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

doi:10.3233/FAIA230453

FATRER: Full-Attention Topic Regularizer for Accurate and Robust Conversational Emotion Recognition

Yuzhao Mao[a], **Di Lu**[b;*], **Yang Zhang**[a] **and Xiaojie Wang**[c]

[a]Artificial Intelligence Center, Dasouche Inc, China
[b]Pattern Recognition Center, Wechat AI, Tencent Inc, China
[c]Beijing University of Posts and Telecommunications, China

Abstract. This paper concentrates on the understanding of interlocutors' emotions evoked in conversational utterances. Previous studies in this literature mainly focus on more accurate emotional predictions, while ignoring model robustness when the local context is corrupted by adversarial attacks. To maintain robustness while ensuring accuracy, we propose an emotion recognizer augmented by a full-attention topic regularizer, which enables an emotion-related global view when modeling the local context in a conversation. A joint topic modeling strategy is introduced to implement regularization from both representation and loss perspectives. To avoid over-regularization, we drop the constraints on prior distributions that exist in traditional topic modeling and perform probabilistic approximations based entirely on attention alignment. Experiments show that our models obtain more favorable results than state-of-the-art models, and gain convincing robustness under three types of adversarial attacks. Code: https://github.com/ludybupt/FATRER

Figure 1: An example of adversarial attacks in ERC

1 Introduction

ERC (Emotion Recognition in Conversations) [27, 34] aims to identify hidden emotion states by mining utterance expressions in conversations, and each utterance conveys one type of emotion, such as happiness, anger, sadness, etc. ERC is undeniably essential for enabling a humanoid robot to be empathetic with human emotions [24, 26]. In addition, ERC technologies can be easily transferred to other dialogue understanding tasks [10, 45] for further research.

Emotion dynamics modeling is an important criterion for the ERC task [20]. Specifically, emotion dynamics assumes that the interlocutors' emotions are influenced by two factors in a conversation: self and interpersonal dependencies [29]. Studies [28, 36, 51, 22] formulate the two factors in a way of hierarchical context modeling for more accurate emotional predictions. However, model robustness [44] is rarely touched, especially robustness against adversarial examples. Adversarial examples [18, 35] shown in Figure 1 are small and intentional worst-case perturbations to test samples that have high confidence in fooling a target classifier. Compared to speech recognition errors, robustness to adversarial examples can explain the reliability and safety of a target model.

Adversarial attacks arise mainly from local perturbations, and the fundamental issue with weak robustness stems from an excessive dependence on local context. Therefore, a global view is needed to withstand the impact of adversarial attacks. In addition, the global view should correlate with hidden emotional states so that accuracy is also ensured. To this end, we try to employ topic models [4] to help achieve accurate and robust emotion recognition in conversations. The topic is an important latent semantic factor related to hidden emotional states and is reported to be effective for more accurate emotional predictions in previous studies [42, 41, 51]. However, these topic models represent documents as topic variables, which only indicate the importance of topics, lacking a global view to ensure robustness. Therefore, the key problem lies in how to enable an emotion-related global view through topic modeling, and how to conduct proper joint modeling to fully exploit the advantages of the emotion-related global view to strike a balance between accuracy and robustness. In this paper, we propose FATRER (**F**ull-**A**ttention **T**opic **R**egularizer augmented conversational **E**motion **R**ecognizer) for the ERC task and share our insights on the above issues.

First, our topic modeling interprets document generation as a two-stage process, which is

$$p(w|d) = \sum_z p(w|z)p(z|d). \tag{1}$$

Here, based on the Law of Total Probability, the conditional probability of words given a document (utterance), $p(w|d)$, is decomposed into topic-word conditional probabilities, $p(w|z)$, and document-topic conditional probabilities, $p(z|d)$. Traditional topic models only use $p(z|d)$ to represent a document, while ignoring to exploit $p(w|z)$ that is associated with the global vocabulary. The vocabulary is nat-

* Corresponding Author. Email: dylu@tencent.com.

urally robust to local perturbations. Thus, based on the generation process in equation (1), an utterance can be represented as a combination of topic embedding by $p(z|d)$, and $p(w|z)$ is used to combine the whole word embedding into the topic embedding. Through such a two-stage weighted linear combination, we obtain a topicalized representation of an utterance, which is tightly coupled with the entire word embedding and thus has the global views to ensure robustness.

Secondly, the global view obtained by topic modeling should not only guarantee robustness but also enhance accuracy. To achieve this, the topics discovered should be emotion-related, i.e. the top words of a topic should relate to a specific emotion, which can lead to more accurate emotional predictions. Thus, we design a training loss including a classification loss term and a topic-oriented regularization term to optimize a full-attention model. The classification loss term is used to model emotion labels. This term drives the attention model to assign more weight to words that are related to the target emotion labels. The topic-oriented regularization term guides $p(w|d)$ to be as similar as possible to the observed word distribution of a given utterance. As we use the attention mechanism to approximate $p(z|d)$ and $p(w|z)$, regularization of $p(w|d)$ will propagate to attention alignment for $p(z|d)$ and $p(w|z)$ and increase sparsity of the two distributions. To sum up, the full-attention framework and the two terms in the training loss help the topics discovered to be discriminative and emotion-related, which are properties towards more accurate emotional predictions.

Finally, we use our FATR (Full-Attention Topic Regularizer) to enable a global view when modeling the local context for the ERC task. Regularization from both representation and loss is conducted. To avoid over-regularization, we drop the constraints on prior distributions that exist in traditional topic modeling. Specifically, we perform our topic modeling to represent an utterance from a global view, obtaining the topicalized representation. On the other hand, we conduct context modeling for representing an utterance from the local view, generating contextualized representation. Such multi-view modeling for representing interlocutor-specific utterances constitutes our approach to self-dependency[1] modeling. Then, the multi-view representations of interlocutors at each conversation turn form a time-series sequence. Through deep layers of sequence interactions for modeling interpersonal dependency[2], the multi-view representations are fully balanced and mixed in the final representation. Together with our training loss, the full-attention model is guided toward accurate and robust emotional predictions.

Experimental results show that our models achieve new SOTA (State-Of-The-Art) on four benchmark datasets in terms of Micro F1, and obtain convincing robustness under three types of adversarial attacks. Analysis and visualizations are conducted to better understand our models.

Our contributions can be summarized as:

- We propose a novel full-attention topic regularizer to enable an emotion-related global view when modeling local context for recognizing emotions in conversations.
- Our joint topic modeling strategy provides regularization from both representation and loss perspectives for accurate and robust emotional predictions.
- We conduct a series of experiments in terms of not only generalization but also robustness to evaluate the performance of our proposed models and baselines on the ERC task.

