Emotion Inference in Multi-Turn Conversations with Addressee-Aware Module and Ensemble Strategy

Dayu Li[1], Xiaodan Zhu[2], Yang Li[3], Suge Wang[1,*],
Deyu Li[1], Jian Liao[1], Jianxing Zheng[1]
[1]School of Computer and Information Technology, Shanxi University, China
[2]ECE & Ingenuity Labs Research Institute, Queen's University, Canada
[3]School of Finance, Shanxi University of Finance and Economics, China
wsg@sxu.edu.cn, xiaodan.zhu@queensu.ca

Abstract

Emotion inference in multi-turn conversations aims to predict the participant's emotion in the next upcoming turn without knowing the participant's response yet, and is a necessary step for applications such as dialogue planning. However, it is a severe challenge to perceive and reason about the future feelings of participants, due to the lack of utterance information from the future. Moreover, it is crucial for emotion inference to capture the characteristics of emotional propagation in conversations, such as persistence and contagiousness. In this study, we focus on investigating the task of emotion inference in multi-turn conversations by modeling the propagation of emotional states among participants in the conversation history, and propose an addressee-aware module to automatically learn whether the participant keeps the historical emotional state or is affected by others in the next upcoming turn. In addition, we propose an ensemble strategy to further enhance the model performance. Empirical studies on three different benchmark conversation datasets demonstrate the effectiveness of the proposed model over several strong baselines.

1 Introduction

In this paper, we investigate the task of emotion inference in multi-turn conversations, which aims to explore how the conversation history affects the participant's future emotion, and predict the participant's emotion in the next upcoming turn, without knowing the participant's response yet. An example of the task is shown in Figure 1. Emotion inference is a necessary step for applications such as dialogue planning, dialogue generation, and interpretability, among others (Lin et al., 2008; Hasegawa et al., 2013; Gaonkar et al., 2020). For example, in a human-machine conversation scenario, if a chatbot tries to say something to cheer you up when you feel down, then the chatbot must predict the emotional consequence first, and avoid words that may offend you or elicit negative emotion on you.

Previous studies on emotion analysis in conversations have mainly focused on recognizing the emotion of a given utterance, including bc-LSTM (Poria et al., 2017), DialogueRNN (Majumder et al., 2019), DialogueGCN (Zhong et al., 2019), COSMIC (Ghosal et al., 2020), etc., while the emotion inference task is to predict the emotion of the next upcoming utterance, in which the utterance in the next turn is not given. Hasegawa et al. (2013) studied a similar task to the emotion inference, however they only took two turns as context and the multi-party multi-turn scenario was not considered. Bothe et al. (2017) and Wang et al. (2020) estimate the sentiment polarity (*negative* or *positive*) of the next utterance, while our work anticipates the fine-grained emotion, such as *happy*, *sad*, *angry*, *excited*, and *frustrated*, etc.

Although extensive related work has been conducted, emotion inference in multi-turn conversations is still an understudied and challenging task, due to the lack of utterance information from the future and the complexity to capture the characteristics of emotional propagation in multi-turn conversations, such as persistence and contagiousness. To address these issues, an addressee-aware module is designed for both a sequence-based and a graph-based model to capture the propagation of emotional state in conversations and automatically learn whether the participant keeps the historical emotional state or is affected by others.

In addition, we propose an ensemble strategy to further enhance the model performance. Since the exact response of the participant in the next upcoming turn is unknown, there may be multiple potential emotional reactions. We run the models with different random seeds to generate multiple candidate results, and then train a fusion classifier

[*] Corresponding author.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3935–3941
November 7–11, 2021. ©2021 Association for Computational Linguistics

