ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230537

Do Topic and Causal Consistency Affect Emotion Cognition? A Graph Interactive Network for Conversational Emotion Detection

Geng Tu[a,b], Bin Liang[a,b,c;*], Xiucheng Lyu[a,b], Lin Gui[d] and Ruifeng Xu[a,b,e;*]

[a]Harbin Institute of Technology, Shenzhen, China
[b]Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
[c]The Chinese University of Hong Kong, Hong Kong, China
[d]Department of Computer Science, University of Warwick, Coventry CV4 7AL, U.K.
[e]Peng Cheng Laboratory, Shenzhen, China

Abstract. Emotion recognition in conversations (ERC) typically requires modeling both intra- and inter-speaker context dependencies. However, when modeling inter-speaker dependencies, it may not capture differences among other participants in the conversation. Recent ERC research has attempted to improve utterance representations by utilizing speakers' commonsense knowledge. Nonetheless, these studies ignore the causal consistency in knowledge between the two participants, which contradicts the above modeling of speaker-sensitive context dependencies. Additionally, it is observed that historical utterances from various topics are blindly leveraged in context modeling, which fails the inter- and intra-topic coherence. To address these issues, we propose the topic- and causal-aware interactive graph network (TCA-IGN). Specifically, we suggest a graph encoder to model topic-level context dependencies, achieving inter- and intra-topic coherence. The topics of utterances are derived from a context-sensitive neural topic model. Then, we present a causal-aware graph attention to keep the speaker's causal consistency in commonsense knowledge, improving speaker-level context modeling. Finally, considering the defect of modeling inter-speaker or inter-topic context dependencies, we employ supervised contrastive learning to sweeten it. Experimental results show that TCA-IGN outperforms state-of-the-art methods on three public conversational datasets.

1 Introduction

Emotion recognition in conversations (ERC) has received considerable attention [14, 37, 38] due to its potential applications in several areas, like recommendation systems and dialogue generation [36].

According to the emotion generation theory [10] and emotional dynamics of conversations [31], existing efforts focus on modeling speaker-sensitive context dependencies, including recurrent-based network [12, 23, 15], transformer-based network [21], and graph-based network [8]. However, all these methods rely on future utterances to determine the emotion of the current utterance, which is not feasible to achieve in a real-life scenario. And they ignore the differences among other participants in the modeling context. This can be intuitively understood as the impact of others in a conversation on the current speaker is necessarily different.

More importantly, these context- and speaker-sensitive methods are not able to work like a human because of the lack of commonsense knowledge [36]. Recently, the birth of if-then commonsense knowledge generated by the COMET [2] brings vitality to enriching the utterance representation in ERC [6]. The COMET is an generative model trained on the ATOMIC dataset [32] (containing 9 if-then relationships shown in Fig. 1, i.e. {*xEffect*, *xWant*, *xReaction*, *xIntent*, *xAttribute*, *xNeed*} $\in \mathcal{X}$ of speakers. And {*oEffect*, *oReaction*, *oWant*} $\in \mathcal{O}$ of listeners.). For example, Ghosal et al. [6] proposed the COSMIC to model the speaker's psychological state by leveraging the COMET, for better utterance representations. Then Li et al. [19] further built the SKAIG to model the speaker's structured psychological interactions. Unfortunately, these methods only utilize external commonsense knowledge to improve the utterance representations and fail to meet the emotional dynamics of conversations [31], as they do not take into account the speaker's causal consistency (SCC) in knowledge. The SCC means when two utterances

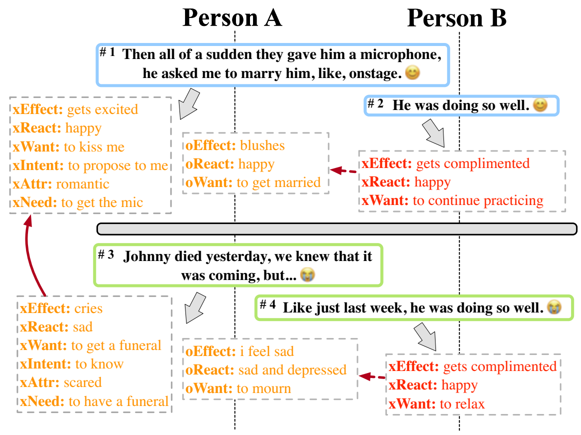

Person A
Person B
# 1 Then all of a sudden they gave him a microphone, he asked me to marry him, like, onstage.
xEffect: gets excited
xReact: happy
xWant: to kiss me
xIntent: to propose to me
xAttr: romantic
xNeed: to get the mic
# 2 He was doing so well.
oEffect: blushes
oReact: happy
oWant: to get married
xEffect: gets complimented
xReact: happy
xWant: to continue practicing
# 3 Johnny died yesterday, we knew that it was coming, but...
xEffect: cries
xReact: sad
xWant: to get a funeral
xIntent: to know
xAttr: scared
xNeed: to have a funeral
# 4 Like just last week, he was doing so well.
oEffect: i feel sad
oReact: sad and depressed
oWant: to mourn
xEffect: gets complimented
xReact: happy
xWant: to relax

Figure 1. Utterances from [48] related to specific topics convey distinct emotions. Utterances on the same topic use a consistent color scheme.

[*] Corresponding authors. Email: bin.liang@cuhk.edu.hk and xuruifeng@hit.edu.cn.

are from different speakers, {*xEffect*, *xWant*, *xReaction* } of the former and {*oEffect*, *oWant*, *oReaction* } of the latter need to be consistent. Whereas if the two utterances are from the same speaker, $\mathcal{X}$ of the former and latter need to be consistent, blending with speaker-level context modeling. From Fig. 1, we can see that the SCC can effectively differentiate between utterance 1 and utterance 3, prompting context modeling.

Additionally, topics, the basic knowledge of conversations, play a vital role in dialogue modeling [40] and generation [46]. In particular, modeling inter- and intra-topic coherences in conversations is practical for modeling dialogue context [41]. In Fig. 1, the phrase "He was doing so well" conveyed different emotions depending on the topic it was associated with, which is unsolvable for a topic-agnostic ERC model. But there is scarcely work centralizing on topic modeling in ERC, which causes the mixture of utterances with various topics in context modeling and fails the inter- and intra-topic coherences. For example, Zhu et al. [48] borrow the encoder-decoder architectures leveraging the pre-trained language model to model high-level syntactic features as the topic but still tend to neglect the inter- and intra-topic propagation.

Based on the above, we propose a topic- and causal-aware interactive graph network (TCA-IGN) to keep the SCC and achieve inter- and intra-topic coherences. More concretely, we present a causal-aware graph attention network (CAGAT) to model the SCC in commonsense knowledge to meet the emotional dynamics of conversations, promoting the speaker-level graph context encoder (SGCE). Moreover, we explore a context-sensitive neural topic model [24] (CNTM) to obtain topics, incorporating the concept-level knowledge retrieved from SenticNet [4] to enrich short utterances. Then we employ the topics to model inter- and intra-topic context dependencies in the topic-level graph context encoder (TGCE). In addition, to fill the gap of insensitive context modeling to different speakers and topics, we introduce supervised contrastive learning (SCL) to capture correlations and differences between utterance representations according to their speakers and topics, enhancing the modeling of inter-speaker and inter-topic context dependencies. To sum up, the main contributions of this paper can be summarized as follows:

- We are the first to model topic-level context dependencies and keep the SCC in commonsense knowledge to supplement speaker-level context dependencies in ERC.
- We introduce SCL to differentiate utterance representations with different topics and speakers, respectively, improving the inter-speaker and inter-topic context modeling.
- Experimental results demonstrate that our proposed method outperforms state-of-the-art ERC methods.