[1] Self dependency is known as emotion inertia manifested as the interlocutor's tendency to maintain their emotional state during the conversation.

[2] Interpersonal dependency is an emotional factor from other interlocutors in a conversation that tries to change the emotion of the current interlocutor.

2 Related Work

Emotion Recognition in Conversations. Prior research on ERC mostly focuses on exploring different context models, such as Bi-directional LSTM [32], memory network [14], Transformer [28], etc. Later, emotion dynamics [20] is summarized to guide the context modeling in ERC. Following this criterion, many hierarchical structures [27, 11, 28] are putting forward to capture the self and inter-personal dependencies in conversations. Inspired by the success of the pre-training paradigm [7], approaches [13, 36] try to implicitly memorize external knowledge into a large number of parameters via the language model pre-training. Other studies [50, 9] explicitly extract commonsense knowledge via the ConceptNet [37]. Very recently, [51] proposes a two-stage learned topic-driven knowledge-aware Transformer whose topic module is removed in the fine-tuning stage. Differently, we perform a joint topic modeling strategy with consistent topic guidance. Our strategy not only boosts accuracy but also guarantees robustness.

Topic Model. Typically, a statistical topic model [15, 4] captures topics in the form of latent variables with probability distributions over the entire vocabulary and performs approximate inference over document-topic and topic-word distributions through Variational Bayes [3]. However, such a learning paradigm requires an expensive iterative inference step performed on every document in a corpus [30]. The efficiency is boosted after the introduction of VAE-based (Variational AutoEncoder) neural topic model [2, 49] because variational inference can be performed through a single forward pass [19]. The VAE-based topic models often hypothesize a prior distribution [38, 8] that is used to approximate the posterior for latent document-topic distribution by maximizing the Evidence Lower Bound (ELBO) using the reparametrization trick. Differently, our strategy of topic modeling does not have prior constraints. Recently, pre-training models are used to augment neural topic models via enhanced text representation [2], knowledge distilling [16], or embedding clustering [39]. Besides VAEs, there are other frameworks of neural topic models including autoregressive models [12], and graph neural networks [47]. Our topic modeling is fully based on attention alignments.

Neural Topic Enhanced Supervised Model. Fitting unsupervised topic representations is sub-optimal for the supervised task. Research has been conducted to build a joint learning framework that integrates neural topic models with supervised learning models [6, 31, 42, 43, 41]. [6] proposes a simple feed-forward net with a max-margin loss to approximate topic-word distribution and a classification loss for the downstream task. [31] proposes an attention-based topic module without a regularization term. [41] uses document-topic distribution as just a vector to guide the attentive pooling of classification vectors (hidden layers of recurrent nets). [42] enables the classification model to be aware of topic information through different kinds of topic inference in auxiliary tasks, but there are no connections to link topic and classification models. [43] links a VAE-based topic model to a recurrent net for better representations. Our joint learning framework produces regularization from both representation and loss perspectives and can lead to more accurate and robust predictions.

3 Methodology

3.1 Topicalized Representation

The topicalized representation of an utterance is obtained via the full-attention topic modeling depicted in the bottom right of Figure 2. We first represent documents (utterances), topics, and words in the

embedding space. Then, we use attention alignments to approximate parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ for document-topic and topic-word multinomial distributions, respectively. Finally, we deploy $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as the weights to construct a topicalized representation of an utterance.

A document is represented as a linear combination of topics, and a topic is a linear combination of words. Specifically, let K be the topic number, V be the vocabulary size, and H be the embedding dimension. The topicalized representation of a document is calculated as $\sum_{k=1}^{K}\alpha_k\boldsymbol{z}_k$, or $\boldsymbol{\alpha}\boldsymbol{Z}^\top$ in matrix form, where $\boldsymbol{\alpha}=\{\alpha_1,\cdots,\alpha_K\}$ is the document-topic distribution over K topics given a document. $\boldsymbol{Z}\in\mathbb{R}^{K\times H}$ is the topic embedding matrix stacked from $\{\boldsymbol{z}_k\}_{k=1}^{K}$. The k-th topic embedding $\boldsymbol{z}_k\in\mathbb{R}^H$ is represented as $\sum_{v=1}^{V}\beta_{k,v}\boldsymbol{e}_v$, or $\boldsymbol{\beta}_k\boldsymbol{E}^\top$ in matrix form, where $\boldsymbol{\beta}_k=\{\beta_{k,1},\cdots,\beta_{k,V}\}$ is the topic-word distribution over V words given the k-th topic. $\boldsymbol{e}_v\in\mathbb{R}^H$ is the v-th word embedding from the embedding matrix $\boldsymbol{E}\in\mathbb{R}^{V\times H}$, where $\boldsymbol{E}$ is initialized from the embedding layer of BERT and is fine-tuned during training. In the following, z denotes the topic variable, and bold $\boldsymbol{z}$ or $\boldsymbol{Z}$ indicates the topic embedding or embedding matrix.

Let $p(z|d)\sim$ *Multinomial*$(\boldsymbol{\alpha})$, $p(w|z)\sim$ *Multinomial*$(\boldsymbol{\beta})$, our topic modeling output the document-word posterior distributions via probability inference in equation (1), which is

$$\begin{aligned} p(w|d) &= \sum_{z} p(z|d)p(w|z) \\ &= \sum_{k=1}^{K}\boldsymbol{\alpha}\boldsymbol{\beta}_k^\top = \boldsymbol{\alpha}\boldsymbol{\mathcal{B}}^\top. \end{aligned} \tag{2}$$

Here, $\boldsymbol{\mathcal{B}}=\{\boldsymbol{\beta}_k\}_{k=1}^{K}$ is the topic-word probability matrix. Note that, the attention alignment is natural conditional probabilities of the values given the queries. Thus, $\boldsymbol{\alpha}$ can be approximated via attention alignments between a document as the query and topics as the values. Similarly, $\boldsymbol{\beta}$ is from alignments between topics as the query and words as the values. The process is defined as,

$$\boldsymbol{\alpha} := att(\boldsymbol{r}^{bow}\rightarrow\{\boldsymbol{z}_k\}_{k=1}^{K}) = softmax(\frac{\boldsymbol{r}^{bow}\boldsymbol{Z}^\top}{\sqrt{H}}), \tag{3}$$

$$\boldsymbol{\beta_k} := att(\boldsymbol{z}_k\rightarrow\{\boldsymbol{e_v}\}_{v=1}^{V}) = softmax(\frac{\boldsymbol{z}_k\boldsymbol{E}^\top}{\sqrt{H}}). \tag{4}$$

Here, $\boldsymbol{r}^{bow}=\boldsymbol{\gamma}\boldsymbol{E}^\top$ is the bag-of-word combination of local word embedding. $\boldsymbol{\gamma}\in\mathbb{R}^V$ is the observed word distribution of a document. $\frac{1}{\sqrt{H}}$ is the scaling factor [40]. Through topic modeling, the representation of a document evolves from a combination of local terms, $\boldsymbol{\gamma}_i\boldsymbol{E}^\top$, to a combination of global topics, $\boldsymbol{\alpha}_i\boldsymbol{Z}^\top$.

