HIERARCHICAL AND MULTI-VIEW DEPENDENCY MODELLING NETWORK FOR CONVERSATIONAL EMOTION RECOGNITION

Yu-Ping Ruan, Shu-Kai Zheng, Taihao Li, Fen Wang, Guanxiong Pei*

Zhejiang Lab, Hangzhou, P.R. China
{ypruan, zhengsk, lith, fenwang, pgx}@zhejianglab.com

ABSTRACT

This paper proposes a new model, called hierarchical and multi-view dependency modelling network (HMVDM), for the task of emotion recognition in conversations (ERC). The modelling of conversational context plays an important role in ERC, especially for the multi-turn and multi-speaker conversations which hold complex dependency between different speakers. In our proposed HMVDM[1], we model the dependency between different speakers at both token-level and utterance-level. Specifically, the HMVDM model has a hierarchical structure with two main modules: 1) token-level dependency modelling module (TDM), which aims to learn the long-range token-level dependency between different utterances in a speaker-aware manner and output the utterance representation; 2) utterance-level dependency modelling module (UDM), which accepts the utterance representation from TDM as inputs and aims to learn the utterance-level dependency from intra-, inter-, and global-speaker(s) view simultaneously. Extensive experiments are conducted on four ERC benchmark datasets with state-of-the-art models employed as baselines for comparison. The empirical results demonstrate the superiority of our proposed HMVDM model and confirm the importance of hierarchical and multi-view context dependency modelling for ERC.

***Index Terms*—** emotion recognition, conversation, dependency, context modelling

1. INTRODUCTION

Emotion recognition is a long-standing research problem in the field of artificial intelligence (AI). With the popularity of conversational AI research, there has been a growing interest in the topic of emotion recognition in conversations (ERC) [1, 2, 3, 4]. The ERC task aims to identify the emotion of each utterance in a conversation, which acts an essential role in building human-like chatbots [5] and has many potential applications in several areas, such as education, health-care [6], and opinion mining in social media [7]. The emotion of a query utterance is likely to be influenced by many factors such as the utterances spoken by the same or different speaker(s) and the global history conversation context. Indeed, previous studies have shown that the modelling of conversational context lies at the heart of the ERC task [8]. Different from the plain text, the structured information in conversational context are more complicated, especially for the multi-turn and multi-speaker conversations which holds complex dependency between different speakers.

Many attempts have been devoted to the modelling of conversational context. One representative class of methods are based on the recurrent neural networks (RNNs) [3, 9, 10, 11], among which both the ICON [10] and CMN [11] utilize gated recurrent unit (GRU) [12] to model the intra- and inter-speaker dependency in a dialogue. Further, DialogRNN [3] and COSMIC [9] tries to derive better context representation by introducing more RNNs to learn more comprehensive dependency between different speakers. However, above RNN-based models tend to be struggling in modelling the long distance dependency. Another representative class of methods are graph-based [2, 13, 1, 14], which can concurrently gather information from surrounding utterances regardless of the distance and so has the potential to overcome the deficiency of the RNN-based models. For example, the DialogueGCN [1] constructs the graph by using the utterances as nodes and the relations between speakers as the edges. Both the graph-based and RNN-based methods have achieved impressive results on the ERC task, however, they all encode each utterance independently and then model the speaker dependency on utterance-level, which ignored the long-range dependency on the token-level between different utterances.

Actually, for a multi-turn and multi-speaker conversation, the context dependency between different speakers are reflected in two aspects, i.e., the token-level and the utterance-level. In this paper, we propose a hierarchical and multi-view dependency modelling network (HMVDM) for the ERC task, which models the speaker dependency at both token-level and utterance-level in a hierarchical manner. Specifically, the HMVDM model has two main modules, i.e., the token-level dependency modelling module (TDM) and the utterance-level dependency modelling module (UDM). The TDM aims to learn long-range token-level dependency between different utterances in a speaker-aware manner, in which each utterance in the dialogue is encoded with viewing previous history utterances and speakers identity information. The UDM accepts the utterance representation from TDM as inputs and aims to learn the utterance-level dependency from intra-, inter-, and global-speaker(s) view simultaneously by utilizing the Transformer [15], which is an efficient and popular graph neural network. We performed extensive experiments on four benchmark datasets. The experimental results demonstrate the superiority of our HMVDM model and confirm the importance of hierarchical and multi-view context dependency modelling for ERC.

