TSAM: A Two-Stream Attention Model for Causal Emotion Entailment

Duzhen Zhang[*], **Zhen Yang, Fandong Meng, Xiuyi Chen**[†], **Jie Zhou**
Pattern Recognition Center, WeChat AI, Tencent Inc, Beijing, China
{bladedancer957,hugheren.chan}@gmail.com
{zieenyang,fandongmeng,withtomzhou}@tencent.com

Abstract

Causal Emotion Entailment (CEE) aims to discover the potential causes behind an emotion in a conversational utterance. Previous works formalize CEE as independent utterance pair classification problems, with emotion and speaker information neglected. From a new perspective, this paper considers CEE in a joint framework. We classify multiple utterances synchronously to capture the correlations between utterances in a global view and propose a Two-Stream Attention Model (TSAM) to effectively model the speaker's emotional influences in the conversational history. Specifically, the TSAM comprises three modules: Emotion Attention Network (EAN), Speaker Attention Network (SAN), and interaction module. The EAN and SAN incorporate emotion and speaker information in parallel, and the subsequent interaction module effectively interchanges relevant information between the EAN and SAN via a mutual BiAffine transformation. Extensive experimental results demonstrate that our model achieves new State-Of-The-Art (SOTA) performance and outperforms baselines remarkably.

1 Introduction

With the recent proliferation of open conversational data on social media platforms, such as Twitter and Facebook, Emotion Analysis in Conversations (EAC) has become a popular research topic in the field of Natural Language Processing (NLP). Most of the existing works on EAC mainly focus on Emotion Recognition in Conversations (ERC), i.e., recognizing emotion labels of utterances (e.g., happy, sad, etc.) (Poria et al., 2017, 2019b; Wang et al., 2020; Zhang et al., 2020). However, Poria et al. (2021) point out that these studies lack further reasoning about emotions, such as understanding the stimuli and the cause of the emotion. Since Recognizing Emotion Cause in Conversations (RECCON) holds the potential to improve the interpretability and performance of affect-based models, Poria et al. (2021) put forward a new promising task, named RECCON, which includes two different sub-tasks: Causal Span Extraction (CSE) at word/phrase level and Causal Emotion Entailment (CEE) at utterance level. Due to the simplicity and sufficiency describing emotion causes at the utterance level, we focus on the CEE task in this paper, whose goal is to predict which particular utterances in the conversational history contain the cause of non-neutral emotion in the target utterance.

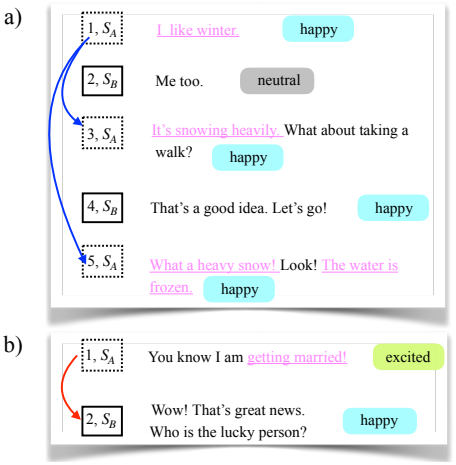

Figure 1: Example conversations sampled from the benchmark dataset (Poria et al., 2021)

Compared to the Emotion Cause Extraction (ECE) in news articles (Gui et al., 2016a; Xia and Ding, 2019), CEE is particularly challenging due to the informal expression style and the intermingling dynamic among interlocutors. Poria et al. (2021) consider CEE as a set of independent utterance pair classification problems and neglect the emotion and speaker information in the conversational history. Thus, they can neither capture the correlations between contextual utterances in a global view nor

[*] This work was done when Duzhen Zhang was interning at Pattern Recognition Center, WeChat AI, Tencent Inc, China.
[†] Corresponding author.

Proceedings of the 29th International Conference on Computational Linguistics, pages 6762–6772
October 12–17, 2022.

model the speaker's emotional influences, namely the intra-speaker and inter-speaker emotional influences.[1] Intra-speaker emotional influences mean that the cause of the emotion is primarily due to the speaker's stable mood induced from previous dialogue turns. As shown in Figure 1 (a), utterance 1 establishes the concept that Speaker A (S_A) likes winter, which triggers a happy mood for future utterances 3 and 5. Inter-speaker emotional influences mean that the emotion of the target speaker is induced from an event mentioned or emotion revealed by other speakers. As Figure 1 (b) shows, S_B's happy emotion may be triggered by the event "getting married" mentioned by S_A, or by the fact that S_A is excited about getting married.

To remedy this defect, we tackle CEE in a joint framework. We classify multiple utterances synchronously to capture the correlations between contextual utterances and propose a TSAM to effectively model the speaker's emotional influences in the conversational history. Specifically, the TSAM contains three modules: EAN, SAN, and interaction module. The EAN provides utterance-to-emotion interactions to incorporate emotion information by performing attention over emotion embeddings. The SAN represents different speaker relations between utterances in a graph, which provides utterance-to-utterance interactions to incorporate speaker information by performing attention over the speaker relation graph. These two modules incorporate emotion and speaker information in parallel. Moreover, inspired by (Li et al., 2021a; Tang et al., 2020), the interaction module interchanges relevant information between the EAN and SAN through a mutual BiAffine transformation. Finally, the entire TSAM can be stacked in multiple layers to refine iteratively and interchange emotion and speaker information.

