Disentangled Variational Autoencoder for Emotion Recognition in Conversations

Kailai Yang, Sophia Ananiadou

National Centre for Text Mining,
Department of Computer Science, The University of Manchester
kailai.yang@postgrad.manchester.ac.uk, sophia.ananiadou@manchester.ac.uk

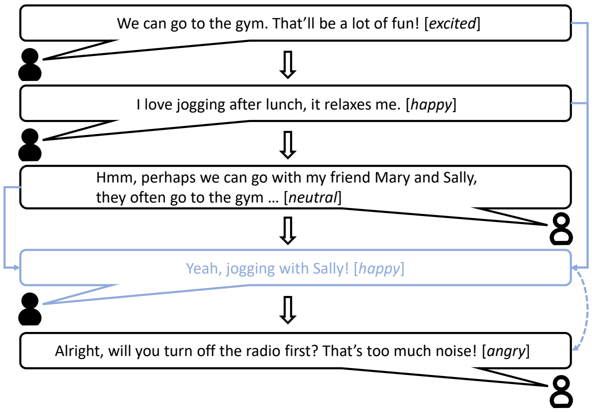
Abstract

In Emotion Recognition in Conversations, the emotions of target utterances are closely dependent on their context. Therefore, existing works train the model to reconstruct the response of the target utterance, which aims to recognize emotions considering diverse contexts. However, adjacent response generation ignores long-range dependencies and provides limited affective information in many cases. In addition, most ERC models learn a unified distributed representation for each utterance, which lacks interpretability and robustness. To address these issues, we propose a **VAD**-disentangled **V**ariational **A**uto**E**ncoder (VAD-VAE), which first introduces a target utterance reconstruction task based on Variational Autoencoder, then disentangles three affect representations Valance-Arousal-Dominance (VAD) from the latent space. We also enhance the disentangled representations by introducing VAD supervision signals from a sentiment lexicon and minimising the mutual information between VAD distributions. Experiments show that VAD-VAE outperforms the state-of-the-art model on three datasets. Further analysis proves the effectiveness of each proposed module and the quality of disentangled VAD representations. Our Code will be available online upon acceptance.

Introduction

Emotion Recognition in Conversations (ERC) aims to identify the emotion of each utterance within a dialogue from pre-defined emotion categories (Poria et al. 2019b). As an extension of traditional emotion detection from text, ERC attracts increasing research interest from the NLP community, since it is more suitable for usage in real-world scenarios such as empathetic dialogues systems (Ma et al. 2020), emotion-related social media analysis (Nandwani and Verma 2021; Chatterjee et al. 2019) and opinion mining from customer reviews (Zad et al. 2021; Wang et al. 2020a)

Unlike sentence-level emotion recognition, the emotion of each utterance is dependent on contextual information in ERC. Some works enhance the context modelling ability by training the model to reconstruct the dialogue. For example, Hazarika et al. (2021); Chapuis et al. (2020) pretrain the utterance encoders on large-scale conversation data and transfer the pre-trained weights to ERC. More recent works utilise Pre-trained Language Models (PLMs) (Qiu et al. 2020) to model the dialogue and avoid pre-training from scratch (Shen et al. 2021a; Xie et al. 2021). Li, Yan, and Qiu (2022) combine both methods by fine-tuning pre-trained BART (Lewis et al. 2020) with an auxiliary response generation task on the dialogue, which trains the model to generate the next sentence given the target utterance. This task aims to force the model to recognize emotions considering diverse contexts.



We can go to the gym. That'll be a lot of fun! [*excited*]

I love jogging after lunch, it relaxes me. [*happy*]

Hmm, perhaps we can go with my friend Mary and Sally, they often go to the gym … [*neutral*]

Yeah, jogging with Sally! [*happy*]

Alright, will you turn off the radio first? That's too much noise! [*angry*]

Figure 1: An example dialogue. Solid lines show the influence of previous utterances on the emotion of target utterance (marked blue), and dashed lines denote the sudden change of topic. As illustrated, the emotion of target utterance is dependent on long-range history. As the topic changes, response generation provides limited affective information.

However, response generation only mines the dependencies between two adjacent utterances, while the influence of long-range history on the target utterance is ignored. Reconstructing the next sentence also provides limited affective information for the target utterance in many cases, such as the sudden change of topic. An example is illustrated in Figure 1 to prove the above points. We argue that a context-aware reconstruction of the target utterance itself is more appropriate since ERC is centered on the target utterance representations. In addition, current ERC methods mostly learn a unified distributed representation for each target utterance. Though achieving impressive results, entangled features lack interpretability and robustness (Bengio, Courville, and Vincent 2013). The success of affective text generation models (Ma et al. 2020; Goswamy et al. 2020) also proves the viability of disentangling emotion features from content.

To address these issues, we propose a **VAD**-disentangled **V**ariational **A**uto**E**ncoder (VAD-VAE) for ERC. Firstly, instead of reconstructing the response, we introduce a target utterance reconstruction task based on the Variational Autoencoder (VAE) (Kingma and Welling 2014) generative model. We devise a PLM-based context-aware encoder to model the dialogue, and sample the latent representations from a Gaussian distribution estimated from the utterance representations. The Gaussian distribution also aims to regularise the latent space. Then another PLM-based decoder is leveraged to reconstruct the target utterance from the latent representations. VAD-VAE outperforms the state-of-the-art model on three ERC datasets.

Secondly, we utilise Disentangled Representation Learning (Higgins et al. 2018) techniques to disentangle key features from the utterance representations. Studies in affect representation models in psychology point out that Valance-Arousal-Dominance (VAD) are both orthogonal and bipolar, which are appropriate to define emotion states (Russell and Mehrabian 1977; Mehrabian 1995). Therefore, we propose to disentangle the three VAD features from the latent space of the VAE, where we also sample each of the VAD representations from a corresponding Gaussian distribution estimated from the utterance representations. Then the disentangled features are combined for both ERC and target utterance reconstruction tasks.

Thirdly, two techniques are used to enhance the disentangled VAD representations. We boost their *informativeness* (Eastwood and Williams 2018) by introducing supervision signals from NRC-VAD (Mohammad 2018), a sentiment lexicon that contains human ratings of VAD for all emotions. To enforce the *independence* (Higgins et al. 2018) of latent spaces, we minimise the Mutual Information (MI) between VAD representations. During training, we estimate and minimise the variational Contrastive Log-ratio Upper-Bound (vCLUB) (Cheng et al. 2020a) of MI. Further analysis proves the quality of disentangled VAD representations.