2 Related Work

2.1 Emotion Recognition in Conversations

Context-sensitive Models: The generation of emotions is influenced by contextual information according to the emotion generation theory [10]. Therefore, RNN-based models [29] are commonly employed to capture context dependencies. However, they may not be capable of distinguishing between various contexts [21], that is, historical utterances. To address this limitation, memory networks [12, 15] have gained more attention. Moreover, the role of participants in ERC plays a crucial role in determining the speaker's emotional state [31]. To model the speaker-level context, researchers have focused more on speaker-specific models [23], graph-based models [26], and so on. For example, Majumder et al. [23] utilized three GRUs to track global context, speaker state, and emotional state in conversations. Shen et al. [33] employed a graph-based model to model self- and inter-speaker dependencies.

Knowledge-sensitive Models: While previous studies have achieved impressive performance in ERC, they still fall short of human-like conversational abilities due to the lack of commonsense knowledge [47]. To address this, Ghosal et al. [6] employed GRUs to model participants' psychological states in conversations and generated commonsense knowledge from COMET. Li et al. [19] introduced the SKAIG model to further capture the structural psychological states. Recently, Zhao et al. [45] developed a causal-aware model that utilizes generated commonsense knowledge to capture contextual information. However, these models fail to meet the emotional dynamics of conversations [31], as they do not consider the SCC in commonsense knowledge.

2.2 Topic Modeling

Traditional approaches utilize topic models for inferring latent semantics of a document, such as probabilistic latent semantic analysis [13] and latent Dirichlet allocation (LDA) [1]. However, recently, the neural topic model (NTM) [24] has gained popularity due to the success of neural variational inference [17]. NTM can infer a latent distribution to capture the underlying semantics of a document. In previous research, topic models were utilized to aid in text classification. For instance, Zeng et al. [44] attempted to address the issue of data sparsity in short text categorization by using a topic model. In ERC, Zhu et al. [48] utilized the pre-trained language model to model high-level syntactic features as topics. However, these methods only employ the topic model as a sentence encoder. In contrast, we propose an NTM for annotating utterances in conversations to model inter- and intra-topic coherences.

2.3 Contrastive Learning

Chen et al. [5] introduced SimCLR, a contrastive learning (CL) network that uses diverse image augmentation techniques to generate positive and negative samples for visual representation. In NLP, Yan et al. [42] proposed a self-supervised CL method for fine-tuning BERT in response to poor performance in semantic text similarity tasks. Kim et al. [16] investigated a CL approach without data augmentation, utilizing BERT with both frozen and fine-tunable parameters to generate positive and negative samples. Gunel et al. [11] extended the self-supervised CL method to a fully-supervised CL setting to enhance performance in few-shot learning scenarios. In ERC, Li et al. [20] attempted to employ SCL to separate utterances with different emotions to improve emotion identification. No relevant work has been conducted on unsupervised CL in ERC.

3 Methodology

Our proposed TCA-IGN consists of three primary components: utterance-level context encoder, speaker- and topic-level GCE, and SCL modules. The overall architecture is as shown in Fig. 2.

3.1 Problem Definition

Let a conversation $\mathcal{C} = \{u_1, u_2..., u_n\}$, where n is the number of utterances. Each utterance u_i is uttered by one of speakers $\mathcal{S} = \{s_1, s_2..., s_m\}$ and comprises n_k tokens. Given a pre-defined emotion label set $\mathcal{Y} = \{y_1, y_2, ..., y_n\}$, the ERC task aims to predict the emotion label y_i of each utterance u_i in conversations.

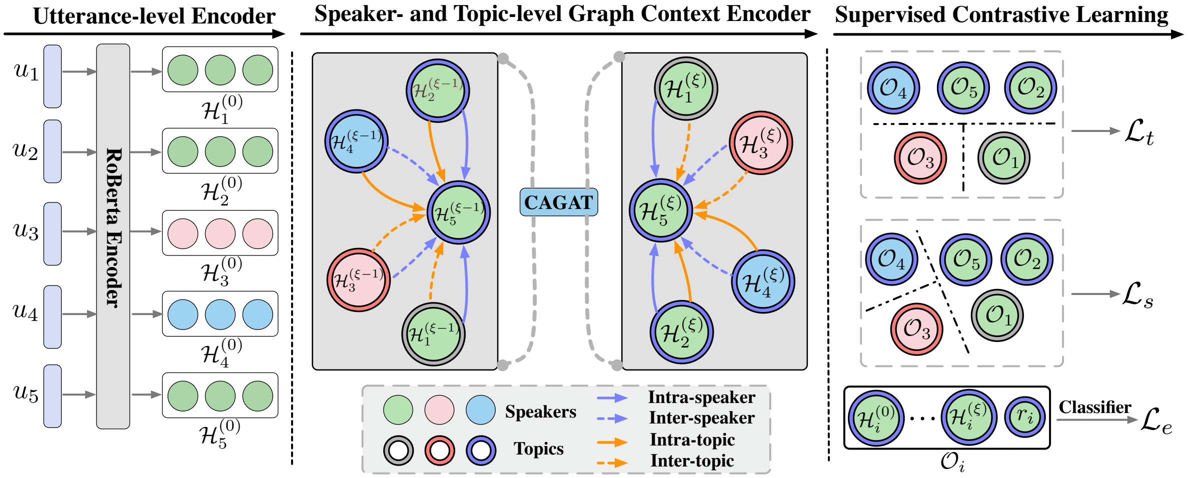


Figure 2. The proposed TCA-IGN for ERC. Different colors of the circles represent different speakers, while the different colors of the circle edges represent different topics. The topics utilized in TGCE and SCL modules are both derived from CNTM.

3.2 Utterance-level Context Encoder

In ERC, it is critical to extract contextual information (i.e. surrounding utterances) of the current utterance u_i [29]. However, using future utterances to determine the emotion of the current utterance is not practical in a real-life scenario. Therefore, we define the historical utterances $u_j, \forall j < i$ can be considered as its context. Then, we utilize the pre-trained model Roberta [22] to encode u_i and employ GRUs to obtain the utterance representation $\mathcal{H}_1^{(\xi)} \in \mathbb{R}^{d_h}, \forall \xi \geq 1$:

$$r_i = Roberta(u_i) \tag{1}$$

$$\mathcal{H}_i^{(0)} = ReLU(Linear(r_i)) \tag{2}$$

$$\mathcal{H}_1^{(\xi)} = \overrightarrow{GRU}_u(\mathcal{H}_1^{(\xi-1)}) + \overleftarrow{GRU}_u(\mathcal{H}_1^{(\xi-1)}) \tag{3}$$

where $r_i \in \mathbb{R}^{d_e}$ is the d_e-dimension hidden states of u_i after applying the Roberta extractor. $\mathcal{H}_i^{(0)} \in \mathbb{R}^{d_h}$ represents the output of a linear transformation with ReLU activation to u_i.