- For the first time, we tackle CEE in a joint framework to capture the correlations between contextual utterances in a global view.
- We propose a TSAM to model the speaker's emotional influences in the conversational history, which contains EAN, SAN, and interaction module to incorporate and interchange emotion and speaker information.
- Experimental results on the benchmark dataset (Poria et al., 2021) demonstrate the effectiveness of our proposed model, surpassing the SOTA model significantly.

[1]The speaker's emotional influences are predominant types of emotion causes in the dataset as shown in Table 1.

2 Related Work

ECE Early works mainly exploit rule-based methods (Lee et al., 2010a,b; Chen et al., 2010) to identify the potential causes for certain emotion expressions in the text. Gui et al. (2016a) first release a public annotated dataset for ECE, and based on which some feature based (Gui et al., 2016b) and neural based methods (Gui et al., 2017; Li et al., 2018; Ding et al., 2019; Xia et al., 2019; Yan et al., 2021; Li et al., 2021b) appear. To extract emotion and its corresponding cause jointly, Xia and Ding (2019) first put forward the Emotion-Cause Pair Extraction (ECPE) task and tackle it by a two-step method. Subsequently, many improved methods are proposed to tackle ECPE in an end2end manner (Ding et al., 2020a,b; Yuan et al., 2020; Fan et al., 2020; Wei et al., 2020; Cheng et al., 2020; Chen et al., 2020a,b). However, these works mentioned above use news articles as the target corpus for ECE, which largely reduces reasoning complexity. By contrast, CEE is more challenging due to the intermingling dynamic among interlocutors and the informal expression style.

ERC Recently, due to the proliferation of publicly available conversational datasets (Zhou et al., 2018; Chen et al., 2019; Poria et al., 2019a; Chatterjee et al., 2019; Xie et al., 2022), there is a growing number of studies on ERC (Hazarika et al., 2018a,b; Majumder et al., 2019; Zhong et al., 2019; Jiao et al., 2019; Ghosal et al., 2020b; Ishiwatari et al., 2020; Ghosal et al., 2020a; Shen et al., 2021; Zhu et al., 2021; Hu et al., 2021; Guibon et al., 2021; Zhao et al., 2022; Peng et al., 2022). Although substantial progress has been made in ERC, these studies lack further reasoning about emotions, such as understanding the stimuli or the cause of an emotion expressed by a speaker (Poria et al., 2021).

RECCON For further reasoning about emotions, Poria et al. (2021) propose a new task named RECCON, which contains two different sub-tasks: CSE at word/phrase level and CEE at utterance level. Poria et al. (2021) formalize CEE as a set of independent utterance pair classification problems, neglecting the emotion and speaker information in the conversational history. Specifically, they pair a target utterance with each utterance in its conversational history and determine whether the utterance

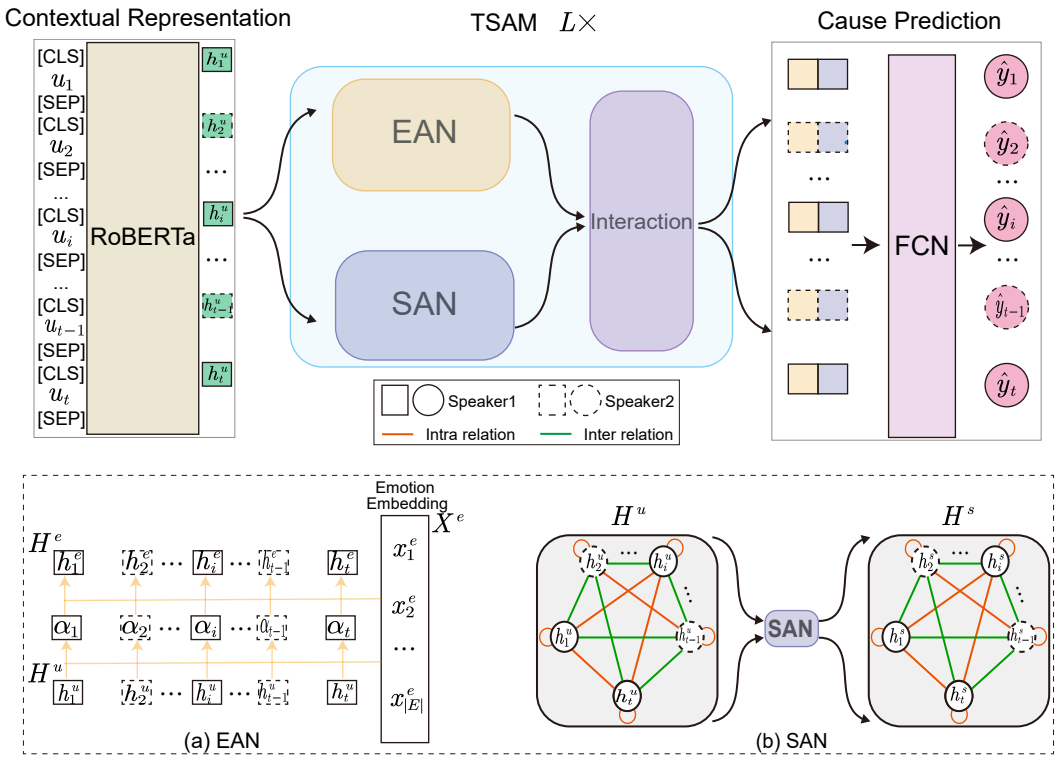

Figure 2: The top is the proposed model's entire architecture, and the bottom is the detailed architecture of model components: (a) EAN, (b) SAN. First, we obtain the contextual representation for each utterance with RoBERTa. Then, the TSAM is utilized to model the speaker's emotional influences in the conversational history. Finally, the cause prediction module is used to output the predictions.