2. METHODOLOGY

2.1. Task Definition

In ERC, a conversation is defined as a sequence of utterances $[(u_1, p_1), (u_2, p_2), ..., (u_N, p_N)]$, where N is the dialogue length and p_i denotes the speaker of the utterance u_i. Each utterance u_i

*Corresponding author.

[1]Code available at https://github.com/ypruan/HMVDM.

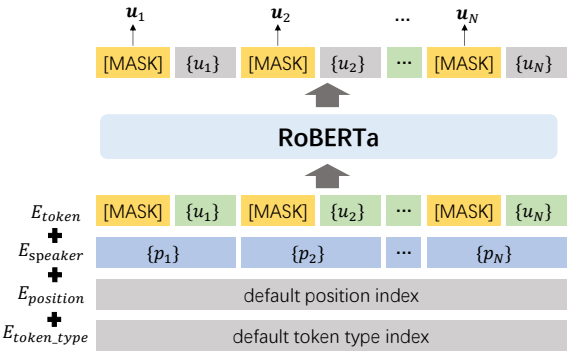

Fig. 1. The structure of the TDM module. The TDM accepts the summation of the token embeddings, speaker embeddings, position embeddings, and token type embeddings as inputs and outputs the sequential representation $[\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_N]$ for the utterances. The encoder is built upon the pretrained RoBERTa-large model.

consists of n_i tokens, namely $u_i = [w_1^i, w_2^i, ..., w_{n_i}^i]$. A discrete value $y_i \in \mathcal{E}$ is used to denote the emotion label of u_i, where $\mathcal{E}$ is the set of emotion labels. The objective of this task is to predict the emotion label for a given query utterance u_t based on the history dialog context $[(u_1, p_1), (u_2, p_2), ..., (u_t, p_t)]$.

2.2. Model Architecture

The proposed HMVDM model has two main modules, i.e., the TDM module and the UDM module, which aims to model the context dependency between different speakers on token-level and utterance-level respectively. Followings are details about the HMVDM model.

2.2.1. The TDM module

As introduced in Section 1, the TDM module encodes each utterance in a dialogue with viewing previous history utterances and speaker identity information, which aims to learn the long-range dependency between different speakers on the token-level. The structure of the TDM module is illustrated in Figure 1.

As shown in Figure 1, the TDM module is built based on the pretrained RoBERTa-large [16], which is also used in many state-of-the-art ERC models [9, 17, 18] and have achieved impressive results. We modified the embedding layers in the original RoBERTa-large by adding a speaker embedding layer, whose parameters are randomly initialized. The input tokens for TDM are the concatenation of the utterance sequence $[u_1, u_2, ..., u_N]$ with the special token "[MASK]" as the delimiter, in which $\{u_i\}$ represents the token sequence of utterance u_i, i.e, $[w_1^i, w_2^i, ..., w_{n_i}^i]$. Except for the token embeddings E_{token}, position embeddings $E_{position}$, and token type embeddings E_{token_type} used in original RoBERTa, the TDM uses another speaker emebddings $E_{speaker}$ to indicate the speaker identity for each utterance token. The $\{p_i\}$ in Figure 1 represents the speaker index sequence whose length is equal to $len(\{u_i\}) + 1$. The final input embeddings for the TDM is the summation of E_{token}, $E_{speaker}$, $E_{position}$, and E_{token_type}.

The TDM module defines a customized attention mask to control the attention flow in RoBERTa. Specifically, the tokens in utterance u_i can only view tokens in previous utterance $[u_1, u_2, ..., u_i]$, which do not include future tokens and so is more compatible with applications in real scenarios. The $i-th$ delimiter token "[MASK]" can only view the tokens of its immediate latter utterance, i.e., $\{u_i\}$. Finally, the encodings for "[MASK]" output from the RoBERTa are used as the utterances representation $[\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_N]$.

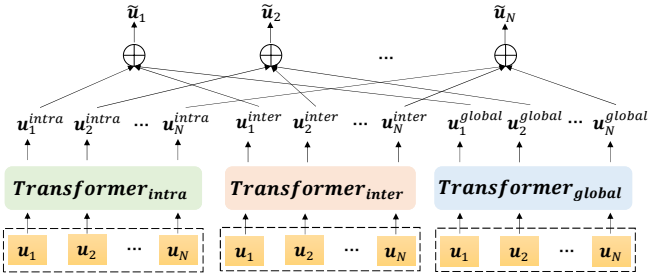

Fig. 2. The structure of the UDM module. The UDM accepts the sequential representation $[\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_N]$ of the utterances as inputs and model the context dependency on utterance level from intra-, inter-, and global-speaker(s) view simultaneously.

