EmoTrans: Emotional Transition-based Model for Emotion Recognition in Conversation

Zhongquan Jian[1,2,†], **Ante Wang**[2,4,†], **Jinsong Su**[2,4], **Junfeng Yao**[1,2,3,4,5], **Meihong Wang**[2,*], **Qingqiang Wu**[1,2,3,4,5,*]
[1]Institute of Artificial Intelligence, Xiamen University, China
[2]School of Informatics, Xiamen University, China
[3]School of Film, Xiamen University, China
[4]Key Laboratory of Digital Protection and Intelligent Processing of Intangible Cultural Heritage of Fujian and Taiwan, Ministry of Culture and Tourism, Xiamen University, China
[5]Xiamen Key Laboratory of Intelligent Storage and Computing, School of Informatics, Xiamen University, China
{jianzq, wangante}@stu.xmu.edu.cn, {jssu, yao0010, wangmh, wuqq}@xmu.edu.cn

Abstract

In an emotional conversation, emotions are causally transmitted among communication participants, constituting a fundamental conversational feature that can facilitate the comprehension of intricate changes in emotional states during the conversation and contribute to neutralizing emotional semantic bias in utterance caused by the absence of modality information. Therefore, emotional transition (ET) plays a crucial role in the task of Emotion Recognition in Conversation (ERC) that has not received sufficient attention in current research. In light of this, an **Emo**tional **Trans**ition-based Emotion Recognizer (**EmoTrans**) is proposed in this paper. Specifically, we concatenate the most recent utterances with their corresponding speakers to construct the model input, known as samples, each with several placeholders to implicitly express the emotions of contextual utterances. Based on these placeholders, two components are developed to make the model sensitive to emotions and effectively capture the ET features in the sample. Furthermore, an ET-based Contrastive Learning (ETCL) is developed to compact the representation space, making the model achieve more robust sample representations. We conducted exhaustive experiments on four widely used datasets and obtained competitive experimental results, especially, new state-of-the-art results obtained on MELD and IEMOCAP, demonstrating the superiority of EmoTrans.

Keywords: Natural Language Processing, Emotion Recognition in Conversation, Contrastive Learning, Emotional Transition

1. Introduction

Social interaction is a requirement of human beings (Tomova et al., 2020), people can have their emotions acknowledged and psychologically comforted through conversation. In an emotional conversation, utterances without emotions ups and downs tend to be dull and tasteless, accurately identifying the emotional state expressed in the utterance and empathetically responding can effectively promote the sustainability of the conversation (Ma et al., 2020; Zhu et al., 2022). As a result, a crucial task known as Emotion Recognition in Conversation (ERC) seeks to recognize the emotional states present during each conversational exchange. As illustrated in Figure 1, a conversation consists of several utterances, each of which is uttered by a speaker and exhibits a certain emotion. Hence, ERC aims to identify the emotions expressed by speakers in the conversation based on conversational features, which is beneficial for the development of intelligent conversation systems.

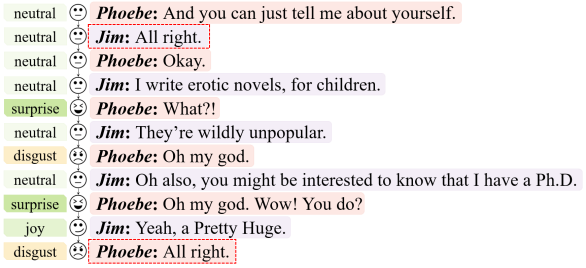

Figure 1: Example of a conversation, with utterance emotions shown on the left.

The most important conversational features are the contextual utterances and their corresponding speakers. As a result, numerous studies have been conducted to leverage these features to recognize the emotions expressed in utterances, including: 1) **recurrence-based approaches** (Majumder et al., 2019; Jiao et al., 2019) encode utterances turn by turn, describing the flow of semantics in temporal sequence; 2) **graph-based techniques** (Ghosal et al., 2019; Sheng et al., 2020; Shen et al., 2021b) use the directed graph to model the relationships

[†]Equal contribution
[*]Corresponding authors

LREC-COLING 2024, pages 5723–5733
20-25 May, 2024.

of utterances and speakers, then aggregate the surrounding information by using GAT or GCN; 3) **PLM-based methods** (Shen et al., 2021a; Lee and Lee, 2022; Li et al., 2022) employ the pre-trained language model (PLM) as the backbone, which greatly improves the semantic understanding of phrases and leads to notable results. Additionally, some works devote themselves to discovering other potentially useful conversational features (Ide and Kawahara, 2022; Bao et al., 2022; Ong et al., 2022; Song et al., 2022b) for the ERC task.

Along this line, we attempt to explore the role of emotional transition in the conversation. As seen in Figure 1, the emotions transmitted among speakers make up an emotional transition that transfers emotions from *"neutral"* to *"disgust"*. Typically, emotions in a conversation are significantly influenced by preceding emotional states and transmitted causally based on participants' personalities. In this conversation, the second and final utterances are identical in content, yet they were articulated by different speakers, conveying distinct emotions, where the former extends the *"neutral"* emotion from the previous speaker, while the latter exhibits *"disgust"* in response to the vulgar semantic of the previous utterance. These observations suggest that modeling previous emotional states is as important as considering the contextual utterances when identifying the emotional state of the current utterance. Therefore, emotional transition (ET) emerges as another valuable conversational feature that is beneficial for capturing the emotional atmosphere in the conversation and offering auxiliary information to identify the speaker's emotion correctly.