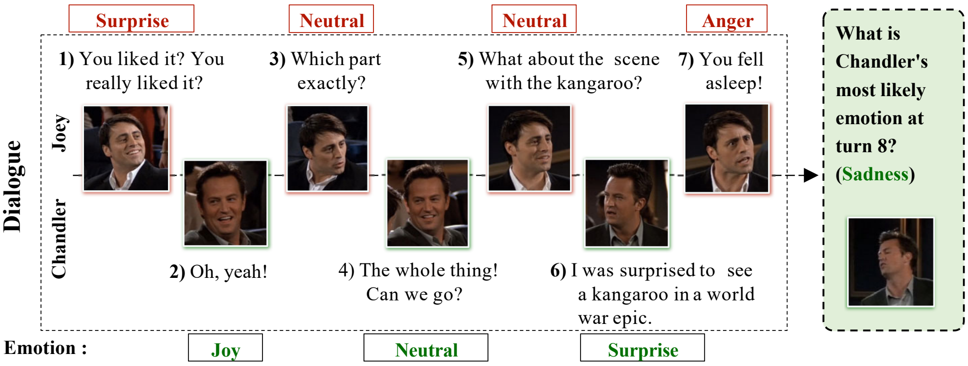
Surprise
Neutral
Neutral
Anger
1) You liked it? You really liked it?
3) Which part exactly?
5) What about the scene with the kangaroo?
7) You fell asleep!
Joey
Dialogue
Chandler
2) Oh, yeah!
4) The whole thing! Can we go?
6) I was surprised to see a kangaroo in a world war epic.
Emotion :
Joy
Neutral
Surprise
What is Chandler's most likely emotion at turn 8? (Sadness)

Figure 1: A dialogue example in the MELD dataset (Poria et al., 2019). The task of emotion inference in multi-turn conversations is to predict Chandler's emotion in the next upcoming turn (8) based on the previous 7 turns of the dialogue.

to automatically select the final result most suitable for the current context and dialogue scene from the candidates.

The main novelty and contribution of this work is that we propose an addressee-aware module for the emotion inference task to model the emotional characteristics and anticipate the emotion trend in multi-turn conversations. Moreover, an ensemble strategy is proposed to further enhance the model performance. The experiments on three benchmark conversational datasets show that our model achieves the new state-of-the-art F1 score.

2 Task Definition

Given a multi-party multi-turn conversation history $\mathcal{D}$ along with the participants information, the emotion inference task aims to infer and anticipate the participant's emotion in the next upcoming turn. Formally, conversation history $\mathcal{D}$ = $[(U_1, p_1^w)$, (U_2, p_2^w), $\cdots$, (U_m, p_m^w), $p_{m+1}^a]$ is a sequence of utterances, where U_m is the utterance at time m and consists of N words, i.e., $U_m = (w_1^m, w_2^m, \cdots, w_N^m)$, p_m^w is the writer/speaker of utterance U_m at timestamp m. And p_{m+1}^a is the addressee/listener in the next upcoming turn $m + 1$.

The task is to infer the addressee p_{m+1}^a's emotion E_{m+1}^a at time $m+1$ based on the utterances of previous m turns along with the participants information, without knowing the utterance information at time $m + 1$ yet:

$$E_{m+1}^a \sim P(E_{m+1}^a | (U_1, p_1^w), \cdots, (U_m, p_m^w), p_{m+1}^a). \quad (1)$$

3 Methodology

Feature Extraction: First, we employ both a GloVe-based CNN encoder (Kim, 2014; Pennington et al., 2014) and a RoBERTa Large encoder (Liu et al., 2019) to encode each utterance in the dataset. Following Ghosal et al. (2019, 2020), we fine-tune each encoder for the context-independent utterance-level emotion label recognition task on the training set, and then extract the emotional representation of each utterance from the last layer of the encoder, and obtained a 100-dimensional and a 1024-dimensional vector for each utterance from the GloVe-based encoder and the RoBERTa-based encoder respectively. The encoding process can be simplified as:

$$u_1, u_2, \cdots, u_m = \text{CNN/RoBERTa}(U_1, U_2, \cdots, U_m), \quad (2)$$

where $(U_1, U_2, \cdots, U_m)$ is the conversation history, U_t is the utterance at time t and $u_t \in \mathbb{R}^H$ is the corresponding utterance representation encoded by CNN/RoBERTa, $H = 100/1024$.

