conversational graph-based convolutional neural network. On the one hand, our approach represents each utterance and each speaker as a node. On the other hand, the context-sensitive dependence is represented by an undirected edge between two utterances nodes from the same conversation and the speaker-sensitive dependence is represented by an undirected edge between an utterance node and its speaker node. In this way, the entire conversational corpus can be symbolized as a large heterogeneous graph and the emotion detection task can be recast as a classification problem of the utterance nodes in the graph. The experimental results on a multi-modal and multi-speaker conversation corpus demonstrate the great effectiveness of the proposed approach.

1 Introduction

Emotions play an important role in our daily life and emotion detection in text has become a longstanding goal in Natural Language Processing (NLP). In the literature, emotion detection has mainly focused on non-conversation text, such as sentence-level text [Li et al., 2015] and document-level text [Wang et al., 2016]. More recently, emotion detection in conversations has attracted increasing attention in NLP due to its applications in many emerging tasks such as public opinion mining over chat history [Cambria et al., 2017], social media analysis in Facebook, YouTube, Twitter, etc [Majumder et al., 2019], and intelligent systems like smart homes and chatbots [Young et al., 2018].

Different from non-conversation cases, nearby utterances in a conversation are closely semantic-related and thus the emotion categories of nearby utterances are closely related. It is important and also challenging to effectively model the context-sensitive dependence in nearby utterances in a conversation. For example, as shown in Figure 1(a), the emotion of the first utterance (said by Phoebe) is disgust. After the second utterance (said by Monica) for comforting Phoebe, the third utterance (said by Phoebe) becomes joyful. From this example, we can see that the emotion categories could be changed from a negative one to a positive one when meeting a comforting utterance. A few recent studies, such as [Poria et al., 2017] and [Hazarika et al., 2018b], have realized this
challenges and proposed some approaches to handling this. However, all existing studies in emotion detection focus on two-speaker conversations where only two speakers are involved. In this study, we aim to tackle emotion detection in conversations where three or more speakers are involved. In brief, we refer to the conversation in this scenario as multi-speaker conversation. Compared with two-speaker conversations, emotion detection in multi-speaker conversations needs to model the interactive influence of multiple speakers for a better performance. This is mainly because each speaker has a specific personality and characteristic of uttering which has a significant impact on emotion expressing. Therefore, besides the context-sensitive dependence, it is also important and challenging to effectively model another dependence, namely speaker-sensitive dependence, in utterances for emotion detection in multi-speaker conversations. For example, as shown in Figure 1(b), given a neutral utterance (said by Policeman), Rachel gives a neutral reply “Oh well” while Ross gives an angry reply “I can handle the stick”. The difference of the replying emotions is mainly due to the difference of the two speakers. Specifically, according to the role setting, it is not difficult to notice that Rachel’s personality is afraid of trouble while Ross is a bit irritable and impatient.

In this paper, we aim to overcome above two challenges in properly modeling both the context-sensitive and speaker-sensitive dependence in multi-speaker conversations. Specifically, we propose a graph-based convolutional neural network towards conversations, namely ConGCN, to model both context-sensitive and speaker-sensitive dependence for emotion detection. On the one hand, each utterance of the whole conversation corpus is represented as a node in a graph, with an edge between the two utterances in the same conversation to symbolize the contextual dependence. On the other hand, each speaker of the whole corpus is represented as a node, and we bridge the specific-speaker dependence between each utterance and its speaker with an undirected edge. On this basis, the entire conversational corpus can be symbolized as a large heterogeneous graph and the emotion detection task can be recast as a classification problem of the utterance nodes in the graph. Experimentation on a multi-modal and multi-speaker conversation corpus shows that our approach could effectively capture the contextual dependence of the utterances in a conversation and the specific-speaker dependence of its corresponding utterance simultaneously. Furthermore, it also shows that our approach is superior in both uni-modality and multi-modality.

2 Related Work

As an interdisciplinary research field, emotion detection has been drawing more and more attention in natural language processing and multi-modal communication [Picard, 2010] with focus on exploring various types of features for different-level emotion classification, such as document-level [Aim et al., 2005], sentence-level [Li et al., 2015], and short text-level [Felbo et al., 2017]. Compared with above studies in non-conversational text, the studies in conversations are much less and limited to two-speaker conversations. In the following, we give an overview of emotion detection in two-speaker conversations, together with related studies on graph-based neural networks.

