Utterance-level Emotion Recognition in Conversation with Conversation-level Supervision

Ximing Li[1,2], **Yuanchao Dai**[1,2], **Zhiyao Yang**[1,2*], **Jinjin Chi**[1,2], **Wanfu Gao**[1,2], **Lin Yuanbo Wu**[3]

[1]College of Computer Science and Technology, Jilin University, China
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, China
[3] Swansea University, United Kingdom
{liximing86, yuanchaodai, yangzy9529, chijinjin616}@gmail.com, gaowf@jlu.edu.cn, l.y.wu@swansea.ac.uk

Abstract

Emotion Recognition in Conversations (ERC) involves automatically identifying the emotion of each utterance in conversations. The emotion of an utterance is contingent to the conversation context, and thus, annotating each utterance in ERC entails repetitive screening the whole conversation from annotators. Such a requirement leads to prohibitive cost in fine-grained labeling on utterance. In this paper, we propose an efficient coarse-grained labeling strategy for ERC, which assigns a set of emotions for each conversation. In specific, we reformulate the ERC predictors with conversation-level emotion sets as weakly-supervised learning to optimise a potential candidate for ERC, which is termed as Dataless ERC (DERC). To validate this, we propose a simple-yet-flexible DERC framework with Progressive Learning (DERC-PL). We jointly update pseudo-utterance-level emotions and the ERC predictor in a self-training manner, where we progressively update the ERC predictor from training subsets with lower noise densities to the ones with higher noise densities. We implemented several versions of DERC-PL by incorporating various off-the-shelf ERC methods. Extensive experimental results demonstrate that the proposed DERC-PL can be on par with existing weakly-supervised learning baselines and supervised learning ERC methods.

Introduction

Emotion **R**ecognition in **C**onversations (**ERC**) refers to the task of identifying emotions, such as sadness, happiness, or anger within conversations, which specifically focuses on the emotion conveyed in each utterance (Zhang, Chen, and Chen 2023; Hu et al. 2023; Li et al. 2023; Qin et al. 2023; Zhang et al. 2023b). Recently, ERC has garnered significant attention from the natural language processing community due to its applicability in diverse real-world scenarios, including medical conversation analysis (Barnes 2019; Priya, Firdaus, and Ekbal 2023), social media analysis (Brambilla et al. 2022), and dialogue system construction (Ma et al. 2020; Liu et al. 2021b), among others.

Typically, addressing ERC requires the collection of fine-grained training datasets where each utterance within a conversation is manually labeled. However, unlike traditional emotion recognition (Mittal et al. 2020; Alhuzali and Ananiadou 2023; Ouyang et al. 2024), the emotion of an utterance in ERC is heavily influenced by the conversational context (Zhang et al. 2023a). As shown in Fig. 1, the same utterance may convey different emotions depending on the context, necessitating annotators to repeatedly review the entire conversation. This makes the fine-grained labeling process highly time-consuming. According to labeling statistics from ERC studies (Zahiri and Choi 2018; Chen et al. 2018; Poria et al. 2019) and our preliminary experiments, we estimate that the average time required to label a single utterance in ERC is approximately 10 seconds, which can be prohibitively expensive for certain real-world applications.

Inspired by the cognitive observation that humans can rapidly grasp emotions from lengthy texts even with quick skimming (Duggan and Payne 2011), we propose a more efficient labeling approach for ERC. Instead of labeling each utterance individually, we suggest that annotators assign a set of emotions to each conversation. To validate this approach, we conducted preliminary experiments, manually labeling conversation-level emotion sets across benchmark ERC datasets. The results indicate that this method maintains accuracy while significantly enhancing efficiency, reducing the average labeling time for a conversation with 5-10 utterances to just 13 seconds. This leads us to pose the question "**Is it possible, and if so, how, to utilize these efficient yet coarse-grained datasets with conversation-level emotion sets to train ERC predictors that can accurately identify utterance-level emotions in a weakly-supervised manner**". We formally introduce this new weakly-supervised learning task as **D**ataless **ERC** (**DERC**), and demonstrate the distinction between conventional ERC and DERC through a comparative example in Fig. 1.

To address this question, we present, to the best of our knowledge, the first attempt to tackle DERC. Drawing inspiration from prior weakly-supervised learning techniques (Liu et al. 2021a; Matsuo et al. 2023; Li et al. 2024; Yang et al. 2023), we propose a straightforward solution based on pseudo-labeling. Specifically, we initialize pseudo-utterance-level emotions using conversation-level emotion sets and iteratively update them alongside the ERC predictor in a self-training framework. Typically, pseudo-utterance-level emotions are refined based on the current predictions of the ERC predictor. To enhance accuracy, we divide all

*Corresponding author

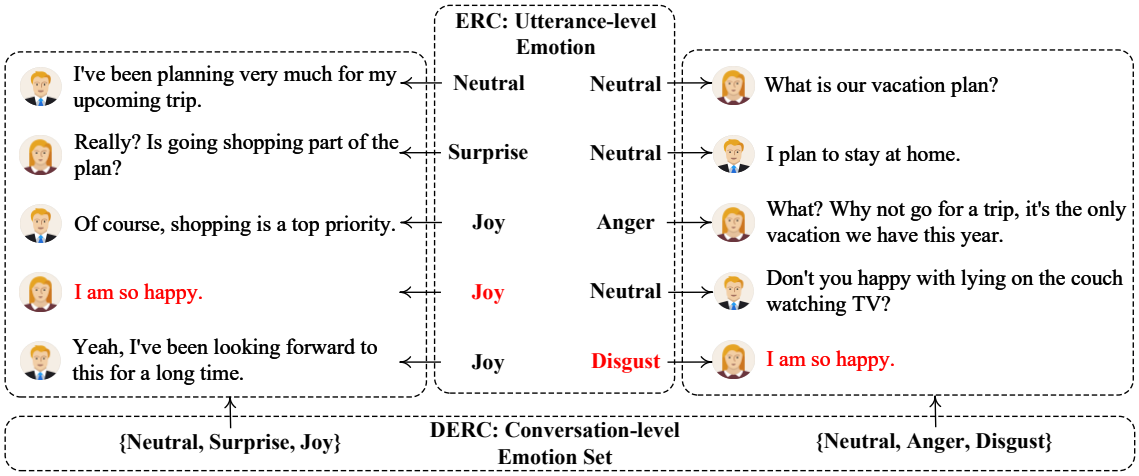

Figure 1: ERC and DERC refer to utterance-level and conversation-level supervision, respectively. This figure illustrates that the same utterance, such as "I am so happy," can be associated with distinct emotions like "`Joy`" and "`Disgust`" in different conversations.

training conversations into subsets categorized by noise density, determined by the number of emotions assigned within each conversation-level emotion set. We then progressively train the ERC predictor, starting with subsets of lower noise density and gradually moving to those with higher noise, thereby achieving more precise pseudo-utterance-level emotions during the initial training stages. Building on these ideas, we propose a simple yet flexible DERC framework called **P**rogressive **L**earning (**DERC-PL**), within which most existing ERC methods can be directly integrated as ERC predictors.