contains the cause of emotion in the target utterance. Thus, they cannot capture the correlations between contextual utterances in a global view and fail to model the speaker's emotional influences in the conversational history. From a new perspective, we tackle CEE in a joint framework. We encode and classify multiple utterances synchronously to capture the correlations in a global view and propose a TSAM to model the speaker's emotional influences effectively.

3 Task Definition

We first define the task of CEE formally. For a target utterance u_t, i.e., the t^{th} utterance in the conversation, the goal of CEE is to predict which particular utterances in the conversational history $L(u_t) = (u_1, u_2, ..., u_t)$ are responsible for the non-neutral emotion in the target utterance. u_i is set as a positive example if it contains the cause of non-neutral emotion in the target utterance and a negative example otherwise, where $i = 1, ..., t$. The independent utterance pair classification framework (Poria et al., 2021) performs t independent classifications, each of which takes (u_t, u_i) as input. Therefore, it fails to capture the correlations between contextual utterances in a global view. On the contrary, the proposed joint classification framework only performs one joint classification with $L(u_t)$ as input, which makes it possible to capture the correlations between contextual utterances.

4 Method

The proposed model consists of three components: the contextual utterance representation, the TSAM, and the cause prediction modules. The whole architecture of our model is illustrated in Figure 2.

4.1 Contextual Utterance Representation

The pre-trained RoBERTa is utilized as the utterance encoder, and we extract the contextual utterance representations by feeding the whole of the conversational history $L(u_t)$ into the RoBERTa (Liu et al., 2019). Specifically, each utterance in $L(u_t)$ is expanded to start with the token "[CLS]" and end with the token "[SEP]". The input representation for each token is the sum of its corresponding token and position embeddings. The contextual representation $\boldsymbol{h}_i^u \in \mathbb{R}^{d_h}$ for utterance u_i is the output

of the corresponding "[CLS]" token, where d_h denotes the dimension of the utterance representation. The contextual representation for all utterances is denoted as $\boldsymbol{H}^u \in \mathbb{R}^{t \times d_h}$. The RoBERTa we utilized is fine-tuned with the training process.

4.2 TSAM

The TSAM models the speaker's emotional influences with three modules: EAN, SAN, and Interaction. We first illustrate the calculation process of each module in one-layer TSAM and then generalize it to multiple successive layers.

4.2.1 EAN

The EAN provides utterance-to-emotion interactions to explicitly incorporate emotion information by performing attention over emotion embeddings.

Emotion Representation Given the set of candidate emotion labels $\mathcal{E} = \{e_1, ..., e_{|\mathcal{E}|}\}$, each emotion label e_k is represented using an embedding vector (Cui and Zhang, 2019):

$$\boldsymbol{x}_k^e = \text{Embed}(e_k) \in \mathbb{R}^{d_h} \tag{1}$$

where $k = 1, ..., |\mathcal{E}|$, d_h denotes the dimension of the emotion embedding. Embed represents an emotion embedding lookup table, which is initialized by contextual embeddings from RoBERTa and tuned during model training. The embedding for the set of the whole emotion labels is denoted as $\boldsymbol{X}^e \in \mathbb{R}^{|\mathcal{E}| \times d_h}$.

EAN Inference With the emotion labels represented as embeddings, we extract the emotion information $\boldsymbol{H}^e \in \mathbb{R}^{t \times d_h}$ by performing dot-product attention over contextual utterance representations and emotion embeddings, which is calculated as:

$$\boldsymbol{H}^e = \textit{attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \boldsymbol{\alpha}\boldsymbol{V} \tag{2}$$

$$\boldsymbol{\alpha} = \textit{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_h}}) \tag{3}$$

where $\boldsymbol{Q} = \boldsymbol{H}^u, \boldsymbol{K} = \boldsymbol{V} = \boldsymbol{X}^e$, $\boldsymbol{\alpha} \in \mathbb{R}^{t \times |\mathcal{E}|}$ is an attention matrix consisting of potential emotion distributions for all utterances. Compared to the standard attention mechanism above, it may be beneficial to use multi-head attention (Vaswani et al., 2017) to capture multiple emotion distributions in parallel and obtain richer emotion information:

$$\boldsymbol{H}^e = \textit{concat}(head_1, ..., head_m) \tag{4}$$

$$head_j = \textit{attention}(\boldsymbol{Q}\boldsymbol{W}_j^Q, \boldsymbol{K}\boldsymbol{W}_j^K, \boldsymbol{V}\boldsymbol{W}_j^V) \tag{5}$$

where $\boldsymbol{W}_j^Q, \boldsymbol{W}_j^K, \boldsymbol{W}_j^V \in \mathbb{R}^{d_h \times \frac{d_h}{m}}$ are learnable parameters and m is the number of parallel heads.