Topic Embedding Chain Rule. Notice that, computing $\boldsymbol{z}$ needs $\boldsymbol{\beta}$ and computing $\boldsymbol{\beta}$ also needs $\boldsymbol{z}$. Thus, we deploy a chain rule that connects $\boldsymbol{z}$ and $\boldsymbol{\beta}$ along the training steps. Let T be the number of total training steps, and the chain rule can be formulated as

$$\{\boldsymbol{\beta}^0\mapsto\overbrace{\boldsymbol{z}^0\mapsto\boldsymbol{\beta}^1\mapsto\underbrace{\boldsymbol{z}^1}_{2nd\ training\ step}}^{1st\ training\ step}\mapsto\boldsymbol{\beta}^2\mapsto\boldsymbol{z}^2\cdots\overbrace{\boldsymbol{z}^{T-1}\mapsto\boldsymbol{\beta}^T\mapsto\boldsymbol{z}^T}^{Tth\ training\ step}\}, \tag{5}$$

Here, the computation of $\boldsymbol{z}^t$ at the t-th training step relies on $\boldsymbol{\beta}^t$, and $\boldsymbol{\beta}^t$ at the t-th training step is computed based on $\boldsymbol{z}^{t-1}$ from previous training step. $\boldsymbol{\beta}$ can be randomly initiated at the 0-th step, just like the attention alignments that are (almost) equally distributed on all keys at the early training stage. $\boldsymbol{z}$ cannot be random initialized first since computing $\boldsymbol{z}$ depends on $\boldsymbol{E}$ as well. We formulate the learning process in a recursive form,

$$\boldsymbol{\beta} = att(\tilde{\boldsymbol{z}}\rightarrow\{\boldsymbol{e}\}_{v=1}^{V}), \tag{6}$$

$$\boldsymbol{z} = \boldsymbol{\beta}\boldsymbol{E}^\top. \tag{7}$$

Here, the symbol $\tilde{}$ indicates that the variable comes from the previous training step, e.g., $\tilde{\boldsymbol{Z}}=\{\tilde{\boldsymbol{z}}_{\boldsymbol{k}}\}_{k=1}^{K}$ denotes the previously preserved topic embedding matrix.

Representation Underlying Different Topic Hypotheses. We offer two hypotheses, multi- and single-topic, to explain the generation process of a document. The two hypotheses correspond to the two versions of our models.

Multi-Topic: Each document describes a mixture of topics from an underlying topic set. In the multi-topic hypothesis, a document is represented as a weighted sum of the entire topic embedding. Since we have two topic embedding, $\tilde{\boldsymbol{Z}}$ and $\boldsymbol{Z}$, the model outputs two kinds of document-topic distribution, which is $\tilde{\boldsymbol{\alpha}}$, by attention alignments between $\boldsymbol{r}^{bow}$ and $\tilde{\boldsymbol{Z}}$, and $\boldsymbol{\alpha}$, by attention alignments between $\boldsymbol{r}^{bow}$ and $\boldsymbol{Z}$. To keep the learned representation stable, we define the following process to represent a document,

$$\boldsymbol{r}^{topic} = \sigma(\tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{Z}}^\top)\odot\xi(\boldsymbol{\alpha}\boldsymbol{Z}^\top) + \sigma(\boldsymbol{\alpha}\boldsymbol{Z}^\top)\odot\xi(\tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{Z}}^\top). \tag{8}$$

Here, $\tilde{\boldsymbol{Z}}$ can be understood as the topic embedding learned during the training steps (long-term info), $\boldsymbol{Z}$ is the updated topic embedding that affected by the current document (short-term info). The topicalized representation $\boldsymbol{r}^{topic}$ is learned in the form of additive gate unit [1] that balances the long-term and short-term info. σ and ξ are sigmoid and Leaky ReLU [46], respectively. $\odot$ denotes the element-wise product.

Single-Topic: Each document describes a single topic out of an underlying topic set. Under the single-topic hypothesis, a document is represented as one single topic embedding sampled from document-topic distribution. Let $\hat{\boldsymbol{\alpha}}\sim$*Multinomial*$(\boldsymbol{\alpha})$, $\hat{\tilde{\boldsymbol{\alpha}}}\sim$*Multinomial*$(\tilde{\boldsymbol{\alpha}})$, the topicalized representation is calculated as:

$$\boldsymbol{r}^{topic} = \sigma(\hat{\tilde{\boldsymbol{\alpha}}}\tilde{\boldsymbol{Z}}^\top)\odot\xi(\hat{\boldsymbol{\alpha}}\boldsymbol{Z}^\top) + \sigma(\hat{\boldsymbol{\alpha}}\boldsymbol{Z}^\top)\odot\xi(\hat{\tilde{\boldsymbol{\alpha}}}\tilde{\boldsymbol{Z}}^\top). \tag{9}$$

Here, $\hat{\boldsymbol{\alpha}}_i,\hat{\tilde{\boldsymbol{\alpha}}}_i\in\mathbb{R}^K$ are one-hot vectors in which the position of the sampled topic is set to one during training. For inference, we always choose the topic having the highest probability.

Finally, our topic modeling process is shortly denoted as:

$$\boldsymbol{\alpha},\boldsymbol{\mathcal{B}},\boldsymbol{r}^{topic} = f_{topic}(d,\Omega), \tag{10}$$

where d is the utterance and Ω is the entire vocabulary. The topicalized representation is the result of full attention on the overall word embedding and thus is less sensitive to local perturbations.