2.1 Emotion Detection in Two-speaker Conversations

Various studies have attributed emotional dynamics as an interactive phenomenon, rather than being within-person and one-directional [Richards, 2010], and modeled them by observing transition properties. Recently, several studies employ memory networks to model the self-speaker and inter-speaker dynamics in two-speaker conversations [Hazarika et al., 2018b; Hazarika et al., 2018a]. Although Majumder et al. [2019] claim that their approach can be easily extended to multiple speakers in a conversation, the characteristics of multiple speakers in the whole conversation corpus cannot be effectively highlighted.

Unlike above studies, we distinguish multiple speakers of the whole corpus and model context relationship between utterances of the same conversation with a graph-based neural network. To our best knowledge, this is the first to consider both context-sensitive and speaker-sensitive dependence in multi-speaker conversations.

2.2 Graph-based Neural Networks

Graph-based Neural Networks have received growing attentions recently [Cai et al., 2018]. As a pioneer, Kipf and Welling [2017] present a simplified graph neural network model, called graph convolutional networks (GCN), and achieve the state-of-the-art classification results on a number of benchmark graph datasets. GCN is also exported to several NLP tasks such as sentiment role labeling [Marcheggiani and Titov, 2017], text classification [Yao et al., 2019] and machine translation [Bastings et al., 2017], where GCN is used to encode the syntactic structure of sentences.

However, to our best knowledge, there are no studies in the literature adopting a graph-based approach to bridge the contextual relationship between two utterances in a conversation and the speaker-dependent relationship between an utterance and its speaker.

3 Emotion Detection in Multi-speaker Conversations with Conversational GCN

Figure 2 shows the architecture of our proposed ConGCN for emotion detection in multi-speaker conversations.

3.1 Multi-modal Feature Extraction

In this study, we focus on a multi-modal scenario for emotion detection in multi-speaker conversations where both the linguistic context (text) and acoustic characteristics (audio) are employed to infer the emotion of each utterance. Both networks for textual and acoustic feature extraction are trained at utterance level with the emotion labels separately.

Textual Features

The textual features are generated by following [Hazarika et al., 2018a] where these textual features achieve the state-of-the-art performance in multi-modal emotion detection. First,
the input is obtained with the pre-trained word embeddings extracted from the 300-dimensional GloVe embeddings [Pennington et al., 2014]. Then, a convolutional layer consists of three filters with sizes $f^1; f^2; f^3$ and each filter has $f_{out}$ feature maps. Finally, a max-pooling layer is employed and the pooled features are propagated into two stacked bi-directional LSTMs layers. Formally, textual representation of an utterance is denoted as $u_t \in \mathbb{R}^{d_t}$ of dimension $d_t$.

**Acoustic Features**

The acoustic features are generated by following [Gu et al., 2018] where these acoustic features achieve the state-of-the-art performance in multi-modal emotion detection. First, each utterance-video is formatted as a 16-bit PCM WAV file. Then, the synchronized frames of static, delta, and double delta feature maps are extracted from the 3D-array MFSCs map of audio file to form the low-level acoustic feature vectors $A_j = [s_j, \Delta s_j, \Delta \Delta s_j]$. Finally, two stacked bi-directional LSTMs are employed to compute the bidirectional acoustic contextual states $(\hat{h}_t^j, \tilde{h}_t^j)$. Formally, acoustic representation of an utterance is denoted as $u_a \in \mathbb{R}^{d_a}$ of dimension $d_a$.

### 3.2 Conversational GCN

Graph neural networks have proved to be effective on tasks that have rich relational structure and should well preserve global structure information in a graph with graph embeddings. In this work, we employ GCN to perform emotion detection in multi-speaker conversations, which considers both the context-sensitive and speaker-sensitive dependence.