We have developed several versions of DERC-PL by incorporating various ERC methods, and our extensive experimental results demonstrate that DERC-PL not only surpasses existing weakly-supervised learning baselines but also competes closely with supervised ERC methods. Importantly, our empirical findings suggest that the coarse-grained DERC may serve as a viable alternative to traditional fine-grained ERC approaches. The key contributions of this paper are as follows:

- We evaluate the accuracy and efficiency of manually labeling conversation-level emotion sets and introduce a new weakly-supervised learning task for ERC, termed **DERC**.
- We propose a simple yet flexible DERC framework named DERC-PL, which progressively learns from training subsets with varying noise densities.
- We conduct extensive experiments to validate the effectiveness of DERC-PL and provide empirical evidence that coarse-grained DERC can be a strong candidate for fine-grained ERC.

Manual Labeling Evaluations of ERC and DERC

In this section, we conduct preliminary experiments to empirically evaluate the accuracy and efficiency of manual labeling for ERC and DERC. To this end, we invited 20 undergraduates from diverse backgrounds at our university as volunteers and evenly divided them into 2 groups to separately conduct manual labeling for ERC and DERC. In the following, we introduce the manual labeling rules and empirical results.

Manual labeling rules. Referring to (Zahiri and Choi 2018; Chen et al. 2018; Poria et al. 2019), we build the labeling rules of ERC and DERC for annotators. We employ the benchmark dataset **MELD** (Poria et al. 2019), and divide it into 4 sets of conversations whose utterance numbers belong to [1, 5], [6, 10], [11, 15], and [16,20], respectively. We randomly select 50 conversations from each set, leading to a collection of 200 conversations containing 1,964 utterances for manual labeling evaluation. The labeling range (*i.e.,* the label space $\mathcal{Y}$) contains 7 emotions, including {`Neutral, Joy, Sadness, Fear, Anger, Disgust, Surprise`}. In terms of ERC and DERC, annotators are asked to assign emotions and emotion sets to utterances and conversations, respectively. They should meet the required *labeling type*, *labeling order*, and *labeling demand*. We build rules of *golden label* to determine the assigned emotions, and also rules of *conflict handling* to deal with controversial utterances and conversations. All details of labeling rules are listed in Table 1.

Empirical results. First, we evaluate the labeling accuracy of our emotion annotations. We apply the original utterance-level emotions of **MELD** as the ground-truth emotions, and add them into the corresponding conversation-level emotion sets as the ground-truth emotion sets. The labeling accuracy scores of ERC and DERC are defined as the proportions of our emotion annotations matching ground-truth emotions and ground-truth emotion sets, respectively. The results are shown in Table 2. We can observe that the labeling accuracy scores of both ERC and DERC are above 90%, indicating the emotion annotations from our evalua-

	ERC rule	DERC rule
labeling range	`Neutral, Joy, Sadness, Fear, Anger, Disgust, Surprise`	`Neutral, Joy, Sadness, Fear, Anger, Disgust, Surprise`
label type	One emotion label for each **utterance**.	One emotion set for each **conversation**.
labeling order	From short conversations to long conversations.	From short conversations to long conversations.
labeling demand	Thinking at least **3 seconds** before assigning the emotion.	Reading the conversation at least **3 times** before assign the emotion set.
golden label	For each utterance, we count the assigned numbers of emotions by annotators. If the majority emotion is assigned by over half of the annotators, it will be the assigned emotion finally. Otherwise, the utterance will be marked with "conflict".	For each conversation, we count the assigned numbers of emotions by annotators. If any emotion is assigned by over half of the annotators, it will be added to the emotion set. If the emotion set is empty, the conversation will be marked with "conflict".
conflict handling	For each utterance marked with "conflict", we revise the *labeling range* to the top-3 voted emotions and then re-label it. If the re-labeling result cannot meet the rule of *golden label*, we will drop this utterance.	For each conversation marked with "conflict", we re-label it over all assigned emotions. If the re-labeling result cannot meet the rule of *golden label*, we will drop this conversation.

Table 1: Details of labeling rules for ERC and DERC.

tions are almost consistent with the ones from the public **MELD** dataset. That is, our manual labeling results are convincing so that the corresponding efficiency results can be believable. Further, we evaluate the labeling efficiency of our emotion annotations. We present the average labeling time, including labeling and conflict handling time, for one utterance and conversation, respectively. As shown in Table 2, we can observe that labeling conversation-level emotion sets is as efficient as labeling utterance-level emotions. Specifically, the time cost of labeling conversations containing 1∼5 and 6∼10 utterances is almost the same as the time cost of labeling utterances; and the time cost of labeling conversations containing 11∼15 and 16∼20 utterances is only about $2 \sim 3$ times than that of labeling utterances. In summary, the empirical results indicate that one can collect accurate and efficient conversation-level emotion sets, to support the potential task of DERC.

The Proposed DERC-PL

In this section, we detail the proposed DERC framework named **DERC-PL**.

