Topic-Enriched Variational Transformer for Conversational Emotion Detection

Jiamin Luo, Jingjing Wang(✉), and Guodong Zhou

School of Computer Science and Technology, Soochow University, Suzhou, China
20204027003@stu.suda.edu.cn, {djingwang,gdzhou}@suda.edu.cn

Abstract. Conversational Emotion Detection (CED), spanning across multiple modalities (e.g., textual, visual and acoustic modalities), has been drawing ever-more interest in the multi-modal fields. Previous studies consistently consider the CED task as an emotion classification problem utterance by utterance, which largely ignore the global topic information of each conversation, especially the multi-modal topic information inside multiple modalities. Obviously, such information is crucial for alleviating the emotional information deficiency problem in a single utterance. With this in mind, we propose a **T**opic-enriched **V**ariational **T**ransformer (TVT) approach to capture the conversational topic information inside different modalities for CED. Particularly, a modality-independent topic module in TVT is designed to mine topic clues from either the discrete textual content, or the continuous visual and acoustic contents in each conversation. Detailed evaluation shows the great advantage of TVT to the CED task over the state-of-the-art baselines, justifying the importance of the multi-modal topic information to CED and the effectiveness of our approach in capturing such information.

Keywords: Multi-modal Topic Information · Topic-enriched Variational Transformer · Conversational Emotion Detection

1 Introduction

Conversational Emotion Detection (CED) has emerged as a crucial area of research, which aims to predict the emotions expressed by speakers during conversational interactions. CED holds significant potential for diverse applications, such as opinion mining, health care, and intelligent assistants. Thus, CED has garnered interest among researchers for exploring this promising task in the fields of natural language processing (NLP) [1,2] and multi-modal processing [6,7].

Previous studies have primarily focused on leveraging RNN-variants (e.g., LSTM or GRU) [2] or graph-based [3] models to encode contextual information during conversations. However, these approaches ignore the global topic information present within these utterances, though such information obviously could mitigate the challenge of RNN-variants or graph-based models in modeling long-range dependencies. As illustrated in Fig. 1, there is a conversation

D. F. Wong et al. (Eds.): NLPCC 2024, LNAI 15363, pp. 3–15, 2025.
https://doi.org/10.1007/978-981-97-9443-0_1

Fig. 1. A conversational example generated by two speakers to illustrate the importance of multi-modal topic information, including each utterance (text, visual frame, and acoustic spectrum) and its emotion.

between two speakers. "*hernia*", "*doctor*" and "*stomach*" words indicate a conversational topic related to *disease*, which often reflects negative emotions (i.e., *fear*, *sad*). Besides, the visual frames (*fear faces*) and acoustic spectrums (*low tones*) also express negative emotions. There, we argue that incorporating the multi-modal topic information from text, visual frames and acoustic spectrums could potentially enable the capture of the overall emotional orientation during the conversational interactions, thereby contributing to improved emotion detection performance.

However, conventional latent topic models [4,5] have primarily focus on processing the text modality, comprising discrete textual signals (i.e., words), which assume that each topic corresponds to a multinominal distribution over the vocabulary. On this basis, we consider leveraging such topic models to mine topic information inside visual and acoustic modalities, which are composed of continuous signals. Therefore, an appropriate topic model should be able to capture not only the textual topic information present in discrete text, but also the visual and acoustic topic information present in continuous visual frames and acoustic spectrums for CED.

To tackle the above challenges, we propose a **T**opic-enriched **V**ariational **T**ransformer (TVT) approach to capture the multi-modal topic information. Specifically, we design a modality-independent topic model capable of capturing topic information in both discrete textual signals and continuous visual and acoustic signals. Furthermore, we incorporate these topic representation with the conversation representation, and employ a transformer architecture to model contextual information during conversations. Finally, we jointly train three modality-independent topic models and the transformer model using a multi-task learning framework. Experimental results demonstrate that our TVT app-