Since the emotion labels of the utterances in the conversational history are known, we can also simply use the embedding of emotion label corresponding to the utterance as the extracted emotion information:

$$\boldsymbol{H}^e = \tilde{\boldsymbol{X}}^e \tag{6}$$

where $\tilde{\boldsymbol{X}}^e \in \mathbb{R}^{t \times d_h}$ is the embedding of emotion labels corresponding to all utterances in the history. We refer to the method as Direct Application Emotional Embedding (DAEE). Compared with DAEE, the potential advantages of the EAN are as follows: (1) The EAN can provide utterance-to-emotion interactions and capture multiple potential emotion distributions through multi-head attention to obtain more comprehensive and richer emotion information; (2) The soft emotion distributions can model the mutual impact among different emotions for further enhancement of emotion embeddings, while each emotion embedding is relatively independent of each other in DAEE; (3) The EAN can avoid emotion annotation errors to a certain extent. We apply EAN in our model to incorporate emotion information by default and compare the EAN and DAEE in the part of experiments.

4.2.2 SAN

The SAN provides utterance-to-utterance interactions to incorporate speaker information by performing attention over the speaker relation graph.

Graphical Structure We define a conversational history with t utterances as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$, with nodes (utterances) $v_i \in \mathcal{V}$ and labeled edges (relations) $(v_i, r, v_j) \in \mathcal{E}$, where $r \in \mathcal{R}$ is a relation type. We also add a self-loop edge to every node, as the cause may be present within the target utterance itself. The representation of node v_i is initialized with the contextual utterance representation $\boldsymbol{h}_i^u \in \mathbb{R}^{d_h}$, i.e., the i^{th} embedding in $\boldsymbol{H}^u$. There are two relation types of edges: (1) **Intra** relation type: how the utterance influences other utterances (including itself) expressed by the same speaker; (2) **Inter** relation type: how the utterance influences ones expressed by other speakers.

SAN Inference The representation of a node $\boldsymbol{h}_i$ is updated by aggregating representations of its neighborhood $\mathcal{N}^r(i)$ under the relation type r. The graph attention mechanism (Veličković et al., 2018)

is used to attend to the neighborhood's representations. The output of a node $\boldsymbol{h}_i^s \in \mathbb{R}^{d_h}$ is calculated as the sum of the hidden features $\boldsymbol{h}_{ir} \in \mathbb{R}^{d_h}$ under relation r. The propagation is defined as follows:

$$\alpha_{ijr} = \textit{softmax}_i(\textit{LRL}(\boldsymbol{a}_r^T[\boldsymbol{W}_r\boldsymbol{h}_i^u; \boldsymbol{W}_r\boldsymbol{h}_j^u])) \quad (7)$$

$$\boldsymbol{h}_{ir} = \sum_{j \in \mathcal{N}^r(i)} \alpha_{ijr}\boldsymbol{W}_r\boldsymbol{h}_j^u \quad (8)$$

$$\boldsymbol{h}_i^s = \sum_{r \in \mathcal{R}} \boldsymbol{h}_{ir} \quad (9)$$

where α_{ijr} denotes the edge weight from utterance u_i to its neighborhood u_j under relation type r, $\boldsymbol{W}_r \in \mathbb{R}^{d_h \times d_h}$ and $\boldsymbol{a}_r \in \mathbb{R}^{d_h}$ denote a learnable weight matrix and a vector for each relation type r respectively. *LRL* denotes *LeakyReLU* activation function. The updated representation of all nodes is denoted as $\boldsymbol{H}^s \in \mathbb{R}^{t \times d_h}$.

4.2.3 Interaction Module

To effectively interchange relevant information between the EAN and SAN, we apply a mutual Bi-Affine transformation as a bridge. The calculation process is formulated as follows:

$$\boldsymbol{A}_1 = \textit{softmax}(\boldsymbol{H}^e\boldsymbol{W}_1(\boldsymbol{H}^s)^T) \quad (10)$$

$$\boldsymbol{A}_2 = \textit{softmax}(\boldsymbol{H}^s\boldsymbol{W}_2(\boldsymbol{H}^e)^T) \quad (11)$$

$$\boldsymbol{H}^{e'} = \boldsymbol{A}_1\boldsymbol{H}^s \quad (12)$$

$$\boldsymbol{H}^{s'} = \boldsymbol{A}_2\boldsymbol{H}^e \quad (13)$$

where $\boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathbb{R}^{d_h \times d_h}$ are trainable parameters and $\boldsymbol{A}_1, \boldsymbol{A}2 \in \mathbb{R}^{t \times t}$ are temporary alignment matrices projecting from $\boldsymbol{H}^s$ to $\boldsymbol{H}^e$ and $\boldsymbol{H}^e$ to $\boldsymbol{H}^s$, respectively. Here, $\boldsymbol{H}^{e'} \in \mathbb{R}^{t \times d_h}$ can be viewed as a projection from $\boldsymbol{H}^s$ to $\boldsymbol{H}^e$, and $\boldsymbol{H}^{s'} \in \mathbb{R}^{t \times d_h}$ follows the same principle.