	DD	MELD	IEMOCAP	EmoryNLP
#Dial.	13,118	1,432	151	897
train	11,118	1,038	100	713
dev	1,000	114	20	99
test	1,000	280	31	85
#Utt.	102,979	13,708	7,433	12,606
train	87,170	9,989	4,810	9,934
dev	8,069	1,109	1,000	1,344
test	7,740	2,610	1,623	1,328

Table 1. The statistics of four ERC datasets used in experiments.

2.2.2. The UDM module

The UDM module performs high-level context dependency modelling based on the encoding results output from the TDM module, which learns the low-level dependency, i,e., token-level. Specifically, the UDM aims to learn the utterance-level dependency from intra-, inter-, and global-speaker view simultaneously to derive a more comprehensive context representation. As Figure 2 shows, the UDM uses three different Transformer network, denoted as $Transformer_{intra}$, $Transformer_{inter}$, and $Transformer_{global}$ respectively, to make a multi-view dependency modelling.

The UDM accepts the sequential representation $[\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_N]$ of utterances in the dialogue as inputs. The attention flow in $Transformer_{intra}$, $Transformer_{inter}$, and $Transformer_{global}$ are controlled by customized attention mask, which makes the encoding process of utterance u_t can only view previous utterances belong to the same speaker of u_t (intra-view), other speakers (inter-view), and all speakers (global-view) respectively as follows,

$$\mathbf{u}_t^{intra} = Transformer_{intra}(\{\mathbf{u}_i\}, \mathbf{u}_t), \quad i \le t, p_i = p_t, t \in [1, N] \tag{1}$$

$$\mathbf{u}_t^{inter} = Transformer_{inter}(\{\mathbf{u}_i\}, \mathbf{u}_t), \quad i \le t, p_i \ne p_t, t \in [1, N] \tag{2}$$

$$\mathbf{u}_t^{global} = Transformer_{global}(\{\mathbf{u}_i\}, \mathbf{u}_t), \quad i \le t, t \in [1, N] \tag{3}$$

Models	DailyDialog		MELD		IEMOCAP		EmoryNLP	
	Macro F1	Micro F1	W-Avg. F1	Micro F1	W-Avg. F1	Micro F1	W-Avg. F1	Micro F1
DialogueRNN	-	-	57.03	-	62.75	-	-	-
+ RoBERTa	-	57.32	63.61	-	64.76	-	37.44	-
COSMIC	51.05	58.48	65.21	-	65.28	-	38.11	-
DialogueGCN	49.95	53.71	58.37	56.17	60.85	60.63	34.29	33.13
DialogXL	-	54.93	62.41	-	65.94	-	34.73	-
RGAT	-	54.31	60.91	-	65.22	-	34.42	-
KET	-	53.48	58.18	-	59.56	-	34.39	-
DAG-ERC*	-	59.33	63.65	-	**68.03**	-	39.02	-
TODKAT*	52.56	58.47	**68.23**	64.75	61.33	61.11	**43.12**	42.68
w/o KB	50.03	53.44	63.97	61.11	58.96	57.38	33.79	32.62
HMVDM	**53.48**	**68.42**	65.92	**66.31**	67.96	**67.88**	38.46	**42.91**

Table 2. The overall performance of different models on the four datasets. The two models annotated with *, i.e., DAG-ERC and TODKAT, are the most recent state-of-the-art methods on the ERC task.

Then the final representation for utterance u_t is the concatenation of $\mathbf{u}_t^{intra}$, $\mathbf{u}_t^{inter}$, and $\mathbf{u}_t^{global}$,

$$\widetilde{\mathbf{u}}_t = Concat([\mathbf{u}_t^{intra}, \mathbf{u}_t^{inter}, \mathbf{u}_t^{global}]), \tag{4}$$

2.2.3. Classification module

The final classification module in the HMVDM model is a fully connected network with one linear hidden layer. It accepts the output $\widetilde{\mathbf{u}}_t$ from UDM and makes the final emotion prediction,

$$\begin{aligned} probs_t &= softmax(W\widetilde{\mathbf{u}}_t + \mathbf{b}), \\ \hat{y}_t &= \arg\max_k(probs_t[k]). \end{aligned} \tag{5}$$

in which k is number of emotion categories.