3.3 Speaker-level Graph Context Encoder

3.3.1 Graph Construction:

Unlike the previous graph construction [7, 19, 38], they utilize future utterances to update the current utterance node. Following the approach described in [33], we build a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R}, \mathcal{A})$ to model context dependencies. $u_i \in \mathcal{V}$ and $r_k \in \mathcal{R}$ represent the utterance node and edge type, respectively. $e_{i,j} = (u_i, r_k, u_j) \in \mathcal{E}, \forall j > i$ denotes the edge between node i and j. $\alpha_{i,j} \in \mathcal{A}$ is the weight of $e_{i,j}$. Specifically, the utterance-level context encoder initializes each utterance node u_i. The construction of edges $\mathcal{E}$ is based on a hypothesis that utterances are all dependent on n_w window-size historical utterances and themselves in a conversation. And each window contains an utterance spoken by the current speaker s_i as most. $\mathcal{R}$ of an edge $e_{i,j}$ is set depending upon speaker and topic of utterances. $\mathcal{W}$ is assigned based on SGCE and TGCE.

3.3.2 CAGAT:

For each node at each layer, the graph attention network (GAT) aggregates the information of its neighboring nodes as follows.

$$\alpha_{\dagger,i,j}^{(\xi)} = Softmax_{j \in N_r}(\mathcal{W}_e^{(\xi)}\ [\mathcal{H}_i^{(\xi-1)} \oplus \mathcal{H}_j^{(\xi)}]) \tag{4}$$

$$\mathcal{Q}_{\dagger,i}^{(\xi)} = \textstyle\sum_{j \in N_r} \mathcal{M}_x(\alpha_{\dagger,i,j}^{(\xi)} \mathcal{W}_c^{(\xi)} \mathcal{H}_j^{(\xi)}) + \mathcal{M}_o(\alpha_{\dagger,i,j}^{(\xi)} \mathcal{W}_s^{(\xi)} \mathcal{H}_j^{(\xi)}) \tag{5}$$

where $\mathcal{Q}_{\dagger,i}^{(\xi)}$ is the speaker-level context representation of u_i. N_r denotes the set of neighboring nodes. $\oplus$ represents the concatenation operation. $\mathcal{W}_e^{(\xi)} \in \mathbb{R}^{d_h \times 2d_h}$, $\mathcal{W}_c^{(\xi)} \in \mathbb{R}^{d_h \times d_h}$, and $\mathcal{W}_s^{(\xi)} \in \mathbb{R}^{d_h \times d_h}$ represents projection parameters of the model. $\mathcal{M}_x = \mathbb{1}_{[s_i = s_j]}$ and $\mathcal{M}_o = \mathbb{1}_{[s_i \neq s_j]}$ denote the indicator function, used to model the intra- and inter-speaker context dependencies.

To enrich utterance representations, existing methods [6, 19, 45, 48] typically incorporate knowledge in the following forms.

$$\widehat{\alpha}_{\dagger,i,j}^{(\xi)} = Softmax_{j \in N_r}(\mathcal{W}_e^{(\xi)}\ [\mathcal{H}_i^{(\xi-1)} \oplus (\mathcal{H}_j^{(\xi-1)}; \mathcal{K}_j)]) \tag{6}$$

$$\widehat{\alpha}_{\dagger,i,j}^{(\xi)} = Softmax_{j \in N_r}(\mathcal{W}_e^{(\xi)}\ [(\mathcal{H}_i^{(\xi-1)}; \mathcal{K}_i) \oplus (\mathcal{H}_j^{(\xi-1)}; \mathcal{K}_j)]) \tag{7}$$

where $\mathcal{K}_i$ is the subset of 9 if-then relationships for the i-th utterance. ; stands for the aggregation method, including concatenation, summation, or attention-based methods. Obviously, the forms did not consider the SCC in commonsense knowledge, which goes against the emotional dynamics of conversations. Therefore, we develop the GAT to CAGAT, blending with speaker-level context modeling.

$$\alpha_{\pm,i,j}^{(\xi)} = Softmax_{j \in N_r}^{s_i = s_j}(\widehat{\mathcal{W}}_x^{(\xi)}\ [\widehat{\mathcal{X}}_i^{(\xi-1)} \oplus \widehat{\mathcal{X}}_j^{(\xi)}]) \tag{8}$$

$$\alpha_{\ddagger,i,j}^{(\xi)(s)} = Softmax_{j \in N_r}^{s_i = s_j}(\mathcal{W}_x^{(\xi)}\ [\mathcal{X}_i^{(\xi-1)} \oplus \mathcal{X}_j^{(\xi)}]) \tag{9}$$

$$\alpha_{\ddagger,i,j}^{(\xi)(d)} = Softmax_{j \in N_r}^{s_i \neq s_j}(\mathcal{W}_o^{(\xi)}\ [\mathcal{X}_i^{(\xi-1)} \oplus \mathcal{O}_j^{(\xi)}]) \tag{10}$$

$$\mathcal{Q}_{\pm,i}^{(\xi)} = \sum\nolimits_{j \in N_r} \alpha_{\pm,i,j}^{(\xi)} \widehat{\mathcal{W}}_p^{(\xi)} \mathcal{H}_j^{(\xi)} \tag{11}$$

$$\mathcal{Q}_{\ddagger,i}^{(\xi)(s/d)} = \sum\nolimits_{j \in N_r} \alpha_{\ddagger,i,j}^{(\xi)(s/d)} \mathcal{W}_p^{(\xi)(s/d)} \mathcal{H}_j^{(\xi)} \tag{12}$$

where $\mathcal{Q}_{\ddagger,i}^{(\xi)}$ and $\mathcal{Q}_{\pm,i}^{(\xi)}$ are causal-aware context representation and initialized to 0. $\alpha_{\ddagger,i}^{(\xi)(s/d)}$ and $\alpha_{\pm,i}^{(\xi)}$ are attention weights of CAGAT. $\mathcal{X}^{(\xi)} \in \mathbb{R}^{3d_x}$ and $\widehat{\mathcal{X}}^{(\xi)} \in \mathbb{R}^{3d_x}$ contains 3 if-then relation types: {*xEffect*, *xWant*, *xReaction* } of speakers and {*xIntent*, *xAttribute*, *xNeed* } of speakers, respectively. $\mathcal{O}^{(\xi)} \in \mathbb{R}^{3d_x}$ includes the {*oEffect*, *oReaction*, and *oWant* } of listeners. $\mathcal{W}_x^{(\xi)} \in \mathbb{R}^{d_h \times 6d_x}$, $\widehat{\mathcal{W}}_x^{(\xi)} \in \mathbb{R}^{d_h \times 6d_x}$, $\mathcal{W}_o^{(\xi)} \in \mathbb{R}^{d_h \times 6d_x}$, $\mathcal{W}_p^{(\xi)(s/d)} \in \mathbb{R}^{d_h \times d_h}$, and $\widehat{\mathcal{W}}_p^{(\xi)} \in \mathbb{R}^{d_h \times d_h}$ represent projection parameters.

