Empathy Intent Drives Empathy Detection

Liting Jiang[1,2,3], **Di Wu**[1,2,3], **Bohui Mao**[1,2,3], **Yanbing Li**[1,2,3], **Wushour Slamu**[1,2,3]*
[1]School of Computer Science and Technology, Xinjiang University, Urumqi, China
[2]Resource Monitoring and Research Center on Minority Languages, Urumqi, China
[3]Xinjiang Laboratory of Multi-language Information Technology, Urumqi, China
{107556521210, wd11, 107552204050}@stu.xju.edu.cn
{liyb, wushour}@xju.edu.cn

Abstract

Empathy plays an important role in the human dialogue. Detecting the empathetic direction expressed by the user is necessary for empathetic dialogue systems because it is highly relevant to understanding the user's needs. Several studies have shown that empathy intent information improves the ability to response capacity of empathetic dialogue. However, the interaction between empathy detection and empathy intent recognition has not been explored. To this end, we invite 3 experts to manually annotate the healthy empathy detection datasets IEMPATHIZE and TwittEmp with 8 empathy intent labels, and perform joint training for the two tasks. Empirical study has shown that the introduction of empathy intent recognition task can improve the accuracy of empathy detection task, and we analyze possible reasons for this improvement. To make joint training of the two tasks more challenging, we propose a novel framework, Cascaded Label Signal Network, which uses the cascaded interactive attention module and the label signal enhancement module to capture feature exchange information between empathy and empathy intent representations. Experimental results show that our framework outperforms all baselines under both settings on the two datasets. [1]

1 Introduction

Empathy is essential in human social interaction. In the process of human dialogue, empathy enables listeners to establish rapport with speakers by understanding their emotional and cognitive states, arousing their interest, and comforting them (Kim et al., 2022). Therefore, it is worthwhile to detect the empathetic direction of dialogue utterances. In recent years, researchers have studied empathy detection in various fields, such as mental health support (Sharma et al., 2020; Zhou and Jurgens, 2020), empathetic expression understanding in newswire (Buechel et al., 2018), medical and healthcare (Khanpour et al., 2017; Hosseini and Caragea, 2021a; Chen et al., 2020; Wijaya et al., 2023), human-computer interaction (Virvou and Katsionis, 2004; Xie and Pu, 2021; Gao et al., 2021; Samad et al., 2022), etc.

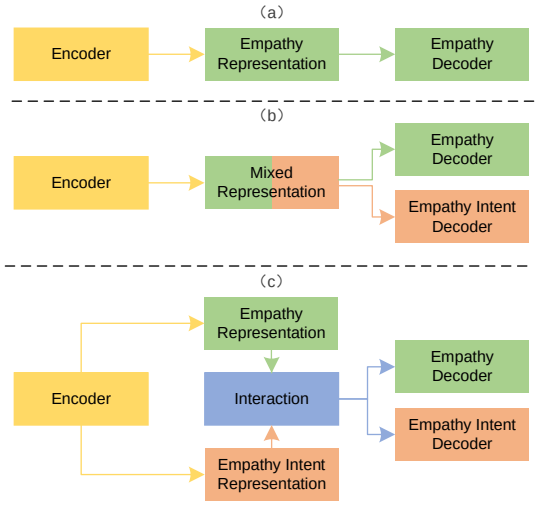

Figure 1: Different modeling approaches.

Currently, millions of people seek psychological support by expressing their emotions in online health communities and look forward to receiving support from peers who may have had similar experiences and can understand their feelings. Therefore, some researchers have performed studies on the direction of empathy expression in people's utterances. For example, the utterance: "*I lost my mom to cancer in April and just miss her so much. There are so many pieces to work on and I find it so hard to work on bc my grief is so strong*". The model needs to detect the direction of empathy expressed by the user, '*Seek*'. Hosseini and Caragea (2021a) provided an online healthy dataset and a baseline for empathy detection (ED), but their methods are monotonous, as shown in Figure 1 (a). They did not consider that humans have potential empa-

*Corresponding author.
[1]https://github.com/JiangT7/CLSN

thy intent information while expressing empathy. Chen et al. (2022) learned the distribution of potential empathy intent and then combined implicit and explicit representations of empathy intent to generate responses with empathy intent. However, they do not explore the interaction between empathy and empathy intent representations.

In this paper, we invite 3 experts to manually annotate the empathy intent of each utterance in two empathy detection datasets IEMPATHIZE (Hosseini and Caragea, 2021b) and TwittEmp (Hosseini and Caragea, 2021a) according to the lexicon of empathy intent example utterance provided by Welivita and Pu (2020). On this basis, we use a simple joint training method to test the feasibility of empathy intent recognition (EIC) as an auxiliary task, as shown in Figure 1 (b). The experimental results show that the joint training of ED and EIC leads to higher accuracy of ED, and we find that the categories of ED are easier to distinguish after the introduction of EIC due to the obvious correspondence between empathy and empathy intent labels. To make joint training tasks more challenging, we propose a novel framework, called **C**ascaded **L**abel **S**ignal **N**etwork (CLSN). First, it uses the BERT (Kenton and Toutanova, 2019) model to obtain the semantic features of utterances and extracts different representations through two linear layers. Then, the cascaded interactive attention module is used to implement feature interaction and control knowledge flow. Finally, the label signal enhancement module is used to further extract interactive features from the label information of the two tasks and feed them to different decoders to complete ED and EIC. We model the interaction information flow between both tasks, as shown in Figure 1(c).

Overall, our contributions are as follows: (1) We invite 3 experts to manually annotate empathy intent labels on the ED datasets IEMPATHIZE and TwittEmp, provide the necessary datasets for the joint ED and EIC study; (2) To the best of our knowledge, we are the first to attempt joint training of the ED and EIC tasks to improve the accuracy of the ED task. We also explore the possible reasons for the accuracy improvement; (3) To make joint training of the two tasks more challenging, we propose a novel framework, CLSN, which explicitly controls the knowledge transfer between the two tasks. Experimental results show that our framework outperforms all baselines under both settings in the two datasets.