3. EXPERIMENTS

3.1. Datasets

Four benchmark ERC datasets are used in our experiments. The statistics of them are shown in Table 1.

- **DailyDialog** [19] collets real conversations from communication of English learners. The utterances in this dataset are annotated with one of the 7 emotion labels: *anger, disgust, fear, happiness, sadness, surprise* and *neutral*. Since it has no speaker information, we consider utterance turns as speaker turns by default.
- **MELD** [20] is a multi-modal ERC dataset constructed from the scripts of TV show *Friends*. There are 7 emotion labels including *neutral, anger, disgust, fear, happiness, sadness*, and *surprise*.
- **IEMOCAP** [21] is a multi-modal dataset. Each dialogue in IEMOCAP comes from the performance based on script by two actors. Its emotion labels are *happy, sad, angry, excited, frustrated* and *neutral*.
- **EmoryNLP** [22] is also built with the scripts from TV show *Friends*, but with a slightly different annotation scheme. The emotion labels include *neutral, sad, mad, scared, powerful, peaceful*, and *joyful*.

3.2. Baselines

We compare the performance of HMVDM model with the following baselines:

- **RNN-based models**: DialogueRNN [3], DialogRNN-RoBERTa [9], and COSMIC which utilize external commonsense knowledge [9].
- **Graph-based models**: DialogueGCN [1], DialogXL [23], RGAT [14], DAG-ERC [17], KET [13], and TODKAT [18], in which both KET and TODKAT integrates external commonsense knowledge.

3.3. Implementation Details

The PyTorch implementation of `RoBERTa-large` released by Huggingface was employed in our experiments[2]. The transformer layers, number of heads for the sub-module $Transformer_{intra}$, $Transformer_{inter}$, and $Transformer_{global}$ in the UDM are all set to 1, and 16 respectively. For the training process, we mostly followed the default settings. Specifically, the learning rate was $2e-05$, and the batch size was set to 8.

We utilize only the textual modality of above datasets in our experiments. Following previous work [18, 17], we choose the Micro F1 and Macro F1 excluding the majority class (neutral) as the evaluation metrics for DailyDialog. For the MELD, IEMOCAP, and EmoryNLP datasets, we use the weighted-average F1 (denoted as W.-Avg. F1 in following pages) and Micro F1 as the evaluation metrics. The reported results of our models are averaged of five runs, which are computed at best validation scores.

3.4. Results

3.4.1. Overall performance

The overall evaluation results of different models on four dataset are present in Table 2. From Table 2, we can find that our HMVDM model outperformed all baseline models except the DAG-ERC and TODKAT[3] on all datasets. Both DAG-ERC and TODKAT

[2] `https://huggingface.co/roberta-large`

[3] These two models are the most recent state-of-the-art models and have just been accepted on the ACL 2021.

	W-Avg. F1	Micro F1
original	67.96	67.88
w/o speaker embedding	64.02	64.43
w/o long-range token dep.	62.00	61.71

Table 3. The results of ablation studies on IEMOCAP dataset for the TDM module in HMVDM without introducing the speaker embeddings and long-range token dependency respectively.

	W-Avg. F1	Micro F1
original	67.96	67.88
w/o intra-speaker dep.	64.81	64.73
w/o inter-speaker dep.	64.78	64.67
w/o global-speaker dep.	62.49	62.58

Table 4. The results of ablation studies on IEMOCAP dataset for the UDM module in HMVDM without considering the utterance-level dependency from intra-, inter-, and global-speaker view respectively.

utilizes the RoBERTa-large for feature extraction, and the TODKAT incorporates external commonsense knowledge.

Compared with the graph-based DAG-ERC, our proposed HMVDM model achieved obvious better performance on both DailyDialog and MELD datasets with over 9% and 2% increase respectively. On the IEMOCAP and EmoryNLP datasets, the HMVDM performed slightly worse than the DAG-ERC, but the performance gap is relatively small, i.e., 0.07% and 0.56%, especially on the IEMOCAP dataset. For the TODKAT model, which incorporated external commonsense knowledge, it performed worse than our HMVDM model on both DailyDialog and IEMOCAP dataset with obvious performance gap. On the MELD and EmoryNLP datasets, our HMVDM achieved better results than TODKAT on the Micro F1 metric but worse results on the weighted-average F1 metric. Further, compared with the TODKAT without using external commonsense knowledge, the HMVDM achieved significant better performance scores on all datasets. The overall comparison results in Table 2 demonstrate our HMVDM model owns better performance.