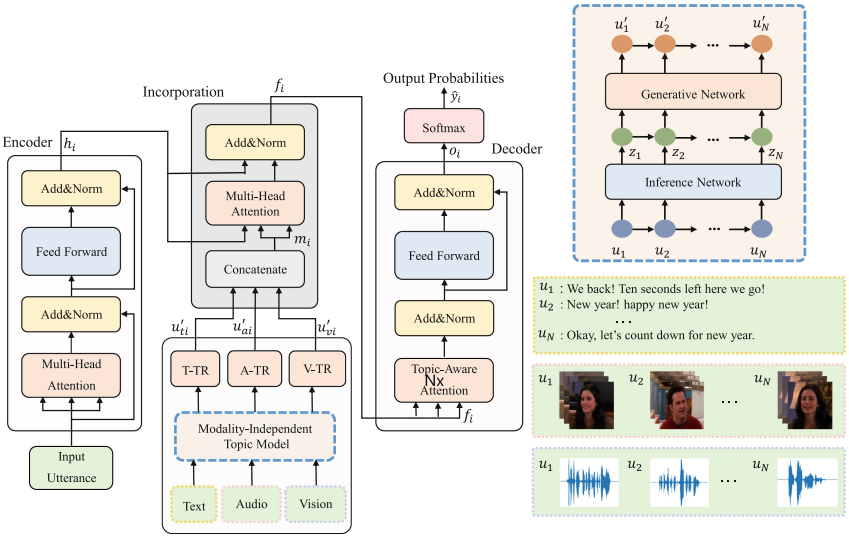

Fig. 2. The overall structure of our Topic-enriched Variational Transformer (TVT) approach. Here, TR is the abbreviation of Topic Representation.

roach can significantly outperforms the state-of-the-art CED baselines, including those focused solely on textual modality.

2 Topic-Enriched Variational Transformer (TVT)

In this section, we formulate the CED task as follows. Let $U = [u_1, u_2, ..., u_N]$ be a conversation, where N is the number of utterances. Each utterance u_i contains sources of utterance-aligned data corresponding to multiple modalities, i.e., text ($\boldsymbol{t}$), vision ($\boldsymbol{v}$) and acoustic ($\boldsymbol{a}$). The task of CED aims to predict the emotion label $\hat{y}_i$ for each utterance u_i. In this paper, we propose a Topic-enriched Variational Transformer (TVT) approach with three modality-independent topic models to mining multi-modal topic information. Figure 2 shows the overall architecture of our TVT approach, consisting of four major blocks: 1) Encoder; 2) Modality-Independent Topic Model; 3) Incorporation; 4) Decoder. Before introducing our TVT approach, we firstly have an overview of the basic transformer.

2.1 Basic Transformer

Transformer [20], comprised of an encoder and decoder, is often utilized to capture long-range dependencies, as is commonly encountered in neural machine translation. In this sense, transformer is well-suited to modeling the contextual information necessary for learning conversational representations.

Encoder is built by stacking identical layers, each consisting of two sub-layers. The first comprises a self-attention module, while the second comprises a fully connected feed-forward network. To improve the stability of training, residual connections are applied to both sub-layers, and layer normalization is performed. Using these sub-layers and operations, we can obtain the contextual representation h_i of each utterance u_i with the following formulas:

$$h_i = Encoder(u_i) \tag{1}$$

Decoder in transformer architecture shares the same identical stack of sub-layers as the encoder. However, unlike the encoder, the first sub-layer in decoder is a topic-aware attention[1] module. With this modification, the decoder is capable of generating output representation o_i after the incorporation representation f_i with the following formulas:

$$o_i = Decoder(f_i) \tag{2}$$

where f_i represents the incorporation of contextual representation h_i and multi-modal topic representation m_i (details in Sect. 2.3).

2.2 Modality-Independent Topic Model

Unlike traditional neural topic model [4,5] that primarily generate a discrete bag-of-words representation for input text, our modality-independent topic model focuses on generating intermediate modal topic representations, namely textual-, visual- and acoustic-topic representation (T-TR,V-TR,A-TR). By encoding different modal sequences into a unified type of representation, our proposed topic model can be considered modality-independent. Similar to Miao et al. [17], our topic model also employs the variational auto-encoder architecture. The workflow of our topic model, comprising the inference network and the generative work, is illustrated in Fig. 2.