2 Related Work

2.1 Empathy Detection

Empathy detection, one of a series of empathy tasks, has been widely studied by many researchers (Hosseini and Caragea, 2021a; Chen et al., 2020). Currently, the task involves two types of research: analyzing empathy from text (Yang et al., 2019; Buechel et al., 2018; Sedoc et al., 2020; Ghosh et al., 2022), and from spoken dialogues (Fung et al., 2016; Kim et al., 2022; Alam et al., 2018; Pérez-Rosas et al., 2017; Ayshabi and Idicula, 2021). Empathy plays an important role in online health communities (Medeiros and Bosse, 2016) as it can facilitate the healing process by reducing psychological distress and increasing optimism through empathetic dialogue (Yalcin and DiPaola, 2018; Williams and Cano, 2005). Sharma et al. (2020) used a RoBERTa-based bi-encoder model to identify empathy in conversations on online mental health platforms. Khanpour et al. (2017) proposed a model based on Convolutional Networks and Long Short-Term Memory to identify empathetic messages in online health communities. Alam et al. (2018) proposed an Italian spoken empathetic dialogue system consisting only of paired conversations between patient and therapist in different audio. Bi et al. (2023) defined an empathy planner to capture and reason about multi-source information that considers cognition and affection. They also introduced a dynamic integrator module that allows the model dynamically select the appropriate information to generate empathetic responses.

2.2 Empathy Intent Recognition

Some empathetic dialogue generation studies incorporate empathy intent information, such as, Welivita and Pu (2020) have manually labeled 500 response intents. Using lexical and machine learning methods, they automatically analyzed utterances of the entire dataset with identified response intents and 32 emotion categories, and the information visualization method is used to summarize the emotional dialogue exchange model and its temporal evolution. Welivita et al. (2021) curated a novel large-scale silver dialogue dataset, EDOS (Emotional Dialogues in OpenSubtitles). It contains 1 million movie subtitles for emotional dialogue, with 32 fine-grained emotions, 8 intent categories, and a neutral category. Saha and Ananiadou (2022) proposed a fusion model of the Transformer model and the Hierarchical Encoder Decoder, called the

Hierarchical Transformer Network, to capture the speaker's emotion and dialogue context. To generate intent controlled empathetic responses, they used the results of Reinforcement Learning to implicitly optimize rewards.

3 Empirical Study

In this section, we introduce EIC as an auxiliary task in the ED task to verify its effectiveness. In addition, we explore the reasons for the influence of the EIC task on ED.

3.1 Problem Definition

Both ED and EIC can be regarded as classification tasks. Given an utterance $U = (u_1, u_2, \cdots, u_s)$, the ED task detects the direction of empathy based on the U, and the EIC task identifies the potential empathy intent of the user based on the U.

3.2 Dataset and Annotation

IEMPATHIZE[2] containing sentences from online cancer survivors. It contains 5007 sentences, 3 empathy labels: '*Seek*', '*provide*' and '*None*'. In total, 1046 sentences are annotated as '*Seek*', 966 as '*provide*', and 2995 as '*None*'. To introduce the EIC task, we annotate the intent labels on the empathy detection dataset. Inspired by Welivita and Pu (2020); Chen et al. (2022), we identify 7 empathy intent labels: '*Acknowledging*', '*Consoling*', '*Questioning*', '*Sympathizing*', '*Wishing*', '*Positive*', '*Negative*' and '*Neutral*' (all other unmentioned intents) by observing the IEMPATHIZE dataset. We recruit 3 experts to assign an empathy intent label to each utterance in the IEMPATHIZE dataset in combination with the lexicon of intent labels example utterance provided by Welivita and Pu (2020). The first round of annotation is completed by 2 experts, and Cohen's kappa coefficient is 88.4% for the empathy intent of the IEMPATHIZE dataset. For samples with inconsistent annotation, the 3rd expert decided which category the empathy intent belonged to. The distribution of the dataset is shown in Figure 2 (a).

Following previous work (Hosseini and Caragea, 2021a,b), we use both settings to create classifiers. The binary setting is used to identify utterances that seek empathy or provide empathy. For example, To create the seek-classifier, we set '*Seek*' as positive samples, '*None*' and '*Provide*' as negative samples. The provide-classifiers is created in similar way as seek-classifier. The multi-class setting considers three classes. In experiments, we split the dataset, keeping 60% of the data for the training set, 20% for the validation set, and 20% for the test set. See Appendix A for detailed statistics.

[2] https://github.com/Mahhos/Empathy

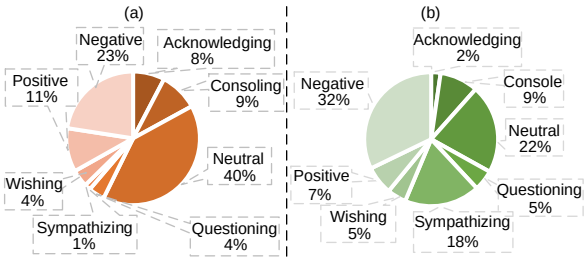

Figure 2: Distribution of empathy intent labels in the IEMPATHIZE dataset (a) and the TwittEmp dataset (b).

3.3 Empirical Study Results

Following the work of (Hosseini and Caragea, 2021b), BERT is used as a shared encoder to train the two tasks together and decode the hidden states separately to get the results of both tasks. We pre-experiment with IEMPATHIZE under both settings.

Table 1 shows a significant improvement in ED accuracy when the EIC task is introduced. The F1-score of the ED improved by 1.74%, 2.79%, and 5.61% under the both settings, demonstrating that implementing joint learning between EIC and ED can improve the performance of the ED task.

Settings	P(%)	R(%)	F1(%)
Seek	+2.11	+8.50	+2.79
Provide	+3.36	+6.63	+5.61
Multi-class	+1.06	+3.98	+1.74

Table 1: Experimental results of the empirical study. Multi-class denotes multi-class setting, seek and provide denote binary settings.

3.4 Analysis of Reasons for Improvement

The accuracy of the ED task has improved substantially, and two questions are asked to explore the reasons for the improvement. Question (1): After the introduction of the EIC task, why did the accuracy of the ED task improve significantly? Question (2): How does the introduction of the EIC task help to improve the accuracy of ED?

For Question (1): In terms of whether the introduction of the EIC task improves the performance of the model, we performed visualization on the test set of the IEMPATHIZE dataset (multi-class). The T-SNE (Van der Maaten and Hinton, 2008) results are shown in Figure 3.