To summarise, this work mainly makes the following contributions:

- We propose a VAE-based target utterance reconstruction auxiliary task for ERC, which improves model performance and regularises the latent spaces.
- For the first time in ERC, We explicitly disentangle the three VAD features from the utterance representations. Analysis shows it benefits interpretability and bears potential in the affective text generation task.
- We enhance the *informativeness* of the disentangled representations with VAD supervision signals from the lexicon NRC-VAD, and minimise the vCLUB estimate of their mutual information to improve *independence*.

Related Work

Emotion Recognition in Conversations

For ERC, the emotion of a dialogue participant is largely influenced by the dialogue history, which makes context modelling a key challenge. Early works utilise Recurrent Neural Networks (RNN) to model each participant's dialogue flow as a sequence and revise them as memories at each time step (Hazarika et al. 2018a,b). Considering multi-party relations, Majumder et al. (2019) leverage another global-state RNN to model inter-speaker dependencies and emotion dynamics. To avoid designing complex model structures, more recent works leverage the strong context-modelling ability of PLMs to model the conversation as a whole (Li et al. 2020; Shen et al. 2021a). Some other works (Shen et al. 2021b; Li et al. 2021) build a graph upon the dialogue with each utterance as a node, and leverage graph neural networks to model ERC as a node-classification task.

Enhancing the utterance representations is also crucial for ERC. Some works manage to incorporate task-related information. For example, commonsense knowledge is introduced (Ghosal et al. 2020; Xie et al. 2021) to enrich the semantic space. To enhance the conversation modelling ability, some methods pre-train the model on large-scale conversation data and transfer the weights to ERC (Chapuis et al. 2020; Hazarika et al. 2021). Multi-task learning is also leveraged to introduce topic information (Wang et al. 2020b; Zhu et al. 2021), discourse roles (Ong et al. 2022) and speaker-utterance relations (Li et al. 2020) to aid emotion reasoning. Park et al. (2021); Mukherjee et al. (2021) incorporate VAD information to introduce fine-grained sentiment supervisions. Contrastive learning (Li, Yan, and Qiu 2022) is also devised to distinguish utterances with similar emotions.

Disentangled Representation Learning

Disentangled Representation Learning (DRL) aims to map key features of data into distinct and independent low-dimensional latent spaces (Higgins et al. 2018). Current DRL methods are mainly divided into unsupervised and supervised disentanglement. Early unsupervised methods mainly design constraints on the latent space to enforce the independence of each dimension, such as information capacity (Burgess et al. 2018) and mutual information gap (Chen et al. 2018). Supervised methods focus on introducing supervision signals to different parts of the latent space to enforce *informativeness*. Some works utilise ground-truth labels of the corresponding generative factors such as syntactic parsing trees (Bao et al. 2019) and style labels (Cheng et al. 2020b), while other works use weakly-supervised signals, including pairwise similarity between representations (Chen and Batmanghelich 2020) and semi-supervised ground-truth labels (Vasilakes et al. 2022). Still, supervised methods devise techniques such as mutual information minimisation (Vasilakes et al. 2022) and adversarial learning (Bao et al. 2019; John et al. 2019) to enforce *independence* and *invariance* (Shu et al. 2020) of the disentangled representations.

Methodology

Task Definition

ERC task is defined as follows: a dialogue D contains n utterances $\{u_1, u_2, ..., u_n\}$, with the corresponding ground-truth emotion labels $\{y_1, y_2, ..., y_n\}$, where $y_i \in E$, E is the pre-defined emotion label set. Each u_i contains m_i tokens: $\{u_i^1, u_i^2, ..., u_i^{m_i}\}$. The dialogue is also accompanied by a speakers list $S(D) = \{S(u_1), S(u_2), ..., S(u_n)\}$, where u_i is uttered by $S(u_i) \in S$, and S is the set of dialogue participants. With the above information, ERC aims to identify the emotion of each target utterance u_i, which is formalised as: $\hat{y}_i = f(u_i, D, S(D))$.

Target Utterance Reconstruction

This section introduces the target utterance reconstruction auxiliary task. Based on the context-aware utterance encoder, we also disentangle VAD latent representations from the utterance representations and build a VAE-based generative model to reconstruct the target utterance, which is the backbone of VAD-VAE, as illustrated in Figure 2.

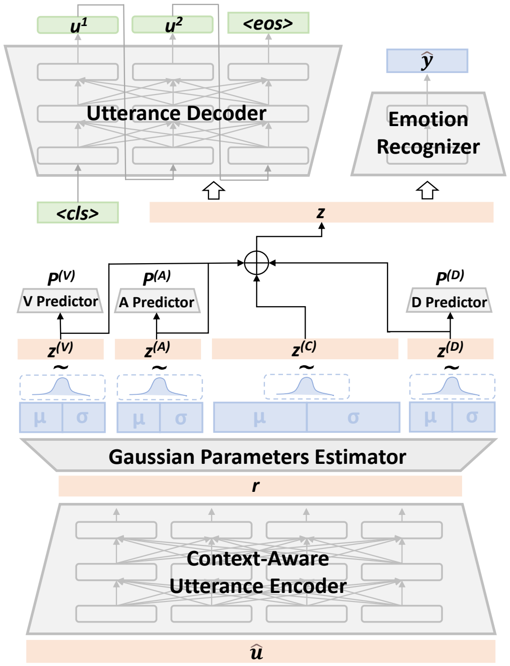


Figure 2: Main components of VAD-VAE. Each latent representation $z^{(\mathcal{R})}$ is sampled from a Gaussian distribution estimated from the context-aware utterance representation r. The VAD prediction $P^{(\mathcal{R})}$ is obtained from $z^{(\mathcal{R})}$ via the VAD predictors. The concatenated representation z is utilised for both ERC and target utterance reconstruction. In the utterance decoder, "$\langle cls \rangle$" denotes the start-of-sentence token, and "$\langle eos \rangle$" denotes the end-of-sentence token.