3.2 Contextualized Representation

The contextualized representation is obtained in the process of hierarchical emotion dynamics modeling. Let $\{\mathcal{D}_n\}_{n=1}^{N}$ be a corpus of N conversations. Each conversation $\mathcal{D}_n=\{(d_i^{\lambda_i},y_i)\}_{i=1}^{M_n}$ contains M_n utterance-emotion pair, where $d_i^{\lambda_i}$ is the i-th utterance spoken by interlocutor λ_i, and y_i is the corresponding emotion. The self dependency modeling captures emotional influence within an interlocutor. We use $f_{self}(d_i^{\lambda_i},c_i^{\lambda_i})$ to denote the self dependency modeling where $c_i^{\lambda_i}=\{d_\tau^{\lambda_\tau}|\tau\in[1,i),\lambda_\tau=\lambda_i\}$ is λ_i's historical utterances. The interpersonal dependency modeling, f_{inter}, combines the emotional influence across interlocutors. The complete emotion dynamics modeling forms a hierarchical structure, which is

$$f_{inter}\left(f_{self}(d_1^{\lambda_1},c_1^{\lambda_1}),f_{self}(d_2^{\lambda_2},c_2^{\lambda_2}),\cdots\right). \tag{11}$$

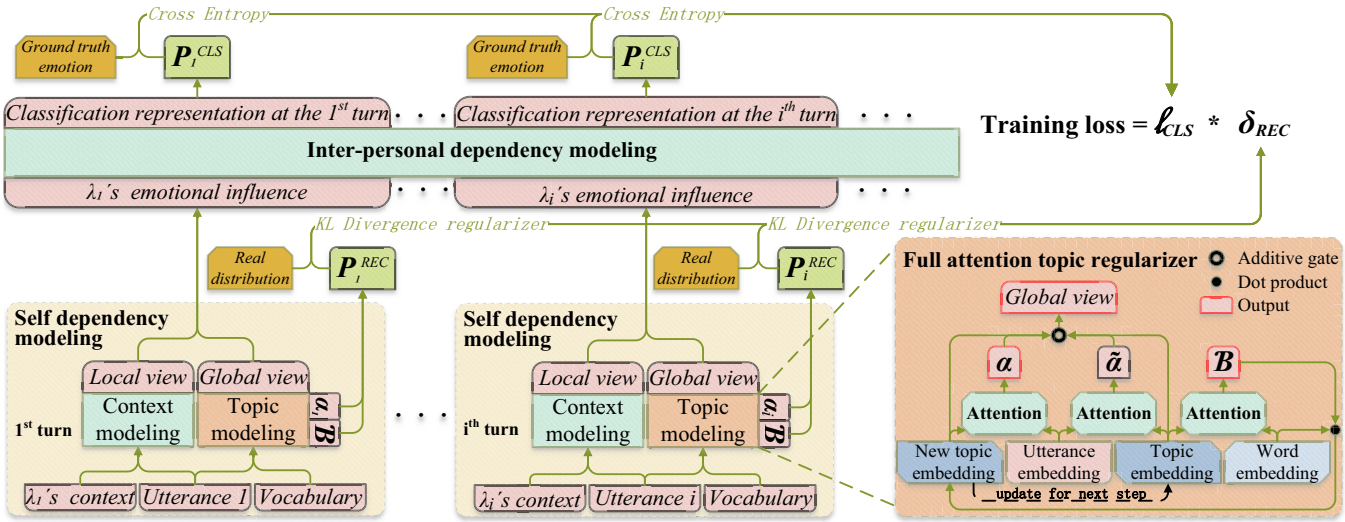


Figure 2: Joint modeling framework

We drop the superscript λ_i for simple in the following part.

Here, we use BERT [7] as the branches of the hierarchical structure to model f_{self}, then the branches are connected to a Transformer [40] backbone to model f_{inter}. Given an utterance of L words, $d = w_1 \cdots w_L$, and its historical context with L' words, $c = \omega_1 \cdots \omega_{L'}$, the input of the BERT can be denoted as,

$$\boldsymbol{X} = [\text{CLS}]\, w_1 \cdots w_L\, [\text{SEP}]\, \omega_1 \cdots \omega_{L'}\, [\text{SEP}], \tag{12}$$

where [CLS] and [SEP] are two special words with specific purposes in BERT. After feeding BERT with $\boldsymbol{X}$, it outputs the contextualized representations from the last hidden layer at [CLS] position, which is

$$\boldsymbol{r}^{context} = f_{self}(d, c) = \text{BERT}(\boldsymbol{X}). \tag{13}$$

The BERT encodes a M_n-turn conversation into a sequence of contextualized representations $\boldsymbol{R}=\{\boldsymbol{r}_i^{context}\}_{i=1}^{M_n}$. By feeding $\boldsymbol{R}$ to a TRM (TRansFormer), the interpersonal dependency is modeled as,

$$\begin{aligned} &f_{inter}(f_{self}(d_1, c_1), \cdots, f_{self}(d_i, c_i)) \\ &\quad = \text{TRM}(\boldsymbol{R}, \boldsymbol{\eta}_i); (\boldsymbol{\eta}_i = \underbrace{11\cdots1}_{i}\underbrace{00\cdots0}_{M_n-i}) \end{aligned} \quad . \tag{14}$$

Here, we use the last hidden layer of TRM at the i-th position as the classification representation for the i-th turn utterance. $\boldsymbol{\eta}_i$ masks the future information.

3.3 Joint Modeling

Given the i-th utterance d_i, corresponding context c_i of the same interlocutor, and a vocabulary set Ω, the process of topic modeling links d_i and Ω such that the topicalized representation has a global view, and the process of context modeling connects d_i with c_i to enable the contextualized representation with a local view. The regularization from the representation perspective is enforced by simply concatenating the two representations. Specifically, the modeling of emotion dynamics can be reformulated as,

$$f_{inter}\left(f'_{self}(d_1, c_1, \Omega), f'_{self}(d_2, c_2, \Omega), \cdots\right) \tag{15}$$

where f'_{self} is for self-dependency modeling that concatenates output of f_{topic} and the original f_{self} in equation (13) to form representations with both global and local views. Let r^{concat} be the concatenated representation output by f'_{self}, the interpersonal dependency is modeled via feeding the TRM with a representation sequence, $\boldsymbol{R} = \{r_i^{concat}\}_{i=1}^{M_n}$. After interpersonal dependency modeling, the global and local views can be fully balanced and mixed through deep layers of multi-head attention within the TRM. $\{\boldsymbol{h}_i\}_{i=1}^{M_n}$ are the output representation sequence for emotional predictions.