Inference Network is leveraged to map the utterance u_i to a low-dimension latent space representation z_i. Specifically, the inference network comprises two different fully connected layers f_μ and f_σ, which estimate the mean μ_i and standard deviation σ_i of a Gaussian distribution for each time step in the utterance u_i, denoted as follows:

$$\mu_i = f_\mu(u_i), \sigma_i = \log f_\sigma(u_i) \tag{3}$$

The mean μ_i and standard deviation σ_i are then used to parameterize the diagonal Gaussian distribution $q(z_i|u_i)$, which serves as a variational approximation to the actual posterior distribution of the latent representation z_i given the

[1] Topic-aware attention is a multi-head attention mechanism that employs the incorporation of contextual representation and multi-modal topic representation as query, key and value. It is named topic-aware attention because it incorporates rich topic information.

utterance u_i. We can sample $\hat{z}_i$ from $q(z_i|u_i)$ using a reparameterization trick, i.e., $\hat{z}_i = \mu_i + \epsilon\sigma_i$. Here, ϵ is sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

Generative Network is leveraged to generative the modal topic representation $u_i^{'}$ from the latent representation z_i at each time step. Specifically, the generative network takes the sampled latent representation z_i as input and generates the reconstructed $u_i^{'}$ of the utterance u_i as the modal topic representation. Furthermore, the generative network simultaneously learns to approximate the conditional probability distribution $p(u_i^{'}|z_i)$. Similar to the inference network, the generative network also employs Gaussian distributions for both generative prior and the variational distribution.

To minimize the reconstruction error between the original utterance u_i and the reconstructed $u_i^{'}$, we introduce Kullback-Leibler (KL) divergence to minimize the difference between the prior topic distribution $q(z_i|u_i)$ and conditional probability distribution $p(u_i^{'}|z_i)$, denoted as follows:

$$\mathcal{L} = KL[q(z_i|u_i)||p(u_i^{'}|z_i)] \tag{4}$$

2.3 Incorporation

Based on the proposed modality-independent topic model, we extend our TVT approach by incorporating three topic models, aiming at capturing the topic information present in textual, visual and acoustic modalities. Specifically, the textual/visual/acoustic sequence is first encoded into a textual/visual/acoustic utterance, i.e., u_{ti}, u_{vi} and u_{ai}, which is then fed into a modality-independent topic model to obtain the textual/visual/acoustic topic representation, i.e., $u_{ti}^{'}$, $u_{vi}^{'}$ and $u_{ai}^{'}$. After obtaining the topic representation for each modality, we compute the multi-modal topic representation m_i by combining the modal topic representation of three modality-independent topic models, denoted as follows:

$$m_i = u_{ti}^{'} \oplus u_{vi}^{'} \oplus u_{ai}^{'} \tag{5}$$

2.4 Training

We employ the joint loss function to simultaneously optimize the emotion detection task and three modality-independent topic models. Specifically, we leverage cross-entropy loss for CED, i.e., $\mathcal{L}_{CED}$, denoted as follows:

$$\mathcal{L}_{CED} = -\frac{1}{C}\sum_{i=1}^{C}\sum_{j=1}^{N} y_{ij} \log \hat{y}_{ij} \tag{6}$$

where C and N represent the total number of conversations and utterances, respectively. y_{ij} and $\hat{y}_{ij}$ represent the ground-truth and predicted emotion label of utterance j in conversation i, respectively.

In addition, the loss function for our topic model is defined as shown in Eq. 4. For clarity, we present the individual losses for three modality-independent topic

Table 1. Comparison of our TVT approach with several state-of-the-art CED approaches on IEMOCAP dataset. TF and TM are the abbreviation of transformer and topic model, respectively. Bold font denotes the best performance.