4.2.4 The Whole Process

We generalize the TSAM to multiple successive layers to iteratively refine and interchange emotion and speaker information. The detailed procedures are as follows:

$$\boldsymbol{H}_l^e = \textbf{EAN}(\boldsymbol{E}_l, \boldsymbol{X}^e) \quad (14)$$

$$\boldsymbol{H}_l^s = \textbf{SAN}(\boldsymbol{S}_l) \quad (15)$$

$$\boldsymbol{H}_l^{e'}, \boldsymbol{H}_l^{s'} = \textbf{Interaction}(\boldsymbol{H}_l^e, \boldsymbol{H}_l^s) \quad (16)$$

$$\boldsymbol{E}_{l+1}, \boldsymbol{S}_{l+1} = \boldsymbol{H}_l^{e'}, \boldsymbol{H}_l^{s'} \quad (17)$$

where $\boldsymbol{E}_0 = \boldsymbol{S}_0 = \boldsymbol{H}^u$. The TSAM can be stacked in L layers and $l \in [0, L-1]$.

Statistics			RECCON-DD
Data Distributions	Train	Positive	7269
		Negative	20646
	Dev	Positive	347
		Negative	838
	Test	Positive	1894
		Negative	5330
Cause Type Distributions	No Context		43%
	Inter		32%
	Intra		9%
	Hybrid		11%
	Unmentioned		5%

Table 1: Statistics of the RECCON-DD dataset. *No Context*: The cause is present within the target utterance itself; *Inter*: Inter-speaker emotional influences; *Intra*: Intra-speaker emotion influences (Self-Contagion); *Hybrid*: Inter and Intra can jointly cause the emotion of an utterance; *Unmentioned*: Some instances have no explicit emotion causes in the conversational history.

4.3 Cause Prediction

We obtain the final utterance representation for u_i by concatenating the output $(\boldsymbol{E}_L, \boldsymbol{S}_L)$ of the L-layer TSAM. Finally, the concatenated vector is classified using a Fully-Connected Network (FCN):

$$\boldsymbol{l}_i = \textit{ReLU}(\boldsymbol{W}_1[\boldsymbol{e}_i^L; \boldsymbol{s}_i^L] + \boldsymbol{b}_1) \quad (18)$$

$$\hat{y}_i = \textit{sigmoid}(\boldsymbol{W}_2\boldsymbol{l}_i + b_2) \quad (19)$$

where $\hat{y}_i$ is the probability for utterance u_i containing the cause of emotion in the target utterance, $\boldsymbol{e}_i^L, \boldsymbol{s}_i^L \in \mathbb{R}^{d_h}$ denote the i^{th} embedding in $\boldsymbol{E}_L$ and $\boldsymbol{S}_L$, respectively, $\boldsymbol{W}_1 \in \mathbb{R}^{d_h \times 2d_h}$, $\boldsymbol{W}_2 \in \mathbb{R}^{1 \times d_h}$, $\boldsymbol{b}_1 \in \mathbb{R}^{d_h}$ and b_2 are learnable parameters of FCN.

5 Experimental Settings

5.1 Dataset and Evaluation Metrics

We evaluate the proposed model on a benchmark dataset for RECCON, named RECCON-DD (Poria et al., 2021), which is constructed based on DailyDialog dataset (Li et al., 2017).[2] Some statistics about RECCON-DD are reported in Table 1. Following (Poria et al., 2021), the macro-averaged F1 score is utilized as the evaluation metric in this paper. We also report the F1 score for both positive and negative samples, denoted as Pos. F1 and Neg. F1 respectively.

[2]DailyDialog is a natural human communication dataset which is usually used in ERC task. It contains utterance-level emotion labels and covers various topics related to daily lives.

#	Model	W/O CH			W/ CH		
		Pos. F1	Neg. F1	macro F1	Pos.F1	Neg.F1	macro F1
0	INDEP$_{base}$	56.64	85.13	70.88	64.28	88.74	76.51
1	INDEP$_{large}$	50.48	87.35	68.91	66.23	87.89	77.06
2	JOINT$_{base}$	-	-	-	66.61	89.11	77.86
3	JOINT$_{large}$	-	-	-	68.30	89.16	78.73
4	Ours$_{base}$	-	-	-	68.59	89.75	79.17
5	Ours$_{large}$	-	-	-	**70.00**†	**90.48**†	**80.24**†

Table 2: Performance of our model and baselines on the test set of RECCON-DD. Bold font denotes the best performance. "Ours" denotes the proposed model without removing any module ("Ours" = "JOINT" + TSAM). "†" denotes that Ours$_{large}$ is statistically significant (Koehn, 2004) better than INDEP$_{large}$ W/ CH (p-value < 0.05).

5.2 Baselines

For a comprehensive performance evaluation, we compare our model with the following baselines:

(1) INDEP$_{base}$ (Poria et al., 2021) It tackles CEE in an independent classification framework (INDEP) and uses the RoBERTa-Base model (Liu et al., 2019) as the utterance encoder. The input is formated as "[CLS]u_t[SEP]u_i[SEP]" and the classification is performed from the final representation of the token "[CLS]".

(2) INDEP$_{large}$ (Poria et al., 2021) Compared to (1), it uses the RoBERTa-Large model as utterance encoder;

(3) JOINT$_{base}$ It's one of the variants of our model, where the TSAM is removed. It tackles RECCON in a joint classification framework (JOINT) and uses the RoBERTa-Base model as the utterance encoder. Moreover, its input format is "[CLS]u_1[SEP][CLS]$u2$[SEP],...,[CLS]u_t[SEP]" and the classifications are performed synchronously from the corresponding contextual utterance representations of the [CLS] tokens;

(4) JOINT$_{large}$ Compared to (3), it uses RoBERTa-Large model as the utterance encoder.