Context-Aware Utterance Encoder To explicitly introduce speaker information, we first prepend the speaker $S(u_j)$ to each utterance u_j. Then the target utterance u_i is concatenated with *both past and future dialogues* to obtain the context-aware input $\hat{u}_i$. Utilising an encoder, we obtain the context-aware utterance embeddings:

$$r_i = Encoder(\hat{u}_i) \tag{1}$$

where $Encoder$ denotes the RoBERTa-Large (Liu et al. 2019) utterance encoder, $r_i \in \mathbb{R}^{S \times D_h}$ is the utterance representations, S denotes the sequence length, and D_h is the hidden states dimension. We leverage the embedding of the start-of-sentence token at position 0: $r_i^{[CLS]} \in \mathbb{R}^{D_h}$ as the utterance-level representation of u_i.

VAE-based Generative Model We build a VAE-based generative model, and disentangle three latent features Valance-Arousal-Dominance (VAD) from the utterance representation, where Valance reflects the pleasantness of a stimulus, Arousal reflects the intensity of emotion provoked by a stimulus, and Dominance reflects the degree of control exerted by a stimulus (Warriner, Kuperman, and Brysbaert 2013). We also define a "Content" feature that controls the content generation of the target utterance.

A VAE is utilised to estimate this model, which imposes a standard Gaussian prior distribution on each latent space Z. The deterministic utterance representation is replaced with an approximation of the posterior $q_\phi(z|x)$, which is parameterised by a neural network. We utilise four feed-forward neural networks to map $x = r_i^{[CLS]}$ to four sets of Gaussian distribution parameters (μ, σ), which parameterise the latent distributions of Valance, Arousal, Dominance, and Content, denoted as $\mathcal{R} \in \{V, A, D, C\}$. For each feature, we sample the latent representation $z^{(\mathcal{R})}$ from the Gaussian distribution defined by the corresponding $(\mu^{(\mathcal{R})}, \sigma^{(\mathcal{R})})$ using the reparameterisation trick (Kingma and Welling 2014):

$$z_i^{(\mathcal{R})} = \mu^{(\mathcal{R})} \odot \sigma^{(\mathcal{R})} + \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{2}$$

where $z_i^{(\mathcal{R})} \in \mathbb{R}^{D_{(\mathcal{R})}}$, $D_{(\mathcal{R})}$ is the pre-defined latent space dimension. Then the latent representations are concatenated: $z_i = [z_i^V; z_i^A; z_i^D; z_i^C]$. z_i is used to initialise the decoder and reconstruct the target utterance:

$$u_i^j = Softmax(Decoder(z_i, u_i^{<j})) \tag{3}$$

where $Softmax$ denotes the softmax operation, u_i^j denotes the j-th generated tokens, and $u_i^{<j}$ denotes the previously generated tokens. We utilise BART-Large decoder (Lewis et al. 2020) as $Decoder$, since it shares the vocabulary with RoBERTa-Large in Huggingface (Wolf et al. 2019) implementations and is proved powerful in many generative tasks. As in standard VAE, we include a KL-divergence term for each latent space to keep the approximate posterior close to the prior distribution. During training, we utilise the Evidence Lower BOund (ELBO) as the training objective:

$$\begin{aligned} \mathcal{L}_{ELBO}(\phi, \theta) = -\mathbb{E}_{q_\phi(z_i|x)}\left[log\, p_\theta(x|z_i)\right] + \\ \sum_{\mathcal{R} \in \{V,A,D,C\}} \alpha_{\mathcal{R}} KL\left[q_\phi^{(\mathcal{R})}(z_i^{(\mathcal{R})}|x) || p(z_i^{(\mathcal{R})})\right] \end{aligned} \tag{4}$$

where ϕ and θ denote the parameters of the encoder and decoder, each $\alpha_{\mathcal{R}}$ weights the corresponding KL-divergence term, and standard Gaussian prior is used for each $p(z_i^{(\mathcal{R})})$.

Enhancing Disentangled VAD Representations

We aim to enhance the disentangled VAD representations considering the following two aspects: (a). *informativeness*: the representation should include enough information to predict the corresponding generative factor well (Higgins et al.

2017; Eastwood and Williams 2018). (b). *Independence*: for each generative factor the representation should lie in an independent latent space (Higgins et al. 2018). Therefore, we introduce supervision signals from a sentiment lexicon to enforce *informativeness* and a mutual information minimisation objective to enforce *independence*.

Informativeness To enhance the representation's ability to predict the corresponding generative factor, we introduce supervision signals from NRC-VAD (Mohammad 2018), a VAD sentiment lexicon that contains reliable human ratings of VAD for 20,000 English terms. All the terms in NRC-VAD denote or connote emotions, and are selected from commonly used sentiment lexicons and tweets. Each term is strictly annotated via best-worst scaling with crowdsourcing annotators, and an aggregation process calculates the VAD for each term ranging from 0 to 1. For example, the emotion *happiness* is assigned $vad_{happiness} = \{0.960, 0.732, 0.850\}$. More details about NRC-VAD are in Appendix A. With the pre-defined categorical emotion set E, we extract the VAD score $vad_{e_j} = \{vad^V_{e_j}, vad^A_{e_j}, vad^D_{e_j}\}$for each of the emotion $e_j \in E$ from NRC-VAD, where $j \in [1, |E|]$. Since fine-grained VAD supervision signals are introduced, we expect to improve both the *informativeness* of VAD representations and model performance on ERC.

Specifically, for each $\hat{\mathcal{R}} \in \{V, A, D\}$, we compute the corresponding prediction from the latent representation using a feed-forward neural network predictor:

$$P_i^{(\hat{\mathcal{R}})} = \frac{1}{1 + e^{-(z_i^{(\hat{\mathcal{R}})} W^{(\hat{\mathcal{R}})} + b^{(\hat{\mathcal{R}})})}} \tag{5}$$

where $W^{(\hat{\mathcal{R}})}$ and $b^{(\hat{\mathcal{R}})}$ are parameters of the predictor corresponding to $\hat{\mathcal{R}}$. As the training objective, we compute the mean squared error loss between the predictions and the supervision signals:

$$\mathcal{L}_{INFO}(\phi, \lambda) = \frac{1}{N} \sum_{i=1}^{N} \sum_{\hat{\mathcal{R}} \in \{V,A,D\}} (P_i^{(\hat{\mathcal{R}})} - vad_{y_i}^{(\hat{\mathcal{R}})})^2 \tag{6}$$

where ϕ and λ denote the parameters of the encoder and the predictor, y_i denotes the emotion label of i-th utterance, and N denotes the batch size.