Definition of DERC. Formally, the training dataset of DERC is composed of N coarse-grained labeled samples $\mathcal{D} = \{(\mathcal{C}_i, \mathbf{y}_i)\}_{i=1}^{i=N}$, where each $\mathcal{C}_i$ denotes a conversation, and $\mathbf{y}_i \in \{0,1\}^{|\mathcal{Y}|}$ is its corresponding emotion set. We review that $\mathcal{Y}$ denotes the emotion space. To be specific, each $\mathcal{C}_i$ consists of M rounds of utterances $\{\mathbf{u}_{ij}\}_{j=1}^{j=M}$,[1] where each $\mathbf{u}_{ij}$ is the raw content but its ground-truth emotion is unknown. The goal of DERC is to apply the coarse-grained training dataset $\mathcal{D}$ to induce an ERC predictor $\mathcal{F}_\theta(\cdot)$, which can predict the emotion of each utterance in future conversations.

[1]To make notations simple, we suppose that all conversations contain M utterances.

	ERC	DERC
Accuracy	92.9%	94.5%
Time$_{\text{Utt.}}$	10s ∼15s	–
Time$_{\text{Con.[1,5]}}$	–	5s ∼ 11s
Time$_{\text{Con.[6,10]}}$	–	10s ∼ 16s
Time$_{\text{Con.[11,15]}}$	–	18s ∼ 25s
Time$_{\text{Con.[16,20]}}$	–	31s ∼ 44s

Table 2: Accuracy and time cost (seconds) of manual labeling evaluations. **Time$_{\text{Utt.}}$** and **Time$_{\text{Con.[a,b]}}$** denote the average labeling time for one utterance and one conversation whose utterance number belongs to [a,b], respectively.

DERC-PL Synopsis

Our DERC-PL is based on pseudo-labeling. For each utterance $\mathbf{u}_{ij}$, we initialize its pseudo-utterance-level emotion $\hat{\mathbf{y}}_{ij} \in [0,1]^{|\mathcal{Y}|}$ using the emotion set $\mathbf{y}_i$ of the conversation from which it comes:

$$\hat{\mathbf{y}}_{ij} = \frac{\mathbf{y}_i}{|\mathbf{y}_i|}, \ i = [N], \ j = [M], \tag{1}$$

where $|\mathbf{y}_i|$ denotes the number of assigned emotions in $\mathbf{y}_i$; and $\sum_{h=1}^{|\mathcal{Y}|} \hat{\mathbf{y}}_{ijh} = 1$. Therefore, we can initialize a pseudo-ERC dataset $\hat{\mathcal{D}} = \{(\mathcal{C}_i, \{\hat{\mathbf{y}}_{ij}\}_{j=1}^{j=M})\}_{i=1}^{i=N}$. We begin with $\hat{\mathcal{D}}$ and jointly update $\hat{\mathbf{y}}$ and a base ERC predictor $\mathcal{F}_\theta(\cdot)$, parameterized by θ, in a self-training manner. We update each $\hat{\mathbf{y}}_{ij}$ by using the predictions of the current ERC predictor $\mathbf{p}_{ij} = \mathcal{F}_\theta(\mathbf{u}_{ij})$. To achieve more precise $\hat{\mathbf{y}}$ at the early training periods, we divide $\hat{\mathcal{D}}$ into several training subsets with different noise densities, and then progressively update $\mathcal{F}_\theta(\cdot)$ from training subsets with lower noise densities to the ones with higher noise densities. In the following, we introduce the processes of **generating training subsets** and **progressive updating** in more detail.

Generating Training Subsets

As each pseudo-utterance-level emotion $\hat{\mathbf{y}}_{ij}$ must be covered by its corresponding conversation-level emotion set $\mathbf{y}_i$, we can use the number of assigned emotions $|\mathbf{y}_i|$ to express the noise densities of utterances from $\mathcal{C}_i$. Accordingly, we can divide $\hat{\mathcal{D}}$ into several disjoint training subsets $\hat{\mathcal{D}} = \hat{\mathcal{D}}_1 \cup \cdots \cup \hat{\mathcal{D}}_{|\mathcal{Y}|}$,[2] where each subset $\hat{\mathcal{D}}_g$ is defined by:

$$\hat{\mathcal{D}}_g = \{(\mathcal{C}_i, \{\hat{\mathbf{y}}_{ij}\}_{j=1}^{j=M})\}_{i=1}^{i=N_g}, \quad \forall\, |\mathbf{y}_i| = g, \tag{2}$$

where N_g is the number of conversations in this subset. To some extent, the noise density corresponds to the learning difficulty.

Progressive Updating

Given the training subsets, we progressively update the pseudo-utterance-level emotions $\hat{\mathbf{y}}$ and the ERC predictor $\mathcal{F}_\theta(\cdot)$ jointly from $\hat{\mathcal{D}}_1$ to $\hat{\mathcal{D}}_{|\mathcal{Y}|}$. Specifically, we perform the following learning process:

1: Initialize a training pool $\widetilde{\mathcal{D}}$ by $\hat{\mathcal{D}}_1$, and jointly update $\hat{\mathbf{y}}$ and $\mathcal{F}_\theta(\cdot)$ by using $\widetilde{\mathcal{D}}$ with T epochs.

2: For $t = 2, \cdots, |\mathcal{Y}|$

3: Add $\hat{\mathcal{D}}_t$ into $\widetilde{\mathcal{D}} = \widetilde{\mathcal{D}} \cup \hat{\mathcal{D}}_t$

4: Continue to jointly update $\hat{\mathbf{y}}$ and $\mathcal{F}_\theta(\cdot)$ by using $\widetilde{\mathcal{D}}$ with T epochs.

Given any $\widetilde{\mathcal{D}}$, we apply the following training objective with respect to $\mathbf{y}$ and θ:

$$\mathcal{L}(\widetilde{\mathcal{D}}; \hat{\mathbf{y}}, \theta) = -\frac{1}{|\widetilde{D}|M} \sum_{i=1}^{|\widetilde{D}|} \sum_{j=1}^{M} (1 - \mathbf{p}_{ij})^\gamma \ell_{ce}(\mathbf{p}_{ij}, \hat{\mathbf{y}}_{ij}), \tag{3}$$

where ℓ_{ce} is the cross-entropy loss; and γ is a hyper-parameter used to mitigate the data imbalance (Lin et al. 2017). In terms of the parameter θ, we can update it by using gradient-based methods. We update $\hat{\mathbf{y}}_{ij}$ by refining the current predictions $\mathbf{p}_{ij}$ as follows:

$$\hat{\mathbf{y}}_{ij} = \begin{cases} \frac{\hat{\mathbf{p}}_{ij} \circ \mathbf{y}_i}{|\hat{\mathbf{p}}_{ij} \circ \mathbf{y}_i|_1}, & \text{if } \max(\frac{\hat{\mathbf{p}}_{ij} \circ \mathbf{y}_i}{|\hat{\mathbf{p}}_{ij} \circ \mathbf{y}_i|_1}) > \alpha \\ \hat{\mathbf{y}}_{ij}, & \text{otherwise} \end{cases} \tag{4}$$

where $\circ$ denotes the operation of element-wise product; $|\cdot|_1$ is the ℓ_1 norm; $\max(\cdot)$ is the maximum operation for a vector; α is the confidence threshold; and $\hat{\mathbf{p}}_{ij}$ is the sharpened version of $\mathbf{p}_{ij}$ computed with a temperature parameter τ as follows:

$$\hat{\mathbf{p}}_{ij} = \frac{\mathbf{p}_{ij}^{1/\tau}}{\sum_{k=1}^{|\mathcal{Y}|} \mathbf{p}_{ik}^{1/\tau}}. \tag{5}$$

The above updating strategy with respect to $\hat{\mathbf{y}}$ indicates that (1) each of $\hat{\mathbf{y}}_{ij}$ must be covered by $\mathbf{y}_i$, and (2) only the high-confidence prediction $\mathbf{p}_{ij}$, measured by α, will be used to update $\hat{\mathbf{y}}_{ij}$. In DERC-PL, we can apply any off-the-shelf ERC method as the **base ERC predictor** $\mathcal{F}_\theta(\cdot)$. The computational steps of DERC-PL is outlined in **Algorithm 1**.

[2]Note that some subsets may be empty.

Algorithm 1: Computation of DERC-PL

Input: Training dataset $\mathcal{D}$, parameters γ, α, number of epochs T.
Output: An ERC predictor.

```
Apply an off-the-shelf ERC method as the base ERC predictor F_θ(·)
θ ← Initialize the predictor parameter randomly
ŷ ← Initialize the pseudo-utterance-emotions by Eq. 1
{D̂_g}_{g=1}^{g=|Y|} ← Divide D̂ into training subsets
D̃ ← Initialize a training pool by D̂_1
Update θ with D̃ by minimizing Eq.3 over T epochs
Update ŷ by Eq.4 per-epoch
for t ∈ 2, ..., |Y| do
    D̃ ← Add D̂_t into D̃
    Update θ by D̃ by minimizing Eq.3 over T epochs
    Update ŷ by Eq.4 per-epoch
end for
```

Related Works

Emotion Recognition in Conversations

The primary challenge of ERC is how to capture the contextual information of conversations. The body of ERC methods can be categorised into three streams: content-based, knowledge-assisted, and relation-based methods. The idea of content-based methods is straightforward, where, as the name suggests, they integrate utterances with the historical content from the same conversation before feeding them into various text encoders (Jiao et al. 2019; Lu et al. 2020; Li et al. 2020; Hu, Wei, and Huai 2021; Tu et al. 2022; Song et al. 2022; Zhang et al. 2023a; Wei et al. 2023). In parallel, the knowledge-assisted methods promote utterance representations by applying external knowledge tools such as knowledge bases (Zhong, Wang, and Miao 2019; Zhang et al. 2020; Ghosal et al. 2020; Jiang et al. 2022; Li et al. 2023)), and pieces of auxiliary information such as speaker background (Majumder et al. 2019; Chen et al. 2023; Hu et al. 2023; Zhang et al. 2023b) and discourse role (Ong et al. 2022)). Additionally, the relation-based methods capture the contextual information of conversations by transforming each conversation into a graph, whose nodes are utterances and edges are generated by various relations and dependencies, *e.g.,* speaker-utterance relations (Song et al. 2023), speaker dependency (Ghosal et al. 2019; Ishiwatari et al. 2020; Zhang, Chen, and Chen 2023), and discourse dependency (Zhang, Chen, and Chen 2023). They then apply graph neural networks to generate discriminative utterance representations.

The aforementioned ERC methods are all built on fine-grained training datasets with utterance-level emotions, which are expensive to collect. In contrast, in this work, we investigate DERC, a weakly-supervised learning task of ERC, and suggest a new DERC framework named DERC-PL, which is built on coarse-grained training datasets with conversation-level emotion sets. Despite its superior performance, DERC-PL can be treated as an efficient candidate for ERC methods in real-world scenarios.

Weakly-supervised Learning with Bag-level Supervision

To some extent, DERC can be considered a special case of Multi-Instance Multi-Label learning (MIML) (Zhou et al. 2012), a prevalent paradigm of weakly-supervised learning in which the bag of instances is associated with a bag-level label set, instead of instance-level labels. Early MIML methods use traditional machine learning methods such as ensembling (Wu, Huang, and Zhou 2014), boosting (Zhang and Zhang 2006) and maximum margin (Zhang and Zhou 2008). Recently, MIML works utilize many cutting-edge technologies such as self-training (Wang et al. 2023) and contrastive learning (Liu et al. 2023). MIML has been applied to many fields such as sentiment analysis (Li et al. 2020; Ji et al. 2020; Ouyang et al. 2024; Yang et al. 2023), offensive detection (Liu et al. 2022) and relation extraction (Surdeanu et al. 2012).

In parallel, another relevant paradigm of weakly-supervised learning to DERC is Learning from Label Proportions (LLP), where the bag of instances is associated with the bag-level proportion of labels. They mainly concentrate on the bag-level learning objectives, such as proportion loss (Liu et al. 2021a), forward correction loss (Zhang, Wang, and Scott 2022), consistency regularization (Tsai and Lin 2020), contrastive loss (Yang, Zhang, and Lam 2021) and so on (Dulac-Arnold et al. 2019). Such methods integrate instance-level prediction results into bag-level outputs and calculate the empirical risk loss with truth bag-level label proportion.

In contrast to MIML and LLP, one distinction of DERC is that each conversation can be considered as a coupled bag of utterances, *i.e.,* the utterances of one conversation are dependent. Besides, in many real-world scenarios, the training dataset of DERC may contain auxiliary information such as speakers' backgrounds, which can be leveraged to promote the predictor induction.