To this end, we propose an Emotional Transition-based Emotion Recognizer (**EmoTrans**) to capture and leverage ET features to alleviate the issue of information scarcity in the ERC task. Specifically, for the target utterance, we first concatenate the most recent utterances with their corresponding speakers to construct the model input, known as a sample. Each contextual utterance in the constructed sample is associated with a placeholder to implicitly express the emotion. Hence, the contextual utterances' emotions constitute the sample's ET information (a subset of the conversation's ET), and our goal is to capture ET features from ET information to enhance the sample representation. To achieve this, EmoTrans is equipped with an emotional perception enhancement component and a unidirectional LSTM-based component, where the former is designed to improve the model's ability to discern the emotions inherent in the sample, and the latter can capture ET features that serve as the auxiliary emotional information intended to enhance the sample representation, which initially comprised only textual features. Furthermore, following the acquisition of sample representations enriched with emotional attributes through the information integration of contextual utterance, speakers, and ET, we introduce an ET-based Contrastive Learning (ETCL) to compress the representation space, making the model achieve more robust sample representations for the ERC task.

In summary, our contributions are three-fold:

- We develop a sample construction method to construct samples that contain the partial or whole ET information of the conversation and propose EmoTrans to capture and leverage ET features to address the ERC task.
- With the application of emotional perception enhancement, capture ET features, and representation alignment, EmoTrans can perceive the contextual utterances' emotions in the sample and generate more robust sample representations.
- We assess EmoTrans's performance on four widely used datasets and achieve new state-of-the-art results on MELD and IEMOCAP. These successes demonstrate that the conversation's ET information plays a vital role in emotion recognition.

The remainder of this paper is organized as follows: some relevant research contents are introduced in Section 2. The methodology of EmoTrans is described in detail in Section 3. Sections 4 and 5 each provide information on the experimental design and outcome analysis. Section 6, offers concluding observations.

2. Related Work

2.1. Emotion Recognition in Conversation

In the ERC task, the most important conversational features are the contextual utterances and their corresponding speakers. As such, numerous works attempt to pinpoint the intrinsic relationships between these features. Initially, Recurrent Neural Networks were used to model utterances or entire conversations sequentially. Jiao et al. (2019) proposed using hierarchical GRU to capture word-level and utterance-level information in conversations, yielding representations with extended context for ERC. Similarly, Majumder et al. (2019) utilized a GRU unit to generate the global state by considering utterance and speaker state simultaneously, where the speaker state is updated by another GRU unit depending on the current utterance and context. However, researchers found that the relationships between utterances in the conversation are more complex than just sequential.

To reflect the intrinsic dependencies of speakers and utterances, DialogGCN (Ghosal et al., 2019) used the graph network to describe the self and inter-speaker dependencies and propagated the context information by using GCN. Based on DialogGCN, Sheng et al. (2020) emphasized the conversation's emotional fluctuation by making references to both global topic-related emotional phrases and local dependencies. Moreover, DAG-ERC (Shen et al., 2021b) argued that utterances are non-sequential but directed, hence, a directed acyclic graph is designed to better model the factual structure of the conversation.

Based on the strong representation ability of PLM, DialogXL (Shen et al., 2021a) applied XLNet (Yang et al., 2019) on ERC and designed an enhanced memory module to store historical context and modified the original self-attention mechanism to capture intra- and inter-speaker dependencies. CoMPM (Lee and Lee, 2022) also utilized a PLM to construct the pre-trained memory based on the speaker's previous utterances and then concatenated the context embedding obtained by another PLM to generate the final emotional representation for ERC. To enable the model to recognize emotions with similar semantics in diverse contexts, CoG-BART (Li et al., 2022) introduced BART (Lewis et al., 2020) to understand the contextual context and generate the next utterance as an auxiliary task. Furthermore, SCL is employed to enhance the difference between the representations of utterances with distinct emotions.

Additionally, numerous works are devoted to discovering deeper information beyond the context and speakers in the conversation. SKAIG (Li et al., 2021) concentrated on the structural psychological interactions among utterances, and four kinds of speaker relations were designed to model the speaker's action and intention. Bao et al. (2022) proposed a novel SGED framework to model intra- and inter-speaker dependencies jointly in a dynamic manner. Gao et al. (2022) introduced an auxiliary task of emotion shift detection, which will introduce emotional information to enhance the final sample representations. Despite these techniques achieving great success, scarcity of information remains the biggest problem for ERC.

2.2. Contrastive Learning in ERC

Contrastive Learning (CL) aims to learn the specific representation by regulating the distances of different sample pairs, where positive samples are expected to be gathered in the representation space while negative samples are expected to be pushed away as much as possible. Conventional CL was applied in the form of self-supervised representation learning, Khosla et al. (2020) extended the self-supervised batch contrastive approach to the fully-supervised setting, known as SCL, to make full use of label information. In the ERC task, CoG-BART (Li et al., 2022) applied SCL to strengthen the difference of utterance representations with distinct emotions. However, positive samples may not exist in the batch even with a large batch size due to the imbalance of emotional categories. To address this issue, CoG-BART creates a duplicate batch to guarantee the presence of positive samples. However, this duplication does not provide additional information and further constrains the batch size. As a result, SPCL (Song et al., 2022a) collected the temporary prototype vector of each emotion category and added them into the batch, which ensures that each sample has at least one positive sample and enables the model to train in a large batch.