Independence We improve the independence of all disentangled latent spaces by making their distributions as dissimilar as possible. A common method is to minimise the Mutual Information (MI) between each pair of spaces (Poole et al. 2019). However, MI is hard to calculate in high-dimensional spaces. The conditional distribution between each pair of latent variables is also unavailable in our cases. Therefore, we utilise the variational Contrastive Log-ratio Upper-Bound (vCLUB) (Cheng et al. 2020a) to estimate the MI. Since no extra supervision signals are introduced, we expect the model to still achieve comparable performance as a trade-off for more *independence* of each latent space.

Specifically, we separately use a feed-forward neural network as an estimator to approximate the conditional distribution between each pair in VAD variables: $P(\hat{\mathcal{R}}_i | \hat{\mathcal{R}}_j)$ where $i \neq j$, and the parameters are updated along with VAD-VAE at each time step. An unbiased vCLUB estimate between each pair $\hat{\mathcal{R}}_i$, $\hat{\mathcal{R}}_j \in \{V, A, D\}$ are summed to get the MI minimisation loss as the training objective:

$$\begin{aligned}\mathcal{L}_{MI}(\phi, \delta) = \frac{1}{N} \sum_{k=1}^{N} \sum_{i,j} \Big[& log\ q_{\delta_{ij}}(z_k^{(\hat{\mathcal{R}}_i)} | z_k^{(\hat{\mathcal{R}}_j)}) - \\ & \frac{1}{N} \sum_{l=1}^{N} log\ q_{\delta_{ij}}(z_l^{(\hat{\mathcal{R}}_i)} | z_k^{(\hat{\mathcal{R}}_j)}) \Big]\end{aligned} \tag{7}$$

where δ_{ij} denotes the parameters of the corresponding estimator. The detailed proof of Eqn. 7 is in Appendix B.

Model Training

For ERC task, the concatenated latent representation z_i is utilised to compute the final classification probability:

$$\hat{y}_i = Softmax(z_i W_0 + b_0) \tag{8}$$

where W_0 and b_0 are learnable parameters. Then we compute the ERC loss using standard cross-entropy loss:

$$\mathcal{L}_{ERC} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{|E|} y_i^j log\ \hat{y}_i^j \tag{9}$$

where y_i^j and $\hat{y}_i^j$ are j-th element of y_i and $\hat{y}_i$. Finally, we combine all proposed modules and train in a multi-task learning manner:

$$\mathcal{L} = \mathcal{L}_{ERC} + \mu_E \mathcal{L}_{ELBO} + \mu_I \mathcal{L}_{INFO} + \mu_{MI} \mathcal{L}_{MI} \tag{10}$$

where the μs are the pre-defined weight coefficients.

Experimental Settings

Benchmark Datasets

We evaluate our model on the following four benchmark datasets, where the statistics are listed in Table 1:

Dataset	Conv.(Train/Val/Test)	Utter.(Train/Val/Test)	Utter./Conv
IEMOCAP	100/20/31	4,778/980/1,622	49.2
MELD	1,038/114/280	9,989/1,109/2,610	9.6
EmoryNLP	713/99/85	9,934/1,344/1,328	14.1
DailyDialog	11,118/1,000/1,000	87,170/8,069/7,740	7.9

Table 1: Statistics of the datasets.

IEMOCAP (Busso et al. 2008): An acted two-party multi-modal conversation dataset. The pre-defined categorical emotion labels are *neutral, sadness, anger, happiness, frustrated, excited.*

MELD (Poria et al. 2019a): A multi-modal dataset collected from the scripts of the TV show *Friends*. The pre-defined emotion labels are *neutral, sadness, anger, disgust, fear, happiness, surprise*.

EmoryNLP (Zahiri and Choi 2018): Another dataset collected from the scripts of *Friends*, but with a different emotion category set. The pre-defined emotion labels are *neutral, sad, mad, scared, powerful, peaceful, joyful*.

DailyDialog (Li et al. 2017): From human-written daily conversations with only two parties involved. The pre-defined emotion labels are *neutral, happiness, surprise, sadness, anger, disgust, fear*.

Baseline Models

We select the following baseline models for comparison:

TL-ERC (Hazarika et al. 2021): The method pre-trains an encoder-decoder architecture on large-scale conversation data, then the weights are transferred to ERC. **BERT-Large** (Devlin et al. 2019): Initialised from pre-trained weights of BERT-Large, then fine-tuned for ERC. **DialogXL** (Shen et al. 2021a): The PLM-based model uses dialog-aware self-attention to model speaker dependencies. **DAG-ERC** (Shen et al. 2021b): Based on RoBERTa-Large, this model builds a Directed Acyclic Graph (DAG) on the dialogue. **SKAIG** (Li et al. 2021): This work builds a graph on the dialogue and utilises psychological knowledge to enrich edge representations. **COSMIC** (Ghosal et al. 2020): With the dialogue sequence-based structure, this work introduces utterance-level mental state knowledge to model the mental states of speakers. **Dis-VAE** (Ong et al. 2022): This work utilises a VAE to model discourse information in an unsupervised manner. **SGED** (Bao et al. 2022): This method proposes a speaker-guided encoder-decoder framework to exploit speaker information for ERC. **CoG-BART** (Li, Yan, and Qiu 2022): Based on BART-Large, this work utilises contrastive learning and a response generation task to enhance utterance representations.

Implementation Details

We conduct all experiments using a single Nvidia Tesla A100 GPU with 80GB of memory. We initialise the pre-trained weights of all PLMs and use the tokenization tools both provided by Huggingface (Wolf et al. 2019). We leverage AdamW optimiser (Loshchilov and Hutter 2019) to train the model. All hyper-parameters are tuned on the validation set. We use the weighted-F1 measure as the evaluation metric for MELD, EmoryNLP, and IEMOCAP. Since "neutral" occupies most of DailyDialog, we use micro-F1 for this dataset, and ignore the label "neutral" when calculating the results as in the previous works (Shen et al. 2021b; Li, Yan, and Qiu 2022). All reported results are averages of five random runs. More details are in Appendix D.