3.4 *Topic-level Graph Context Encoder*

3.4.1 *CNTM*

Encoder-decoder: We utilize the Bag-of-Words (BoW) approach to represent documents, where each utterance in a conversation is considered a document. To enrich short documents, we retrieve emotion concept-level knowledge from SenticNet [4] that provides a set of semantics associated with natural language concepts. Our document is represented as a BoW vector w, where each element w_i represents the frequency of the i-th word of the document in the vocabulary. The encoder of our CNTM is a parallel structure of a multi-layer perceptron (MLP) and LSTM network, which maps w to an output layer that contains k units. The encoder ensures fast convergence while capturing contextual topic information, which to some extent guarantees the coherence of the topic in a conversation. We then apply the softmax function to derive the document-topic vector $\theta \in \mathbb{R}^K$. The encoder serves as the recognition network, enabling inference: $\mathbf{Q}(\theta|w)$, which is approximately equal to a prior distribution of w, $\mathbf{P}(\theta|w)$. A deterministic encoder can be used instead of a variational autoencoder, which is both conceptually and computationally simpler. $\mathbf{Q}(\theta|w)$ is a Dirac Delta distribution. Given θ, the decoder structure is similar to the encoder and maps θ to an output layer that contains n_v units. We then apply the softmax function to obtain a probability distribution $\widehat{w} \in \mathbb{R}^{n_v}$ over the words.

Training: In the autoencoder, the reconstruction loss is calculated as the negative cross-entropy loss between the BoW w and its estimated value $\widehat{w}$ obtained from the decoder. Mathematically, we have:

$$\mathbf{C}(w,\widehat{w}) = -\sum\nolimits_{j=1}^{n_v} w_i \, log(\widehat{w}_j) \tag{13}$$

For distribution matching, following [25], we employ maximum mean discrepancy (MMD) [9] with information diffusion kernel [18] to match high-dimensional Dirichlet distributions [1].

$$\begin{aligned}\widehat{MMD}(\mathbf{Q},\mathbf{P}) = {} & \tfrac{1}{m(m-1)} \sum_{i \neq j} \mathbf{K}(\theta_i,\theta_j) \\ & + \tfrac{1}{m(m-1)} \sum_{i \neq j} \mathbf{K}({\theta'}_i,{\theta'}_j) - \tfrac{2}{m^2} \sum_{i,j} \mathbf{K}(\theta_i,{\theta'}_j)\end{aligned} \tag{14}$$

$$\mathbf{K}(\theta,\theta') = \exp\left(-\arccos^2\left(\sum\nolimits_{k=1}^{K} \sqrt{\theta_k {\theta'}_k}\right)\right) \tag{15}$$

where m samples $\{\theta_1, ..., \theta_m\}$ and $\{\theta'_1, ..., \theta'_m\}$ are sampled from $\mathbf{Q}$ and $\mathbf{P}$, respectively. In practice, the reconstruction loss $\mathbf{C}(w,\widehat{w})$ may be significantly greater than the regularization term $\widehat{MMD}(\mathbf{Q},\mathbf{P})$. To balance the two factors without introducing an additional hyper-parameter, we assume the document with length d_l, including d_l unique words and the decoder output is completely uninformative, i.e. $\widehat{w}_i = 1/n_v, i = 1, ..., n_v$. Then the reconstruction loss can be expressed as $d_l log(n_v)$. And we can normalize the reconstruction loss to 1 by scaling it by $1/\left(d_l log(n_v)\right)$.

3.4.2 *Topic-aware Context Modeling:*

To model the topic-level context dependencies, we aggregate the neighboring nodes with the same topic as follows.

$$\mathcal{Q}_{\S,i}^{(\xi)} = \sum\nolimits_{j \in N_r} \mathcal{M}_t(\alpha_{\dagger,i,j}^{(\xi)} \mathcal{W}_t^{(\xi)} \mathcal{H}_j^{(\xi)}) + \mathcal{M}_d(\alpha_{\dagger,i,j}^{(\xi)} \mathcal{W}_d^{(\xi)} \mathcal{H}_j^{(\xi)}) \tag{16}$$

where $\mathcal{W}_t^{(\xi)} \in \mathbb{R}^{d_h \times d_h}$ and $\mathcal{W}_d^{(\xi)} \in \mathbb{R}^{d_h \times d_h}$ are projection parameters. $\mathcal{M}_t = \mathbb{1}_{[z_i = z_j]}$ and $\mathcal{M}_d = \mathbb{1}_{[z_i \neq z_j]}$ denote the indicator function, where $z_i \in \mathcal{Z} \in \mathbb{R}^K$ is the topic of the u_i. Then we can obtain the new utterance representation $\mathcal{H}_i^{(\xi)}$ of the ξ layer.

$$\mathcal{Q}_i^{(\xi)} = \mathcal{W}_q^{(\xi)}(\mathcal{Q}_{\dagger,i}^{(\xi)} \oplus \mathcal{Q}_{\ddagger,i}^{(\xi)(s)} \oplus \mathcal{Q}_{\ddagger,i}^{(\xi)(d)} \oplus \mathcal{Q}_{\pm,i}^{(\xi)}) + \mathcal{Q}_{\S,i}^{(\xi)} \tag{17}$$

$$\mathcal{H}_i^{(\xi)} = \overrightarrow{GRU}_{\varrho}(\mathcal{Q}_i^{(\xi)}, \mathcal{H}_i^{(\xi-1)}) + \overleftarrow{GRU}_{\varrho}(\mathcal{H}_i^{(\xi-1)}, \mathcal{Q}_i^{(\xi)}) \tag{18}$$

where $\mathcal{Q}_i^{(\xi)}$ is the merger of $\mathcal{Q}_{\dagger,i}^{(\xi)}$, $\mathcal{Q}_{\ddagger,i}^{(\xi)(s)}$, $\mathcal{Q}_{\ddagger,i}^{(\xi)(d)}$, $\mathcal{Q}_{\pm,i}^{(\xi)}$, and $\mathcal{Q}_{\S,i}^{(\xi)}$. And $\mathcal{W}_q^{(\xi)} \in \mathbb{R}^{d_h \times 4d_h}$ are projection parameters.

3.5 *Model Training*

We utilize a linear unit to predict the emotion distributions:

$$\mathcal{O}_i = \mathcal{H}_i^{(0)} \oplus \mathcal{H}_i^{(1)} \oplus ... \oplus \mathcal{H}_i^{(\xi)} \oplus r_i \tag{19}$$

$$\widehat{\mathcal{Y}}_i = Argmax(Softmax(\mathcal{W}_r \mathcal{O}_i + b_r)) \tag{20}$$

where $\mathcal{W}_r \in \mathbb{R}^{d_h \times ((\xi+1)d_h + de)}$ and b_r are the projection parameter and bias. $\widehat{\mathcal{Y}} \in \mathbb{R}^n$ is the predicting emotional label set of utterances.

$$\mathcal{L}_e = CrossEntropy(\widehat{\mathcal{Y}}, \mathcal{Y}) + \beta \left\|\Theta\right\|_2 \tag{21}$$

where $\mathcal{L}_e$ is the classification loss. Θ is a set of projection parameters. β represents the coefficient of L_2-regularization. To differentiate different speakers and topics respectively, we introduce $\mathcal{L}_s$ and $\mathcal{L}_t$ in modeling inter-speaker and inter-topic context dependencies.

$$\mathcal{L} = \mathcal{L}_e + \Psi_s \mathcal{L}_s + \Psi_t \mathcal{L}_t \tag{22}$$

$$\mathcal{L}_s = -\frac{1}{N_b} \sum \log \ell(\mathcal{S}) \qquad \mathcal{L}_t = -\frac{1}{N_b} \sum \log \ell(\mathcal{Z}) \tag{23}$$

$$\ell(\zeta) = \frac{\sum_{j=1}^{N_b} \mathbb{1}_{[i \neq j]} \mathbb{1}_{[\zeta_i = \zeta_j]} \mathcal{F}(\mathcal{O}_i, \mathcal{O}_j, \tau)}{\sum_{k=1}^{N_b} \mathbb{1}_{[i \neq k]} \mathcal{F}(\mathcal{O}_i, \mathcal{O}_k, \tau)} \tag{24}$$

where N_b is the size of a mini-batch sample $\mathcal{B}$. Ψ_s and Ψ_t are tuned hyper-parameters. $\mathcal{F}(\star, \star, \tau) = e^{simi(\star,\star)/\tau}$. τ is the temperature parameter, $simi(.)$ denotes the cosine similarity function.