2.3. Large Language Model for ERC

The advent of Large Language Models (LLMs) has catalyzed a paradigm shift in the field of Natural Language Processing (NLP) (Shen et al., 2024). These models, characterized by their vast parameter space and trained on extensive corpora, have significantly enhanced capabilities in natural language understanding and generation. As a result, they have set new benchmarks in a range of NLP tasks, such as question answering, named entity recognition, sentiment analysis, and beyond. However, LLMs often exhibit limitations in processing tasks that require nuanced sentiment analysis (Zhang et al., 2023), among which the ERC task represents a particularly challenging domain. InstructERC (Lei et al., 2023) attempts to address this issue by constructing the prompt with a semantic similar context and then using the LLM to extract the emotional information from the prompt. After fine-tuning with the ERC dataset, InstructERC didn't achieve significant improvement, even worse than the performance of traditional ERC models, demonstrating the need to model potential conversational features for ERC even in the LLM era, as most researchers have done for ERC in recent years.

Hence, we focus on exploring the potential conversation feature to adapt the characteristics of the ERC task, which is compatible with LLM and can enhance the LLMs' capabilities in solving them.

3. Methodology

The overview of EmoTrans is depicted in Figure 2, which is mainly composed of three components: a) capturing ET features; b) emotional perception enhancement; c) representation alignment. To support the execution of the model component, we first designed a sample construction method that allows each sample to implicitly express their emotions by

placeholders. In this section, we first describe the ERC task briefly and introduce each component of EmoTrans in detail. Bold typeface is used to denote sets that contain multiple elements.

3.1. Problem Definition

In general, an utterance $\boldsymbol{u_i}$ is composed of n_i tokens, denoted as $\boldsymbol{u_i} = \{w_{i,1}, w_{i,2}, ..., w_{i,n_i}\}$. Several sequential utterances uttered by different speakers constitute a conversation $\boldsymbol{conv} = \{\langle \boldsymbol{u_1}, s_1 \rangle, \langle \boldsymbol{u_2}, s_2 \rangle, ..., \langle \boldsymbol{u_n}, s_n \rangle\}$, where $s_i \in \boldsymbol{S}$ represents the speaker of $\boldsymbol{u_i}$ with the expressed emotion as $y_i \in \boldsymbol{Y}$.[1] The ERC task aims to identify the emotion y_t in t-th turn given the previous utterances $\{\langle \boldsymbol{u_1}, s_1 \rangle, ..., \langle \boldsymbol{u_t}, s_t \rangle\}$.

3.2. Emotional Sample Construction

Similar to previous works (Song et al., 2022b,a), we concatenate each utterance $\boldsymbol{u_i}$ with its speaker s_i and add a $\boldsymbol{template}$ with a placeholder to implicitly express the speaker's emotion:

$$\tilde{\boldsymbol{u}}_{\boldsymbol{i}} = [s_i, \boldsymbol{u_i}, \langle /s \rangle, \boldsymbol{template}, \langle /s \rangle] \quad (1)$$

$$\boldsymbol{template} = s_i \text{ expresses } \langle mask \rangle \quad (2)$$

where $\langle /s \rangle$ denotes the separator used to distinguish different contents, $\langle mask \rangle$ is the placeholder (known as the masked token) used for learning the emotion expressed in this utterance. Hence, the constructed utterance not only contains the speaker's information but also possesses a $\boldsymbol{template}$ to express emotion in natural language form. To predict the emotion of t-th utterance, the most recent k utterances are concatenated as the model input:

$$\boldsymbol{x} = [\langle s \rangle, \tilde{\boldsymbol{u}}_{\boldsymbol{t-k}}, ..., \tilde{\boldsymbol{u}}_{\boldsymbol{t-1}}, \tilde{\boldsymbol{u}}_{\boldsymbol{t}}] \quad (3)$$

where $\boldsymbol{x}$ denotes the constructed sample that aims to identify the emotion in the t-th turn of utterances, $\langle s \rangle$ is a special token used to learn the semantic representation of the whole input text. Note that k is dynamically set according to the hyperparameter L deciding the maximum length of $\boldsymbol{x}$. In this way, the partial or whole conversational utterances are contained in the constructed sample. For each training sample, the emotions expressed in the contextual utterances are the ground truth of the masked tokens, expressed as $\boldsymbol{Y^M} = [y_{t-k}, ..., y_{t-1}, y_t]$, and $\boldsymbol{E^M} = [e^m_1, ..., e^m_{k-1}, e^m_k]$ denotes the set of corresponding embedding representations obtained after embedding operation. Some examples of the constructed samples are shown in Table 5.

In line with previous works, we employ the RoBERTa (Liu et al., 2019) as the encoder to extract the textual features of input tokens. As shown

[1] $\boldsymbol{S}$ and $\boldsymbol{Y}$ denote the predefined sets of speakers and emotions, respectively.