Results and Analysis

Overall Performance

We present the performance of VAD-VAE and the baseline models on the four benchmark datasets in Table 2. According to the results, PLM-based methods BERT-Large and DialogXL significantly outperform TL-ERC on all datasets, showing their advantages over RNN-based models that pretrain from scratch. Following this trend, the rest of the baseline models and our VAD-VAE all utilise RoBERTa as the utterance encoder (except CoG-BART which uses BART).

COSMIC explicitly introduces mental state information to enrich the contexts, and the performance improves significantly on simple-context datasets MELD, EmoryNLP, and DailyDialog. Dis-VAE and SGED implicitly introduce discourse roles and speaker information, and perform well on both simple and rich-context datasets. Specifically, SGED achieves the best result 40.24% on EmoryNLP, and both methods achieve over 68% on IEMOCAP. VAD-VAE also introduces NRC-VAD information and outperforms COSMIC by over 4% on IEMOCAP and DailyDialog.

Model	IEMOCAP	MELD	EmoryNLP	DailyDialog
TL-ERC	59.30	57.46*	30.57*	52.46*
BERT-Large	60.98	61.50	34.17	54.09
DialogXL	65.94	62.41	34.73	54.93
COSMIC	65.28	65.21	38.11	58.48
Dis-VAE	68.23	65.34	–	60.95
SGED	68.53	65.46	**40.24**	–
DAG-ERC	68.03	63.65	39.02	59.33
SKAIG	66.96	65.18	38.88	59.75
CoG-BART	66.18	64.81	39.04	56.29
VAD-VAE	**70.22**(±0.85)	64.96(±0.19)	38.35(±0.35)	**62.14**(±0.23)
-vCLUB	69.19(±0.66)	**65.94**(±0.31)	38.90(±0.21)	61.23(±0.77)

Table 2: Test results on IEMOCAP, MELD, EmoryNLP and DailyDialog datasets. "-vCLUB" denotes VAD-VAE trained without the vCLUB loss $\mathcal{L}_{MI}$. Results with "*" are the re-implementations of Shen et al. (2021a). Best values: bold.

To enhance the utterance representations, both DAG-ERC and SKAIG build dialogue-level graphs to introduce priors on context modelling, and perform well on all datasets. The competitive performance of CoG-BART also proves the effectiveness of contrastive learning and response generation. VAD-VAE achieves an impressive 4.04% improvement on IEMOCAP and 5.85% improvement on DailyDialog over CoG-BART, showing the advantage of VAE-based target utterance reconstruction over response generation.

Overall, VAD-VAE achieves new state-of-the-art performance 70.22% on IEMOCAP, 65.94% on MELD, and 62.14% on DailyDialog. However, there is less improvement on EmoryNLP. A possible reason is that EmoryNLP defines fuzzy emotions *powerful* and *peaceful*. Though they appear very positive in NRC-VAD, we notice that many utterances labelled with fuzzy emotions do not yield positive sentiments, and unified VAD supervision signals of the fuzzy emotions are misleading for many samples. We provide case studies in Appendix C to further investigate this hypothesis. In future work, we will consider leveraging other fine-grained information to understand subtle changes in fuzzy emotions, such as intent information, which is successfully applied to empathetic dialog generation (Xie and Pu 2021).

Ablation Study

To investigate the effect of each module, we provide ablation analysis in Table 5. "-" denotes removing a module. "vCLUB" denotes the MI minimisation modules. "VAE Decoder" denotes the VAE decoder module for target utterance reconstruction. "V Sup.", "A Sup.", and "D Sup." denote the NRC-VAD supervision signals corresponding to Valance, Arousal, and Dominance. "Utter. Encoder" directly trains an ERC model on the context-aware utterance encoder.

According to the results, VAD-VAE achieves comparable performance with "-vCLUB", which corresponds with our early hypothesis. With vCLUB loss, the performance improves on IEMOCAP and DailyDialog. A possible reason is that MI minimisation enforces the model to learn dissimilar representations for VAD, which corresponds with the

Target Utterance	Next Utterance	Key Long-Range Context	RG	TUR	Gold
Chandler: It's all finished!	Joey: This was Carol's favorite beer. She always drank it out of can, I should have known. [*sad*]	Joey: Tell Monica we are **done with the bookcase** at last! [*joyful*]	*neutral*	*joyful*	*joyful*
Mary: It's a good idea to live together before you get married.	James: I think so, too. Are you going to get a house or- [*neutral*]	Linda: I have to think about moving now. We are going to **move in together**! [*happy*]	*surprised*	*happy*	*happy*

Table 3: Two cases where Target Utterance Reconstruction (TUR) leverages the key long-range context of the target utterance for emotion reasoning, while the next utterance provides limited information for Response Generation (RG). "RG", "TUR" and "Gold" denotes corresponding predictions for the two methods and the golden labels. Key information in the context: bold.

	IEMOCAP				MELD				EmoryNLP			
Methods	V	A	D	MI	V	A	D	MI	V	A	D	MI
$\mathcal{L}_{ELBO}$	0.434	0.156	0.094	0.732	0.021	0.133	0.152	0.398	0.071	0.029	-0.017	0.356
$+\mathcal{L}_{MI}$	0.306	0.132	0.150	**0.133**	0.044	0.230	0.141	**0.068**	0.115	-0.026	-0.016	**0.094**
$+\mathcal{L}_{INFO}$	**0.882**	0.708	0.743	0.531	0.565	**0.643**	0.557	0.885	0.424	0.348	0.321	0.789
$+\mathcal{L}_{INFO}+\mathcal{L}_{MI}$	0.872	**0.715**	**0.765**	0.355	**0.568**	0.639	**0.558**	0.312	**0.433**	**0.355**	**0.330**	0.226

Table 4: The Pearson's correlation coefficients between predicted VAD scores and supervision signals on IEMOCAP, MELD, and EmoryNLP test sets, and the average vCLUB estimates of MI between VAD latent distributions. Best values: bold.

orthogonal nature of VAD in psychology. With "-VAE Decoder" the performance drops significantly on all datasets, which further indicates the effectiveness of the target utterance reconstruction task.