**Graph Construction**

A large graph is constructed for the entire conversation corpus. Specifically, in this graph, all utterances and all speakers are considered as nodes. The edge between two utterance nodes in a conversation is built by contextual information while the edge between an utterance node and its speaker node is built using speaker-dependent relationship. On this basis, the emotion detection task on utterances becomes a node classification problem. Formally, let $G = (V, E)$ represent a conversational graph, where $V (|V| = n)$ denotes a set of utterance and speaker nodes. $E \subset V \times V$ is a set of relationships containing context-sensitive and speaker-sensitive dependence. Every node is assumed to be connected to itself, i.e., $(v; v) \in E$ for any $v$.

**Node Representations**

(1) For the utterance node, we generate its final representation $u$ by concatenating both textual and acoustic features at multi-modal setting: $u = [u_t; u_a] \in \mathbb{R}^{d_t + d_a}$ of dimension $d_t + d_a$. (2) For the speaker node, we randomly initialize the embedding of different speakers $s \in \mathbb{R}^{d_t + d_a}$ of dimension $d_t + d_a$.

**Edge Weighting**

Two kinds of edges exist in our conversational graph. The first kind of edges represents the context-sensitive dependence when the two utterances are from the same conversation. In our approach, following [Skianis et al., 2018], we adopt the angular similarity to represent the weight of an edge between two utterances, i.e.,

$$A_{ij} = \frac{1 - \arccos (\text{sim}(u_i, u_j))}{\pi}$$

where $u_i$ and $u_j$ denote the vector representations of $i$-th and $j$-th utterances in the same conversation.

The second kind of edges represents the speaker-sensitive dependence between one utterance and its speaker. In our approach, we employ the inverse speaking frequency of the speaker, i.e.,

$$A_{ij} = \frac{c}{\text{Freq}}$$

where either component of $(i, j)$ is an utterance or its speaker. $\text{Freq}$ denotes the utterance number of a speaker in the whole corpus. $c$ is a speaking coefficient to avoid over-fitting, which is a pre-specified hyper-parameter.

**Graph Learning**

Let $X \in \mathbb{R}^{n \times m}$ be a matrix containing all $n$ nodes with their features, where $m$ is the dimension of the feature vectors, and each row $x_v \in \mathbb{R}^m$ is the feature vector for $v$. We introduce
an adjacency matrix \( A \) of \( G \) and its degree matrix \( D \), where
\( D_{ii} = \sum_j A_{ij} \). The diagonal elements of \( A \) are set to be 1 because of self-loops.

For a one-layer GCN, the new \( k \)-dimensional node feature matrix \( X^{(1)} \in \mathbb{R}^{n \times k} \) is computed as:
\[
X^{(1)} = \rho(\tilde{A}XW_0) \tag{3}
\]
where \( \tilde{A} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}} \) is the normalized symmetric adjacency matrix and \( W_0 \in \mathbb{R}^{m \times k} \) is a weight matrix. \( \rho \) is an activation function, e.g., a ReLU \( \rho(x) = \max(0, x) \).

For a multi-layer GCN, the node feature matrix is updated by the following formula, i.e.,
\[
X^{(l+1)} = \rho(\tilde{A}X^{(l)}W_l) \tag{4}
\]
where \( l \) denotes the layer number and \( X^{(0)} = X \). In our approach, we employ a two-layer GCN to perform emotion detection in conversations.

**Model Training**

After transformation by two-layer GCN, the final layer node embeddings are fed into a softmax classifier:
\[
Z = \text{softmax}(\tilde{A}\text{ReLU}(\tilde{A}XW_0)W_1) \tag{5}
\]
where \( \tilde{A} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}} \) is the same as in equation (3), and \( \text{softmax}(x_i) = \frac{1}{Z} \exp(x_i) \) with \( Z = \sum_i \exp(x_i) \). The loss function is defined as the cross-entropy error over all labeled nodes:
\[
\mathcal{L} = -\sum_{d \in y_D} \sum_{f = 1}^{F} Y_{df} \ln Z_{df} \tag{6}
\]
where \( y_D \) is the set of node indices that have labels and \( F \) is the dimension of the output features, which is equal to the number of classes. \( Y \) is the label indicator matrix. The weight parameters \( W_0 \) and \( W_1 \) can be trained via gradient descent. In equation (5), \( R^{(1)}(X) = \tilde{A}XW_0 \) contains the first layer utterance and speaker embeddings and \( R^{(2)}(X) = \tilde{A}\text{ReLU}(\tilde{A}XW_0)W_1 \) contains the second layer utterance and speaker embeddings.