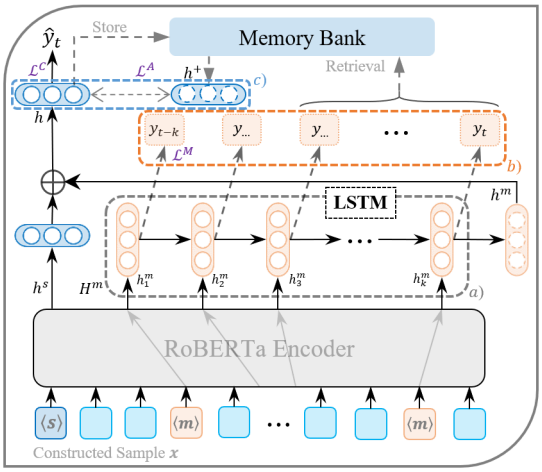

Figure 2: The architecture of EmoTrans, where the paths indicated with dashed lines are executed only in the training stage. i.e. the components of emotional perception enhancement and representation alignment. The defined components are highlighted with dashed boxes.

in Figure 2, we feed the constructed sample into the RoBERTa encoder:

$$\boldsymbol{H^x} = \text{RoBERTa}(\boldsymbol{x}) \quad (4)$$

where $\boldsymbol{H^x} \in \mathbb{R}^{l^x \times d}$ is the set of hidden states with $l^x < L$ length and d dimensions. Then, we extract the hidden states of special tokens for subsequent processing, i.e. h^s and $\boldsymbol{H^m}$, where h^s is the hidden state of the $\langle s \rangle$ token in $\boldsymbol{x}$ (the first tensor in $\boldsymbol{H^x}$), and $\boldsymbol{H^m} \in \mathbb{R}^{k \times d}$ is the set of hidden states of k masked tokens.

3.3. Emotional Perception Enhancement

Inspired by the impressive understanding ability of Masked Language Modeling (MLM) (Fu et al., 2022; Wettig et al., 2023), we enhanced the model's ability to understand the emotional information in the sample by predicting the emotions of the contextual utterances. Specifically, during the training stage, we add an auxiliary task to predict the emotions of the masked tokens, which is formulated as:

$$\mathcal{L}^M = -\frac{1}{k}\sum_{i=1}^{k} \log \frac{\exp(h^m_i * e^m_i)}{\sum_{j=1}^{|\boldsymbol{\mathcal{V}}|} \exp(h^m_i * e_j)} \quad (5)$$

where $h^m_i \in \boldsymbol{H^m}$ denotes the i-th masked token's hidden state, $k = |\boldsymbol{H^m}|$ is the number of masked tokens in the sample (i.e. the number of contextual utterances). $|\boldsymbol{\mathcal{V}}|$ is the size of vocabulary $\boldsymbol{\mathcal{V}}$. e^m_i is the word vector of the corresponding emotion, and e_j is the word vector of the j-th word in the vocabulary of $\boldsymbol{\mathcal{V}}$.

3.4. Emotional Transition Features

The role of emotional perception enhancement is to improve the model's ability to perceive the emotions of contextual utterances in the sample, thereby facilitating the effective extraction of ET features. Due to the characteristics of one-way propagation in the transmission of utterances or emotions, using the unidirectional Long Short-Term Memory (LSTM) approach to extract ET features is both straightforward and efficacious. Therefore, based on the hidden states of masked tokens, the process of capturing ET features is as follows:

$$\boldsymbol{f}_i = \sigma(W_f \cdot [v_{i-1}, h_i^m] + b_f) \tag{6}$$

$$\boldsymbol{i}_i = \sigma(W_i \cdot [v_{i-1}, h_i^m] + b_i) \tag{7}$$

$$\boldsymbol{o}_i = \sigma(W_o \cdot [v_{i-1}, h_i^m] + b_o) \tag{8}$$

$$\hat{\boldsymbol{c}}_{\boldsymbol{t}} = \tanh(W_c \cdot [v_{i-1}, h_i^m] + b_c) \tag{9}$$

$$\boldsymbol{c}_i = \boldsymbol{f}_i \odot \boldsymbol{c}_{i-1} + \boldsymbol{i}_i \odot \hat{\boldsymbol{c}}_{\boldsymbol{i}} \tag{10}$$

$$v_i = \boldsymbol{o}_i \odot \tanh(\boldsymbol{c}_i) \tag{11}$$

where $\boldsymbol{f}$, $\boldsymbol{i}$ and $\boldsymbol{o}$ denote forget gate, input gate and output gate respectively, which serve to control the interactions between memory cells and environments. σ and $\tanh$ are two activation functions that enable non-linear modeling of the network. The symbol "$\cdot$" stands for matrix multiplication and "$\odot$" denotes element-wise multiplication. W_* and b_* are all trainable parameters. v_i denotes the hidden state after assimilating the knowledge of the current emotional state h_i^m, and the process of emotional information aggregation is completed when $i = k$. Then, $h^m = v_k$ is the obtained ET features.