4 EXPERIMENTS

4.1 *Datasets*

We conduct experiments on three datasets: IEMOCAP [3], MELD [30], and EmoryNLP [43]. The statistics of datasets are shown in Table 1. **IEMOCAP** comprises interactive sessions between two individuals, wherein they enact improvisations or scripted scenarios. Each spoken utterance is annotated with one of the six emotions: happy, angry, neutral, sad, excited, or frustrated. **MELD** is a collection of multi-party conversations extracted from the TV show *Friends*. Every utterance in this dataset is tagged with one of the seven emotions: surprise, fear, disgust, anger, sadness, neutral, or joy, and one of the three sentiments: neutral, negative, or positive. **EmoryNLP** consists of multi-party sessions from the *Friends* TV show, where each spoken utterance is labeled with one of the seven emotions: surprise, fear, disgust, anger, sadness, neutral, or joy, and one of the three sentiments: neutral, negative, or positive.

4.2 *Experimental Settings*

We perform a hyper-parameter search for TCA-IGN on each dataset with a validation set, including learning rate, batch size, dropout rate, tuned hyper-parameters $\Psi_{s/t}$, the window-size n_w for IGN,

Table 1. Statistics of experimental datasets.

Dataset	Dialogues			Utterances		
	train	val	test	train	val	test
IEMOCAP	120	12	31	5,810		1,623
MELD	1039	114	280	9,989	1,109	2610
EmoryNLP	659	89	79	7,551	954	984

Dataset	Classes	Metric
IEMOCAP	6	Weighted Avg F1
MELD	3 and 7	Weighted Avg F1 over 3 and 7 classes
EmoryNLP	3 and 7	Weighted Avg F1 over 3 and 7 classes

the window-size $\widehat{n}_w$ for CAGAT, and the number of TCA-IGN layers n_l[1], where $\forall\ \widehat{n}_w \geq n_w$. And we let $d_u = 1024$, $\tau = 0.07$, $d_x = 768$, and $d_h = 300$ on each dataset. Reported results of TCA-IGN are all based on the average score of 5 runs on the test set.[2]

4.3 Baselines

We compare our proposed framework with various ERC baselines, including RNN-based models: DialogueRNN [23], COSMIC [6], CauAIN [45]; Memory network: ICON [12], AGHMN [15], and Graph-based models: DialogueGCN [7], DAG-ERC [33], and SKAIG [19]; Transformer-based model: KET [47], TODKAT [48], and CoG-BART [20]. Additionally, we also compare several models including ChatGPT [28], fine-tuned Roberta-Large [22], and prompt-tuned Curie[3]. Curie, an intermediate-to-large scale language model developed by OpenAI, is equipped with 1.3 billion parameters and possesses the ability to execute the sentiment classification task with improved efficacy and precision [27].

5 Experimental Results and Analysis

5.1 Comparison with Baselines

Table 2 shows experimental results, where ♠ indicates models that use external knowledge, while ♮ and ♯ indicate results from the original papers and [6], respectively. ♭ represents our reproduced result. Our proposed TCA-IGN achieves competitive performance and reaches a new state-of-the-art across all three datasets.

Table 2. Comparison of results against various methods on three public datasets, expressed in percentage form.

Methods	IEMOCAP	MELD	EmoryNLP
ChatGPT ♭	40.07	54.37	37.55
Curie ♭	57.33	65.01	37.40
Roberta ♯	54.55	62.02	37.29
ICON ♯	58.54	-	-
DialogueRNN ♯	62.75	57.03	31.70
AGHMN ♮	63.50	58.10	-
CoG-BART ♮	66.18	64.81	39.04
DialogueGCN ♮	64.18	58.10	-
DAG-ERC ♮	68.03	63.65	39.02
KET ♮ ♠	59.56	58.18	34.39
COSMIC ♯ ♠	65.28	65.21	38.11
SKAIG ♮ ♠	66.96	65.18	38.88
CauAIN ♮ ♠	67.61	65.46	-
TODKAT ♮ ♠	61.33	65.47	38.69
Ours ♠	**68.69**	**66.03**	**39.84**
w/o CAGAT	67.54 (↓ 1.15%)	64.65 (↓ 1.38%)	38.63 (↓ 1.22%)
w/o TGCE	66.94 (↓ 1.75%)	64.52 (↓ 1.51%)	38.17 (↓ 1.67%)
w/o SCL	67.28 (↓ 1.41%)	64.70 (↓ 1.33%)	38.48 (↓ 1.36%)

As shown in Table 2, on IEMOCAP and EmoryNLP datasets, graph-based models generally perform better than recurrence-based models, suggesting that graph-based models excel at capturing crucial local context, especially in long conversation datasets (e.g. almost 70 utterances per dialogue in the IEMOCAP). Our TCA-IGN performs better than other graph methods because the SCC in commonsense knowledge is considered, which can effectively complement the previous speaker-level context modeling. On MELD, because the data is collected from TV shows, there may be instances where two consecutive utterances are not coherent. In such cases, the advantage of graph-based models in the encoding context may not be significant. But our TCA-IGN still has a competitive performance, which may benefit from modeling topic-level context dependencies. In addition, in multi-party conversation datasets (e.g. EmoryNLP), the potential of graph-based methods has not been fully unleashed, so they are not completely superior to other types of methods. This is mainly because they still construct graph networks in the form of dyadic conversations. To compensate for this deficiency, we adopted SCL to differentiate different speakers and topics, respectively.

5.2 Hyper-parameter Analysis

Analysis of n_l, n_w, and $\widehat{n}_w$: Fig. 6 shows the F1 score on validation datasets for different layer and context window sizes. Increasing n_l and n_w improves emotion recognition in long conversation datasets (e.g. IEMOCAP) up to a certain point, after which performance declines, because earlier contexts provide minimal useful information [34] and may add noise to the model. Relatively short conversation datasets (e.g. MELD and EmoryNLP) show a distinct pattern, with peak performance achieved with smaller layer and window sizes followed by fluctuations. However, due to limitations in computational resources, we were unable to explore larger ranges of layer and window sizes. Additionally, increasing $\widehat{n}_w$ almost outperforms the GAT across different window sizes, indicating that keeping the SCC in commonsense knowledge can enhance the GAT. And it performs best with larger window sizes, indicating its capacity to capture a wider range of contextual information and complement the GAT.

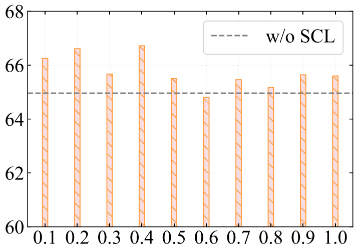


Figure 3. Performance of TCA-IGN on the validation set of the IEMOCAP dataset under different loss weight Ψ_t.