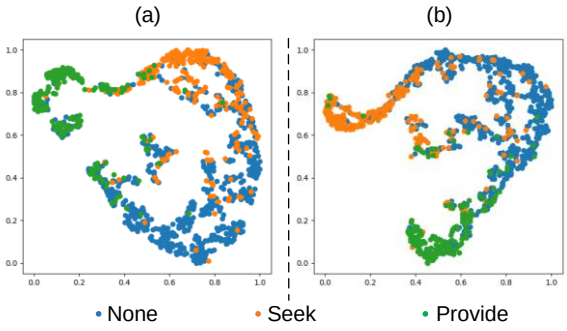

Figure 3: T-SNE visualization of the label representations on the IEMPATHIZE dataset.

The T-SNE results for the introduced EIC task are shown in Figure 3 (b). The categories '*Provides*' and '*Seek*' are more discriminable. In addition, the three categories are more aggregated. **This demonstrates the ability of the model to learn a better representation of the hidden state after the introduction of the EIC task, which is one of the important reasons for the accuracy gains achieved.**

For Question (2): We count the frequency of co-occurrence of different labels in the IEMPATHIZE training dataset, as shown in Table 2.

Labels	*None*	*Seek*	*Provide*
Acknowledging	**141**	38	50
Consoling	42	26	**202**
Neutral	**1126**	38	49
Questioning	**92**	26	4
Sympathizing	3	3	**39**
Wishing	14	10	**104**
Positive	**189**	26	**103**
Negative	187	**463**	29

Table 2: Empathy and empathy intent labels co-occurrences frequency.

From Table 2 we can see that when the empathy intent label is '*Acknowledging*', '*Neutral*', '*Questioning*' or '*Positive*', the empathy label is more likely to be '*None*', the empathy intent label is '*Negative*', the empathy label is more likely to be '*Seek*', the empathy intent label is '*Consoling*', '*Wishing*', '*Sympathizing*', or '*Positive*', the empathy label is more likely to be '*Provide*'. For example: "I am sorry to hear about your mother-in-law's troubles after her diagnosis, I am sympathetic", the word 'sympathetic' reflects the empathy intent of 'Sympathizing' and expresses the direction of empathy as 'Provide'. The EIC channel shares the captured semantics with the ED channel which can help the ED channel to better perceive the empathy information contained in the utterance. **Intuitively, since there are 8 possible empathy intent labels for an utterance and only 3 possible empathy labels, it is harder to accurately identify the empathy intent label of an utterance than it is to identify the empathy label. Once the model correctly predicts the empathy intent label of the utterance, the model is able to more accurately identify the empathy label of the utterance based on the apparent correspondence of the two labels co-occurrence frequency in Table 2.** Figure 4 shows some cases, and we can see that the simultaneous occurrence of the two labels is consistent with the co-occurrence frequency in Table 2. The empirical study is able to correctly predict the empathy labels of the three utterances, demonstrating that the introduction of the EIC task can help the model to improve the performance of ED tasks.

(a) So he is basically stuck at home watching TV.
BERT : Seek; Empirical Study : None, Neutral

(b) Sadly, it never filled in again and it has been 4 years now.
BERT : None; Empirical Study : Seek, Negative

(c) Just wanted to send best wishes for the bone scan too!
BERT : None; Empirical Study : Provide, Wishing

Figure 4: Case study. The red color indicates the wrong label, the green color indicates the correct label

We use the same data annotation approach to annotate TwittEmp[3]. The Cohen's kappa coefficient is 80.0%, The distribution of the dataset is shown in Figure 2 (b). TwittEmp was collected from Twitter on cancer topics and there are a total of 3000 sentences. The category labels are the same as IEMPATHIZE, where 1000 sentences are annotated as "*Seek*", 1000 as "*Provide*", and 1000 as "*None*".

Although the joint training method can improve the accuracy of the ED task, the modeling approach does not explicitly control the information flow between ED and EIC. To address this shortcoming and to make the joint training of the two tasks more challenging, we propose a novel framework in section 4.

4 Methodology

In this section, we provide a detailed description of the CLSN, which can effectively model the interaction between empathy and empathy intent representations. It consists of an encoding layer (§4.1), a cascaded interactive attention module (§4.2), a la-

[3] https://github.com/Mahhos/KDempathy

bel signal enhancement module (§4.3), and two separate classification decoders (§4.4). An overview of our framework is shown in Figure 5.

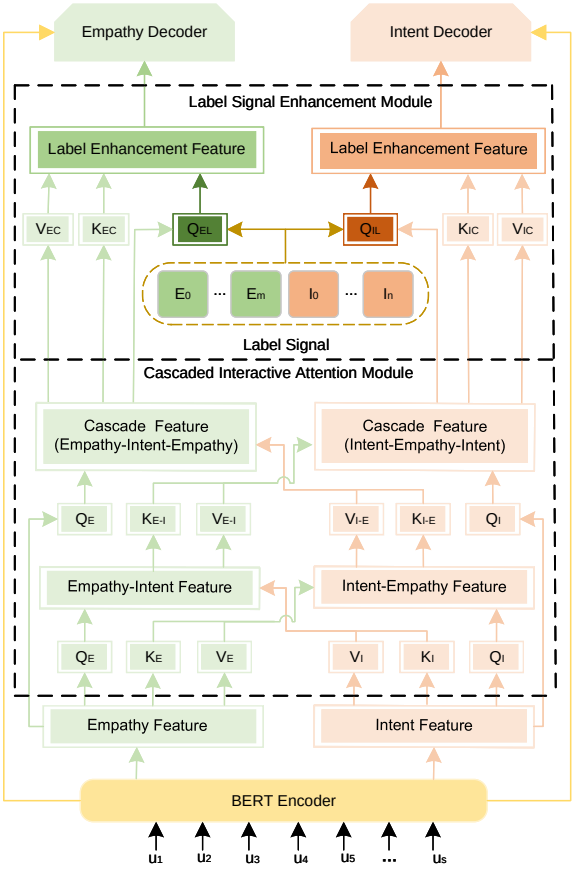

Figure 5: The overall architecture of the CLSN.