After that, a sample representation that embeds rich textual and emotional information is obtained by fusing the textual features and ET features:

$$h = \mathrm{Pooling}(h^s, h^m) \tag{12}$$

where $\mathrm{Pooling}$ denotes feature fusion operation, mean pooling is utilized in our work. h^s represents the textual features comprising information from contextual utterances and their corresponding speakers, and h^m denotes ET features that accumulated the emotional information of contextual utterances in the sample. Subsequently, we feed h to a fully connected layer to predict the emotion:

$$\hat{y} = \mathrm{softmax}(Wh + b) \tag{13}$$

where W and b are learnable parameters and $\hat{y}$ denotes the estimated probability for each emotional category. During the training stage, the cross-entropy loss is calculated:

$$\mathcal{L}^C = y * \log(\hat{y}) \tag{14}$$

where y denotes the ground truth.

3.5. Representation Alignment

Although the sample representations are enriched with sufficient emotional information, they are still suffering from weak robustness due to the vast representation space. To further improve the sample representation, ETCL is developed to compact the representation space by aligning the representations of samples with the same partial ET information. Firstly, for each sample $\boldsymbol{x}$, we retrieve samples that have the same partial ET information as that of $\boldsymbol{x}$ from training data $\boldsymbol{D}$:

$$\boldsymbol{R}^{\boldsymbol{x}} = \mathrm{Retrieval}(\boldsymbol{Y}^{\boldsymbol{M'}}, \boldsymbol{D}) \tag{15}$$

where $\boldsymbol{Y}^{\boldsymbol{M'}} = [y_{t-k'}, ..., y_{t-1}, y_t]$ contains $k' \leq k$ emotions is the partial ET information of $\boldsymbol{x}$. $\mathrm{Retrieval}$ denotes a function that retrieves samples $\boldsymbol{R}^{\boldsymbol{x}}$ with the same ET information as $\boldsymbol{Y}^{\boldsymbol{M'}}$. The exact matching method based on emotional categories is used as the $\mathrm{Retrieval}$ function in our work.

To save repeated calculation costs, the Memory Bank (Wu et al., 2018) is employed to store and extract sample representations. Specifically, for each target sample, we will store the generated sample representation h into the Memory Bank $\boldsymbol{\mathcal{M}}$ and extract the sample representations of the retrieval samples $\boldsymbol{H}^{\boldsymbol{R}} = \boldsymbol{\mathcal{M}}[\boldsymbol{R}^{\boldsymbol{x}}]$. Furthermore, the average of $\boldsymbol{H}^{\boldsymbol{R}}$ is regarded as the positive sample (h^+) of h. Hence, for a batch with n_b samples, the set of paired representation vectors $\boldsymbol{H}^{\boldsymbol{B}} = (h_i, h_i^+)_{i=1}^{n_b}$ is constructed, and the loss of representation alignment is calculated:

$$\mathcal{L}^A = -\frac{1}{n_b}\sum_{i=1}^{n_b} log\frac{e^{\mathrm{sim}(h_i, h_i^+)/\tau}}{\sum_{h_j \in \boldsymbol{H}^{\boldsymbol{B}}} e^{\mathrm{sim}(h_i, h_j)/\tau}} \tag{16}$$

where $\mathrm{sim}(h_i, h_j)$ is the cosine similarity function and τ is a temperature hyperparameter. n_b denotes the batch size.

3.6. Training

As described above, except for the primary emotional classification loss, two ET-related losses are added to force the model to capture and leverage the ET features. Hence, the final loss is a weighted sum of three losses:

$$\mathcal{L} = \lambda\mathcal{L}^C + (1-\lambda)\mathcal{L}^M + \mathcal{L}^A \tag{17}$$

where the role of $\mathcal{L}^C$ and $\mathcal{L}^M$ are enforcing the model to generate ET-based sample representations, and thus, we set a weighting coefficient $\lambda = 0.8$ to balance these two losses. The role of $\mathcal{L}_A$ is encouraging the model to align sample representations with the same partial ET information, making them more robust.

4. Experimental Setup

4.1. Datasets and Metrics

Experiments are conducted on four datasets:

MELD(Poria et al., 2019) is a set of multi-party conversations derived from the American drama *"Friends"*, including more than 1400 conversations and each comprises 1 to 33 utterances, where each utterance has annotated one of seven emotion states, i.e. *Anger*, *Disgust*, *Sadness*, *Joy*, *Neutral*, *Surprise*, and *Fear*.

EmoryNLP (ENLP)(Zahiri and Choi, 2018) is also derived from *"Friends"* TV series with total 12,606 utterances, where each utterance is annotated with one of the seven emotions borrowed from the six primary emotions in the Willcox's feeling wheel(Willcox, 1982), i.e., *Sad*, *Mad*, *Scared*, *Powerful*, *Peaceful*, *Joyful*, and a default emotion of *Neutral*.

IEMOCAP(Busso et al., 2008) is a two-party and multi-modal ERC dataset, built with subtitles from improvised videos. This dataset consists of 151 videos of recorded conversations, each conversation contains a maximum of 167 turns, and utterances are labeled with 6 emotional tags.

DailyDialog (DD)(Li et al., 2017) is a two-party dataset of daily conversations, whose annotation method is Ekman's emotion type (Ekman, 1993), including *neutral*, *happiness*, *surprise*, *anger*, *disgust*, *fear*, and *sadness*.

The statistics results of these datasets are listed in Table 1, where #Conv. lists the number of conversations, #Utt. stands for the total number of utterances, and #CLS. indicates the number of different emotions in the dataset. Following most previous works (Song et al., 2022a,b; Li et al., 2022), weighted average F1 is adopted as the evaluation metrics for MELD, EmoryNLP, and IEMOCAP due to their class-imbalance phenomenon. For DailyDialog, the emotion *"neutral"* accounts for the majority, hence, we only report the Micro-F1 score of other emotions as in the previous works.