Analysis of Ψ_s and Ψ_t: During training, we fixed the loss weight Ψ_s at 1 and varied Ψ_t. Fig. 3 shows the F1 score on validation datasets for different Ψ_t. TCA-IGN almost outperformed the model without SCL. Smaller Ψ_t resulted in better performance, as the two SCL loss items complemented each other. However, increasing Ψ_t beyond a certain threshold decreased performance, possibly due to conflicting effects caused by the excessively large Ψ_t on another SCL loss item.

5.3 Ablation Study

To analyze the individual impact of each component on the performance of TCA-IGN, we conducted an ablation study, and the results

[1] Please refer to the appendix for more detailed hyper-parameter settings.
[2] Code and appendix are available at https://github.com/TuGengs/TCA-IGN.
[3] Please refer to the appendix for the prompt used for ChatGPT and Curie.

are presented in Table 2. It is evident that all components have contributed significantly to the overall improvement, as evidenced by the paired t-test p-value $\ll 0.05$.

Analysis of CAGAT: We use CAGAT to model the SCC in commonsense knowledge, improving context modeling. Random visualization of attention weights, as shown in Fig. 7, shows original weights are too sparse, limiting context to local information because of n_w. However, CAGAT enlarges the receptive field of the GAT, capturing richer contextual information, albeit with some unwanted information. The SCC ensures benefits outweigh drawbacks, as confirmed by performance improvement in Table 2.

Analysis of TGCE: We assessed the topic model using metrics such as topic diversity (td) [35] and coefficient of variation (tc)[4], with the latter being suitable for annotated topics. Our final model had a high td (>0.6) and the highest tc. In Table 3, we listed the topic words for each topic in the IEMOCAP dataset and noted that these words are strongly semantically related within their respective topics. Additionally, the number of topics K had a significant impact on TGCE, as shown in Fig. 4. Increasing K initially improved performance until reaching a peak, after which it fluctuated similarly to tc. However, td decreased as K increased.

Table 3. Part examples of topics in the IEMOCAP dataset.

Topics	Examples
1	'discover', 'booksmart', 'Burden', 'bring', 'forecast', 'row', 'annoyed', 'waited', 'effort', 'times', etc.
2	'newborn', 'infant', 'babe', 'cub', 'child', 'break_silence', 'laugh', 'confidence', 'glasses', 'opposition', etc.
3	'dealer', 'monger', 'bargainer', 'trader', 'principal', 'prevent', 'thwart', 'halt', 'continue', 'explanation', etc.
4	'Really', 'utterly_delightful', 'Hey', 'mother', 'hit', 'preposterous', 'Hello', 'Alright', 'crazy', 'Turn', etc.
5	'silly', 'nonrational', 'preposterous', 'luggage', 'nuts', 'dotard', 'glad', 'forget', 'euphoria', 'contentment', etc.
6	'Alright', 'mom', 'minute', 'kissed', 'mother', 'Hey', 'Turn', 'suppose', 'Wait', 'Calm', etc.

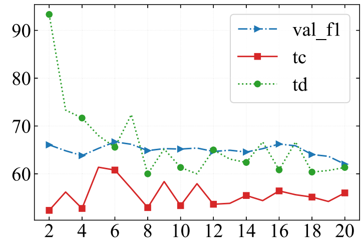


Figure 4. Performance of TCA-IGN on the validation set of the IEMOCAP dataset under a different number of topics.

Analysis of SCL: Fig. 5 shows the t-SNE [39] visualization of the intermediate representation of TCA-IGN and TCA-IGN w/o SCL. It demonstrates the convincingness of SCL, pulling utterances with different topics and speakers away, respectively. And in the dyadic conversational dataset IEMOCAP, the loss item $\mathcal{L}_s$ also plays a role with a 1.21% improvement, clarifying that stress differences between the speaker and listeners are also not insignificant.

5.4 Case Study

To better understand the working mechanism of the TCA-IGN model, we present a case study depicted in Fig. 8, where our model can correctly recognize all emotions of utterances. Firstly, potential topics are extracted through CNTM. Intuitively, incorporating topic-level context information improves ERC, as evidenced in utterances 11 and 10. Without considering the topics, the emotion of utterance 11 cannot be correctly recognized. Additionally, the impact of context information results in the grouping of utterances 7 and 9 into the same topic. In Att, utterances 1 and 5 are the most important, while 2, 3, and 4 are of secondary importance, which is clearly inappropriate. This is because utterances 4 and 5 are semantically similar and should carry more weight. Furthermore, due to the context window limitation, GAT can only extract context from utterances 6 and 7, making it difficult to identify the emotion in the 8 utterance. CAGAT complements GAT by highlighting utterances 4 and 5, where Person A's *xReact* and Person B's *oReact* are both 'happy' for the 8 utterance. Additionally, in the speaker-level context modeling for utterance 3 grouping utterances 1 and 2 into the same class would fail to reflect differences between speakers. Hence, SCL is employed to address disparities between speakers, which also positively impacts the prediction results of utterance 3. And Differences between different topics can also be addressed in such a way.

Table 4. Comparison of results against various SCL loss items.

Methods	IEMOCAP	MELD	EmoryNLP
Ours	**68.69**	**65.88**	**39.84**
w/o $\mathcal{L}_s$	67.48 (↓ 1.21%)	64.91 (↓ 1.12%)	38.89 (↓ 0.95%)
w/o $\mathcal{L}_t$	67.54 (↓ 1.15%)	65.06 (↓ 0.97%)	38.64 (↓ 1.20%)
w/o SCL	67.28 (↓ 1.41%)	64.70 (↓ 1.33%)	38.48 (↓ 1.36%)

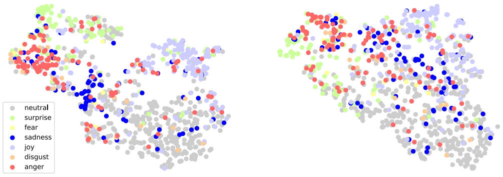


Figure 5. Visualization of intermediate embeddings of TCA-IGN (left) and TCA-IGN w/o SCL (right).

Table 5. Analysis of TCA-IGN on Emotional Shifts.

Methods	IEMOCAP	MELD	EmoryNLP
Ours	**68.69**	**66.03**	**39.84**
w/o Emotion Shift	76.94 (↑ 8.25%)	75.61 (↑ 9.58%)	51.20 (↑ 11.36%)
w/ Emotion Shift	63.25 (↓ 5.44%)	57.85 (↓ 8.18%)	34.19 (↓ 5.65%)

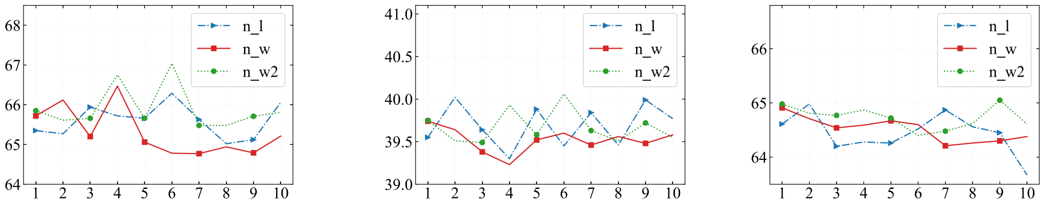
5.5 Error Analysis

Most errors in our analysis of the dataset can be attributed to class imbalance, such as the low F1 score for the 'fear' emotion as low as 8.76. And we may encounter errors because of setting the topic to -1 for very brief utterances with a BOW size of 0 after removing stop words. Furthermore, the commonsense knowledge is heavily reliant on the model's generated responses, which may lead to inaccuracies. Additionally, we are also focusing on solving the emotional shift problem, where consecutive utterances express different emotions, which has been challenging for prior approaches. In Table 5, TCA-IGN still struggles to perform well on samples with emotional shifts compared to those without.