4.1 Encoding Layer

In our framework, ED shares an encoder with EIC. Given an utterance U, we first mark $[CLS]$ and $[SEP]$ at the beginning and end of the utterance and then use the pre-trained language model BERT (bert-base-uncased) to encode the semantic utterance to capture contextual semantic information.

$$\mathbf{H} = \text{BERT}\left([CLS] + U + [SEP]\right), \quad (1)$$

where $\mathbf{H} \in R^d$, and d denotes the hidden dimension. During the empirical study, we have found that the model can learn the task representation by itself, even without the additional label representation as an auxiliary feature guide. Therefore, before modeling the interaction between the two tasks of ED and EIC, we obtain different features of text encoding through two fully connected layers to allow the model to learn empathy and empathy intent representations by itself.

$$\mathbf{H}_\delta = \text{FC}_\delta\left(\mathbf{H}\right), \quad (2)$$

where $\delta \in \{I, E\}$, I denotes empathy intent and E denotes empathy, respectively. FC denotes a fully connected layer, $\mathbf{H}_E$ and $\mathbf{H}_I$ are empathy and empathy intent representations, respectively.

4.2 Cascaded Interactive Attention Module

The approach of simply sharing the hidden state is not sufficient to achieve explicit information transfer between two tasks. Therefore, we design a cascade of interactive attention modules. It can explicitly model the interaction between ED and EIC tasks. By this method, interactive information is continuously extracted features between tasks and finally returns to the original task to complete the cascade operation.

To make the two tasks learn features from each other, we use the attention mechanism (Vaswani et al., 2017) to connect the two tasks. First, we map the matrices $\mathbf{H}_E$ and $\mathbf{H}_I$ by linear projection to obtain the corresponding $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$, respectively. Then the attention mechanism is used to obtain the corresponding empathy representation $\mathbf{H}_{E-I}$ with empathy intent features and empathy intent representation $\mathbf{H}_{I-E}$ with empathy features, respectively. At the same time, we use the empathy and empathy intent representations combined with $\mathbf{H}_{I-E}$ and $\mathbf{H}_{E-I}$ respectively to construct the attention layer to obtain the cascaded interaction features $\mathbf{H}_{E-I-E}$ and $\mathbf{H}_{I-E-I}$.

$$\mathbf{H}_{\gamma-\delta} = \text{Attention}\left(\mathbf{H}_\gamma, \mathbf{H}_\delta, \mathbf{H}_\delta\right), \quad (3)$$

$$\mathbf{H}_{\delta-\gamma-\delta} = \text{Attention}\left(\mathbf{H}_\delta, \mathbf{H}_{\gamma-\delta}, \mathbf{H}_{\gamma-\delta}\right), \quad (4)$$

where γ and $\delta \in \{I, E\}$, if γ denotes E, then δ denotes I, and vice versa.

4.3 Label Signal Enhancement Module

To make the extracted features more effective, we constructed a label signal enhancement module. It combines the empathy labels $\mathbf{E}_M$ and the empathy intent labels $\mathbf{I}_N$ as label signals. Specifically, concatenating all label encodings of the two tasks as the query vectors of the module with the empathy and empathy intent representations for the attention computation can effectively capture, the semantics of the two information with respect to labels. Taking the empathy detection task as an example, the computation with label encoding can not only capture the semantics about empathy, but also capture the semantics about empathy intent for assisting the decoding of empathy, and the same is true for the empathy intent recognition task.The label signal

and the interaction features can obtain the corresponding representation of label perception. We do not explicitly use the predicted results to guide another task, as this could lead to problems with the error cascade. Formally, before decoding, we use the learnable label signal embedding $\mathbf{LS}_\delta$ as input to establish an association with the representations $\mathbf{H}_{E-I-E}$ and $\mathbf{H}_{I-E-I}$, respectively, which can explicitly enhance the importance of features that are easier to classify in feature information. We also add $\mathbf{H}$ for enhanced capture contextual semantic information.

$$\mathbf{LS}_\delta = \mathbf{I}_N \oplus \mathbf{E}_M \oplus \mathbf{H}_{\delta-\gamma-\delta}, \tag{5}$$

$$\mathbf{LH}_\delta = \text{Attention}\left(\mathbf{LS}_\delta, \mathbf{H}_{\delta-\gamma-\delta}, \mathbf{H}_{\delta-\gamma-\delta}\right), \tag{6}$$

$$\mathbf{H}^{'}_\delta = \mathbf{LH}_\delta + \mathbf{H}, \tag{7}$$

where $\mathbf{LH}_E$ and $\mathbf{LH}_I$ denote the empathy and the empathy intent representations after label signal enhancement. $\mathbf{H}^{'}_E$ and $\mathbf{H}^{'}_I$ denote the final empathy and the empathy intent representations.

4.4 Decoder

Finally, we apply the max-pooling operation to $\mathbf{H}^{'}_E$ and $\mathbf{H}^{'}_I$ to obtain the representations $\mathbf{H}_{EM}$ and $\mathbf{H}_{IM}$, respectively. Two separate decoders are used for ED and EIC.

$$y^E = \text{softmax}\left(\mathbf{W}^E\mathbf{H}_{EM} + \mathbf{b}^E\right), \tag{8}$$

$$y^I = \text{softmax}\left(\mathbf{W}^I\mathbf{H}_{IM} + \mathbf{b}^I\right), \tag{9}$$

where y^E, y^I are the predicted distributions for ED and EIC, respectively. $\mathbf{W}^E$ and $\mathbf{W}^I$ are trainable parameters, $\mathbf{b}^E$ and $\mathbf{b}^I$ are bias.

4.5 Joint Training

We use a joint training scheme to consider both ED and EIC and update the parameters by joint optimization. Lin et al. (2018) proposed to add a modulation factor $(1-p_i)^\eta$ to the cross-entropy loss to discriminate between easy/hard examples. p_i represents the probability value of the t-th class from softmax output. It can alleviate the problem of unbalanced category distribution in our data. Focal loss is used for ED and EIC, respectively:

$$\mathcal{L}^{Emp} = -\lambda^E_t\left(1-y^E_t\right)^\eta log\left(y^E_t\right), \tag{10}$$

$$\mathcal{L}^{Intent} = -\lambda^I_t\left(1-y^I_t\right)^\eta log\left(y^I_t\right), \tag{11}$$

where λ^δ_t represents the weight of the t-th class sample, y^δ_t denotes p_i, η is hyper-parameters and $\eta \in [0, 5]$.

The final joint objective is formulated as:

$$\mathcal{L} = \alpha\mathcal{L}^{Emp} + \beta\mathcal{L}^{Intent}, \tag{12}$$

where α and β are hyper-parameters.