Experiments

In this section, we conduct experiments to evaluate **DERC-PL**, and attempt to answer the following questions:

Q1 : Can **DERC-PL** compete with the existing weakly-supervised learning methods in DERC settings?

Q2 : Can **DERC-PL** compete with the existing supervised learning ERC methods?

Experimental Settings

Datasets. We employ three benchmark ERC datasets: MELD (Poria et al. 2019), IEMOCAP (Busso et al. 2008), and EmoryNLP (Zahiri and Choi 2018). Statistics of these datasets are listed in Table 3. For each dataset, we generate its DERC version by directly adding the utterance-level emotions into the corresponding conversation-level emotion sets.

Baseline methods. We consider four ERC methods as the **base ERC predictors**, including naive **BERT$_{\text{base}}$+MLP**, **RGAT** (Ishiwatari et al. 2020), **SACL** (Hu et al. 2023), and **DualGATs** (Zhang, Chen, and Chen 2023). To evaluate the performance on DERC datasets, each of the ERC methods is integrated with three weakly-supervised learning methods including **PLOT** (Liu et al. 2021a), **ItS2CLR** (Liu et al. 2023), and the proposed **DERC-PL**. Besides, to address *Q2*, we supplement each ERC method with an additional supervised learning scenario under ERC datasets for comparison.

Implementation details. Our experiments are conducted on Ubuntu 20.04 with a single RTX-4090 GPU with 24G memory. For BERT-based methods (*i.e.,* BERT$_{\text{base}}$+MLP, RGAT (Ishiwatari et al. 2020)) / RoBERTa-based methods (*i.e.,* SACL (Hu et al. 2023), DualGAT (Zhang, Chen, and Chen 2023)), we use the AdamW optimizer (Loshchilov and Hutter 2019), with learning rates of $2e^{-5}$ and $1e^{-4}$, respectively. The layer dropout rate, batch size, and the number of epochs T are configured to $0.1/0.2$, $16/16$, and $30/20$, respectively. The hyperparameter α is adjusted to 0.8 for IEMOCAP (Busso et al. 2008), 0.3 for EmoryNLP (Zahiri and Choi 2018), and 0.4 for MELD (Poria et al. 2019). The pre-trained BERT$_{\text{base}}$ backbone can be downloaded from Huggingface.[3]

In terms of the two weakly supervised learning baselines PLOT (Liu et al. 2021a) and ItS2CLR (Liu et al. 2023), we effectuate modifications to fit DERC. Specifically, since PLOT requires the conversation-level proportion of emotions, we calculate the proportion by aggregating the utterance-level emotions for each conversation. [4] For ItS2CLR, we first conduct a warm-up stage using the training subset $\hat{\mathcal{D}}_1$, followed by updating the predictor with the entire training dataset. As an MIML method, ItS2CLR outputs the emotion with the highest predicted score.

Evaluation metrics. Inspired by (Zhang and Zhou 2013), we employ two widely-used binary-based metrics to measure the performance, including Macro Averaging F1 (Macro-F1) and Micro Averaging F1 (Micro-F1). In terms of all baseline methods, we run their source codes 5 times for each dataset and report their average results.

Main Results

Results for *Q1*. Table 4 presents the empirical comparison results between two weakly supervised methods (*i.e.,* PLOT and ItS2CLR) and our proposed DERC-PL across the three benchmark datasets, using four different base ERC predictors (*i.e.,* BERT$_{\text{base}}$+MLP, RGAT, SACL, and DualGATs). We can observe that DERC-PL consistently outperforms them with all base ERC predictors across all datasets, where the performance gain is about 3%$\sim$6%. One exception is in EmoryNLP, where PLOT outperforms our method over Micro-F1 evaluation metric using BERT$_{\text{base}}$+MLP ERC predictor. This likely stems from the severe imbalance in the EmoryNLP dataset, and an excessively high proportion of Neutral emotions can adversely affect the calculation of Micro-F1, which is sensitive to category distribution. Furthermore, the t-test (Wu and Zhang 2018) at 0.05 significance level is conducted to analyze whether DERC-PL

[3] https://huggingface.co/google-bert/bert-base-uncased

[4] It means that PLOT actually applies more supervised signals than ItS2CLR and DERC-PL.

Dataset	Conversation				Utterance			
	Total	Train	Validation	Test	Total	Train	Validation	Test
IEMOCAP	151	120		31	7,433	5,810		1,623
EmoryNLP	827	659	89	79	9,489	7,551	95	984
MELD	1,432	1,039	114	280	13,708	9,989	1,109	2,610

Table 3: The statistics of benchmark datasets.

Method	IEMOCAP		EmoryNLP		MELD	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
BERT$_{base}$+MLP						
with PLOT (Liu et al. 2021a)	54.13 •	52.94 •	**34.41** ○	27.52 •	53.67 •	50.00 •
with ItS2CLR (Liu et al. 2023)	56.25 •	53.32 •	32.01 •	26.55 •	56.44 •	50.72 •
with **DERC-PL** (**Ours**)	**60.04**	**58.07**	33.65	**30.38**	**58.74**	**53.54**
supervised learning	64.98	60.15	36.74	33.98	61.84	57.24
RGAT (Ishiwatari et al. 2020)						
with PLOT (Liu et al. 2021a)	56.78 •	55.88 •	28.84 •	27.47 •	55.21 •	52.07 •
with ItS2CLR (Liu et al. 2023)	58.78 •	57.65 •	31.30 •	28.98 •	58.19 •	54.32 •
with **DERC-PL** (**Ours**)	**61.61**	**59.62**	**34.95**	**30.14**	**62.21**	**56.79**
supervised learning	–	65.22	–	34.42	–	60.91
SACL (Hu et al. 2023)						
with PLOT (Liu et al. 2021a)	57.39 •	56.30 •	32.75 •	29.97 •	56.59 •	54.85 •
with ItS2CLR (Liu et al. 2023)	59.58 •	57.14 •	33.43 •	29.68 •	60.77 •	55.52 •
with **DERC-PL** (**Ours**)	**63.74**	**60.15**	**36.38**	**32.48**	**62.24**	**58.12**
supervised learning	69.08	69.22	42.21	39.65	67.51	66.45
DualGATs (Zhang, Chen, and Chen 2023)						
with PLOT (Liu et al. 2021a)	57.45 •	56.39 •	32.68 •	29.63 •	56.97 •	53.28 •
with ItS2CLR (Liu et al. 2023)	62.60 •	57.62 •	33.94 •	28.65 •	59.57 •	56.52 •
with **DERC-PL** (**Ours**)	**64.37**	**61.05**	**37.68**	**32.78**	**63.82**	**59.32**
supervised learning	–	67.68	–	40.69	–	66.90

Table 4: Empirical results of Marco-F1 and Micro-F1 on benchmark datasets, where •/○ indicates whether DERC-PL is significantly superior/inferior to one weakly-supervised learning baseline via paired t-test at 0.05 significance level. The best scores among weakly-supervised learning methods are indicated in bold.

achieves statistically superior performance to other weakly-supervised methods.