Addressee-Aware Module

To infer and anticipate the participant's emotion, it is important to model the emotion shift in conversations. In this work, we consider two basic characteristics of emotion: persistence and contagiousness (Picard, 1995; Hazarika et al., 2018), as the basis of inferring participant's emotion.

- **Persistence.** Participants may be consistently affected by their own mood and keep the existing emotional state for a period of time. For example, if a participant's car breaks down, then the emotion of this participant may be sad for a long period of time in the conversation.

- **Contagiousness.** The emotional states of participants are interactive, influential and contagious to each other. For example, a sad participant can be encouraged or comforted by others to be happy.

Thus, the addressee p_{m+1}^a either keeps her/his own historical emotional state or is affected by others. In this paper, an addressee-aware module is proposed for both a **sequence-based** and a **graph-based** model to model these two kinds of emotion flow simultaneously.

Sequence-based Model: We first categorize each utterance u_t in the conversation history (u_1, u_2, $\cdots$, u_t, $\cdots$, u_m) into two types according to whether the utterance u_t was spoken by the addressee p^a_{m+1} or others. Two different LSTM units, $LSTM_{store}$ and $LSTM_{affect}$, are then employed to control the different emotional information flow. **Persistence:** If the historical utterance u_t at time t was spoken by the addressee p^a_{m+1}, i.e., $p^w_t = p^a_{m+1}$, then we expect the $LSTM_{store}$ unit to open the *input gate* i_t and **store** the u_t into the internal state c^a_t as much as possible. **Contagiousness:** If the utterance u_t was spoken by someone other than the addressee p^a_{m+1}, i.e., $p^w_t \neq p^a_{m+1}$, then we expect that if the utterance u_t is highly contagious and is likely to **affect** the addressee's emotion, then the $LSTM_{affect}$ unit will open the *forget gate* f_t to forget the addressee's own past state c^a_{t-1} and update the current state c^a_t with the other participant's utterance u_t. Otherwise if the utterance u_t is not contagious, then the $LSTM_{affect}$ unit will close the *input gate* i_t and keep the addressee's own historical state c^a_{t-1} into the internal state at time t. This process can be formalized as:

$$\begin{aligned}(h^a_t, c^a_t) =& \lambda^a_t \cdot LSTM_{store}(u_t, (h^a_{t-1}, c^a_{t-1})) \\ &+ (1-\lambda^a_t)\cdot LSTM_{affect}(u_t, (h^a_{t-1}, c^a_{t-1})), \\ \lambda^a_t =& \begin{cases} 1, if\ p^w_t = p^a_{m+1} \\ 0, if\ p^w_t \neq p^a_{m+1} \end{cases}, \end{aligned} \tag{3}$$

where $t = 1, 2, \cdots, m$. $u_t \in \mathbb{R}^H$ is the utterance feature. (h^a_t, c^a_t) are the hidden state and cell state in the LSTM unit, $h^a_t / c^a_t \in \mathbb{R}^F$, $F = 100$. λ^a_t is the information coefficient at time t. p^w_t is the writer/speaker of the utterance u_t at time t. p^a_{m+1} is the addressee/listener at time $m+1$.

Then the last hidden state h^a_m is then fed to a linear classifier to obtain the emotion distribution es^a_{m+1} of the addressee p^a_{m+1} in the next upcoming turn $m+1$:

$$es^a_{m+1} = \mathrm{softmax}\left(\mathbf{W}^T_c\left(\mathrm{ReLU}(\mathbf{W}^T_s h^a_m)\right) + b\right), \tag{4}$$

where $h^a_m \in \mathbb{R}^F$, $\mathbf{W}_s \in \mathbb{R}^{F\times F}$ is the parameter matrix, $\mathbf{W}_c \in \mathbb{R}^{F\times C}$ is the weight of the linear classifier, C is the total number of emotion categories. $es^a_{m+1} \in \mathbb{R}^C$ is the final emotion distribution of the addressee p^a_{m+1}.