### 4 Experimentation

In this section, we systematically evaluate our graph-based approach towards emotion detection in multi-speaker conversations.

#### 4.1 Experimental Settings

**Dataset**

We use a multi-modal and multi-speaker conversational dataset, namely Multi-modal EmotionLines Dataset (MELD)\(^1\) [Poria et al., 2019]. This dataset extends, improves, and further develops the EmotionLines dataset [Hsu et al., 2018] for the multi-modal scenario, including not only textual conversations, but also their corresponding visual and audio counterparts. MELD contains more than 13,000 utterances, 1432 conversations and 304 different speakers. Note that, different from the multi-modal emotion detection datasets, such as IEMOCAP [Busso et al., 2008] and SEMAINE [McKeown et al., 2012], MELD is a multi-speaker conversation dataset where three or more speakers are involved in a conversation. The train and test data split distributed on all different emotions are shown in Table 1. For a better comparison, we perform 10-fold cross-validation in all our experiments.

#### 4.2 Baselines

For comparison, we implement following state-of-the-art baseline approaches to emotion detection in conversations in order to comprehensively evaluate the performance of our proposed approach.

- **Memory Fusion Network (MFN)\(^2\)** [Zadeh et al., 2018], explicitly accounts for uni-modal and multi-modal interactions in a neural architecture and continuously models them through time. This approach considers neither context-sensitive nor speaker-sensitive dependence. This is the state-of-the-art of multi-modal fusion approach.

- **Bidirectional Contextual LSTM (BC-LSTM)\(^3\)** [Poria et al., 2017], performs context-dependent fusion of multi-sequence data. This approach only considers the contextual information in a conversation. As a strong baseline of the MELD dataset, this approach holds the competitive performance.

- **Conversational Memory Network (CMN)\(^3\)** [Hazara et al., 2018b], models separate contexts for both self-speaker

---

\(^1\) https://github.com/SenticNet/meld

\(^2\) https://github.com/A2Zadeh/CMU-multi-modalSDK

\(^3\) https://github.com/SenticNet/conv-emotion

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Train+Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>1,261</td>
<td>345</td>
</tr>
<tr>
<td>disgust</td>
<td>293</td>
<td>68</td>
</tr>
<tr>
<td>fear</td>
<td>308</td>
<td>50</td>
</tr>
<tr>
<td>joy</td>
<td>1,906</td>
<td>402</td>
</tr>
<tr>
<td>neutral</td>
<td>5,180</td>
<td>1,256</td>
</tr>
<tr>
<td>sadness</td>
<td>795</td>
<td>208</td>
</tr>
<tr>
<td>surprise</td>
<td>1,355</td>
<td>281</td>
</tr>
</tbody>
</table>

Table 1: Statistics of the MELD dataset.
and other-speaker to an utterance for emotion detection in two-speaker conversations, but ignores the global contextual information of the utterance. To our best knowledge, this approach is a pioneer to explore multi-modal emotion detection in two-speaker conversations. In our implementation, for fair comparison, in the context of a test utterance, the different speakers’ utterances are distinguished with different gated recurrent units (GRUs).

Interactive Conversational Memory Network (ICON)\(^3\) [Hazarika et al., 2018a], models separate contexts for both self-speaker and other-speaker to an utterance for emotion detection in two-speaker conversations, and simultaneously incorporates the global contextual information of the test utterance. This approach represents state-of-the-art of uni-modal and multi-modal emotion detection in conversations. In our implementation, for fair comparison, in the context of a test utterance, the different speakers’ utterances are distinguished with different GRUs.

Attentive RNN (DialogueRNN)\(^3\) [Majumder et al., 2019], employs three GRUs to model the speaker, the context from the preceding utterances, and the emotion of the preceding utterances. This approach considers different speakers of a conversation in a two-speaker scenario. This approach represents the state-of-the-art of uni-modal and multi-modal emotion detection in conversations. In our implementation, for fair comparison, we set the speaker representation with one-hot encoding of all speakers in the whole corpus.

A variation of our approach (ConGCN w/o context), removes the context-sensitive dependence modeling.