4.2. Compared Methods

We categorized the compared methods into two classes according to whether utterances are concatenated to construct new samples, where COSMIC (Ghosal et al., 2020), ToDKAT (Zhu et al., 2021), DialogueCRN (Hu et al., 2021), TUCORE-GCN (Shen et al., 2021b), DAG-ERC (Shen et al., 2021b), CoG-BART (Li et al., 2022), SACL (Hu et al., 2023) are **Utterance-based methods**, and CoMPM (Lee and Lee, 2022), SPCL (Song et al., 2022a) and EmotionFlow (Song et al., 2022b), HiDialog (Liu et al., 2023) are **Sample-based methods**.

Dataset		MELD	ENLP	IEMOCAP	DD
#Conv.	train	1038	713	100	11118
	dev	114	99	20	1000
	test	280	85	31	1000
#Utt.	train	9989	9934	4810	87170
	dev	1109	1344	1000	8069
	test	2610	1328	1523	7740
#CLS.		7	7	6	7

Table 1: Statistics of four ERC datasets.

4.3. Parameters

During the model training, AdamW with the learning rate of 5e-6 is used as the optimizer, the cosine learning rate schedule strategy is employed to facilitate model convergence, and the dropout strategy with the rate 0.3 is applied to alleviate over-fitting. We fine-tune our model for a maximum of 6 epochs and stop training if the metric does not increase for 2 consecutive epochs. We keep the best checkpoint on the valid set, then report the results on the test set using the kept checkpoint. Experiments are executed on a single Nvidia GeForce RTX 3090 GPU with 24GB memory.

As mentioned in Section 3, k' and L are two hyperparameters in EmoTrans, which are closely linked to the characteristics of the dataset. In our experiments, setting $k' = 2$ for all datasets, and $L = 148, 160, 512, 131$ for MELD, EmoryNLP, IEMOCAP and DailyDialog, respectively. We will discuss the rule of hyperparameter settings later and their impact on the model's performance.

5. Results and Analysis

5.1. Main Results

Table 2 records the comparative results of Emotrans with the baseline models on four datasets, the best and second-best results are highlighted in bold and underlined, respectively. From Table 2, some observations are obtained:

First, most of the advanced models utilize RoBERTa as their backbone. Almost all RoBERTa-based methods outperform Glove-based and BART-based methods by a large margin on MELD, and most of them achieve better results on other datasets. The evidence presented strongly suggests that RoBERTa is well-suited for the ERC task, consequently, like many other studies, we also adopt it as the backbone for our approach.

Among the utterance-based methods, despite being bad on the multi-party scene (MELD and EmoryNLP), DAG-ERC demonstrates competitive performance on the two-party scene (IEMOCAP and DailyDialog). In addition, the SGED framework further improves DAG-ERC's performance,

Type	Backbone	Methods	Datasets			
			MELD	ENLP	IEMOCAP	DD
Utt.	RoBERTa-large	COSMIC (Ghosal et al., 2020)	65.21	38.11	65.28	51.05
	RoBERTa-large	♯ ToDKAT (Zhu et al., 2021)	65.47	38.69	61.33	-
	Glove (840B)	♯ DialogueCRN (Hu et al., 2021)	63.42	38.91	66.46	-
	RoBERTa-large	TUCORE-GCN (Lee and Choi, 2021)	65.36	39.24	-	**61.91**
	RoBERTa-large	DAG-ERC (Shen et al., 2021b)	63.65	39.02	68.03	59.33
	RoBERTa-large	SGED+DAG-ERC (Bao et al., 2022)	65.46	40.24	68.53	-
	BART-large	CoG-BART (Li et al., 2022)	64.81	39.04	66.18	56.09
	RoBERTa-large	SACL (Hu et al., 2023)	66.45	39.65	69.22	-
Samp.	RoBERTa-large	CoMPM (Lee and Lee, 2022)	66.52	38.93	69.46	60.34
	RoBERTa-large	♯ EmotionFlow (Song et al., 2022b)	66.50	-	-	-
	RoBERTa-large	♯ SPCL (Song et al., 2022a)	67.25	**40.94**	69.74	-
	RoBERTa-large	HiDialog (Liu et al., 2023)	65.64	38.13	-	61.83
	RoBERTa-large	EmoTrans (Ours)	**67.96**	40.30	**70.75**	61.23

Table 2: Experimental results comparison of advanced methods on four ERC datasets. ♯ denotes the experimental results are excerpted from SPCL (Song et al., 2022a), and the others are derived from their original papers or repositories.

suggesting that the introduction of additional valuable conversational features can effectively bolster the capabilities of existing models.

In summary, the experimental results reveal that sample-based methods outperform utterance-based methods, thus demonstrating the feasibility and effectiveness of concatenating previous utterances as context. Among them, SPCL achieves state-of-the-art results on three benchmarks with a substantial superiority over other approaches, which suggests the benefits of CL in effectively differentiating similar yet independent emotional categories.