For **INDEP** baselines, there are two different settings: With Conversational History (W/ CH) and Without Conversational History (W/O CH). W/ CH means considering the conversational history. When performing utterance pair classification, the conversational history $L(u_t)$ is concatenated after the input to incorporate contextual information, while W/O CH means ignoring the history.

5.3 Implementation Details

Our model's base and large versions use pre-trained RoBERTa-Base and RoBERTa-Large models as the utterance encoders, respectively.[3] The binary cross-entropy loss along with L2-regularization is used during training, where the coefficient of L2 term is 0.01 in the RoBERTa structure and $1e$-5 in other structures. We set the dropout rate to 0.1. The learning rate and the batch size are set as $1e$-5 and 2, respectively. Our model is trained with the Adam optimizer (Kingma and Ba, 2015). We set the dimensions of the contextual utterance representation d_h as 768/1024 in the Base/Large version of the proposed model. We use 4-head attention in EAN, and the number of TSAM layers L is set to 3. We train the model for 40 epochs in total and use the early stopping strategy based on the performance on the development set. Then, the model with the highest macro-averaged F1 score is used to evaluate the test set. Other hyper-parameters are selected according to the performance of the development set. All of the experiments are conducted on an NVIDIA V100 GPU with 32GB of memory.

6 Results and Discussions

6.1 Main Results

Experimental results are reported in Table 2. We directly cite the results for the baselines reported in (Poria et al., 2021). For the performance of each model we implemented, we report the average score of 5 runs. From Table 2, we can find that the proposed model (#5) outperforms all of the baselines and surpasses the best model (#1, W/ CH) in (Poria et al., 2021) with more than 3 points of macro F1 score.

Further comparisons show that models with the large pre-trained utterance encoder are more likely

[3]Our RoBERTa models are adapted from this implementation: https://github.com/huggingface/transformers

to achieve better performance (about 1 point of macro F1 score) than the corresponding models with the base one, except for the models under W/O CH setting in the Table 2. By comparing two different settings W/O CH and W/ CH in Table 2, we can find that the conversational history plays a significant role for INDEP$_{base}$ and INDEP$_{large}$ models. This is mainly because that the conversational history is able to provide the contextual information for prediction. Due to the simultaneous classification of multiple utterances in the conversational history under the joint framework, JOINT$_{base}$ and JOINT$_{large}$ models can naturally incorporate the contextual information. The JOINT$_{base}$ and JOINT$_{large}$ models significantly outperform the INDEP$_{base}$ W/ CH and INDEP$_{large}$ W/ CH models by about 1.5 points of macro F1 scores respectively (comparing #0 with #2, and #1 with #3 in Table 2). There may be two main factors: 1) Simply concatenating the conversational history after the utterance pair to be classified in INDEP W/ CH models may destroy the structure of the conversation; 2) Compared to INDEP W/ CH models, classification of multiple utterances synchronously in JOINT models will have more sufficient supervision signals and can more effectively model the correlations between contextual utterances in a global view, i.e., utterances with similar semantics are supposed to have similar chances being the emotion cause. The comparison between #2 and #4 (or #3 and #5) in Table 2 shows the effectiveness of the proposed TSAM. The model with TSAM (#5) achieves an improvement up to 1.51 points of macro F1 score than the model without TSAM (#3).

6.2 Ablation Study

In this subsection, we conduct ablation studies to analyze the effects of different components based on Ours$_{large}$ mentioned in Table 2.

Effect of Emotion Information We compare three different ways for incorporating the emotion information: no emotion information incorporated, incorporating emotion information with direct application emotional embedding, and incorporating emotion information with EAN. The results are shown in Table 3. We can find that the performance of the proposed model degrades if the emotion information is not incorporated (comparing row 1 with 3 in Table 3). This result shows that the emotion information in the conversational history plays a significant role in the task of CEE. By comparing rows 2 with 3 in Table 3, the result shows that EAN achieves better performance than DAEE since EAN can extract richer emotion information and model the mutual impact among different emotions.

Emotion Information	Pos. F1	Neg. F1	macro F1
No	68.40	89.80	79.10
DAEE	68.90	90.03	79.47
EAN	**70.00**	**90.48**	**80.24**

Table 3: Comparison of different ways of incorporating emotion information. *No*: no emotion information incorporated; *DAEE*: incorporating the emotion information with direct application emotional embedding.

Effect of Speaker Information To evaluate the effects of speaker information, we remove the speaker relations in SAN, resulting in a single edge relation throughout the graph. As Table 4 shows, the performance of our model decreases dramatically if not considering the speaker information. This result presents that modeling the speaker information in the conversational history is very important for the final performance.

Speaker Information	Pos. F1	Neg. F1	macro F1
Not Consider	67.99	89.42	78.71
Consider	**70.00**	**90.48**	**80.24**

Table 4: Results on experiments whether considering speaker information or not in SAN.

Effect of Interaction Module We remove the interaction module in each layer so that the EAN and SAN can't interact. As Table 5 shows, the performance of our model decreases dramatically when the interaction module is removed. This result shows that the effective interchange of relevant information between EAN and SAN is conducive to the final performance.