Graph-based Model: A graph-based model is also proposed to model the conversational data for the emotion inference task. We construct a directed graph for each conversation: $\mathbf{G} = (\mathbf{g}, \mathbf{e}, \alpha)$, with nodes $g_t \in \mathbf{g}$, edges $e_{m,t} \in \mathbf{e}$ and edge weights $\alpha_{m,t} \in \alpha$ between nodes g_m and g_t, where $t = 1, 2, \cdots, m$. Each node g_t in the graph is used to represent a dialogue state in the turn t, and we initialize each node g_t with the utterance representation u_t through a linear transform layer (Eq 5). The edges between nodes in the graph are used to represent the complicated dependencies between the dialogue states. In our emotion inference task setting, each node is connected to all the previous nodes (including itself), and then all the historical information is accumulated into the node g_m, based on the edges and edge weights (Eq 6-7), and then the emotion of the next upcoming turn is predicted based on g_m (Eq 8). We formally describe this process below.

For $t = 1, 2, \cdots, m$, we represent each utterance u_t as a node g_t in the directed graph $\mathbf{G}$ through a linear transform layer:

$$g_t = (\mathbf{W}^T_l u_t + b), \tag{5}$$

where $t = 1, 2, \cdots, m$. $u_t \in \mathbb{R}^H$ is the utterance feature. $g_t \in \mathbb{R}^F$ is the node in the graph, $\mathbf{W}_l \in \mathbb{R}^{F\times H}$ is the weight of the linear transform layer, and $F = 100$ is the dimension of nodes.

We then employ two different attention functions, ATT_{store} and ATT_{affect}, to compute the edge weight between the node g_m and node g_t, which is similar to the sequence-based addressee-aware model. If the historical utterance u_t at time t was spoken by the addressee p^a_{m+1}, i.e., $p^w_t = p^a_{m+1}$, then we employ ATT_{store} to compute the edge weight between g_m and g_t, otherwise ATT_{affect}. The edge weight $\alpha^a_{m,t}$ between node g_m and node g_t can be formalized as:

$$\begin{aligned}\alpha^a_{m,t} =& \mathrm{softmax}(\lambda^a_t \cdot ATT_{store}(g_m, g_t) \\ &+ (1-\lambda^a_t)\cdot ATT_{affect}(g_m, g_t)), \\ \lambda^a_t =& \begin{cases} 1, if\ p^w_t = p^a_{m+1} \\ 0, if\ p^w_t \neq p^a_{m+1} \end{cases}, \\ ATT(g_m, g_t) =& \mathbf{W}^T_a\left(\mathrm{ReLU}\left(\mathbf{W}^T_f[g_m||g_t]\right)\right), \end{aligned} \tag{6}$$

where $\alpha^a_{m,t}$ represents the attention weight between the nodes g_m and g_t. $||$ is the concatenation operation. $\mathbf{W}_a \in \mathbb{R}^F$ and $\mathbf{W}_f \in \mathbb{R}^{2F\times F}$ are the parameter matrices.