Despite notable advancements made by recent methods, EmoTrans outperforms them by a large margin and achieves new state-of-the-art results on MELD and IEMOCAP. Compared to EmotionFlow, improving the sample representations by adding ET features is a more intuitive and effective way, which makes EmoTrans possess a performance advantage of 1.46% on MELD. Compared to SPCL, EmoTrans compacts the representation space by gathering together samples with the same partial ET information, rather than solely based on their labels, leading to 0.71% and 1.01% improvements on MELD and IEMOCAP, respectively. HiDialog only models textual features and performs well on the two-party scene, but it is not suitable for the multi-party scene. In contrast, EmoTrans models the textual and emotional features simultaneously, resulting in good performance in both scenes. These experimental results demonstrate the potential of ET information, contributing to the significant performance of EmoTrans.

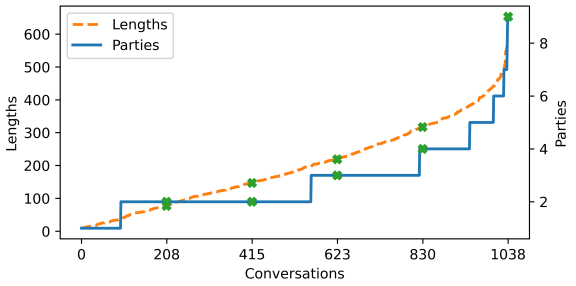

Figure 3: Conversation lengths (left) and parties (right). Setting $L = 148$ means that there are at least 415 conversations with lengths less than 148. Similarly, $k' = 3$ means that 623 conversations parties no more than 3.

k' / L	2 / 78	2 / 148	3 / 219	4 / 317	9 / 512
Metric (Valid Set)	66.38	**67.95**	67.49	67.49	67.69

Table 3: Experimental results on valid set.

5.2. Impact of Hyperparameters

As mentioned in Section 3, there are two hyperparameters in EmoTrans that are selected according to the characteristics of the dataset. Take MELD as an example, we first count the conversation lengths and parties, which are sorted in ascending order. Figure 3 depicts the statistics results, where the left y-axis is the conversation lengths and the right y-axis is the conversation parties. We choose five hyperparameter settings for experiments, which are tabulated in Table 3 as well as the experimental results on the valid set.

From Table 3, it can be found that EmoTrans achieves the optimal results on the valid set when L is set to 148 and k' is set to 2. As the values of hyperparameters increase, the model performance degrades slightly, which indicates that the adjacent utterance information is enough to enable the model to achieve significant results.

Dataset	MELD	IEMOCAP
Baseline	66.50	68.32
+ [P^M]	67.32$_{(\uparrow 0.82)}$	70.44$_{(\uparrow 2.12)}$
+ [P^A]	66.63$_{(\uparrow 0.13)}$	67.46$_{(\downarrow 0.86)}$
+ [P^M, P^A]	67.96$_{(\uparrow 1.46)}$	70.75$_{(\uparrow 2.43)}$

Table 4: Experimental results of ablation studies.

5.3. Ablation Study

To investigate the impacts of individual components and combinations of several components on the overall effect of the model, this section conducts ablation studies on different components in EmoTrans. Table 4 illustrates the results of ablation studies on the test set of MELD and IEMOCAP, where "+" indicates the addition of a single component or multiple components, P^A signifies the representation alignment component employed for the refinement of sample representations, P^M denotes the component combination of emotional perception enhancement and capture ET features, which focus on the semantic understanding and feature extraction upon the masked tokens, and then generate ET features to enhance the sample representations.

From Table 4, it can be found that the model without any components already possesses competitive performance on MELD and IEMOCAP, demonstrating the effectiveness of the sample construction method. Next, EmoTrans without any components functions as the baseline to assess the influence of different components. Notably, the performance of EmoTrans exhibits substantial enhancement on both datasets by the sole adding of P^M, achieving 0.82% and 2.12% improvements on MELD and IEMOCAP, respectively, which indicates that ET information plays a vital role in ERC. However, incorporating the P^A component into the baseline model led to a slight improvement of 0.13% on the MELD dataset, but a decrement of 0.86% on the IEMOCAP dataset, which suggests that the ETCL does not yield significant performance improvements or even has a negative impact. The reason for this phenomenon lies in the input of P^A is the sample representations with textual features instead of with ET features. In other words, the potential of P^A will be activated by the ET information, which is evident by the observations that the performance of EmoTrans is further enhanced by adding P^A and P^M.

5.4. Case Study

In Table 5, we present three cases to show the effectiveness of ET information for the ERC task, where the first column lists the contextual utterances with their speakers and emotions, the second column exhibits the format of the constructed sample, and the last three columns are the predictions of different methods and the ground truth, respectively. As shown in the second column, we concatenate the previous utterances as context, where ⟨***mask***⟩ tokens are used as placeholders to implicitly express the contextual utterance's emotion. Hence, in the constructed sample, the labels of masked tokens are the emotions expressed in contextual utterances, which form the basis of ET feature learning and extraction. The model without ET-related components (w/o ET) serves as the baseline model to assess the effectiveness of EmoTrans (w/ ET).

In the first case, three utterances are uttered by the same person, and emotions are transferred from *"neutral"* to *"joy"*. The baseline exhibits a comprehensive understanding of the texts' semantics, discerning the repeated occurrence of "push" and then regarding the speaker displays *"fear"* emotion. In contrast, EmoTrans can simultaneously perceive the conversation's textual and ET information, and then predict the target utterance's emotion correctly. In this example, ET information effectively neutralizes the bias of text semantics and corrects the wrong emotion prediction.