5 Experiments

5.1 Experimental Settings

We use the BERT pre-trained model to extract vectors as the initialization embedding. The batch size is 32 and 64 on IEMPATHIZE and TwittEmp, respectively. The epoch is set to 100. The learning rate is set to 0.0001 and the dropout ratio is set to 0.2. The output dimension of the cascaded interaction module and the label signal enhancement module is 128. We use Adam (Kingma and Ba, 2015) to optimize the parameters in the model. In the loss function, η is set to 5, λ is set to 1. All experiments are conducted at GeForce RTX 2080Ti.

5.2 Baselines

To validate the performance of our framework, we compare our framework to some baselines, including: (1) **BERT** (Hosseini and Caragea, 2021b): The BERT model was fine-tuned on the IEMPATHIZE dataset; (2) **KD** (Hosseini and Caragea, 2021a): They used the idea of knowledge distillation to combine emotions and sentiments to complete the empathy detection; (3) **Empirical Study**: BERT models as a shared encoder to jointly train of the empathy detection task and the empathy intent recognition task; (4) **Joint ID and SF** (Zhang and Wang, 2016): The model was used for the spoken language understanding task, and we use it for the joint training of ED and EIC; (5) **DCR-Net**(Qin et al., 2020): They propose the DCR-NET to explicitly consider the cross-influence between tasks. The model captures mutual knowledge by stacking of relational layers within itself; (6) **Co-GAT** (Qin et al., 2021): The core module of the model is the co-interactive graph interaction layer, in which cross-utterance connections and cross-task connections are constructed and iteratively updated with each other to establish information transfer between two tasks. Since this study focuses only on a single-turn dialogue task, we replace the cross-utterance connections in the model with cross-word connections.

Models	IEMPATHIZE						TwittEmp					
	ED			EIC			ED			EIC		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1 (%)	P(%)	R(%)	F1(%)
	Multi-class						Multi-class					
BERT	75.88*	73.42*	74.42*	-	-	-	-	-	-	-	-	-
Empirical Study	76.94	77.40	76.16	67.36	66.90	66.15	75.96	75.83	75.78	61.55	60.67	60.75
Joint ID and SF	77.05	76.20	76.24	69.05	68.90	68.38	77.06	76.83	76.82	55.53	57.17	55.94
DCR-NET	77.22	77.70	76.99	68.12	68.20	67.90	76.80	77.00	76.81	58.43	58.17	57.16
Co-GAT	65.21	64.50	64.73	55.61	55.85	55.73	69.64	69.71	69.35	34.23	32.22	30.20
CLSN	**79.70**	**78.80**	**79.13**	**69.81**	**70.40**	**69.69**	**78.32**	**78.33**	**78.31**	**61.90**	**61.50**	**61.01**
	Seek						Seek					
BERT	78.37*	73.40*	76.37*	-	-	-	-	-	-	-	-	-
KD_{EmoNet}	-	-	-	-	-	-	77.32*	61.09*	68.57*	-	-	-
Empirical Study	80.48	81.90	79.16	68.05	67.50	66.83	80.45	80.67	80.52	56.84	58.33	56.97
Joint ID and SF	83.08	84.10	83.15	68.30	68.60	68.42	82.99	83.17	82.71	55.99	60.00	57.16
DCR-NET	82.58	83.70	82.61	67.97	67.40	67.11	80.86	80.17	80.39	57.67	58.50	57.44
Co-GAT	70.59	70.19	70.38	55.81	55.45	55.63	78.80	70.12	71.21	27.81	32.26	29.46
CLSN	**84.84**	**85.40**	**85.02**	**69.55**	**69.40**	**69.31**	**84.14**	**84.33**	**84.17**	**60.37**	**60.50**	**59.50**
	Provide						Provide					
BERT	86.87*	83.37*	84.49*	-	-	-	-	-	-	-	-	-
KD_{EmoNet}	-	-	-	-	-	-	77.21*	82.57*	79.48*	-	-	-
Empirical Study	90.23	90.00	90.10	67.45	67.50	67.83	87.88	87.67	87.74	57.40	58.17	57.19
Joint ID and SF	91.29	90.60	90.85	68.14	68.20	67.17	89.68	89.67	89.67	57.80	59.33	57.82
DCR-NET	90.71	90.90	90.79	67.93	67.00	63.50	90.31	90.33	90.32	56.03	57.67	56.34
Co-GAT	78.38	73.64	75.61	58.20	57.88	58.04	83.21	83.53	83.37	27.18	31.52	28.97
CLSN	**91.64**	**91.60**	**91.62**	**70.92**	**70.70**	**70.65**	**90.82**	**90.83**	**90.83**	**61.19**	**62.00**	**61.11**

Table 3: ED and EIC results on the IEMPATHIZE and TwittEmp datasets. '$*$' and '-' denote original paper results and unreported results, respectively.

5.3 Main Results and Analysis

For both ED and EIC tasks on the two datasets, we use Precision (P), Recall (R), and F1-score (F1) as evaluation metrics. The experimental results are shown in Table 3 and Appendix B. We can see that the performance of CLSN on both datasets is significantly improved compared with all baselines.

The Joint ID and SF uses the bi-directional GRU to extract features and classify them directly, without considering the feature interactions between empathy and empathy intent representations, therefore, the model achieves lower accuracy. The DCR-NET uses the stacked co-interactive relationship layer to implement the interaction between the two tasks. Since the method models the information interaction between two tasks, the accuracy of the model is better than the joint ID and SF. However, DCR-NET has limited information interaction capabilities, resulting in lower accuracy than our framework. It can be seen that the accuracy of Co-GAT is lower, which shows that adopting cross-word connections instead of cross-utterances connections can not fully show the modeling performance of Co-GAT. Compared with all baselines, the cascaded interaction attention mechanism in our framework establishes a deeper information transfer between two tasks. Specifically, after the initial information interaction is completed, the current task can be used as a query to search for single-level interaction features, thus achieving deep feature extraction from the current task to another task. The information flow continuously interacts between two tasks, eventually bringing feature information from another task back to the current task. It not only enriches the feature information of the current task, but also avoids the lack of features due to insufficient interaction. In addition, the label signal enhancement module in our framework uses the label encoding information of the tasks to help the model automatically detect the category features in the model, while avoiding the problem of error cascades due to mutual guidance between tasks. The clear bidirectional flow of modeling information in the two modules enables our framework to achieve competitive results.