6 Conclusion

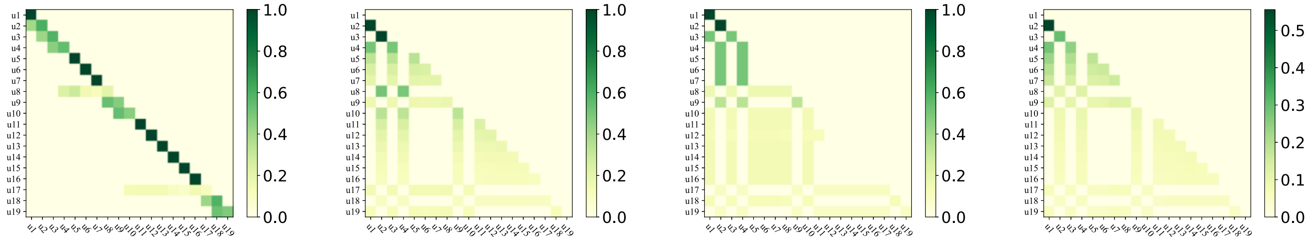
In this paper, we present a TCA-IGN model for ERC that incorporates several novel components. Specifically, we introduce a CAGAT

[4] https://en.wikipedia.org/wiki/Coefficient_of_variation



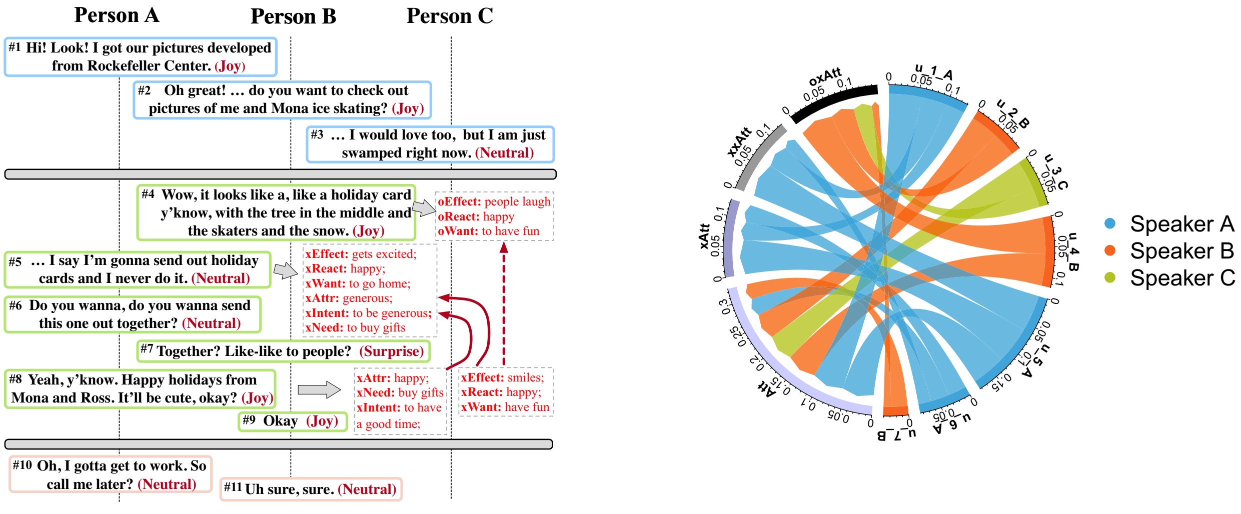
(a) IEMOCAP (b) EmoryNLP (c) MELD

Figure 6. Performance of TCA-IGN on the validation set of different datasets under different n_l (n_l), n_w (n_w), and $\widehat{n}_w$ (n_w2).



(a) Att (b) xAtt (c) oxAtt (d) xxAtt

Figure 7. Visualization of attention weights of TCA-IGN on the MELD dataset. Att, xAtt, oxAtt, and xxAtt represent $\alpha_{\dagger}$, $\alpha_{\ddagger}^{(s)}$, $\alpha_{\ddagger}^{(d)}$, and $\alpha_{\pm}$, respectively.



Person A Person B Person C

#1 Hi! Look! I got our pictures developed from Rockefeller Center. (Joy)

#2 Oh great! … do you want to check out pictures of me and Mona ice skating? (Joy)

#3 … I would love too, but I am just swamped right now. (Neutral)

#4 Wow, it looks like a, like a holiday card y'know, with the tree in the middle and the skaters and the snow. (Joy)

oEffect: people laugh
oReact: happy
oWant: to have fun

#5 … I say I'm gonna send out holiday cards and I never do it. (Neutral)

xEffect: gets excited;
xReact: happy;
xWant: to go home;
xAttr: generous;
xIntent: to be generous;
xNeed: to buy gifts

#6 Do you wanna, do you wanna send this one out together? (Neutral)

#7 Together? Like-like to people? (Surprise)

#8 Yeah, y'know. Happy holidays from Mona and Ross. It'll be cute, okay? (Joy)

xAttr: happy;
xNeed: buy gifts
xIntent: to have a good time;

xEffect: smiles;
xReact: happy;
xWant: have fun

#9 Okay (Joy)

#10 Oh, I gotta get to work. So call me later? (Neutral)

#11 Uh sure, sure. (Neutral)

(a) Examples of utterances with three topics. (b) Various attention weights of the 8-th utterance

Figure 8. A case from the MELD datasets shows that our model provides all correct predictions. Utterances with the same topic are highlighted with the same colored border in (a). In (b), the wider the connecting line, the greater the corresponding attention weight.

module to keep the SCC in commonsense knowledge, leading to improved modeling of speaker-level context and more comprehensive contextual information. And considering the imperfection of insensitive context modeling to different speakers and topics, we employ SCL to enhance it. Additionally, given that different topics may evoke different emotions, we propose a TGCE module that leverages the CNTM to model topic-level dependencies and achieve both intra- and inter-topic coherence, making it works even when two consecutive utterances are not coherent. Through extensive evaluations and an ablation study, we demonstrate the superiority of our TCA-IGN model and the significant impact of its components.

Acknowledgements

We thank the anonymous reviewers for their valuable suggestions to improve the overall quality of this manuscript. This work was partially supported by the National Natural Science Foundation of China (62006062, 62176076), Natural Science Foundation of Guangdong 2023A1515012922, the Shenzhen Foundational Research Funding (JCYJ20220818102415032), the Major Key Project of PCL2021A06, Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies 2022B1212010005.

References

[1] David M Blei, Andrew Y Ng, and Michael I Jordan, 'Latent dirichlet allocation', *Journal of machine Learning research*, **3**, 993–1022, (2003).

[2] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi, 'Comet: Commonsense transformers for knowledge graph construction', in *ACL*, (2019).

[3] 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*, **42**(4), 335–359, (2008).

[4] Erik Cambria, Qian Liu, Sergio Decherchi, Frank Xing, and Kenneth Kwok, ‘Senticnet 7: A commonsense-based neurosymbolic ai framework for explainable sentiment analysis’, in *LREC*, pp. 3829–3839, (2022).

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, ‘A simple framework for contrastive learning of visual representations’, in *ICML*, (2020).

[6] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria, ‘Cosmic: Commonsense knowledge for emotion identification in conversations’, in *Findings of EMNLP*, pp. 2470–2481, (2020).