We then update the nodes. The updated node $g^{'}_m$ is a linear combination of all the connected nodes with the attention coefficient $\alpha^a_{m,t}$:

$$g_m{'} = \sum\nolimits_{g_t \in \mathcal{H}_{g_m}} \alpha^a_{m,t} \cdot g_t, \tag{7}$$

where $g_t \in \mathcal{H}_{g_m}$ represents all the historical nodes g_t connected with g_m. After updating, all the historical information that contributes to the addressee's emotion is accumulated into the node $g_m^{'}$. Then the emotion distribution eg^a_{m+1} of the addressee p^a_{m+1} is obtained:

$$eg^a_{m+1} = \text{softmax}\left(\mathbf{W}_c^T\left(\text{ReLU}(\mathbf{W}_g^T g_m^{'})\right) + b\right), \quad (8)$$

where $g_m^{'} \in \mathbb{R}^F$, $eg^a_{m+1} \in \mathbb{R}^C$. $\mathbf{W}_g \in \mathbb{R}^{F \times F}$ is the parameter matrix, $\mathbf{W}_c \in \mathbb{R}^{F \times C}$ is the weight of the linear classifier.

Ensemble Strategy

We denote the sequence-based and graph-based model as **DialogInfer-S** (Equation 4) and **DialogInfer-G** (Equation 8). And we also integrate the two models through Equation 9 and denote it as **DialogInfer-(S+G)**:

$$ei^a_{m+1} = \text{softmax}\left(\mathbf{W}_c^T\left(\text{ReLU}(\mathbf{W}_i^T(h^a_m + g_m^{'}))\right) + b\right), \quad (9)$$

where $ei^a_{m+1} \in \mathbb{R}^C$, $h^a_m \in \mathbb{R}^F$, $g_m^{'} \in \mathbb{R}^F$. $\mathbf{W}_i \in \mathbb{R}^{F \times F}$ is the parameter matrix.

There may be multiple potential emotional reactions, as the exact response of the participant in the next upcoming turn is unknown. Therefore, different results may be output by the above three different models due to the uncertainty of the emotion inference task. Moreover, even the same model with different parameter initializations may give different results. An ensemble strategy is proposed to address this issue.

We train DialogInfer-S, DialogInfer-G, and DialogInfer-(S+G) 5 times each with different random seeds to generate 15 candidate results, and then train a fusion classifier to automatically select the final result most suitable for the current context and dialogue scene from the candidates:

$$\begin{aligned} ef^a_{m+1} = \text{softmax}(\text{ReLU}(\mathbf{W}_f^T([{es^a_{m+1}}^1||\cdots||{es^a_{m+1}}^5|| \\ {eg^a_{m+1}}^1||\cdots||{eg^a_{m+1}}^5||{ei^a_{m+1}}^1||\cdots||{ei^a_{m+1}}^5]) + b)), \end{aligned} \quad (10)$$

where $ef^a_{m+1} \in \mathbb{R}^C$ is the output emotion probability distribution of the ensemble strategy. es^a_{m+1}, eg^a_{m+1}, ei^a_{m+1} are the output emotion probability distributions of DialogInfer-S, DialogInfer-G, and DialogInfer-(S+G) respectively. The superscripts $1, 2, \cdots, 5$ represent 5 different random initializations. $||$ is the concatenation operation. $\mathbf{W}_f \in \mathbb{R}^{15}$ is the parameter matrix. The ensemble model is denoted as **DialogInfer-Ensemble**.

The final emotion label E^a_{m+1} can be sampled from the output probability distributions of the above 4 types of models:

$$\begin{aligned} E^a_{m+1} \sim P(E^a_{m+1}|(U_1, p^w_1), \cdots, (U_m, p^w_m), p^a_{m+1}) \\ = (es^a_{m+1}/eg^a_{m+1}/ei^a_{m+1}/ef^a_{m+1}). \end{aligned} \quad (11)$$

4 Experiments

4.1 Datasets

We evaluate our model on three multi-turn conversational datasets: **IEMOCAP** (Busso et al., 2008), **MELD** (Poria et al., 2019), and **EmoryNLP** (Zahiri and Choi, 2018). For more dataset details, please refer to their papers.