Our joint learning objective contains two terms, a classification loss term ℓ_{CLS} and a topic-oriented loss regularization term δ_{REC}. Let $\mathcal{Y} = \{1, \cdots, |\mathcal{Y}|\}$ be the emotion set. The classification loss is to minimize the cross entropy between a sequence of the predicted emotion probabilities $\{\boldsymbol{\mathcal{P}}_i^{CLS} \in \mathbb{R}^{|\mathcal{Y}|}\}_{i=1}^{M_n}$ and a sequence of the ground truth $\{y_i \in \mathcal{Y}\}_{i=1}^{M_n}$:

$$\boldsymbol{\mathcal{P}}_i^{CLS} = softmax(\boldsymbol{h}_i \boldsymbol{W}^\top + \boldsymbol{b}), \tag{16}$$

$$\ell_{CLS} = -\sum_{i=1}^{M_n} \log \mathcal{P}_{i,y_i}^{CLS}. \tag{17}$$

The topic-oriented loss regularization term is to minimize the KL (Kullback–Leibler) divergence between the document-word posterior distribution *Multinomial*($\boldsymbol{\mathcal{P}}_i^{REC}$) and the observed distribution *Multinomial*($\boldsymbol{\gamma}_i$):

$$\boldsymbol{\mathcal{P}}_i^{REC} = \boldsymbol{\alpha}_i \boldsymbol{\beta}^\top, \tag{18}$$

$$\delta_{REC} = \frac{1}{M_n} \sum_{i=1}^{M_n} \sum_{j=1}^{V} \gamma_{i,j} \log\left(\frac{\gamma_{i,j}}{\mathcal{P}_{i,j}^{REC}}\right), \tag{19}$$

where $\boldsymbol{\mathcal{P}}_i^{REC}$ is our approximated word probabilities conditioned on the i-th utterance in a conversation.

The regularization from the loss perspective is enforced by the product between the two loss terms, which is

$$\mathcal{L} = \mu \cdot \ell_{CLS} \cdot \delta_{REC}. \tag{20}$$

The values of these two terms serve as the learning rates for each other, which dynamically and interactively adjust the learning weights during the training process. μ is the global learning rate to control the convergent speed of the model.

4 Experimental Setup

Datasets. The benchmark of ERC involves four datasets, including *DailyDialog* [23], *IEMOCAP* [5], *MELD* [33], and *EmoryNLP* [48].

The metrics in **bold** indicate the best values obtained by our models. To focus on the role of topic, we exclude baselines using external resources such knowledge base. * refers to a TodKat variant that only performs topic modeling without using the knowledge base.

Models	Topic	IEMOCAP	MELD	DailyDialog	EmoryNLP
AGHMN *[17]*	×	58.28%	54.52%	51.90%	33.54%
DialGCN *[11]*	×	60.63%	56.17%	53.73%	33.13%
BERT *[7]*	×	-	63.49%	54.85%	41.11%
RoBERTa *[25]*	×	-	63.75%	54.33%	40.81%
DialTRM *[28]*	×	68.41%	64.60%	56.44%	37.13%
CoGBART *[22]*	×	64.10%	63.66%	54.71%	37.57%
TodKat* *[51]*	√	57.38%	61.11%	53.44%	32.62%
VAE (Laplace)	√	68.98%	65.36%	55.11%	39.43%
VAE (LogNormal)	√	68.95%	65.40%	56.27%	37.80%
VAE (Dirichlet)	√	69.19%	65.82%	56.53%	40.35%
VAE (Gamma)	√	68.48%	64.37%	56.34%	37.91%
FATRER-multi	√	**69.69%**	**66.28%**	**56.57%**	39.13%
FATRER-single	√	68.58%	65.32%	56.46%	**41.87%**

Table 1: Generalization results on four datasets

AAA denotes Accuracy After Attack. **U** denotes merely attacking the target utterance. **U+C** means attacking both the target utterance and its context. - indicates being incompatible with context perturbations.

	PWWS		TextFooler		TextBugger	
Models	AAA(**U+C**)	AAA(**U**)	AAA(**U+C**)	AAA(**U**)	AAA(**U+C**)	AAA(**U**)
CoGBART	-	42.57%	-	26.73%	-	44.36%
DialTRM	4.10%	42.77%	7.00%	22.38%	6.50%	38.61%
TodKat	-	28.32%	-	8.12%	-	22.97%
VAE (Laplace)	6.50%	43.56%	4.48%	22.18%	7.05%	39.21%
VAE (LogNormal)	2.63%	40.99%	2.57%	19.21%	6.12%	40.40%
VAE (Dirichlet)	7.52%	41.58%	13.05%	25.35%	15.57%	40.79%
VAE (Gamma)	4.49%	42.38%	13.42%	26.73%	12.54%	43.17%
FATRER-multi	**15.00%**	**46.14%**	**19.13%**	**31.88%**	**22.50%**	**44.75%**
FATRER-single	14.56%	45.74%	14.52%	23.37%	17.52%	43.56%

Table 2: Robustness to Adversarial Examples on IEMOCAP

The first two are dyadic conversational datasets, and the last two comprise multi-party conversations. *IEMOCAP* annotates each utterance with one of 6 emotion labels, and the emotion labels are 7 in the other datasets. We follow the standard split [51] for training, validation, and testing in the benchmark. The '*neutral*' label is not included in *DailyDialog*'s evaluation to avoid category imbalance. [9]

Adversarial Attacks. Three types of adversarial attacks [3], including *PWWS* [35], *TextFooler* [18], and *TextBugger* [21], are employed for robustness evaluation. *PWWS* offers a word-level attack according to the word saliency and the classification probability. *TextFooler* is reported to have effectiveness to attack pre-training models. *TextBugger* can execute both character-level and word-level attacks. We set all the attackers to perturb up to 25% of words per input.

Implementation Details. The single-topic and multi-topic hypotheses yield the *FATRER-single* and *FATRER-multi* versions of our model, respectively. The model described in section 3.2 is named as *Baseline*, in which the BERT is initialized from the $\text{BERT}_{\text{Base}}$ and the TRM is a random initialized, 6-layer, 12-head-attention, and 768-hidden-unit Transformer encoder. Following the spirit of ProdLDA [38], we implement VAE-based topic regularizers underlying different priors. By replacing our topicalized representation with topic variables obtained via variational inference, we have VAE-based variants, including *VAE (Laplace)*, *VAE (LogNormal)*, *VAE (Dirichlet)*, and *VAE (Gamma)*. Based on the same *Baseline*, we can perform a fair comparison between *FATR-based* and *VAE-based* topic regularizers. We use off-the-shelf implementations of *TodKat*[4] and *CoGBART*[5] as well as their trained models to perform robustness evaluations in our experimental settings.

5 Results, Analysis, and Visualization

5.1 Generalization on Benchmark Datasets

The generalization results are presented in Table 1. The listed methods can be divided into two groups according to whether or not latent topics are used. The methods using latent topics can be further categorized as 1) merely pre-trained topics, i.e., *TodKat*, 2) VAE-based joint learned topics, i.e., *VAE (Dirichlet)*, etc., and 3) our *FATR-based* topics. We employ Micro F1 to measure the accuracy of the listed models. The results show that *FATRER-multi* achieves the best performance on the benchmark in terms of Micro F1. Compared with *TodKat* which applies VAE-based topic modeling only in the pre-training stage, we understand the importance of performing joint topic modeling which provides consistent topic guidance for accurate predictions. Compared to adding VAE-based topic regularizers, we can see that our FATR can obtain better generalization results.