A more granular analysis from the perspective of each dataset reveals that DERC-PL outperforms PLOT and ItS2CLR in Micro-F1 (Macro-F1) by approximately 4.8% (4.3%) and 3.1% (3.3%) on IEMOCAP, 3.5% (2.8%) and 3% (3%) on EmoryNLP, and 6.1% (4.4%) and 3% (2.7%) on MELD, respectively. A similar analysis by base ERC predictors shows that DERC-PL's Micro-F1 (Macro-F1) scores surpass those of PLOT and ItS2CLR by approximately 3.4% (3.8%) and 2.6% (3.8%) on BERT$_{base}$+MLP, 6% (3.6%) and 3.5% (1.9%) on RGAT, 5.2% (3.2%) and 2.9% (2.8%) on SACL, and 6.3% (4.6%) and 3.3% (3.5%) on DualGATs. Besides, we would remind that PLOT has applied the conversation-level proportion of emotions, so the performance gain on PLOT further indicates the effectiveness of DERC-PL on the task of DERC.

Results for *Q2*. For each base ERC predictor, we compare the performance with various weakly supervised methods in the DERC scenario against the supervised ERC dataset scenario. The empirical findings are presented in Table 4. Overall, DERC-PL's F1 scores in the weakly supervised scenario are notably close to those in the supervised setting, with performance gaps across all benchmark datasets ranging from 2.1%∼9.1%. Surprisingly, under the BERT$_{base}$+MLP predictor, DERC-PL consistently shows a gap of approximately 3% compared to supervised learning, particularly with the Macro-F1 gap on IEMOCAP being as low as 2%. These competitive results, particularly when compared with supervised learning, suggest that DERC-PL is a strong candidate for ERC methods in real-world applications.

Ablation Study

We conduct the ablation study to evaluate two ablative versions of DERC-PL: (1) without updating the pseudo-utterance-level emotions using Eq.4 (w/o label updating) and (2) without progressively learning from $\hat{\mathcal{D}}_2$ to $\hat{\mathcal{D}}_{|\mathcal{Y}|}$ after initialization with $\hat{\mathcal{D}}_1$ (w/o training subset). Moreover, we also evaluate their combination. The three ablation versions are compared with the full DERC-PL, and the results

Method	IEMOCAP		EmoryNLP		MELD	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
w/o **training subset + label updating**	47.78	46.87	31.02	24.72	53.21	46.24
w/o **label updating**	53.51	52.84	32.53	26.91	54.25	51.12
w/o **training subset**	54.74	53.87	32.60	27.91	55.21	52.07
DERC-PL	**60.04**	**58.07**	**33.65**	**30.28**	**58.73**	**53.54**

Table 5: The empirical results of ablation study based on **BERT$_{\text{base}}$+MLP**. The best scores among different versions of DERC-PL are indicated in bold.

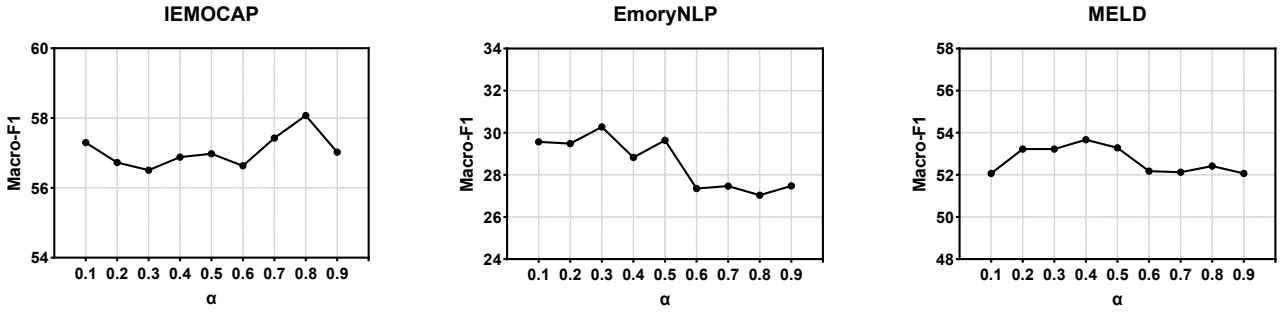

Figure 2: Sensitivity analysis of confidence threshold α of pseudo-utterance-level emotion updating based on **BERT$_{\text{base}}$+MLP**.

are shown in Table 5. Overall, the empirical results demonstrate that the full version of DERC-PL significantly outperforms all ablative versions, proving that both techniques have positive effect on promoting the classification performance. For example, the Macro-F1 scores of DERC-PL are approximately 5.2% higher than those of w/o label updating on IEMOCAP, and about 2.4% higher than those of w/o training subset on MELD. Moreover, the performance gain of DERC-PL over the version w/o both techniques are even about 12%$\sim$ 13% across IEMOCAP. In terms of EmoryNLP and MELD, the gain of Macro-F1 is much higher (*i.e.,* 6.5%$\sim$9.3%) than the gain of Micro-F1 (*i.e.,* 2.6%$\sim$5.5%), likely due to the datasets' imbalance with a high proportion of `Neutral`. All those results further indicate that the two key techniques of DERC-PL can significantly promote the classification performance, and make DERC-PL more robust even for imbalanced datasets.