	Pos. F1	Neg. F1	macro F1
W/O Interaction	68.18	88.93	78.56
W/ Interaction	**70.00**	**90.48**	**80.24**

Table 5: Results on experiments whether removing interaction module or not in TSAM.

Ability on Modeling Emotional Influences To evaluate the proposed model's ability to model the speaker's emotional influences, we collect the positive examples from the test set where the causes

Models	Intra	Inter
W/O TSAM	62.06	72.67
W/ TSAM	**63.82**	**74.81**

Table 6: Accuracy on the collected samples. *Intra*: Intra-speaker emotional influences; *Inter*: Inter-speaker emotional influences.

are induced from the inter-speaker or intra-speaker emotional influences. And we test the prediction accuracy on the collected samples for the proposed Ours$_{large}$ with and without TSAM. As shown in Table 6, W/ TSAM outperforms W/O TSAM by around 2 points on both cause types, which further verifies that the TSAM can effectively model the emotional influences between speakers.

6.3 Impact of the TSAM Layer Number

Since TSAM for modeling speakers' emotional influences is the critical component of our model, we chose the number of TSAM layers L (ranging from 1 to 5) on the development set of RECCON-DD. As shown in Figure 3, our model with three TSAM layers achieves the best performance. On the one hand, emotion and speaker information may not be refined and interchanged well when the number of layers is small. On the other hand, if the number of layers is excessive, the performance will decrease, possibly due to information redundancy.

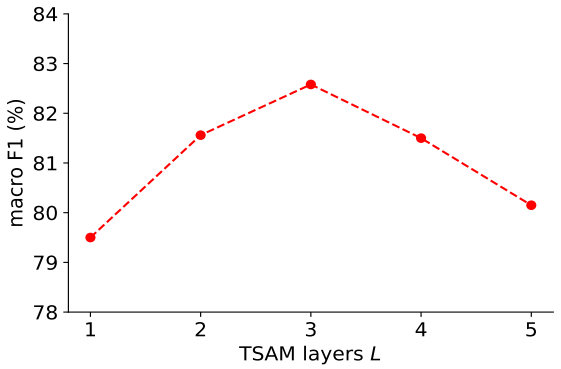

Figure 3: Results of Ours$_{large}$ with various TSAM layers on the development set of RECCON-DD.

6.4 Error Analysis

By analyzing our predicted emotion causes, we find that the following aspects mainly cause the predicted errors. Firstly, our model weakly gives the correct predictions for target utterances with three or more causes.[4] Compared to the utterances with 1 or 2 causes, the proposed model dropped six macro F1 scores on utterances with multiple causes. Secondly, our model cannot predict well when the underlying emotional cause is latent. At this point, recognizing emotion causes may require complex reasoning steps, and commonsense knowledge is an integral part of this process. We take the case below as an example:

- S_A (happy): Hello, thanks for calling 123 Tech Help. I'm Todd. How can I help you?
- S_B (fear): Hello? Can you help me? My computer! Oh, man...

In this case, S_A is happy to help S_B. In this example, the cause of happy emotion is due to the event "greeting" or intention to provide help. On the other hand, S_B is fearful because his or her computer is broken. Both speakers' causes of elicited emotions can only be inferred using commonsense knowledge, which our model does not explicitly consider.

7 Conclusion and Future Work

For the first time, we tackle CEE in a joint framework. We classify multiple utterances synchronously to capture the correlations between contextual utterances in a global view and propose a TSAM to effectively model the speaker's emotional influences. Experimental results on the benchmark dataset show that our model significantly outperforms the SOTA model, and further analysis verifies the effectiveness of each component. This paper points out a new reliable route for follow-up works: incorporating the emotion and speaker information explicitly and modeling the speaker's emotional influences effectively can bring enormous benefits for the tasks similar to CEE.

In the future, we would explore three aspects: (1) Learn emotion recognition and emotion cause recognition in conversations jointly; (2) Apply our model to other similar tasks which need to incorporate the speaker and emotion information; (3) Incorporate commonsense knowledge into the model explicitly to address situations when the underlying emotion cause is latent.

[4]Utterances with 3 or more causes account for approximately 14% of the RECCON-DD dataset

References

Ankush Chatterjee, Umang Gupta, Manoj Kumar Chinnakotla, Radhakrishnan Srikanth, Michel Galley, and Puneet Agrawal. 2019. Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*, 93:309–317.

Sheng Yeh Chen, Chao Chun Hsu, Chuan Chun Kuo, Ting Hao Kenneth Huang, and Lun Wei Ku. 2019. Emotionlines: An emotion corpus of multi-party conversations. In *11th International Conference on Language Resources and Evaluation, LREC 2018*, pages 1597–1601. European Language Resources Association (ELRA).

Xinhong Chen, Qing Li, and Jianping Wang. 2020a. A unified sequence labeling model for emotion cause pair extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 208–218.

Ying Chen, Wenjun Hou, Shoushan Li, Caicong Wu, and Xiaoqiang Zhang. 2020b. End-to-end emotion-cause pair extraction with graph convolutional network. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 198–207.

Ying Chen, Sophia Yat Mei Lee, Shoushan Li, and Chu-Ren Huang. 2010. Emotion cause detection with linguistic constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 179–187.

Zifeng Cheng, Zhiwei Jiang, Yafeng Yin, Hua Yu, and Qing Gu. 2020. A symmetric local search network for emotion-cause pair extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 139–149.