5.4 Ablation Study

This subsection aims to demonstrate the effectiveness of different components in our framework, including the cascaded interactive attention mod-

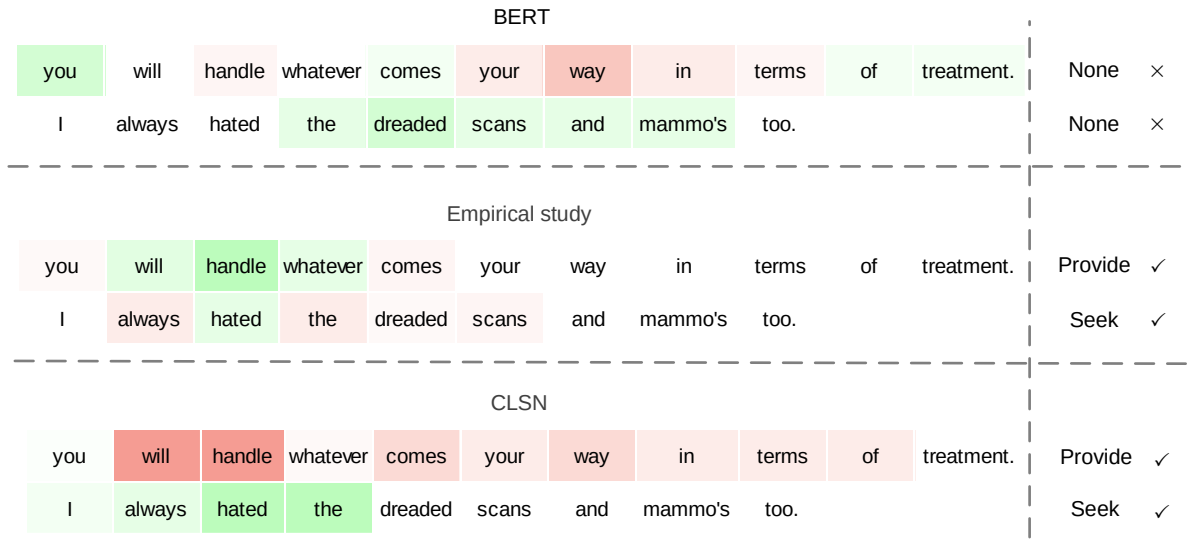

Figure 6: On the empathy detection task, the CLSN and other baselines visualize the distribution of label contributions at the same network layer. Note: Red background color indicates how much the word contributes to '*Seek*', green background color indicates how much the word contributes to '*Provide*'. Words with darker background colors indicate higher contribution values to the corresponding label.

Model	IEMPATHIZE					
	Multi-class		Seek		Provide	
	ED	EIC	ED	EIC	ED	EIC
CLSN	79.13	69.69	85.02	69.31	91.62	70.65
w/o CIA	77.26 ↓	69.42 ↓	82.99 ↓	69.44 ↑	90.93 ↓	68.98 ↓
w/o LSE	78.29 ↓	69.13↓	83.49↓	68.24↓	90.68↓	68.80↓
r.p. CEL	78.33↓	68.72↓	82.57↓	67.62 ↓	91.06↓	67.64↓

Table 4: F1-score of ablation experiments on the IEMPATHIZE dataset.

ule, the label signal enhancement module, and the focal loss.

• **Effectiveness of Cascaded Interactive Attention Module.** We remove the cascaded interactive attention module from our framework (w/o CIA). From Table 4, we can observe that the F1-score of ED decreases by 1.87% after removing the cascaded interaction attention module under the multi-class setting in the IEMPATHIZE dataset. The F1-score of EIC decreased by 0.27%. The F1-score of the two tasks also shows a decreasing trend in the binary classification setting. These results indicate that the cascaded interactive attention module is essential in the proposed CLSN due to its ability to effectively establish the information interaction between the two tasks.

• **Effectiveness of Label Signal Enhancement Module.** We remove this part of the framework (w/o LSE). From Table 4, we can see that the F1-score of both settings in the IEMPATHIZE dataset are degraded after removing this component. Therefore, it is effective for the label signal enhancement component to guide the two task representations. Moreover, our framework achieves optimal results only when both the cascaded interactive attention module and the label signal enhancement module are applied to CLSN, proving that these two modules are complementary.

• **Effectiveness of Focal Loss.** We verify the effectiveness of the focal loss by replacing it with the cross-entropy loss (r.p. CEL). The experimental results show that the use of the focal loss to alleviate the problem of an unbalanced distribution of category labels is effective. We can observe that the use of the cross-entropy loss function is lower than our proposed full framework under both classifier settings, and there is a significant reduction in EIC. This shows that focal loss can indeed help improve the quality of ED by improving the unbalanced samples of empathy intent during training. Furthermore, as can be seen from Tables 3 and 4, when our framework uses cross-entropy as the loss function under both settings, the F1-scores still outperform all baselines in most cases, indicating that the accuracy improvement achieved by CLSN is not only the use of the focal loss function, but the superiority of the model is also a factor.

5.5 Visualisation Analysis

To explore the impact of different modeling approaches on words in utterances, we use the Cap-

tum[4] to visualize the contribution of different words to empathy labels (Kokhlikyan et al., 2020; Chefer et al., 2021). As shown in Figure 6, the BERT does not accurately identify the empathy labels for two cases. The empirical study introduced the EIC task and accurately identifies the categories for the two cases, but the model could not focus well on the words associated with the correct label. In contrast, the CLSN not only accurately identifies the empathy labels in both cases, but is also able to focus more reasonably on the words associated with the empathy labels in the utterance, demonstrating the superiority of our framework.

6 Conclusion

In this paper, to investigate whether the performance of the ED task can be improved by the introduction of the EIC task, we invite 3 experts to manually label the empathy intent labels on two empathy detection datasets, IEMPATHIZE and TwittEmp, as datasets for joint training of ED and EIC. The empirical study shows that the introduction of EIC task is effective in improving the accuracy of the ED task, and we also explore possible reasons for this improvement. In addition, we propose the CLSN, which explicitly models the information flow interaction between the two representations through a cascaded interactive attention module and a labeled signal enhancement module. Experimental results show that the CLSN achieves better accuracy than all baselines under both settings in the two datasets.