Sensitivity Analysis of α

In this experiment, we investigate the confidence threshold α of pseudo-utterance-level emotion updating. We report the Macro-F1 scores of DERC-PL using BERT$_{\text{base}}$+MLP with α ranging from $\{0.1, 0.2, \ldots, 0.9\}$, as shown in Fig.2. It can be seen that the performance trends of different α values are similar across EmoryNLP and MELD, whose the number of utterances per conversation is relatively smaller. The higher scores are achieved at lower α values such as $\{0.2, 0.3, 0.4\}$. Another reason for such phenomenon is the imbalanced emotion distributions in these datasets, where the `Neutral` emotion occupies a high percentage of utterances, making precise predictions for other emotions challenging. In this case, only smaller values of α can maintain the update of pseudo-utterance-level emotions. In contrast, on IEMOCAP, the higher scores can be achieved by larger values of α such as $\{0.7, 0.8\}$. The major reason is that IEMOCAP does not suffer from the imbalanced problem, leading to more precise predictions. Therefore, one needs higher values of α to retain high-confidence pseudo-utterance-level emotions.

Conclusion

In this paper, we investigate a new weakly-supervised learning task of EDR named **DERC**, in which the training dataset is associated with conservation-level emotion sets, instead of utterance-level emotions. To examine the labeling accuracy and efficiency of DERC, we conduct preliminary experiments of manual labeling by inviting 20 volunteers on the public benchmark dataset MELD. Our empirical results demonstrate that labeling conversation-level emotion sets can be simultaneously accurate and efficient, to support the potential task of DERC. We then propose a novel DERC framewokr named **DERC-PL** built on the spirit of pseudo-labeling, which jointly updates pseudo-utterance-level emotions and the ERC predictor in a self-training manner. We conduct loads of experiments to evaluate the performance of DERC-PL. The empirical results demonstrate that DERC-PL consistently outperforms the competitive weakly-supervised learning methods PLOT and ItS2CLR; and it can be on par with the supervised learning methods, further indicating DERC-PL can be an effective candidate for ERC methods.

Acknowledgements

We would like to acknowledge support for this project from the National Science and Technology Major Project (No.2021ZD0112500), and the National Natural Science Foundation of China (No.62276113).

References

Alhuzali, H.; and Ananiadou, S. 2023. Improving textual emotion recognition based on intra-and inter-class variation. *IEEE Transactions on Affective Computing*, 14: 1297–1307.

Barnes, R. K. 2019. Conversation analysis of communication in medical care: description and beyond. *Research on Language and Social Interaction*, 52(3): 300–315.

Brambilla, M.; Javadian Sabet, A.; Kharmale, K.; and Sulistiawati, A. E. 2022. Graph-based conversation analysis in social media. *Big Data and Cognitive Computing*, 6(4): 113.

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: 335–359.

Chen, J.; Huang, P.; Huang, G.; Li, Q.; and Xu, Y. 2023. SDTN: Speaker Dynamics Tracking Network for Emotion Recognition in Conversation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5.

Chen, S.-Y.; Hsu, C.-C.; Kuo, C.-C.; and Ku, L.-W. 2018. Emotionlines: An emotion corpus of multi-party conversations. In *International Conference on Language Resources and Evaluation*.

Duggan, G. B.; and Payne, S. J. 2011. Skim reading by satisficing: evidence from eye tracking. In *SIGCHI Conference on Human Factors in Computing Systems*, 1141–1150.

Dulac-Arnold, G.; Zeghidour, N.; Cuturi, M.; Beyer, L.; and Vert, J.-P. 2019. Deep multi-class learning from label proportions. *arXiv preprint arXiv:1905.12909*.

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

Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; and Gelbukh, A. 2019. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In *Conference on Empirical Methods in Natural Language Processing*, 154–164.

Hu, D.; Bao, Y.; Wei, L.; Zhou, W.; and Hu, S. 2023. Supervised Adversarial Contrastive Learning for Emotion Recognition in Conversations. In *Annual Meeting of the Association for Computational Linguistics*, 10835–10852.

Hu, D.; Wei, L.; and Huai, X. 2021. DialogueCRN: Contextual Reasoning Networks for Emotion Recognition in Conversations. In *Annual Meeting of the Association for Computational Linguistics*, 7042–7052.

Ishiwatari, T.; Yasuda, Y.; Miyazaki, T.; and Goto, J. 2020. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Conference on Empirical Methods in Natural Language Processing*, 7360–7370.

Ji, Y.; Liu, H.; He, B.; Xiao, X.; Wu, H.; and Yu, Y. 2020. Diversified multiple instance learning for document-level multi-aspect sentiment classification. In *Conference on Empirical Methods in Natural Language Processing*, 7012–7023.

Jiang, D.; Wei, R.; Wen, J.; Tu, G.; and Cambria, E. 2022. AutoML-Emo: Automatic Knowledge Selection using Congruent Effect for Emotion Identification in Conversations. *IEEE Transactions on Affective Computing*.

Jiao, W.; Yang, H.; King, I.; and Lyu, M. R. 2019. HiGRU: Hierarchical Gated Recurrent Units for Utterance-Level Emotion Recognition. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 397–406.

Li, C.; Dai, Y.; Feng, L.; Li, X.; Wang, B.; and Ouyang, J. 2024. Positive and Unlabeled Learning with Controlled Probability Boundary Fence. In *International Conference on Machine Learning*.

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 *International Conference on Computational Linguistics*, 4190–4200.

Li, W.; Zhu, L.; Mao, R.; and Cambria, E. 2023. Skier: A symbolic knowledge integrated model for conversational emotion recognition. In *AAAI Conference on Artificial Intelligence*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*, 2980–2988.

Liu, J.; Kong, D.; Huang, L.; Mao, D.; and Xue, H. 2022. Multiple Instance Learning for Offensive Language Detection. In *Findings of the Association for Computational Linguistics*, 7387–7396.

Liu, J.; Wang, B.; Shen, X.; Qi, Z.; and Tian, Y. 2021a. Two-stage Training for Learning from Label Proportions. In *International Joint Conference on Artificial Intelligence*, 2737–2743.

Liu, K.; Zhu, W.; Shen, Y.; Liu, S.; Razavian, N.; Geras, K. J.; and Fernandez-Granda, C. 2023. Multiple instance learning via iterative self-paced supervised contrastive learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3355–3365.