A variation of our approach (ConGCN w/o speaker), removes the speaker-sensitive dependence modeling.

### 4.3 Experimental Results with Uni-modality

Table 2 illustrates the performance of various approaches with uni-modality performed on MELD. From this table, we can see that:

In the text modality, those approaches including CMN, ICON, and DialogueRNN, which deal with the speaker information, are all superior than BC-LSTM, which highlights the importance of using the speaker information. Among all baseline approaches, DialogueRNN averagely performs best, achieving 55.1% in terms of w-average. In comparison, our approach ConGCN performs better than all baseline approaches and the significance test shows that the average improvements over the baseline approaches are all significant ($p$-value $< 0.05$). Moreover, our approach outperforms two ablated approaches ConGCN w/o context and ConGCN w/o speaker. This indicates the necessity to consider both the context-sensitive and speaker-sensitive dependence for emotion detection in multi-speaker conversations.

In the acoustic modality, among the baseline approaches, it seems that speaker information is not always helpful since DialogueRNN performs no better than BC-LSTM. This may be due to the fact that DialogueRNN cannot well deal with

<table>
<thead>
<tr>
<th>Modality</th>
<th>Models</th>
<th>Anger</th>
<th>Disgust</th>
<th>Fear</th>
<th>Joy</th>
<th>Neutral</th>
<th>Sadness</th>
<th>Surprise</th>
<th>w-average</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC-LSTM</td>
<td>38.9</td>
<td>0.0</td>
<td>0.0</td>
<td>45.8</td>
<td>77.0</td>
<td>0.0</td>
<td>47.3</td>
<td>54.3</td>
<td></td>
</tr>
<tr>
<td>CMN</td>
<td>29.8</td>
<td>0.0</td>
<td>0.0</td>
<td>48.3</td>
<td>75.9</td>
<td>21.0</td>
<td>45.7</td>
<td>54.5</td>
<td></td>
</tr>
<tr>
<td>ICON</td>
<td>30.1</td>
<td>0.0</td>
<td>0.0</td>
<td>48.5</td>
<td>76.2</td>
<td>18.9</td>
<td>46.3</td>
<td>54.6</td>
<td></td>
</tr>
<tr>
<td>DialogueRNN</td>
<td>41.5</td>
<td>0.0</td>
<td>5.4</td>
<td>47.6</td>
<td>73.7</td>
<td>23.4</td>
<td>44.9</td>
<td>55.1</td>
<td></td>
</tr>
<tr>
<td>ConGCN w/o context</td>
<td>33.3</td>
<td>0.0</td>
<td>0.0</td>
<td>52.3</td>
<td>71.3</td>
<td>25.6</td>
<td>50.6</td>
<td>54.3</td>
<td></td>
</tr>
<tr>
<td>ConGCN w/o speaker</td>
<td>38.2</td>
<td>0.0</td>
<td>0.0</td>
<td>51.6</td>
<td>77.3</td>
<td>0.0</td>
<td>47.3</td>
<td>55.3</td>
<td></td>
</tr>
<tr>
<td>ConGCN (Ours)</td>
<td>43.2</td>
<td>8.8</td>
<td>6.5</td>
<td>52.4</td>
<td>74.9</td>
<td>22.6</td>
<td>49.8</td>
<td>57.4</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Modality</th>
<th>Models</th>
<th>Anger</th>
<th>Disgust</th>
<th>Fear</th>
<th>Joy</th>
<th>Neutral</th>
<th>Sadness</th>
<th>Surprise</th>
<th>w-average</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC-LSTM</td>
<td>21.9</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>66.1</td>
<td>0.0</td>
<td>16.0</td>
<td>36.4</td>
<td></td>
</tr>
<tr>
<td>CMN</td>
<td>29.6</td>
<td>0.0</td>
<td>0.0</td>
<td>11.8</td>
<td>67.0</td>
<td>0.0</td>
<td>2.8</td>
<td>38.3</td>
<td></td>
</tr>
<tr>
<td>ICON</td>
<td>31.5</td>
<td>0.0</td>
<td>0.0</td>
<td>8.6</td>
<td>66.9</td>
<td>0.0</td>
<td>0.0</td>
<td>37.7</td>
<td></td>
</tr>
<tr>
<td>DialogueRNN</td>
<td>32.1</td>
<td>5.1</td>
<td>0.0</td>
<td>11.2</td>
<td>53.0</td>
<td>8.3</td>
<td>15.6</td>
<td>34.0</td>
<td></td>
</tr>
<tr>
<td>ConGCN w/o context</td>
<td>33.6</td>
<td>4.9</td>
<td>3.5</td>
<td>12.4</td>
<td>61.7</td>
<td>4.5</td>
<td>23.9</td>
<td>39.2</td>
<td></td>
</tr>
<tr>
<td>ConGCN w/o speaker</td>
<td>31.9</td>
<td>5.3</td>
<td>2.4</td>
<td>4.1</td>
<td>59.1</td>
<td>13.5</td>
<td>26.2</td>
<td>37.3</td>
<td></td>
</tr>
<tr>
<td>ConGCN (Ours)</td>
<td>34.1</td>
<td>3.0</td>
<td>4.7</td>
<td>15.5</td>
<td>64.1</td>
<td>19.3</td>
<td>25.4</td>
<td>42.2</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Performance of different approaches to emotion detection on the MELD dataset with uni-modality.