Mod	Approach	Hap	Sad	Neu	Ang	Exc	Fru	Acc	W-F1
T	bc-LSTM	34.43	60.87	51.81	56.73	57.59	58.92	55.21	54.95
	KET	31.11	74.61	53.95	62.17	65.29	60.64	60.20	59.56
	COSMIC	47.40	80.40	63.77	64.16	61.63	66.84	65.19	65.20
	TodKat	16.48	76.92	62.09	51.66	**73.89**	61.95	63.22	61.33
	TVT	43.68	**82.02**	64.64	60.35	71.60	**68.77**	**67.59**	**67.21**
	-w/o BERT	49.10	79.75	65.03	57.77	70.69	57.38	64.26	64.32
	-w/o TF	**55.71**	76.89	62.13	62.13	66.93	64.22	65.00	65.16
	-w/o TM	50.97	80.08	**66.88**	**64.98**	62.75	64.72	66.05	66.00
MM	TFN	33.70	68.60	55.10	64.20	62.40	61.20	58.80	58.50
	bc-LSTM	34.75	76.48	55.19	64.44	61.60	61.74	60.38	60.28
	CMN	32.60	72.90	56.20	64.60	67.90	63.10	61.90	61.40
	DialogueRNN	33.48	**83.33**	56.40	64.44	**73.15**	58.50	63.52	62.85
	ICON	32.80	74.40	60.60	**68.20**	68.40	66.20	64.00	63.50
	TVT	**55.62**	79.35	**68.03**	63.37	72.06	64.83	**67.96**	**68.14**
	-w/o BERT	51.90	79.35	62.45	60.00	64.24	**66.90**	65.31	65.18
	-w/o TF	56.03	77.73	64.29	63.95	67.31	64.24	65.99	66.09
	-w/o TM	54.15	80.08	65.86	64.35	66.80	65.87	66.91	66.99

models, denoted as $\mathcal{L}_{text}$, $\mathcal{L}_{vision}$ and $\mathcal{L}_{acoustic}$. Finally, the joint loss $\mathcal{L}_{total}$ is defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{CED} + \lambda(\mathcal{L}_{text} + \mathcal{L}_{vision} + \mathcal{L}_{acoustic}) \tag{7}$$

where λ represents a weight to achieve a balance between the conversational emotion detection and three modality-independent topic models, and we fine-tune the value to be 0.25.

3 Experimental Settings

3.1 Dataset

We conduct experiments on both IEMOCAP[2] and MELD[3] datasets to verify the effectiveness of TVT. **IEMOCAP** [21] consists of 5,810 and 1,623 utterances in training and test sets, respectively. Utterances are annotated with one of six emotion labels: *happy*, *sad*, *neutral*, *angry*, *excited* and *frustrated*. Since no validation set is provided, we randomly select 10% samples from the training

[2] https://sail.usc.edu/iemocap/iemocap_publication.htm.
[3] https://affective-meld.github.io/.

set as validation set. **MELD** [22] consists of 9,989, 1,109 and 2,610 utterances in training, validation and test sets, respectively. MELD contains textual and acoustic modalities. Besides, each utterance is annotated with one of seven emotion labels: *joy*, *anger*, *fear*, *disgust*, *sadness*, *surprise* and *neutral*.

Table 2. Comparison of our TVT approach with several state-of-the-art CED approaches on MELD dataset.

Mod	Approach	Neu	Sur	Fea	Sad	Joy	Dis	Ang	Acc	W-F1
T	bc-LSTM	76.52	41.79	0	7.73	48.96	0	36.19	58.12	54.26
	COSMIC	79.37	58.24	14.49	**40.20**	**64.84**	20.00	51.98	66.05	65.32
	TodKat	80.10	56.86	**17.91**	38.55	63.32	**25.53**	52.69	67.24	65.47
	TVT	**87.88**	56.58	8.22	33.93	55.54	6.45	**60.00**	**68.39**	**67.91**
	-w/o BERT	83.26	55.29	14.63	31.87	55.36	12.17	53.87	64.52	64.80
	-w/o TF	84.43	**58.67**	3.03	25.08	57.84	14.18	55.38	67.01	65.60
	-w/o TM	84.94	56.86	11.63	24.77	58.93	10.53	57.85	67.28	66.19
MM	bc-LSTM	74.84	50.49	0	21.88	50.85	0	44.23	58.47	56.87
	CMN	74.30	47.20	0	23.40	44.70	0	44.70	-	55.50
	ICON	73.60	50.00	0	23.20	50.20	0	44.80	-	56.30
	DialogueRNN	76.50	48.50	0	24.00	51.18	0	44.99	60.15	57.78
	TVT	**89.33**	56.35	15.79	37.85	55.73	13.01	**61.52**	**69.69**	**69.43**
	-w/o BERT	79.33	**59.53**	**25.88**	**39.89**	**64.17**	**36.07**	53.72	66.78	66.18
	-w/o TF	84.43	58.56	20.41	32.88	59.93	20.75	54.57	68.00	66.93
	-w/o TM	87.91	58.94	18.39	29.66	59.41	13.64	53.41	68.23	67.93