[7] 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*, pp. 154–164, (2019).

[8] 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*, (2020).

[9] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola, ‘A kernel two-sample test’, *The Journal of Machine Learning Research*, **13**(1), 723–773, (2012).

[10] James J Gross and Lisa Feldman Barrett, ‘Emotion generation and emotion regulation: One or two depends on your point of view’, *Emotion review*, **3**(1), 8–16, (2011).

[11] Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov, ‘Supervised contrastive learning for pre-trained language model fine-tuning’, *arXiv preprint arXiv:2011.01403*, (2020).

[12] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann, ‘Icon: Interactive conversational memory network for multimodal emotion detection’, in *EMNLP*, pp. 2594–2604, (2018).

[13] Thomas Hofmann, ‘Probabilistic latent semantic indexing’, in *SIGIR*, pp. 50–57, (1999).

[14] Dazhi Jiang, Runguo Wei, Jintao Wen, Geng Tu, and Erik Cambria, ‘Automl-emo: Automatic knowledge selection using congruent effect for emotion identification in conversations’, *IEEE Transactions on Affective Computing*, 1, (2023).

[15] Wenxiang Jiao, Michael Lyu, and Irwin King, ‘Real-time emotion recognition via attention gated hierarchical memory network’, in *AAAI*, volume 34, pp. 8002–8009, (2020).

[16] Taeuk Kim, Kang Min Yoo, and Sang-goo Lee, ‘Self-guided contrastive learning for bert sentence representations’, in *ACL-IJCNLP*, pp. 2528–2540, (2021).

[17] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling, ‘Semi-supervised learning with deep generative models’, *NIPS*, **27**, (2014).

[18] John Lafferty and Guy Lebanon, ‘Information diffusion kernels’, *NIPS*, 391–398, (2003).

[19] Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang, ‘Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge’, in *Findings of EMNLP*, pp. 1204–1214, (2021).

[20] Shimin Li, Hang Yan, and Xipeng Qiu, ‘Contrast and generation make bart a good dialogue emotion recognizer’, in *AAAI*, volume 36, pp. 11002–11010, (2022).

[21] Zheng Lian, Bin Liu, and Jianhua Tao, ‘Ctnet: Conversational transformer network for emotion recognition’, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **29**, 985–1000, (2021).

[22] 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).

[23] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria, ‘Dialoguernn: An attentive rnn for emotion detection in conversations’, in *AAAI*, volume 33, pp. 6818–6825, (2019).

[24] Yishu Miao, Edward Grefenstette, and Phil Blunsom, ‘Discovering discrete latent topics with neural variational inference’, in *ICML*, pp. 2410–2419. PMLR, (2017).

[25] Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang, ‘Topic modeling with wasserstein autoencoders’, in *ACL*, pp. 6345–6381, (2019).

[26] Weizhi Nie, Rihao Chang, Minjie Ren, Yuting Su, and Anan Liu, ‘I-gcn: Incremental graph convolution network for conversation emotion detection’, *IEEE Transactions on Multimedia*, 4471–4481, (2021).

[27] Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati, ‘Gpt3-to-plan: Extracting plans from text using gpt-3’, *arXiv preprint arXiv:2106.07131*, (2021).

[28] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al., ‘Training language models to follow instructions with human feedback’, *NIPS*, **35**, 27730–27744, (2022).

[29] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency, ‘Context-dependent sentiment analysis in user-generated videos’, in *ACL*, pp. 873–883, (2017).

[30] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea, ‘Meld: A multimodal multi-party dataset for emotion recognition in conversations’, in *ACL*, pp. 527–536, (2019).

[31] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy, ‘Emotion recognition in conversation: Research challenges, datasets, and recent advances’, *IEEE Access*, **7**, 100943–100953, (2019).

[32] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi, ‘Atomic: An atlas of machine commonsense for if-then reasoning’, in *AAAI*, volume 33, pp. 3027–3035, (2019).

[33] Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan, ‘Directed acyclic graph network for conversational emotion recognition’, in *EMNLP-IJCNLP*, pp. 1551–1560, (2021).

[34] Shang-Yu Su, Pei-Chieh Yuan, and Yun-Nung Chen, ‘How time matters: Learning time-decay attention for contextual spoken language understanding in dialogues’, in *NAACL*, pp. 2133–2142, (2018).

[35] Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri, ‘Octis: comparing and optimizing topic models is simple!’, in *EACL*, pp. 263–270, (2021).

[36] Geng Tu, Bin Liang, Dazhi Jiang, and Ruifeng Xu, ‘Sentiment-emotion-and context-guided knowledge selection framework for emotion recognition in conversations’, *IEEE Transactions on Affective Computing*, (2022).

[37] Geng Tu, Bin Liang, Ruibin Mao, Min Yang, and Ruifeng Xu, ‘Context or knowledge is not always necessary: A contrastive learning framework for emotion recognition in conversations’, in *Findings of ACL*, pp. 14054–14067, (2023).

[38] Geng Tu, Jintao Wen, Hao Liu, Sentao Chen, Lin Zheng, and Dazhi Jiang, ‘Exploration meets exploitation: Multitask learning for emotion recognition based on discrete and dimensional models’, *Knowledge-Based Systems*, **235**, 107598, (2022).

[39] Laurens Van der Maaten and Geoffrey Hinton, ‘Visualizing data using t-sne.’, *Journal of machine learning research*, **9**(11), (2008).

[40] Jinxiong Xia, Cao Liu, Jiansong Chen, Yuchen Li, Fan Yang, Xunliang Cai, Guanglu Wan, and Houfeng Wang, ‘Dialogue topic segmentation via parallel extraction network with neighbor smoothing’, in *SIGIR*, pp. 2126–2131, (2022).

[41] J. Xu, H. Wang, Z. Niu, H. Wu, and W. Che, ‘Knowledge graph grounded goal planning for open-domain conversation generation’, *AAAI*, **34**(5), 9338–9345, (2020).

[42] Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu, ‘Consert: A contrastive framework for self-supervised sentence representation transfer’, in *ACL-IJCNLP*, pp. 5065–5075, (2021).

[43] Sayyed M Zahiri and Jinho D Choi, ‘Emotion detection on tv show transcripts with sequence-based convolutional neural networks’, in *Workshops at AAAI*, (2018).

[44] Jichuan Zeng, Jing Li, Yan Song, Cuiyun Gao, Michael R Lyu, and Irwin King, ‘Topic memory networks for short text classification’, in *EMNLP*, pp. 3120–3131, (2018).

[45] Weixiang Zhao, Yanyan Zhao, and Xin Lu, ‘Cauain: Causal aware interaction network for emotion recognition in conversations’, in *IJCAI*, pp. 4524–4530, (2022).

[46] Hanxun Zhong, Zhicheng Dou, Yutao Zhu, Hongjin Qian, and Ji-Rong Wen, ‘Less is more: Learning to refine dialogue history for personalized dialogue generation’, in *NAACL*, pp. 5808–5820, (2022).

[47] Peixiang Zhong, Di Wang, and Chunyan Miao, ‘Knowledge-enriched transformer for emotion detection in textual conversations’, in *EMNLP-IJCNLP*, pp. 165–176, (2019).

[48] Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He, ‘Topic-driven and knowledge-aware transformer for dialogue emotion detection’, in *ACL-IJCNLP*, pp. 1571–1582, (2021).