4.2 Baseline and State-of-the-art Methods

We compare our model with the following related latest neural-network-based methods, and modified them to fit the emotion inference task: **CNN** (Kim, 2014) and **RoBERTa Large** (Liu et al., 2019) model are trained at the utterance level to infer the emotion class of next turn. **sc-LSTM** (Poria et al., 2017) is a simple contextual unidirectional LSTM model. **DialogueRNN** (Majumder et al., 2019) is an RNN-based model, which uses three separate GRU networks to keep track of the individual speaker states. **DialogueGCN** (Ghosal et al., 2019) uses a relational GCN to model the relation between utterances. **COSMIC** (Ghosal et al., 2020) is the state-of-the-art model in emotion recognition in conversations, which incorporates different elements of commonsense. All the baseline methods in our experiments use the same input features (Eq 2) as our proposed methods to ensure a fair comparison (300 dimensional pretrained 840B GloVe vectors (Pennington et al., 2014) for the GloVe-based models, and 1024 dimensional RoBERTa-Large (Liu et al., 2019) for the RoBERTa-based models).

4.3 Experimental Settings

We use the batch size of 16, learning rate of 0.001, and dropout rate of 0.2 to train the inference models. Cross entropy is used as the optimization objective function of the model, and the optimization algorithm is Adam (Kingma and Ba, 2015). The hidden size F is set to 100. All models are trained for 60 epochs and the model checkpoint that achieves the best results on the development set is used for testing. Other hyper-parameters are optimized using the grid search.

	Methods	IEMOCAP	MELD	EmoryNLP
GloVe-based	CNN (2014)	44.09	36.31	20.97
	sc-LSTM (2017)	56.22	36.06	19.75
	DialogueRNN (2019)	58.12	36.93	20.37
	DialogueGCN (2019)	56.48	36.98	19.59
	DialogInfer-S	60.45	38.09	21.08
	DialogInfer-G	59.48	36.62	20.22
	DialogInfer-(S+G)	60.74	38.46	**21.69**
	DialogInfer-Ensemble	**65.31***	**38.48***	20.95
RoBERTa-based	RoBERTa Large (2019)	43.24	36.99	20.46
	sc-LSTM (2017)	58.81	37.71	22.26
	DialogueRNN (2019)	59.53	38.70	21.98
	COSMIC (2020)	61.50	39.49	21.60
	DialogInfer-S	63.63	40.32	23.09
	DialogInfer-G	59.94	38.06	22.81
	DialogInfer-(S+G)	64.70	**40.67**	22.63
	DialogInfer-Ensemble	**66.39***	39.41	**24.09***

Table 1: Performance on three datasets. The weighted macro-F1 is used as the evaluation metric and the best results are in bold. The reported scores are median of five runs, and the asterisk * indicates the statistically significant improvement of our best model over each baseline model (two-tailed paired t-test, $p < 0.05$).

	Methods	IEMOCAP	MELD	EmoryNLP
GloVe-based	**DialogInfer-S**	**60.45**	**38.09**	**21.08**
	w/o addressee-aware	57.04	36.30	18.94
	DialogInfer-G	**59.48**	**36.62**	**20.22**
	w/o addressee-aware	56.42	35.21	20.10
	DialogInfer-(S+G)	**60.74**	**38.46**	**21.69**
	w/o addressee-aware	58.79	37.33	19.70
	DialogInfer-Ensemble	**65.31**	**38.48**	**20.95**
	w/o addressee-aware	58.72	36.75	20.73
RoBERTa-based	**DialogInfer-S**	**63.63**	**40.32**	**23.09**
	w/o addressee-aware	59.23	38.03	22.52
	DialogInfer-G	**59.94**	**38.06**	**22.81**
	w/o addressee-aware	56.43	37.17	21.03
	DialogInfer-(S+G)	**64.70**	**40.67**	**22.63**
	w/o addressee-aware	59.39	38.81	22.16
	DialogInfer-Ensemble	**66.39**	**39.41**	**24.09**
	w/o addressee-aware	61.16	37.85	21.79

Table 2: Ablation analysis on three datasets.