3.2 Baselines

We compare TVT with the following state-of-the-art baselines in CED. **bc-LSTM** [1] uses a bi-directional LSTM to model contextual information. **TFN** [23] designs a multi-dimensional tensor to capture uni-, bi- and tri-modal interactions. **CMN** [7] uses multi-hop memory networks to learn the role of inter-speaker dependency relations. **ICON** [6] extends CMN by simultaneously combining global contextual information with the utterance. **DialogueRNN** [2] models the context of conversation via analyzing emotional status of each speaker. **KET** [24] utilizes external commonsense knowledge to interpret the contextual utterances. **COSMIC** [8] uses commonsense knowledge to learn interactions among interlocutors. **TodKat** [25] designs a topic-augmented language model to capture latent topic information. Particularly, KET, COSMIC and TodKat only consider textual modality in CED presently.

3.3 Implement Details and Metrics

In our experiments, all hyper-parameters are tuned according to the validation set. Specifically, we leverage a pre-trained BERT-base[4] model to encode utterance, which is optimized by the Adam optimizer [27] with the initial learning rate 1e-4. Other parameters of BERT are following [26]. We set the dimension of transformer encoder to be 100 and adopt another Adam optimizer with the initial learning rate 1e−3. Besides, we set the dimension of topic embedding for three modalities to be 20. We train TVT for 60 epochs, and stop if the validation loss does not decrease for 8 consecutive epochs.

Flowing previous works [2], we evaluate performance by *Accuracy* (Acc) and *Weighted-Average F1-score* (W-F1). Besides, we also report the F1-score of each emotion. Moreover, t-test is used to evaluate the significance of performance [28].

4 Results and Analysis

4.1 Experimental Results

Tables 1 and 2 show the comparative results on both IEMOCAP and MELD datasets, respectively. From these two tables, we can see that: 1) Under the multimodal setting, TVT outperforms these state-of-the-art baselines. For example, TVT outperforms ICON by achieving the average improvement of 3.96% (Acc) and 4.64% (W-F1) on IEMOCAP (p-value < 0.05). Besides, TVT significantly outperforms DialogueRNN by achieving the average improvement of 9.54% (Acc) and 10.03% (W-F1) on MELD (p-value < 0.01). Impressively, TVT outperforms TFN by 9.16% (Acc) and 9.64% (W-F1) on IEMOCAP, and TVT outperforms bc-LSTM by 11.22% (Acc) and 12.56% (W-F1) on MELD. Significance test shows that these improvements are all significant (p-value < 0.01). These results demonstrate the effectiveness of TVT, which also highlights the importance of multi-modal topic information in CED. 2) In the realm of the textual-modal setting, TVT outperforms knowledge-based baselines. For example, TVT outperforms TodKat by achieving the average improvement of 4.37% (Acc), 5.88% (W-F1) on IEMOCAP, and 1.15% (Acc), 2.44% (W-F1) on MELD (p-value < 0.05). These results again highlights the importance of topic information in conversations.

Moreover, we also report the performance of each individual emotion. TVT shows great potential in detecting *fear* and *disgust* emotions, which are challenging to distinguish in MELD. These results demonstrate the potential of TVT in enhancing the detection of such intricate emotions.

4.2 Ablation Analysis of TVT

To comprehensively evaluate the effectiveness of TVT, we conduct several ablation analysis, as shown in Tables 1 and 2. According to these results, we can see

[4] https://github.com/google-research/bert.

Table 3. Comparison of the influence of various modal topic used in TVT on IEMOCAP and MELD datasets. ✓ means we use this modal topic.