Limitations

The limitations of this paper are mainly twofold: (1) Although the introduction of the EIC task into the ED task can significantly improve the accuracy of the ED task, this requires manual annotation and is costly, and in future studies we will explore few-shot learning approaches to mitigate this cost; (2) In exploring the reasons for this improvement, we count the frequency co-occurrence matrices of the two labels, and find that there is a clear correspondence between the empathy and empathy intent labels. Although there is an obvious correspondence between the labels, but the proposed CLSN only uses the two implicit states for interaction when modeling information transfer, without explicitly modeling information transfer based on the above label correspondence. How to use this correspondence could be a direction of exploration for our future work.

[4] https://captum.ai/

Ethics Statement

The dataset studied in this paper does not involve ethical issues.

Acknowledgements

We thank the anonymous reviewers for their helpful comments and suggestions. This work was supported by the National Natural Science Foundation of China (61433012) and the National Key R & D Program of China (No. 2014CB340506) and the Excellent Doctoral Student Research Innovation Project of Xinjiang University (No. XJU2022BS077).

References

Firoj Alam, Morena Danieli, and Giuseppe Riccardi. 2018. Annotating and modeling empathy in spoken conversations. *Computer Speech and Language*, 50(C):40–61.

MK Ayshabi and Sumam Mary Idicula. 2021. A multi-resolution mechanism with multiple decoders for empathetic dialogue generation. In *2021 8th International Conference on Smart Computing and Communications (ICSCC)*, pages 240–245. IEEE.

Guanqun Bi, Yanan Cao, Piji Li, Yuqiang Xie, Fang Fang, and Zheng Lin. 2023. Seri: Sketching-reasoning-integrating progressive workflow for empathetic response generation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791.

Mao Yan Chen, Siheng Li, and Yujiu Yang. 2022. Emphi: Generating empathetic responses with human-like intents. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1063–1074.

Zhuohao Chen, James Gibson, Ming-Chang Chiu, Qiaohong Hu, Tara K Knight, Daniella Meeker, James A Tulsky, Kathryn I Pollak, and Shrikanth Narayanan. 2020. Automated empathy detection for oncology

encounters. In *2020 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–8. IEEE Computer Society.

Pascale Fung, Anik Dey, Farhad Bin Siddique, Ruixi Lin, Yang Yang, Yan Wan, and Ho Yin Ricky Chan. 2016. Zara the supergirl: An empathetic personality recognition system. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 87–91.

Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. 2021. Improving empathetic response generation by recognizing emotion cause in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 807–819.

Soumitra Ghosh, Dhirendra Maurya, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Team iitp-ainlpml at wassa 2022: Empathy detection, emotion classification and personality detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 255–260.

Mahshid Hosseini and Cornelia Caragea. 2021a. Distilling knowledge for empathy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3713–3724.

Mahshid Hosseini and Cornelia Caragea. 2021b. It takes two to empathize: One to seek and one to provide. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13018–13026.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Hamed Khanpour, Cornelia Caragea, and Prakhar Biyani. 2017. Identifying empathetic messages in online health communities. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 246–251.

Wongyu Kim, Youbin Ahn, Donghyun Kim, and Kyong-Ho Lee. 2022. Emp-rft: Empathetic response generation via recognizing feature transitions between utterances. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4118–4128.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327.

L Medeiros and T Bosse. 2016. Empirical analysis of social support provided via social media. In *Proceedings of the 8th International Conference on Social Informatics, SocInfo'16*, pages 439–453. Springer Verlag.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and predicting empathic behavior in counseling therapy. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1435.

Libo Qin, Wanxiang Che, Yangming Li, Mingheng Ni, and Ting Liu. 2020. Dcr-net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8665–8672.

Libo Qin, Zhouyang Li, Wanxiang Che, Minheng Ni, and Ting Liu. 2021. Co-gat: A co-interactive graph attention network for joint dialog act recognition and sentiment classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13709–13717.

Tulika Saha and Sophia Ananiadou. 2022. Emotion-aware and intent-controlled empathetic response generation using hierarchical transformer network. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Azlaan Mustafa Samad, Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2022. Empathetic persuasion: Reinforcing empathy and persuasiveness in dialogue systems. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 844–856.

João Sedoc, Sven Buechel, Yehonathan Nachmany, Anneke Buffone, and Lyle Ungar. 2020. Learning word ratings for empathy and distress from document-level user responses. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1664–1673.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

M. Virvou and G. Katsionis. 2004. Relating error diagnosis and performance characteristics for affect perception and empathy in an educational software application. pages 22–27.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899.

Anuradha Welivita, Yubo Xie, and Pearl Pu. 2021. A large-scale dataset for empathetic response generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1264.

Akmal Setiawan Wijaya, Dhomas Hatta Fudholi, and Ahmad R Pratama. 2023. A computational approach in analyzing the empathy to online donations during covid-19. *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, 22(2):185–194.

AC de C Williams and A Cano. 2005. Facing others in pain: the effects of empathy. *Pain*, 118(3):285–288.

Yubo Xie and Pearl Pu. 2021. Empathetic dialog generation with fine-grained intents. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 133–147.

Özge Nilay Yalcin and Steve DiPaola. 2018. A computational model of empathy for interactive agents. *Biologically inspired cognitive architectures*, 26:20–25.

Diyi Yang, Robert E Kraut, Tenbroeck Smith, Elijah Mayfield, and Dan Jurafsky. 2019. Seekers, providers, welcomers, and storytellers: Modeling social roles in online health communities. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–14.

Xiaodong Zhang and Houfeng Wang. 2016. A joint model of intent determination and slot filling for spoken language understanding. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2993–2999.

Naitian Zhou and David Jurgens. 2020. Condolence and empathy in online communities. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 609–626.

A Dataset Statistics

Split	IEMPATHIZE			TwittEmp		
	None	*Seek*	*Provide*	*None*	*Seek*	*Provide*
Train	1794	630	580	616	614	570
Dev	609	191	203	203	180	217
Test	592	225	183	181	206	213
Total	2995	1046	966	1000	1000	1000

Table 5: Dataset Statistics for IEMPATHIZE and TwittEmp on utterance (Utt) and empathy label (*None*, *Seek*, *Provide*) counts in multi-class setting.