4.4 Results and Discussion

We compare the performance of our proposed models with the baselines on the three benchmark conversational datasets, and the results are listed in Table 1. As we can see from the results, our sequence-based and graph-based addressee-aware models surpass the baseline methods, which shows that our models can capture more essential information for inferring the addressee's emotion than other models. In addition, the ensemble model achieves significant improvements in most cases, which also proves the effectiveness of the ensemble strategy for further enhancing the performance of emotion inference in multi-turn conversations.

The performance of utterance level models, CNN and RoBERTa Large, are worse than other models based on conversation history in most cases, which shows that the inference of the addressee's emotion relies heavily on the evidence from the conversation history. Comparing the GloVe-based models with the RoBERTa-based models, most of the results obtained by the RoBERTa-based models are better than those got by the GloVe-based models. This is because the RoBERTa model has been pre-trained on the large-scale unstructured texts and the features extracted from the RoBERTa model are more informative.

Ablation analysis In Table 2, we also report the results of ablation studies by removing the addressee-aware module, and using the same LSTM-unit or attention function in Equation 3 and Equation 6. The results show that the performance of both GloVe-based and RoBERTa-based models drops after removing the addressee-aware module, which proves the effectiveness of our addressee-aware module, and indicates the addressee-aware module can model the persistence and contagiousness of emotion and learn the emotion shift in multi-turn conversations.

5 Conclusion

In this paper, we investigate the emotion inference in multi-turn conversations, which explores how the conversation history affects the participant's future emotion. To model the characteristics of emotion propagation in conversations: persistence and contagiousness, an addressee-aware module is designed for both a sequence-based and a graph-based model. In addition, an ensemble strategy is proposed to further enhance the model performance. The extensive experimental results on three benchmark datasets show that the proposed models achieve the new state-of-the-art F1 score, and the effectiveness of both the addressee-aware module and the ensemble strategy is demonstrated.

Acknowledgements

The research work was supported by the National Natural Science Foundation of China (61632011, 62076158, 62072294, 61906112), the Key Research and Development Program of Shanxi Province (201803D421024), Natural Science Foundation of Shanxi Province, China (201901D211174), Scientific and Technological Innovation Programs of Higher Education Institutions in Shanxi (2019L0008, 2020L0001).

References

Chandrakant Bothe, Sven Magg, Cornelius Weber, and Stefan Wermter. 2017. Dialogue-based neural learning to estimate the sentiment of a next upcoming utterance. In *Artificial Neural Networks and Machine Learning – ICANN 2017*, pages 477–485, Cham. Springer International Publishing.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Radhika Gaonkar, Heeyoung Kwon, Mohaddeseh Bastan, Niranjan Balasubramanian, and Nathanael Chambers. 2020. Modeling label semantics for predicting emotional reactions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4687–4692, Online. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481, Online. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164, Hong Kong, China. Association for Computational Linguistics.

Takayuki Hasegawa, Nobuhiro Kaji, Naoki Yoshinaga, and Masashi Toyoda. 2013. Predicting and eliciting addressee's emotion in online dialogue. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 964–972, Sofia, Bulgaria. Association for Computational Linguistics.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132, New Orleans, Louisiana. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kevin Hsin-Yih Lin, Changhua Yang, and Hsin-Hsi Chen. 2008. Emotion classification of online news articles from the reader's perspective. In *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 220–226. IEEE.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv*, abs/1907.11692.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive RNN for emotion detection in conversations. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6818–6825. AAAI Press.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Rosalind W Picard. 1995. Affective computing-mit media laboratory perceptual computing section technical report no. 321. *Cambridge, MA*, 2139.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Zhongqing Wang, Xiujun Zhu, Yue Zhang, Shoushan Li, and Guodong Zhou. 2020. Sentiment forecasting in dialog. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2448–2458, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *AAAI Workshops*.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 165–176, Hong Kong, China. Association for Computational Linguistics.