Leyang Cui and Yue Zhang. 2019. Hierarchically-Refined Label Attention Network for Sequence Labeling. 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 4115–4128.

Zixiang Ding, Huihui He, Mengran Zhang, and Rui Xia. 2019. From independent prediction to reordered prediction: Integrating relative position and global label information to emotion cause identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6343–6350.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020a. ECPE-2D: Emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3161–3170.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020b. End-to-end emotion-cause pair extraction based on sliding window multi-label learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3574–3583.

Chuang Fan, Chaofa Yuan, Jiachen Du, Lin Gui, Min Yang, and Ruifeng Xu. 2020. Transition-based directed graph construction for emotion-cause pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3707–3717.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020a. COSMIC: COmmonSense knowledge for eMotion Identification in Conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2470–2481.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2020b. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *EMNLP-IJCNLP 2019-2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*.

Lin Gui, Jiannan Hu, Yulan He, Ruifeng Xu, Qin Lu, and Jiachen Du. 2017. A Question Answering Approach for Emotion Cause Extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1593–1602.

Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. 2016a. Event-Driven Emotion Cause Extraction with Corpus Construction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1639–1649.

Lin Gui, Ruifeng Xu, Qin Lu, Dongyin Wu, and Yu Zhou. 2016b. Emotion cause extraction, a challenging task with corpus construction. In *Chinese National Conference on Social Media Processing*, pages 98–109. Springer.

Gaël Guibon, Matthieu Labeau, Hélène Flamein, Luce Lefeuvre, and Chloé Clavel. 2021. Few-Shot Emotion Recognition in Conversation with Sequential Prototypical Networks. In *The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018a. ICON: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2594–2604.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018b. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the conference. Association for Computational Linguistics. North American*

Chapter. Meeting, volume 2018, page 2122. NIH Public Access.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. DialogueCRN: Contextual Reasoning Networks for Emotion Recognition in Conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 7042–7052.

Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto. 2020. 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)*, pages 7360–7370.

Wenxiang Jiao, Haiqin Yang, Irwin King, and Michael R Lyu. 2019. HiGRU: Hierarchical Gated Recurrent Units for Utterance-Level Emotion Recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 397–406.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.

Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. 2010a. A text-driven rule-based system for emotion cause detection. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 45–53.

Sophia Yat Mei Lee, Ying Chen, Shoushan Li, Chu-Ren Huang, et al. 2010b. Emotion cause events: Corpus construction and analysis. In *LREC*. Citeseer.

Ruifan Li, Hao Chen, Fangxiang Feng, Zhanyu Ma, Xiaojie Wang, and Eduard Hovy. 2021a. Dual Graph Convolutional Networks for Aspect-based Sentiment Analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6319–6329.

Xiangju Li, Wei Gao, Shi Feng, Yifei Zhang, and Daling Wang. 2021b. Boundary Detection with BERT for Span-level Emotion Cause Analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 676–682.

Xiangju Li, Kaisong Song, Shi Feng, Daling Wang, and Yifei Zhang. 2018. A co-attention neural network model for emotion cause analysis with emotional context awareness. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4752–4757.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

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

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

Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, and Yajing Sun. 2022. Do You Know My Emotion? Emotion-Aware Strategy Recognition towards a Persuasive Dialogue System. *arXiv preprint arXiv:2206.12101*.

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.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. 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.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, pages 1–16.

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

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed Acyclic Graph Network for Conversational Emotion Recognition. In *Proceedings of the 59th Annual Meeting of the Association*

for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, pages 1551–1560.

Hao Tang, Donghong Ji, Chenliang Li, and Qiji Zhou. 2020. Dependency graph enhanced dual-transformer structure for aspect-based sentiment classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6578–6588.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *International Conference on Learning Representations*.

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

Penghui Wei, Jiahao Zhao, and Wenji Mao. 2020. Effective inter-clause modeling for end-to-end emotion-cause pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3171–3181.

Rui Xia and Zixiang Ding. 2019. Emotion-Cause Pair Extraction: A New Task to Emotion Analysis in Texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012.

Rui Xia, Mengran Zhang, and Zixiang Ding. 2019. RTHN: A RNN-Transformer Hierarchical Network for Emotion Cause Extraction. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5285–5291.

Yuqiang Xie, Yue Hu, Wei Peng, Guanqun Bi, and Luxi Xing. 2022. Comma: Modeling relationship among motivations, emotions and actions in language-based human activities.

Hanqi Yan, Lin Gui, Gabriele Pergola, and Yulan He. 2021. Position Bias Mitigation: A Knowledge-Aware Graph Model for Emotion Cause Extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3364–3375.

Chaofa Yuan, Chuang Fan, Jianzhu Bao, and Ruifeng Xu. 2020. Emotion-cause pair extraction as sequence labeling based on a novel tagging scheme. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3568–3573.

Duzhen Zhang, Xiuyi Chen, Shuang Xu, and Bo Xu. 2020. Knowledge Aware Emotion Recognition in Textual Conversations via Multi-Task Incremental Transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4429–4440.

Weixiang Zhao, Yanyan Zhao, and Xin Lu. 2022. CauAIN: Causal Aware Interaction Network for Emotion Recognition in Conversations. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4524–4530.

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.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1571–1582.