Method	IEMOCAP	MELD	DailyDialog
VAD-VAE	**70.22**	64.96(↓0.98)	**62.14**
-vCLUB	69.19(↓1.03)	**65.94**	61.23(↓0.91)
-VAE Decoder	67.92(↓2.30)	63.99(↓1.95)	60.16(↓1.98)
-V, A, D Sup.	66.85(↓3.37)	63.99(↓1.95)	61.01(↓1.13)
-V Sup.	67.60(↓2.62)	64.20(↓1.74)	61.92(↓0.22)
-A Sup.	68.06(↓2.16)	64.59(↓1.35)	61.38(↓0.76)
-D Sup.	67.16(↓3.06)	64.03(↓1.91)	61.66(↓0.48)
Utter. Encoder	66.52(↓3.70)	63.7(↓2.24)	59.80(↓2.34)

Table 5: Results of ablation study on IEMOCAP, MELD, and DailyDialog datasets. Best values: bold.

"-V, A, D Sup." leads to significant drops in all datasets, which proves our hypothesis that NRC-VAD supervision signals also provide fine-grained information to enhance ERC performance. In further comparisons of separately removing supervision signals for V, A, and D, the performance drops most with "D Sup." removed for IEMOCAP and MELD, and "A Sup." removed for DailyDialog. We notice that the sentiment polarity of emotions is mostly determined by Valance, and similar emotions mainly differ in Arousal and Dominance. Therefore, these results show that our model benefits more from fine-grained information from Arousal and Dominance to distinguish similar emotions.

Case Study for Target Utterance Reconstruction

Target utterance reconstruction enables VAD-VAE to learn long-range dependencies and outperforms response generation-based model CoG-BART. We provide two cases from the test results in Table 3 for further investigation. There are mainly two scenarios where response generation provides limited information and mispredicts the emotion: (1) Sudden change of the topic. In the next utterance of case 1, the topic changes from "Finish assembling the bookcase" to "Carol's favorite beer". (2) Indirect response. In case 2, the next utterance does not directly respond but expresses agreement to the target utterance, which provides little extra information. This scenario is common in multi-party dialogues. In contrast, VAD-VAE is able to learn more inter-speaker influence and key information from the long-range contexts, such as the discussion topic "done with the bookcase" and "move in together".

Disentanglement Evaluation

In this section, we analyse the effects of VAD supervision signals ($\mathcal{L}_{INFO}$) and MI minimisation ($\mathcal{L}_{MI}$) on enhancing VAD disentanglement. In Table 4, we present the Pearson's Correlation Coefficients (PCC) between the predicted VAD scores from latent representations and the supervision signals from NRC-VAD on IEMOCAP, MELD, and EmoryNLP test sets. Higher values indicate more precise predictions, denoting better *Informativeness*. We also provide the average vCLUB estimates of MI between VAD latent distributions on each test set, with lower values denoting lower estimates of MI upper bounds and better *Independence*.

Informativeness According to the results, the model performs poorly on all datasets (PCC below 0.2 in most cases) with standard VAE reconstruction loss ($\mathcal{L}_{ELBO}$) or $\mathcal{L}_{MI}$ introduced, since VAD features could be embedded in Content space without specific supervisions. We observe significant improvement in *Informativeness* for VAD with $\mathcal{L}_{INFO}$, which brings over 0.5 PCC gain for IEMOCAP and 0.3 for MELD and EmoryNLP. These results reflect the effectiveness of NRC-VAD supervision signals. On top of $\mathcal{L}_{INFO}$, $\mathcal{L}_{MI}$ further improves the PCC scores in most cases, which shows that MI minimisation also helps to enhance *Informativeness* of VAD representations to some extent.

Independence For all datasets, The vCLUB estimates remain high with only $\mathcal{L}_{ELBO}$ introduced, since the unified distributed representation in VAE encourages strong correlations between each part. With $\mathcal{L}_{INFO}$, we observe even higher vCLUB in MELD and EmoryNLP. In this case, our model is only optimised for *Informativeness*, and enforces full utilisation of all latent spaces, which leads to high MI. Introducing only $\mathcal{L}_{MI}$ has the lowest vCLUB in all datasets. However, it achieves bad performance in *Informativeness*. With both $\mathcal{L}_{MI}$ and $\mathcal{L}_{INFO}$, VAD-VAE not only achieves the best results in VAD predictions but also greatly decreases vCLUB compared with only $\mathcal{L}_{INFO}$, showing the satisfactory trade-off between *Informativeness* and *Independence*.

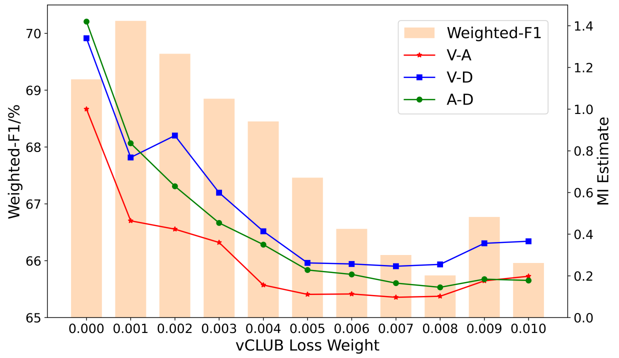


Figure 3: The performance of VAD-VAE on ERC and the MI estimates between each VAD pairs by different vCLUB weight coefficients μ_{MI} on IEMOCAP test set.

ERC-MI Trade-off To further investigate the trade-off between ERC performance and *Independence* of VAD representations, in Figure 3 we present the performance of VAD-VAE on ERC and the MI estimates between each VAD pair by different vCLUB loss coefficients μ_{MI} on IEMOCAP test set. With μ_{MI}=0.001, the model achieves the best results in ERC. All MI estimates decrease rapidly as μ_{MI} increases, while ERC performance decreases at acceptable rates. With μ_{MI}=0.005, all MI estimates drop below 0.3 while VAD-VAE keeps a competitive ERC result over 67%. As μ_{MI} further increases, we observe no apparent decrease in MI estimates but ERC performance drops fast. Overall, the experiments show a best μ_{MI} between 0.001 and 0.005 for IEMOCAP, and the importance of considering the trade-off between ERC and *Independence* performance.