Modal Topic Information			IEMOCAP		MELD	
Textual	Acoustic	Visual	Acc	F1	Acc	F1
✓			67.10	67.16	68.12	68.18
	✓		66.79	66.64	67.55	66.93
		✓	65.31	65.18	-	-
✓	✓		67.96	67.99	**69.69**	**69.43**
✓		✓	66.48	66.49	-	-
	✓	✓	65.93	65.95	-	-
✓	✓	✓	**67.96**	**68.14**	-	-

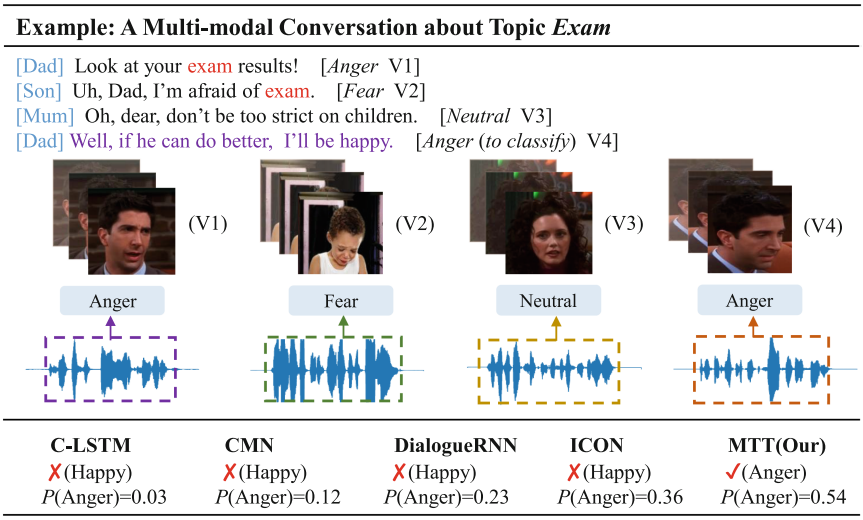

Fig. 3. A multi-modal conversational example, along with each utterance (including text, visual frame and acoustic spectrum) and the probabilities of ground-truth emotion *anger* on V4 predicted by different approaches.

that: 1) Compared to TVT without BERT, TVT achieves the average improvement of 2.96% (W-F1), 3.25% (W-F1) on IEMOCAP and MELD, respectively. This encourage us to use BERT as encoder. 2) The removal of transformer decreases the performance by 2.05% (W-F1), 2.50% (W-F1) on IEMOCAP and MELD, respectively. This demonstrates that leveraging transformer could better model the contextual information. 3) When integrating topic models, TVT consistently improves performance on IEMOCAP and MELD, with 1.15% (W-F1), 2.00% (W-F1) respectively. This demonstrates the importance of multi-modal

topic information, and encourages us to capture such information to enhance the performance of CED.

4.3 Effectiveness Analysis of Various Modal Topics

To provide further insights into the performance of TVT, we conduct experiments to analyze the influence of various combinations of modal topic. Table 3 summarizes the results of TVT integrated with various modal topic (i.e., *textual*/*acoustic*/*visual* topic). From this table, we can see that: 1) Integrating single topic performs better than the state-of-the-art baselines. For example, TVT with textual topic information outperforms ICON by 3.1% (Acc), 3.66% (W-F1) on IEMOCAP. This demonstrates that incorporating single modal topic is helpful to detection emotions. 2) The incorporating of textual topic performs better than acoustic and visual topic. This may caused by using BERT as textual encoder. 3) Incorporating multi-modal topic performs better than single and double modal topic. This again demonstrates the importance of multi-modal topic information, and leverages such information could enhance the performance of CED.

4.4 Qualitative Analysis

Figure 3 shows a multi-modal conversational example around the *exam* topic in daily life, together with the probabilities of the ground-truth emotion *anger* via different approaches. From this figure, we can see that: 1) Though speaker *Dad* in V4 has positive emotions (e.g., "*better*", "*happy*"), the whole conversation expresses negative emotions (e.g., "*afraid*", *fear face*, *angry tones*). This encourages us to consider topic information during the conversation. 2) Compared with other state-of-the-art baselines, TVT can give correct emotion prediction (i.e., *anger*) on V4. This again demonstrates the effectiveness of TVT, and encourages us to incorporate multi-modal topic information for CED.