Liu, S.; Zheng, C.; Demasi, O.; Sabour, S.; Li, Y.; Yu, Z.; Jiang, Y.; and Huang, M. 2021b. Towards Emotional Support Dialog Systems. In *Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing*, 3469–3483.

Loshchilov, I.; and Hutter, F. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Lu, X.; Zhao, Y.; Wu, Y.; Tian, Y.; Chen, H.; and Qin, B. 2020. An iterative emotion interaction network for emotion recognition in conversations. In *International Conference on Computational Linguistics*, 4078–4088.

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.; and Cambria, E. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *AAAI conference on artificial intelligence*, 6818–6825.

Matsuo, S.; Bise, R.; Uchida, S.; and Suehiro, D. 2023. Learning From Label Proportion with Online Pseudo-Label Decision by Regret Minimization. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.

Mittal, T.; Bhattacharya, U.; Chandra, R.; Bera, A.; and Manocha, D. 2020. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *AAAI Conference on Artificial Intelligence*, 2, 1359–1367.

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 Conference on Artificial Intelligence*, 11121–11129.

Ouyang, J.; Yang, Z.; Liang, S.; Wang, B.; Wang, Y.; and Li, X. 2024. Aspect-Based Sentiment Analysis with Explicit Sentiment Augmentations. In *AAAI Conference on Artificial Intelligence*, 18842–18850.

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

Priya, P.; Firdaus, M.; and Ekbal, A. 2023. A multi-task learning framework for politeness and emotion detection in dialogues for mental health counselling and legal aid. *Expert Systems with Applications*, 224: 120025.

Qin, X.; Wu, Z.; Zhang, T.; Li, Y.; Luan, J.; Wang, B.; Wang, L.; and Cui, J. 2023. BERT-ERC: Fine-tuning BERT is enough for emotion recognition in conversation. In *AAAI Conference on Artificial Intelligence*, 13492–13500.

Song, R.; Giunchiglia, F.; Shi, L.; Shen, Q.; and Xu, H. 2023. SUNET: Speaker-utterance interaction Graph Neural Network for Emotion Recognition in Conversations. *Engineering Applications of Artificial Intelligence*, 123: 106315.

Song, X.; Huang, L.; Xue, H.; and Hu, S. 2022. Supervised Prototypical Contrastive Learning for Emotion Recognition in Conversation. In *Conference on Empirical Methods in Natural Language Processing*, 5197–5206.

Surdeanu, M.; Tibshirani, J.; Nallapati, R.; and Manning, C. D. 2012. Multi-instance multi-label learning for relation extraction. In *Joint Conference on Empirical Methods in Batural Language Processing and Computational Natural Language Learning*, 455–465.

Tsai, K.-H.; and Lin, H.-T. 2020. Learning from label proportions with consistency regularization. In *Asian Conference on Machine Learning*, 513–528.

Tu, G.; Liang, B.; Jiang, D.; and Xu, R. 2022. Sentiment-Emotion-and Context-guided Knowledge Selection Framework for Emotion Recognition in Conversations. *IEEE Transactions on Affective Computing*.

Wang, Y.; Zhao, Y.; Wang, Z.; and Wang, M. 2023. Robust self-supervised multi-instance learning with structure awareness. In *AAAI Conference on Artificial Intelligence*, 8, 10218–10225.

Wei, J.; Hu, G.; Tuan, L. A.; Yang, X.; and Zhu, W. 2023. Multi-Scale Receptive Field Graph Model for Emotion Recognition in Conversations. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.

Wu, J.-S.; Huang, S.-J.; and Zhou, Z.-H. 2014. Genome-wide protein function prediction through multi-instance multi-label learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11(5): 891–902.

Wu, X.; and Zhang, M.-L. 2018. Towards Enabling Binary Decomposition for Partial Label Learning. In *International Joint Conference on Artificial Intelligence*, 2868–2874.

Yang, H.; Zhang, W.; and Lam, W. 2021. A two-stage training framework with feature-label matching mechanism for learning from label proportions. In *Asian Conference on Machine Learning*, 1461–1476.

Yang, Z.; Wang, B.; Li, X.; Wang, W.; and Ouyang, J. 2023. S3map: Semisupervised aspect-based sentiment analysis with masked aspect prediction. *Knowledge-based Systems*, 269: 110513.

Zahiri, S. M.; and Choi, J. D. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Workshops at the AAAI Conference on Artificial Intelligence*.

Zhang, D.; Chen, F.; and Chen, X. 2023. Dualgats: Dual graph attention networks for emotion recognition in conversations. In *Annual Meeting of the Association for Computational Linguistics*, 7395–7408.

Zhang, D.; Chen, X.; Xu, S.; and Xu, B. 2020. Knowledge aware emotion recognition in textual conversations via multi-task incremental transformer. In *International Conference on Computational Linguistics*, 4429–4440.

Zhang, J.; Wang, Y.; and Scott, C. 2022. Learning from label proportions by learning with label noise. *Neural Information Processing Systems*, 35: 26933–26942.

Zhang, M.; Zhou, X.; Chen, W.; and Zhang, M. 2023a. Emotion Recognition in Conversation from Variable-Length Context. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5.

Zhang, M.-L.; and Zhou, Z.-H. 2008. M3MIML: A maximum margin method for multi-instance multi-label learning. In *IEEE International Conference on Data Mining*, 688–697.

Zhang, M.-L.; and Zhou, Z.-H. 2013. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8): 1819–1837.

Zhang, T.; Chen, Z.; Zhong, M.; and Qian, T. 2023b. Mimicking the Thinking Process for Emotion Recognition in Conversation with Prompts and Paraphrasing. In *International Joint Conference on Artificial Intelligence*, 6299–6307.

Zhang, Z.-L.; and Zhang, M.-L. 2006. Multi-instance multi-label learning with application to scene classification. *Neural Information Processing Systems*, 1609–1616.

Zhong, P.; Wang, D.; and Miao, C. 2019. Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations. In *Conference on Empirical Methods in Natural Language Processing*, 165–176.

Zhou, Z.-H.; Zhang, M.-L.; Huang, S.-J.; and Li, Y.-F. 2012. Multi-instance multi-label learning. *Artificial Intelligence*, 176(1): 2291–2320.