[3] https://github.com/QData/TextAttack

[4] https://github.com/something678/TodKat

[5] https://github.com/whatissimondoing/CoG-BART

Models	PWWS		TextFooler		TextBugger	
	AAA(**U+C**)	AAA(**U**)	AAA(**U+C**)	AAA(**U**)	AAA(**U+C**)	AAA(**U**)
CoGBART	-	28.50%	-	11.22%	-	23.96%
DialTRM	11.09%	28.91%	4.95%	11.29%	23.38%	35.54%
TodKat	-	28.29%	-	17.03%	-	43.96%
VAE (Laplace)	10.01%	27.72%	5.54%	11.68%	20.72%	33.47%
VAE (LogNormal)	12.28%	27.72%	5.74%	9.31%	22.82%	37.23%
VAE (Dirichlet)	13.47%	28.71%	8.71%	14.46%	25.32%	37.62%
VAE (Gamma)	15.05%	27.52%	11.29%	14.85%	24.92%	36.44%
FATRER-multi	**17.23%**	**31.68%**	**12.28%**	**17.43%**	**26.28%**	**38.61%**
FATRER-single	14.56%	30.89%	8.32%	15.05%	23.19%	37.43%

Table 3: Robustness to Adversarial Examples on MELD

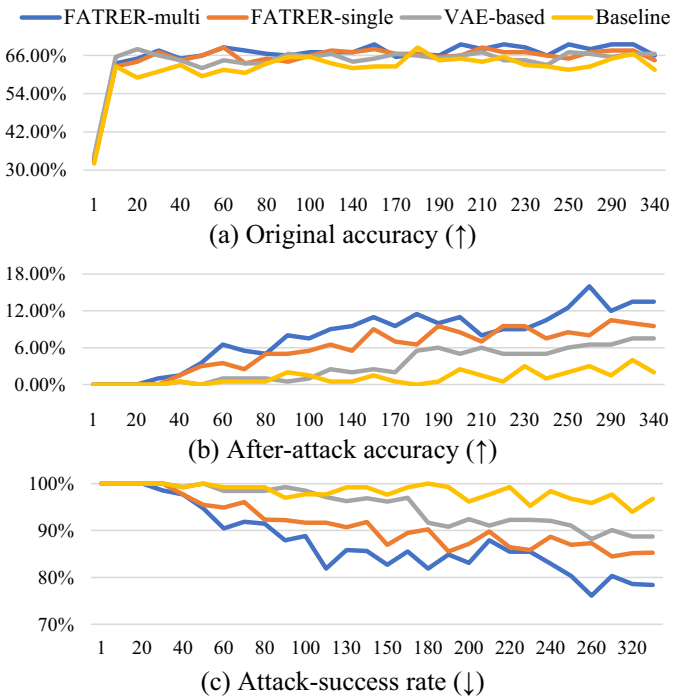


Figure 3: Results over training epochs on IEMOCAP

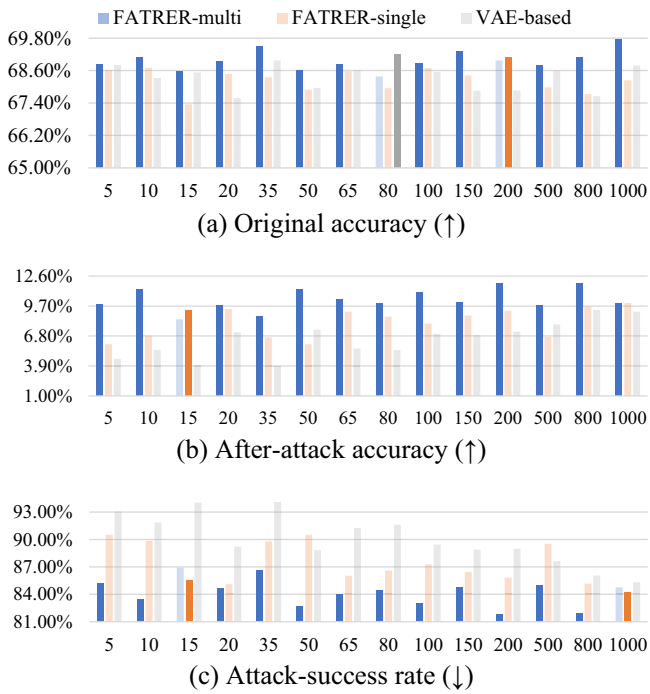


Figure 4: Results over the number of topics on IEMOCAP

5.2 *Robustness to Adversarial Examples*

Based on the fully converged models, the robustness results are presented in Table 2, 3, and Figure 3. The evaluation uses AAA (After-Attack Accuracy) and ASR (Attack-Success Rate) as the metrics, where AAA means the accuracy achieved by the target model on the adversarial samples crafted from the test samples. Specifically, AAA(**U**) denotes AAA of merely attacking target utterance, and AAA(**U+C**) means AAA of attacking both target utterance and its context. ASR is the proportion of adversarial samples that can successfully change the originally correctly predicted labels.

Table 2 and 3 present the robustness results on IEMOCAP and MELD, respectively. For a fair comparison, the ANQ (Average Number of attack Queries) for each model is tuned to be as equal as possible in this experiment. Our models achieve the best results under the three types of adversarial attacks on the two datasets. We conclude four points from the results:

1) Topic-oriented models, e.g., the VAE-based and our *FATRERs*, get better robustness performance, which means the joint modeling of topic and context already provides a certain degree of robustness.
2) Our *FATRERs* are superior to VAE-based models, which indicates that our topicalized representation improves robustness beyond the simple use of topic variables.
3) The robustness of our *FATRERs* is more evident when the conversational context is disrupted, further demonstrating the benefits of the global view behind our topicalized representations.
4) *TodKat* only applies VAE-based topic modeling in the pre-training stage and gets relatively weak performance, which means that without joint topic modeling, it is difficult to guarantee robustness.

In a word, the superior robustness of our *FATRERs* mainly comes from the topicalized representation guided by a global view and the full-attention-based joint modeling framework.

In Figure 3, we track the trend of after-attack accuracy, attack-success rate, and original accuracy (accuracy before the attack) over 340 training epochs. The best performed *VAE (Dirichlet)* is taken as the representative of VAE-based variants. In this experiment, to exert maximum pressure on the models, we let the attacker PWWS determine the number of attack queries and use **U+C** attack strategy. From the trend in Figure 3a, we understand that the generalization ability converges quickly within 10 epochs. By contrast, the model robust-

ness in Figure 3b and 3c keeps getting better until 300 epochs, which means the model robustness needs a long period of training. From the 200th to 300th training epoch, the ANQ of our models is around 800 which is 1.09 to 1.26 times that of the other models. Higher ANQ means the model is more difficult to be successfully attacked. The results show that our models obtain convincing robustness with greater pressures at every epoch of testing.