Table 3: Performance of different approaches to emotion detection on the MELD dataset with multi-modality.
### 4.4 Experimental Results with Multi-modality

Table 3 illustrates the performance of various approaches with multi-modality performed on MELD. From this table, we can see that the performance of MFN is apparently lower than the other approaches. For example, the F1 score of MFN is averagely 2.1% lower than BC-LSTM. This indicates that either context or speaker information is helpful when multi-modality features are employed. Among all approaches, our approach performs best. Furthermore, compared with the two ablated approaches ConGCN w/o context or ConGCN w/o speaker, ConGCN is apparently superior. This highlights the importance of employing both the context-sensitive and speaker-sensitive dependence.

### 5 Analysis and Discussion

#### 5.1 Impact of Multi-layer GCN

Table 4 illustrates the results of our approach when using different numbers of GCN layers. It shows that, whether using uni-modality features or multi-modality features, as the number of GCN layers increases, the performance of our proposed ConGCN first rises and then decreases. It also shows our approach yields the best results when using a two-layer GCN for emotion detection in multi-speaker conversations. This indicates that multi-layer GCN is indeed better than one-layer to some extent, but more than two layers of GCN in our approach seems not a good choice.

#### 5.2 Effectiveness of Speaker Prediction

Our approach treats both the utterances and speakers as nodes for prediction. In this section, we compare the predicted label of a speaker with the distribution of his/her utterances’ emotion categories to check whether they are consistent. Table 5 illustrates the number of one speaker’s utterances in each emotion category with the predicted probability distribution over all emotion categories. From this table, we can see that: (1) Some speakers tend to say some words with particular emotion categories. For instance, the speaker Doug tends to say ‘Joy’ and ‘Neutral’ utterances while Emily tends to say ‘Neutral’ and ‘Anger’ utterances. (2) Our approach performs well in predicting the emotion categories of a speaker. For instance, the speaker Doug is classified to be joy, which is consistent to the number distribution of Doug’s utterances.

### 6 Conclusion

In this paper, we propose a conversational graph-based neural network, namely ConGCN, to model both context-sensitive and speaker-sensitive dependence, for emotion detection in multi-speaker conversations. In our graph-based approach, both utterances and speakers are modeled as nodes while both the context-sensitive dependence and the speaker-sensitive dependence are modeled as edges. Empirical evaluation on a multi-modal and multi-speaker dataset shows that our approach significantly outperforms several state-of-the-art approaches. This indicates the importance of both the context- and speaker-sensitive dependence to emotion detection in multi-speaker conversations and the effectiveness of our graph-based approach in well modeling such dependence.

In our future work, we would like to improve the performance of emotion detection by using unlabeled data since our graph-based neural network approach is easy to add unlabeled data. Moreover, we would like to apply our approach to other applications in conversations, such as sentiment classification and topic classification.

#### Acknowledgments

The research work is partially supported by the Key Project of NSFC No.61836007 and two NSFC grants No.61672366, No.61673290. This work is also supported by the joint research project of Alibaba and Soochow University.
References