Visualisation of VAD Spaces To conduct a more intuitive analysis of the disentangled representations, we present the UMAP (McInnes, Healy, and Melville 2018) visualisations of VAD representations in IEMOCAP test set for four representative emotions in Figure 4, and their corresponding NRC-VAD assignments in Table 6. As shown, for Valance and Dominance, positive and negative emotions are well separated, while emotions within one polarity overlap. In visualisation of Arousal, “happy”, “excited” and “frustrated” lie close while “sad” separates away. These observations largely correspond with the NRC-VAD assignments, which further indicates the quality of the learnt VAD representations. In addition, the distribution of each emotion shows continuity and completeness conditions. In future work, we will explore the potential of VAD-VAE in the affective text generation task. Different from previous works which control categorical emotions, our model is able to control more fine-grained sentiments by separately adjusting Valance, Arousal, and Dominance.

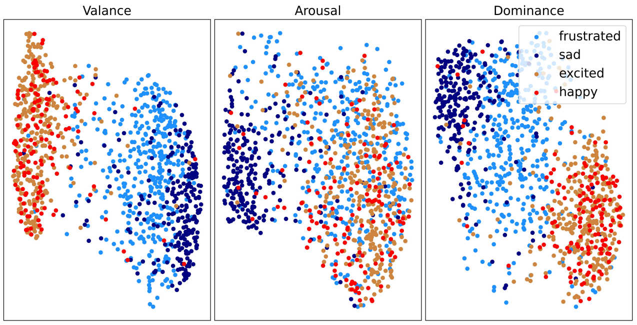


Figure 4: UMAP visualisations of Valance, Arousal and Dominance representations in IEMOCAP test set.

VAD	frustrated	sad	excited	happy
Valance	0.0600	0.0520	0.9080	0.9600
Arousal	0.7300	0.2880	0.9310	0.7320
Dominance	0.2800	0.1640	0.7090	0.8500

Table 6: The NRC-VAD assignments to the four emotions.

Conclusion

In this paper, we propose a VAD-disentangled Variational Autoencoder for emotion recognition in conversations. We first introduce an auxiliary target utterance reconstruction task via the VAE framework. Then we disentangle three key features Valance, Arousal, and Dominance from the latent space. VAD supervision signals and a mutual information minimisation task are also utilised to enhance the disentangled representations. Experiments show that VAD-VAE outperforms the state-of-the-art model on three ERC datasets, and ablation studies prove the effectiveness of proposed modules. The analysis also shows that VAD-VAE learns decent disentangled VAD representations. In the future, we will leverage more fine-grained information such as intent to understand subtle changes in fuzzy emotions, and explore fine-grained emotion control for affective text generation by separately adjusting Valance, Arousal, and Dominance.

References

Bao, Y.; Ma, Q.; Wei, L.; Zhou, W.; and Hu, S. 2022. Speaker-Guided Encoder-Decoder Framework for Emotion Recognition in Conversation. In *IJCAI*, 4051–4057. ijcai.org.

Bao, Y.; Zhou, H.; Huang, S.; Li, L.; Mou, L.; Vechtomova, O.; Dai, X.-y.; and Chen, J. 2019. Generating Sentences from Disentangled Syntactic and Semantic Spaces. In *ACL*, 6008–6019. Florence, Italy: Association for Computational Linguistics.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8): 1798–1828.

Burgess, C. P.; Higgins, I.; Pal, A.; Matthey, L.; Watters, N.; Desjardins, G.; and Lerchner, A. 2018. Understanding disentangling in beta-VAE. In *NeurIPS Workshops*. Curran Associates, Inc.

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4): 335–359.

Chapuis, E.; Colombo, P.; Manica, M.; Labeau, M.; and Clavel, C. 2020. Hierarchical Pre-training for Sequence Labelling in Spoken Dialog. In *Findings of EMNLP*, 2636–2648. Online: Association for Computational Linguistics.

Chatterjee, A.; Gupta, U.; Chinnakotla, M. K.; Srikanth, R.; Galley, M.; and Agrawal, P. 2019. Understanding Emotions in Text Using Deep Learning and Big Data. *Computers in Human Behavior*, 93: 309–317.

Chen, J.; and Batmanghelich, K. 2020. Weakly Supervised Disentanglement by Pairwise Similarities. In *AAAI*, 3495–3502. AAAI Press.

Chen, R. T. Q.; Li, X.; Grosse, R. B.; and Duvenaud, D. K. 2018. Isolating Sources of Disentanglement in Variational Autoencoders. In *NeurIPS*, volume 31. Curran Associates, Inc.

Cheng, P.; Hao, W.; Dai, S.; Liu, J.; Gan, Z.; and Carin, L. 2020a. CLUB: A Contrastive Log-ratio Upper Bound of Mutual Information. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, 1779–1788. PMLR.

Cheng, P.; Min, M. R.; Shen, D.; Malon, C.; Zhang, Y.; Li, Y.; and Carin, L. 2020b. Improving Disentangled Text Representation Learning with Information-Theoretic Guidance. In *ACL*, 7530–7541. Online: Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Eastwood, C.; and Williams, C. K. I. 2018. A Framework for the Quantitative Evaluation of Disentangled Representations. In *ICLR*. OpenReview.net.

Ghosal, D.; Majumder, N.; Gelbukh, A.; Mihalcea, R.; and Poria, S. 2020. COSMIC: COmmonSense knowledge for eMotion Identification in Conversations. In *Findings of EMNLP*, 2470–2481. Online: Association for Computational Linguistics.

Goswamy, T.; Singh, I.; Barkati, A.; and Modi, A. 2020. Adapting a Language Model for Controlled Affective Text Generation. In *COLING*, 2787–2801. Barcelona, Spain: International Committee on Computational Linguistics.

Hazarika, D.; Poria, S.; Mihalcea, R.; Cambria, E.; and Zimmermann, R. 2018a. ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection. In *EMNLP*, 2594–2604. Brussels, Belgium: Association for Computational Linguistics.

Hazarika, D.; Poria, S.; Zadeh, A.; Cambria, E.; Morency, L.-P.; and Zimmermann, R. 2018b. Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos. In *NAACL*, 2122–2132. New Orleans, Louisiana: Association for Computational Linguistics.

Hazarika, D.; Poria, S.; Zimmermann, R.; and Mihalcea, R. 2021. Conversational transfer learning for emotion recognition. *Information Fusion*, 65: 1–12.