5.3 Analysis

Ablation Study. To better understand our models, we yield several variants of our model by removing some key components. Firstly, *FATRER-muti* and *FATRER-single* are two basic variants of our model. - *rep regularizer* indicates removing the topicalized representation and thus is identical to the *Baseline*. - *loss regularizer* means removing the topic-oriented regularization term in the training loss while preserving the attention network for topic modeling, and the topic relative distributions are driven by the classification loss. From the results in Table 4, we understand that removing either representation or loss regularizer leads to a drop in accuracy and robustness. The model robustness is more sensitive to the removal of the two regularizers than the accuracy.

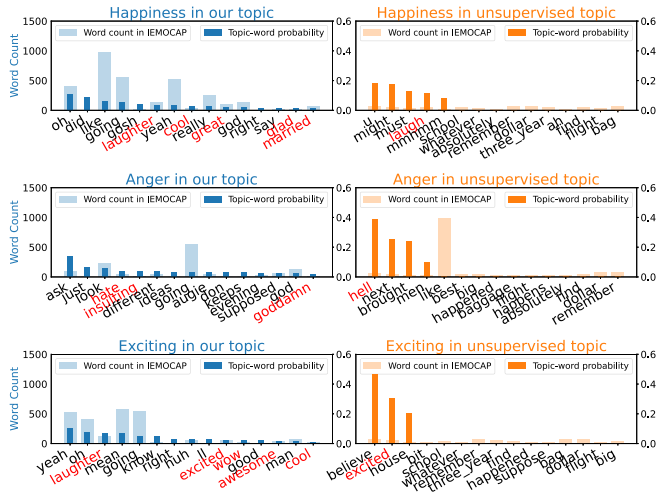
Datasets	Models	AAA(**U+C**)	AAA(**U**)	Micro F1
IEMOCAP	FATRER-multi	15.00%	46.14%	69.69%
	-rep regularizer	4.10%	42.77%	68.41%
	-loss regularizer	7.04%	45.15%	69.38%
	FATRER-single	14.56%	45.74%	68.58%
	-rep regularizer	4.10%	42.77%	68.41%
	-loss regularizer	9.03%	44.36%	68.58%
MELD	FATRER-multi	17.23%	31.68%	66.28%
	-rep regularizer	11.09%	28.91%	64.60%
	-loss regularizer	14.06%	31.09%	64.98%
	FATRER-single	14.65%	30.89%	65.32%
	-rep regularizer	11.09%	28.91%	64.60%
	-loss regularizer	13.27%	30.69%	65.19%

Table 4: Ablation study on IEMOCAP and MELD

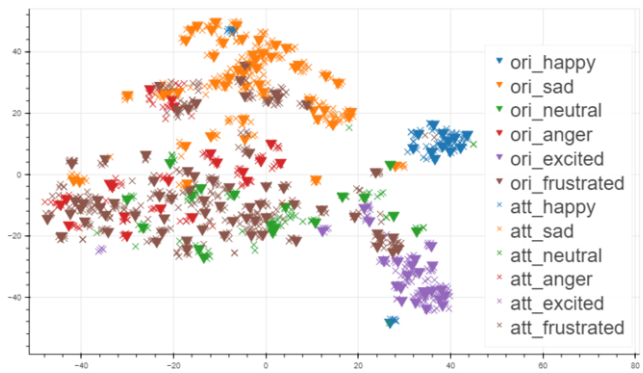
Effects of Topic Number. We explore the number of topics from 5 to 1000 to analyze their impact on our models. From the results shown in Figure 4, we can see that our models have better robustness performance than the *VAE-based* model over the selected numbers of topics. Our model and the VAE-based model achieve the best original accuracy at 80 and 1,000 topics, respectively. Note that, both the number and the content of the adversarial examples are changed for each epoch of testing. Thus, the robustness results shown in Figure 4b and 4c are based on the average of the scores obtained from the 200th to 300th epoch.

5.4 Visualization

Topic words visualization. We evaluate our topics' quality by comparison with unsupervised topics learned from LDA [4]. By inspecting the top 15 keywords, we depict three topics that are most relative to a kind of emotion in Figure 5a. We mark the emotional words in red. Benefiting from the supervision of emotion labels, our topic model can discover more emotional words than LDA. Such emotion-related topics can improve accuracy. The top words of our topics have perceptible probabilities while LDA's topics have less than 5



(a) Topic words visualization.



(b) Representation visualization.

Figure 5: Topic visualization.

valid keywords. Both frequent and rare words are learned in our topics. Such wide-range and emotion-related global views can guarantee robustness.

Representation visualization. To give an intuitive understanding of the adversarial attacks, we plot representations of the **ori**ginal and the **att**acked samples output from *FATRER-multi* in Figure 5b. We randomly choose 5 out of 800+ attacked samples crafted from each original sample. As shown in the figure, the representations of the attacked samples wrap tightly around the original samples. By the way, we notice that our model still has room for the discrimination of neutral, angry, and frustrated emotions, which is a clue for improving our model in the future.

6 Conclusion

We propose a topic-augmented conversational emotion recognizer, namely FATRER, for the task of ERC. To cope with the impact of local perturbations, the discovered topic is enabled with an emotion-related global view based on a full-attention framework. By jointly performing topic and context modeling, our full-attention topic regularizer augmented models can achieve accurate and robust conversational emotion recognition. Experimental results on standard benchmark datasets and adversarial examples have shown the generalization and robustness of our models. Ablation studies are conducted to help understand the effectiveness of our topic regularizers. We also provide visualization to give an intuitive explanation of our models.

References

[1] John Arevalo, Thamar Solorio, Manuel Montes-y Gomez, and Fabio A González, 'Gated multimodal networks', Neural Computing and Applications, 1–20, (2020).

[2] Federico Bianchi, Silvia Terragni, and Dirk Hovy, 'Pre-training is a hot topic: Contextualized document embeddings improve topic coherence', arXiv:2004.03974, (2020).

[3] David M Blei and Jon D McAuliffe, 'Supervised topic models', NeurIPS, (2008).

[4] David M Blei and Andrew Y et.al. Ng, 'Latent dirichlet allocation', the Journal of machine Learning research, **3**, 993–1022, (2003).

[5] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, et al., 'Iemocap: Interactive emotional dyadic motion capture database', Language resources and evaluation, **42**(4), 335, (2008).

[6] Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji, 'A novel neural topic model and its supervised extension', in AAAI, volume 29, (2015).