5 Related Work

5.1 Conversational Emotion Detection

Previous studies of CED focus on capturing contextual information, which can be broadly categorized into RNN-, graph- and knowledge-based approaches. **RNN-based.** bc-LSTM [1] uses a bi-directional LSTM to capture contexts in conversations. Subsequently, CMN [7] leverages a memory-based sequential learning model for multi-view emotion detection, which is further enhanced by ICON [6] considering the role of inter-speaker relations. Besides, Majumder et al. [2] introduce party and global states to capture the dynamics of emotions. **Graph-based.** DialogGCN [3] leverages self- and inter-speaker dependency to model conversational contexts. RGAT [9] incorporates relational position encoding to enhance the relation-aware graph attention network. Besides, MMGCN [10] models multi-modal dependencies and incorporates speaker information. **Knowledge-based.** KET [24] utilizes hierarchical self-attention and

external knowledge to interpret contextual utterances. COSMIC [8] explores commonsense knowledge to better understand aspects of conversations. Besides, DialogXL [13] leverages the pre-trained language model to incorporate external knowledge. However, these approaches are limited to using textual modality.

In summary, all the above studies ignore the multi-modal topic information in conversations, which obviously could enhance the performance of CED.

5.2 Neural Topic Models

Recently, conventional topic models [15] have been leveraged to capture latent semantic topics. Inspired by the architecture of VAE [16], Miao et al. [17] propose the neural topic models to mine topic information inside texts. For example, Wang et al. [4] apply adversarial training for avoiding improper prior over latent topic space. Jin et al. [5] jointly reconstruct sentence and document word counts bag-of-words topical and pre-trained semantic embeddings. Besides, Wang et al. [18] learn topic-enriched representations in customer service via capturing textual topic information. An et al. [19] focus on depression detection and capture topic information in both images and texts.

Different from all above studies, we focus on conversational task and leverage three topic models to capture topic information from both discrete textual modality and continuous visual and acoustic modalities.

6 Conclusion

In this paper, we propose a **T**opic-enriched **V**ariational **T**ransformer (TVT) approach to capture multi-modal topic information in CED. Specifically, we leverages three modality-independent topic models to mine the topic information from textual, visual and acoustic modalities. Experimental results demonstrate the importance of multi-modal topic information and effectiveness of TVT. In the future, we would like to extend TVT to other multi-modal tasks, such as multi-modal recommendation, retrieval, etc. Furthermore, we would like to mine more useful information, such as speaker personality, to assist in CED tasks.

Acknowledgements. This work was supported by three NSFC grants, i.e., No.62006166, No.62376178 and No.62076175. This work was also supported by a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). Also, we would like to thank the anonymous reviewers for their helpful comments.