Higgins, I.; Amos, D.; Pfau, D.; Racaniere, S.; Matthey, L.; Rezende, D.; and Lerchner, A. 2018. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C. P.; Glorot, X.; Botvinick, M. M.; Mohamed, S.; and Lerchner, A. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *ICLR*. OpenReview.net.

John, V.; Mou, L.; Bahuleyan, H.; and Vechtomova, O. 2019. Disentangled Representation Learning for Non-Parallel Text Style Transfer. In *ACL*, 424–434. Florence, Italy: Association for Computational Linguistics.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *ICLR*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *ACL*, 7871–7880. Online: Association for Computational Linguistics.

Li, J.; Ji, D.; Li, F.; Zhang, M.; and Liu, Y. 2020. HiTrans: A Transformer-Based Context- and Speaker-Sensitive Model for Emotion Detection in Conversations. In *COLING*, 4190–4200. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Li, J.; Lin, Z.; Fu, P.; and Wang, W. 2021. Past, Present, and Future: Conversational Emotion Recognition through Structural Modeling of Psychological Knowledge. In *Findings of EMNLP*, 1204–1214. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Li, S.; Yan, H.; and Qiu, X. 2022. Contrast and Generation Make BART a Good Dialogue Emotion Recognizer. In *AAAI*, 11002–11010. AAAI Press.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *IJCNLP*, 986–995. Taipei, Taiwan: Asian Federation of Natural Language Processing.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *ICLR*. OpenReview.net.

Ma, Y.; Nguyen, K. L.; Xing, F. Z.; and Cambria, E. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64: 50–70.

Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A. F.; and Cambria, E. 2019. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. In *AAAI*, 6818–6825. AAAI Press.

McInnes, L.; Healy, J.; and Melville, J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Mehrabian, A. 1995. Framework for a comprehensive description and measurement of emotional states. *Genetic, social, and general psychology monographs*.

Mohammad, S. 2018. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In *ACL*, 174–184. Melbourne, Australia: Association for Computational Linguistics.

Mukherjee, R.; Naik, A.; Poddar, S.; Dasgupta, S.; and Ganguly, N. 2021. Understanding the Role of Affect Dimensions in Detecting Emotions from Tweets: A Multi-task Approach. In *SIGIR*, 2303–2307. ACM.

Nandwani, P.; and Verma, R. 2021. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1): 1–19.

Ong, D.; Su, J.; Chen, B.; Luu, A. T.; Narendranath, A.; Li, Y.; Sun, S.; Lin, Y.; and Wang, H. 2022. Is Discourse Role Important for Emotion Recognition in Conversation? In *AAAI*, 11121–11129. AAAI Press.

Park, S.; Kim, J.; Ye, S.; Jeon, J.; Park, H. Y.; and Oh, A. 2021. Dimensional Emotion Detection from Categorical Emotion. In *EMNLP*, 4367–4380. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Poole, B.; Ozair, S.; van den Oord, A.; Alemi, A. A.; and Tucker, G. 2019. On Variational Bounds of Mutual Information. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, 5171–5180. PMLR.

Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; and Mihalcea, R. 2019a. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *ACL*, 527–536. Florence, Italy: Association for Computational Linguistics.

Poria, S.; Majumder, N.; Mihalcea, R.; and Hovy, E. 2019b. Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances. *IEEE Access*, 7: 100943–100953.

Qiu, X.; Sun, T.; Xu, Y.; Shao, Y.; Dai, N.; and Huang, X. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10): 1872–1897.

Russell, J. A.; and Mehrabian, A. 1977. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3): 273–294.

Shen, W.; Chen, J.; Quan, X.; and Xie, Z. 2021a. DialogXL: All-in-One XLNet for Multi-Party Conversation Emotion Recognition. In *AAAI*, 13789–13797. AAAI Press.

Shen, W.; Wu, S.; Yang, Y.; and Quan, X. 2021b. Directed Acyclic Graph Network for Conversational Emotion Recognition. In *ACL*, 1551–1560. Online: Association for Computational Linguistics.

Shu, R.; Chen, Y.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Weakly Supervised Disentanglement with Guarantees. In *ICLR*. OpenReview.net.

Vasilakes, J.; Zerva, C.; Miwa, M.; and Ananiadou, S. 2022. Learning Disentangled Representations of Negation and Uncertainty. In *ACL*, 8380–8397. Dublin, Ireland: Association for Computational Linguistics.

Wang, J.; Wang, J.; Sun, C.; Li, S.; Liu, X.; Si, L.; Zhang, M.; and Zhou, G. 2020a. Sentiment Classification in Customer Service Dialogue with Topic-Aware Multi-Task Learning. In *AAAI*, 9177–9184. AAAI Press.

Wang, J.; Wang, J.; Sun, C.; Li, S.; Liu, X.; Si, L.; Zhang, M.; and Zhou, G. 2020b. Sentiment Classification in Customer Service Dialogue with Topic-Aware Multi-Task Learning. In *AAAI*, 9177–9184. AAAI Press.

Warriner, A. B.; Kuperman, V.; and Brysbaert, M. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45: 1191–1207.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Xie, Y.; and Pu, P. 2021. Empathetic Dialog Generation with Fine-Grained Intents. In *CoNLL*, 133–147. Online: Association for Computational Linguistics.

Xie, Y.; Yang, K.; Sun, C.; Liu, B.; and Ji, Z. 2021. Knowledge-Interactive Network with Sentiment Polarity Intensity-Aware Multi-Task Learning for Emotion Recognition in Conversations. In *Findings of EMNLP*, 2879–2889. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Zad, S.; Heidari, M.; Jones, J. H. J.; and Uzuner, O. 2021. Emotion Detection of Textual Data: An Interdisciplinary Survey. In *AIIoT*, 0255–0261.

Zahiri, S. M.; and Choi, J. D. 2018. Emotion Detection on TV Show Transcripts with Sequence-Based Convolutional Neural Networks. In *AAAI Workshops*, volume WS-18 of *AAAI Technical Report*, 44–52. AAAI Press.

Zhu, L.; Pergola, G.; Gui, L.; Zhou, D.; and He, Y. 2021. Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection. In *ACL*, 1571–1582. Online: Association for Computational Linguistics.