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 'Bert: Pre-training of deep bidirectional transformers for language understanding', in NAACL-HLT, pp. 4171–4186, (2019).

[8] Ran Ding, Ramesh Nallapati, and Bing Xiang, 'Coherence-aware neural topic modeling', in EMNLP, pp. 830–836, (2018).

[9] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, et al., 'COSMIC: COmmonSense knowledge for eMotion identification in conversations', in Findings of EMNLP, (2020).

[10] Deepanway Ghosal, Navonil Majumder, Rada Mihalcea, and Soujanya Poria, 'Utterance-level dialogue understanding: An empirical study', arXiv:2009.13902, (2020).

[11] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, et al., 'DialogueGCN: A graph convolutional neural network for emotion recognition in conversation', in EMNLP-IJCNLP, (November 2019).

[12] Pankaj Gupta, Yatin Chaudhary, Florian Buettner, and Hinrich Schütze, 'Document informed neural autoregressive topic models with distributional prior', in AAAI, volume 33, pp. 6505–6512, (2019).

[13] Devamanyu Hazarika, Soujanya Poria, et al., 'Emotion recognition in conversations with transfer learning from generative conversation modeling', arXiv:1910.04980, (2019).

[14] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, et al., 'Icon: interactive conversational memory network for multimodal emotion detection', in EMNLP, pp. 2594–2604, (2018).

[15] Thomas Hofmann, 'Probabilistic latent semantic indexing', in ACM SIGIR, pp. 50–57, (1999).

[16] Alexander Hoyle, Pranav Goel, and Philip Resnik, 'Improving neural topic models using knowledge distillation', arXiv:2010.02377, (2020).

[17] Wenxiang Jiao, Michael R Lyu, and Irwin King, 'Real-time emotion recognition via attention gated hierarchical memory network', arXiv:1911.09075, (2019).

[18] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits, 'Is bert really robust? natural language attack on text classification and entailment', arXiv:1907.11932, **2**, (2019).

[19] Diederik P Kingma and Max Welling, 'Auto-encoding variational bayes', arXiv:1312.6114, (2013).

[20] Peter Kuppens and Philippe Verduyn, 'Emotion dynamics', Current Opinion in Psychology, **17**, 22–26, (2017).

[21] Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang, 'Textbugger: Generating adversarial text against real-world applications', arXiv:1812.05271, (2018).

[22] 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).

[23] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu, 'Dailydialog: A manually labelled multi-turn dialogue dataset', in IJCAI, pp. 986–995, (2017).

[24] Zhaojiang Lin, Peng Xu, et al., 'Caire: An end-to-end empathetic chatbot', in AAAI, volume 34, pp. 13622–13623, (2020).

[25] Yinhan Liu, Myle Ott, Naman Goyal, et al., 'Roberta: A robustly optimized bert pretraining approach', arXiv:1907.11692, (2019).

[26] Yukun Ma, Khanh Linh Nguyen, Frank Z Xing, et al., 'A survey on empathetic dialogue systems', Information Fusion, **64**, 50–70, (2020).

[27] 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).

[28] Yuzhao Mao, Guang Liu, Xiaojie Wang, Weiguo Gao, and Xuan Li, 'Dialoguetrm: Exploring multi-modal emotional dynamics in a conversation', in Findings of EMNLP, pp. 2694–2704, (2021).

[29] Michael W Morris and Dacher Keltner, 'How emotions work: The social functions of emotional expression in negotiations', Research in organizational behavior, **22**, 1–50, (2000).

[30] Madhur Panwar, Shashank Shailabh, Milan Aggarwal, and Balaji Krishnamurthy, 'Tan-ntm: Topic attention networks for neural topic modeling', ACL, (2020).

[31] Gabriele Pergola, Lin Gui, and Yulan He, 'Tdam: A topic-dependent attention model for sentiment analysis', Information Processing & Management, **56**(6), 102084, (2019).

[32] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, et al., 'Context-dependent sentiment analysis in user-generated videos', in ACL, pp. 873–883, (2017).

[33] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, et al., 'Meld: A multimodal multi-party dataset for emotion recognition in conversations', in ACL, pp. 527–536, (2019).

[34] 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).

[35] Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che, 'Generating natural language adversarial examples through probability weighted word saliency', in ACL, pp. 1085–1097, Florence, Italy, (July 2019). Association for Computational Linguistics.

[36] Weizhou Shen, Junqing Chen, Xiaojun Quan, et al., 'Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition', in AAAI, volume 35, pp. 13789–13797, (2021).

[37] R Conceptnet Speer, '5.5: An open multilingual graph of general knowledge/r. speer', in AAAI, (2017).

[38] Akash Srivastava and Charles Sutton, 'Autoencoding variational inference for topic models', ICLR, (2017).

[39] Laure Thompson and David Mimno, 'Topic modeling with contextualized word representation clusters', arXiv:2010.12626, (2020).

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al., 'Attention is all you need', in NeurIPS, pp. 5998–6008, (2017).

[41] Chang Wang and Bang Wang, 'An end-to-end topic-enhanced self-attention network for social emotion classification', in Proceedings of The Web Conference 2020, pp. 2210–2219, (2020).

[42] Jiancheng Wang, Jingjing Wang, Changlong Sun, et al., 'Sentiment classification in customer service dialogue with topic-aware multi-task learning', in AAAI, volume 34, pp. 9177–9184, (2020).

[43] Xinyi Wang and Yi Yang, 'Neural topic model with attention for supervised learning', in International Conference on Artificial Intelligence and Statistics, pp. 1147–1156. PMLR, (2020).

[44] Xuezhi Wang, Haohan Wang, and Diyi Yang, 'Measure and improve robustness in nlp models: A survey', in NAACL-HLT, pp. 4569–4586, (2022).

[45] Sixing Wu, Ying Li, Dawei Zhang, et al., 'Topicka: Generating commonsense knowledge-aware dialogue responses towards the recommended topic fact', in IJCAI, pp. 3766–3772, (2021).

[46] Bing Xu, Naiyan Wang, et al., 'Empirical evaluation of rectified activations in convolutional network', arXiv:1505.00853, (2015).

[47] Liang Yang, Fan Wu, Junhua Gu, Chuan Wang, Xiaochun Cao, Di Jin, and Yuanfang Guo, 'Graph attention topic modeling network', in Proceedings of The Web Conference 2020, pp. 144–154, (2020).

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

[49] He Zhao, Dinh Phung, Viet Huynh, Trung Le, and Wray Buntine, 'Neural topic model via optimal transport', International Conference on Learning Representations, (2021).

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

[51] Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He, 'Topic-driven and knowledge-aware transformer for dialogue emotion detection', ACL, (2021).