Split	IEMPATHIZE		TwittEmp	
	None	*Seek*	*None*	*Seek*
Train	2374	630	1186	614
Dev	812	191	420	180
Test	775	225	394	206
Total	3961	1046	2000	1000

Table 6: Dataset Statistics for IEMPATHIZE and TwittEmp on utterance (Utt) and empathy label (*None*, *Seek*, *Provide*) counts in binary setting (seek)

Split	IEMPATHIZE		TwittEmp	
	None	*Provide*	*None*	*Provide*
Train	2424	580	1230	570
Dev	800	203	383	217
Test	817	183	387	213
Total	4041	966	2000	1000

Table 7: Dataset Statistics for IEMPATHIZE and TwittEmp on utterance (Utt) and empathy label (*None*, *Seek*, *Provide*) counts in binary setting (provide)

B class-wise Performance

Figure 7, Figure 8, and Figure 9 show the results of the CLSN for each category in different settings in the ED and EIC tasks, respectively.

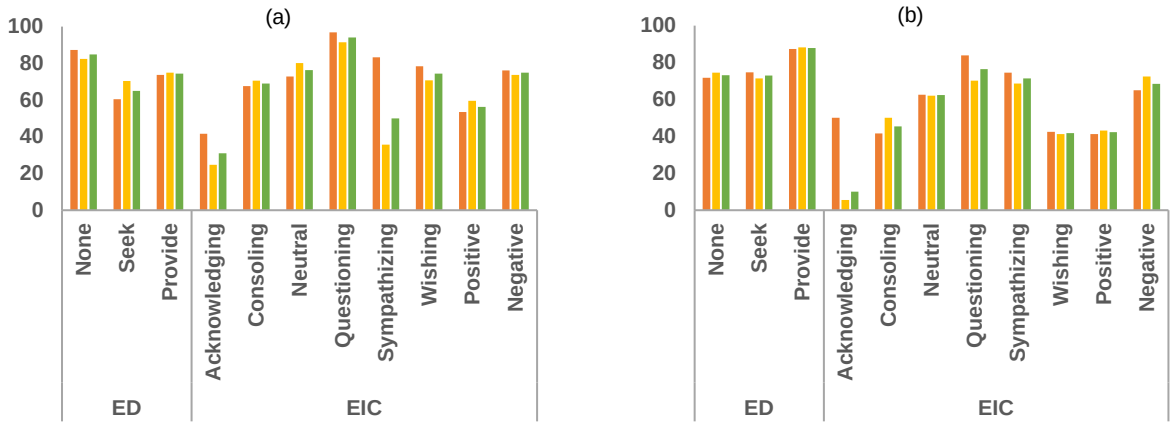

Figure 7: Results of the CLSN model on the ED and EIC tasks in the multi-class setting of the IEMPATHIZE (a) and TwittEmp (b) datasets.

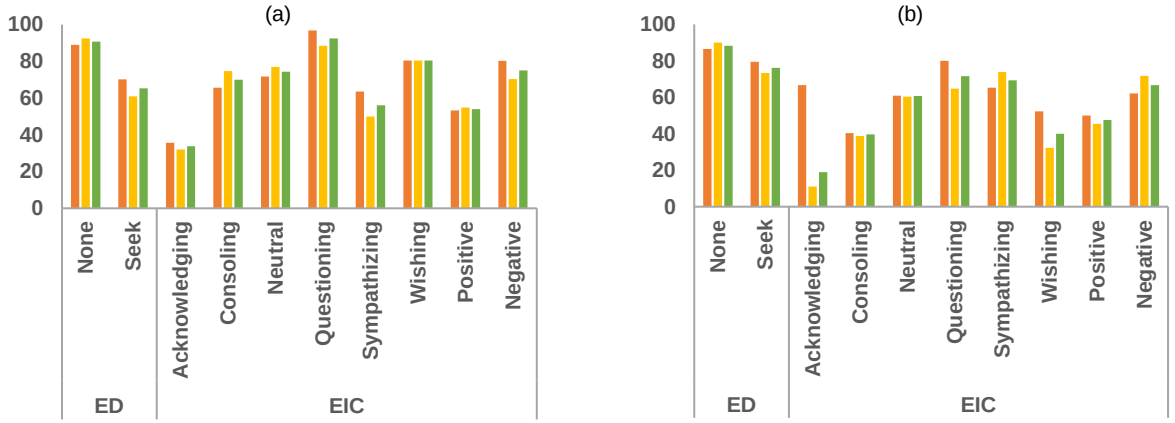

Figure 8: Results of the CLSN model on the ED and EIC tasks in the binary setting (seek) of the IEMPATHIZE (a) and TwittEmp (b) datasets.

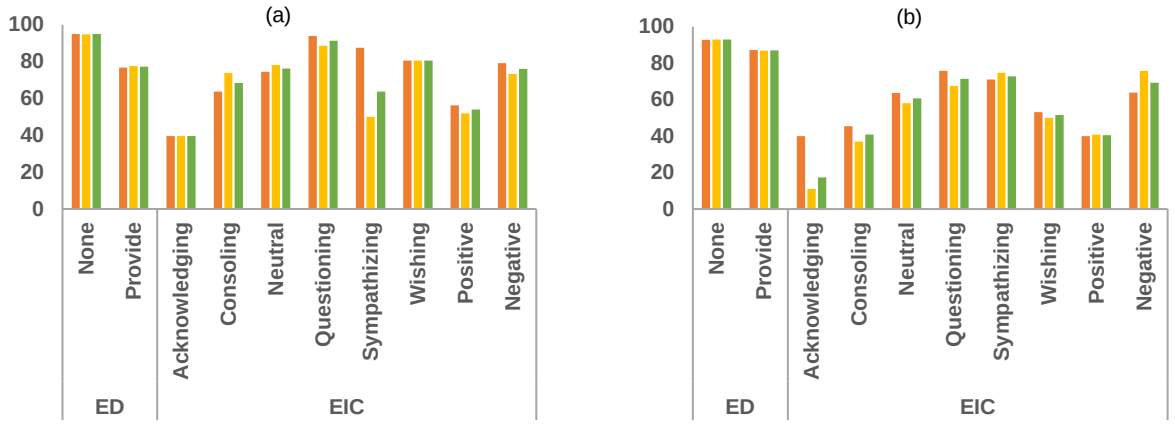

Figure 9: Results of the CLSN model on the ED and EIC tasks in the binary setting (provide) of the IEMPATHIZE (a) and TwittEmp (b) datasets.