References

1. Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., Morency, L.P.: Context-dependent sentiment analysis in user-generated videos. In: Proceedings of ACL 2017, pp. 873–883. Vancouver (2017)
2. Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A.F., Cambria, E.: DialogueRNN: an attentive RNN for emotion detection in conversations. In: Proceedings of AAAI 2019, pp. 6818–6825. Honolulu, Hawaii (2019)
3. Ghosal, D., Majumder, N., Poria, S., Chhaya, N., Gelbukh, A.F.: DialogueGCN: a graph convolutional neural network for emotion recognition in conversation. In: Proceedings of EMNLP 2019, pp. 154–164. Hong Kong (2019)
4. Wang, R., et al.: Neural topic modeling with bidirectional adversarial training. In: Proceedings of ACL 2020, pp. 340–350. Online (2019)
5. Jin, Y., Zhao, H., Liu, M., Du, L., Buntine, W.L.: Neural attention-aware hierarchical topic model. In: Proceedings of EMNLP 2021, pp. 1042–1052. Virtual Event/Punta Cana, Dominican Republic (2021)
6. Hazarika, D., Poria, S., Mihalcea, R., Cambria, E., Zimmermann, R.: ICON: interactive conversational memory network for multimodal emotion detection. In: Proceedings of EMNLP 2018, pp. 2594–2604. Brussels (2018)
7. Hazarika, D., Poria, S., Zadeh, A., Cambria, E., Morency, L.P., Zimmermann, R.: Conversational memory network for emotion recognition in dyadic dialogue videos. In: Proceedings of NAACL 2018, pp. 2122–2132. Louisiana (2018)
8. Ghosal, D., Majumder, N., Gelbukh, A.F., Mihalcea, R., Poria, S.: COSMIC: COmmonSense knowledge for eMotion Identification in Conversations. In: Findings of EMNLP 2020, pp. 2470–2481. Online Event (2020)
9. Ishiwatari, T., Yasuda, Y., Miyazaki, T., Goto, J.: Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In: Proceedings of EMNLP 2020, pp. 7360–7370. Online (2020)
10. Hu, J., Liu, Y., Zhao, J., Jin, Q.: MMGCN: multimodal fusion via deep graph convolution network for emotion recognition in conversation. In: Proceedings of ACL 2021, pp. 5666–5675. Virtual Event (2021)
11. Shen, W., Wu, S., Yang, Y., Quan, X.: Directed acyclic graph network for conversational emotion recognition. In: Proceedings of ACL 2021, pp. 1551–1560. Virtual Event (2021)
12. Zhao, W., Zhao, Y., Qin, B.: MuCDN: mutual conversational detachment network for emotion recognition in multi-party conversations. In: Proceedings of COLING 2022, pp. 7020–7030. Gyeongju, Republic of Korea (2022)
13. Shen, W., Chen, J., Quan, X., Xie, Z.: DialogXL: all-in-one XLNet for multi-party conversation emotion recognition. In: Proceedings of AAAI 2021, pp. 13789–13797. Virtual Event (2021)
14. Hu, D., Wei, L., Huai, X.: DialogueCRN: contextual reasoning networks for emotion recognition in conversations. In: Proceedings of ACL 2021, pp. 7042–7052. Virtual Event (2021)
15. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. In: Proceedings of NeurIPS 2001, pp. 601–608. Vancouver (2001)
16. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. In: Proceedings of ICLR 2014. Banff (2014)
17. Miao, Y., Grefenstette, E., Blunsom, P.: Discovering discrete latent topics with neural variational inference. In: Proceedings of ICML 2017, pp. 2410–2419. Sydney (2017)

18. Wang J.C., et al.: Sentiment classification in customer service dialogue with topic-aware multi-task learning. In: Proceedings of AAAI 2020, pp. 9177–9184. New York (2020)
19. An, M., Wang, J., Li, S., Zhou, G.: Multimodal topic-enriched auxiliary learning for depression detection. In: Proceedings of COLING 2020, pp. 1078–1089. Online Event (2020)
20. Vaswani, A., et al.: Attention is all you need. In: Proceedings of NeurIPS 2017, pp. 5998–6008. Long Beach, CA (2017)
21. Busso, C., et al.: IEMOCAP: interactive emotional dyadic motion capture database. Lang. Resour. Eval. **42**(4), 335–359 (2008)
22. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R.: MELD: a multimodal multi-party dataset for emotion recognition in conversations. In: Proceedings of ACL 2019, pp. 527–536. Florence (2019)
23. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P.: Tensor fusion network for multimodal sentiment analysis. In: Proceedings of EMNLP 2017, pp. 1103–1114. Copenhagen (2017)
24. Zhong, P., Wang, D., Miao, C.: Knowledge-enriched transformer for emotion detection in textual conversations. In: Proceedings of EMNLP 2019, pp. 165–176. Hong Kong (2019)
25. Zhu, L., Pergola, G., Gui, L., Zhou, D., He, Y.: Topic-driven and knowledge-aware transformer for dialogue emotion detection. In: Proceedings of ACL 2021, pp. 1571–1582. Virtual Event (2021)
26. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL 2019, pp. 4171–4186. Minneapolis (2019)
27. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: Proceedings of ICLR 2015. San Diego, CA, USA (2015)
28. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: Proceedings of SIGIR 1999, pp. 42–49. Berkeley, CA, USA (199)