3.4.2. Ablation studies

Effects of long-term token-level dependency

To inspect the effects of the speaker-aware long-range dependency modelling on token-level in the TDM module, we perform the ablation studies on IEMOCAP dataset by omitting the speaker identity information and long-range token dependency modelling respectively. Specifically, we removed the speaker embedding layer in TDM module to leave out the speaker information. And we modified the customized attention mask in TDM to make each utterance is encoded independently to exclude the long-range dependency between different utterance tokens.

The results are reported in Table 3. We can find that the performance of HMVDM declined obviously when without using speaker embeddings, which indicates the speaker identity information is important for the token-level dependency modelling in TDM. For the long-range dependency between different utterance tokens, we can find that the HMVDM without modelling it had a sharp decline with more than 5%, which demonstrates the importance of modelling the long-range dependency between different utterance tokens for the ERC task.

Effects of multi-view utterance-level dependency

To investigate the effects of multi-view utterance-level dependency modelling in the UDM module, several ablation studies are conducted on IEMOCAP dataset by removing the $Transformer_{intra}$, $Transformer_{inter}$, and $Transformer_{global}$ in UDM respectively. The results are present in Table 4.

From Table 4, it can be found that the HMVDM without intra-, inter-, or global-speaker(s) dependency modelling all had an obvious decline of more than 3%, which indicates that modelling the utterance-level dependency from intra-, inter-, and global-speaker view are all important for the ERC task, especially for modelling the dependency from global-speaker view.

4. CONCLUSION

In this paper, we have proposed a model named hierarchical and multi-view dependency modelling network (HMVDM) for the emotion recognition in conversations (ERC). The HMVDM model has a hierarchical structure with two main modules, i.e., the token-level dependency modelling module (TDM) and the utterance-level dependency modelling module (UDM). The TDM aims to learn long-range token-level dependency between different utterances in a speaker-aware manner, the UDM aims to learn the utterance-level dependency from intra-, inter-, and global-speaker view simultaneously to drive a more comprehensive context representation. The experimental results demonstrate the superiority of our HMVDM model and confirm the importance of hierarchical and multi-view context dependency modelling for the ERC task. Applying the HMVDM model to the multi-modal emotion recognition in conversations will be our future work.

5. REFERENCES

[1] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh, "Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation," in *EMNLP-IJCNLP*, 2019, pp. 154–164.

[2] Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou, "Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations.," in *IJCAI*, 2019, pp. 5415–5421.

[3] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 6818–6825.

[4] Yan Wang, Jiayu Zhang, Jun Ma, Shaojun Wang, and Jing Xiao, "Contextualized emotion recognition in conversation as sequence tagging," in *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2020, pp. 186–195.

[5] Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria, "Mime: Mimicking emotions for empathetic response generation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 8968–8979.

[6] Tim Althoff, Kevin Clark, and Jure Leskovec, "Large-scale analysis of counseling conversations: An application of natural language processing to mental health," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 463–476, 2016.

[7] Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal, "Semeval-2019 task 3: Emocontext contextual emotion detection in text," in *Proceedings of the 13th international workshop on semantic evaluation*, 2019, pp. 39–48.

[8] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100943–100953, 2019.

[9] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria, "Cosmic: Commonsense knowledge for emotion identification in conversations," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2020, pp. 2470–2481.

[10] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann, "Icon: Interactive conversational memory network for multimodal emotion detection," in *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2018, pp. 2594–2604.

[11] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*. NIH Public Access, 2018, vol. 2018, p. 2122.

[12] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[13] Peixiang Zhong, Di Wang, and Chunyan Miao, "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)*, 2019, pp. 165–176.

[14] Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto, "Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7360–7370.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[17] Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan, "Directed acyclic graph network for conversational emotion recognition," *arXiv preprint arXiv:2105.12907*, 2021.

[18] Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He, "Topic-driven and knowledge-aware transformer for dialogue emotion detection," *arXiv preprint arXiv:2106.01071*, 2021.

[19] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," *arXiv preprint arXiv:1710.03957*, 2017.

[20] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.

[21] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[22] Sayyed M Zahiri and Jinho D Choi, "Emotion detection on tv show transcripts with sequence-based convolutional neural networks," in *Workshops at the thirty-second aaai conference on artificial intelligence*, 2018.

[23] Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie, "Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 13789–13797.