In the second example, two parties are involved, and discerning the speaker's emotion solely from the semantic aspect proves challenging. Nevertheless, as indicated by the conversation's emotional information exhibited in the first column, the conversation continues in a "joyful" atmosphere, which serves as a valuable characteristic of the conversation. EmoTrans successfully captures the conversation's atmosphere and achieves accurate predictions, thus highlighting the efficacy of ET information in the ERC task.

The third example presents a more significant challenge as it involves a shift in the speaker's emotion from *"anger"* to *"sad"*. Based on contextual semantics, the baseline model identifies the speaker's emotion as "anger", underscoring the challenge it faces in disentangling the biases present in the text. Capturing ET features refers to the comprehension of emotional changes according to the underlying textual semantics. This capability enables EmoTrans to holistically assess the influence of previous utterances and the emotional states of speakers, thereby effectively addressing the challenge posed by diverse emotional fluctuations.

Contextual Utterance	Constructed Sample	w/o ET	w/ ET	Label
Joey$_{neutral}$: "Push 'em out, push 'em out, harder, harder." **Joey**$_{joy}$: "Push!" **Joey**$_{joy}$: "Come on, Lydia, you can do it."	$\langle s \rangle$ *Joey: Push 'em out, push 'em out, harder, harder.* $\langle /s \rangle$ *Joey expresses* $\langle$***mask***$\rangle$ $\langle /s \rangle$ *Joey: Push!* $\langle /s \rangle$ *Joey expresses* $\langle$***mask***$\rangle$ $\langle /s \rangle$ *Joey: Come on, Lydia, you can do it.* $\langle /s \rangle$ *Joey expresses* $\langle$***mask***$\rangle$ $\langle /s \rangle$	fear	joy	joy
Phoebe$_{joy}$: "Okay!" **Gray**$_{joy}$: "Oh yeah? Well maybe you and I should take a walk through a bad neighborhood." **Phoebe**$_{joy}$: "Yeah! Sure! Yep! Oh, y'know what? If I heard a shot right now, I'd throw my body on you."	$\langle s \rangle$ *Phoebe: Okay!* $\langle /s \rangle$ *Phoebe expresses* $\langle$***mask***$\rangle$ $\langle /s \rangle$ *Gary: Oh yeah? Well maybe you and I should take a walk through a bad neighborhood.* $\langle /s \rangle$ *Gary expresses* $\langle$***mask***$\rangle$ $\langle /s \rangle$ *Phoebe: Yeah! Sure! Yep! Oh, y'know what? If I heard a shot right now, I'd throw my body on you.* $\langle /s \rangle$ *Phoebe expresses* $\langle$***mask***$\rangle$ $\langle /s \rangle$	anger	joy	joy
Jessica Lockhart$_{anger}$: "Oh, my baby!" **Dina**$_{anger}$: "What are you going to do? Kill him? Like you did with Charles?!" **Jessica Lockhart**$_{anger}$: "Oh yes there is!" **Dina**$_{sad}$: "I'm going to keep dating him Mother, and there's nothing you can do about it!"	$\langle s \rangle$ *Jessica Lockhart: Oh, my baby!* $\langle /s \rangle$ *Jessica Lockhart expresses* $\langle$***mask***$\rangle$ $\langle /s \rangle$ *Dina: What are you going to do? Kill him? Like you did with Charles?!* $\langle /s \rangle$ *Dina expresses* $\langle$***mask***$\rangle$ $\langle /s \rangle$ *Jessica Lockhart: Oh yes there is!* $\langle /s \rangle$ *Jessica Lockhart expresses* $\langle$***mask***$\rangle$ $\langle /s \rangle$ *Dina: I'm going to keep dating him Mother, and there's nothing you can do about it!* $\langle /s \rangle$ *Dina expresses* $\langle$***mask***$\rangle$ $\langle /s \rangle$	anger	sad	sad

Table 5: Case studies show that the conversational feature of emotional transition (ET) empowers the model to rectify incorrect emotion predictions.

6. Conclusion

ET information in the conversation proves to be a valuable feature, which can facilitate the model's comprehension of intricate emotional changes and mitigate the utterance's emotional semantic bias. In this paper, we first develop a sample construction method to achieve the implicit representation of ET information within each sample and then propose EmoTrans to capture and leverage ET features from this information to enhance the utterance's emotional semantics for the ERC task. The superiority of EmoTrans is verified on four widely-used benchmarks, with new state-of-the-art results achieved on MELD and IEMOCAP, second-best results on EmoryNLP and third-best results on DailyDialog. Motivated by the formidable comprehension capabilities of Large Language Models (LLMs) and the effectiveness of ET information, we will explore the possibility of integrating ET information into LLMs to further improve the ERC performance in the future.

7. Ethical Considerations

To consider ethical concerns, we describe the following: (1) We conduct all experiments on existing datasets derived from public scientific research. (2) Our work does not involve any sensitive tasks or data. (3) Our analysis is consistent with the experimental results. (4) Our code is available at https://github.com/jian-projects/emotrans.

8. Limitations

While EmoTrans consists of multiple components, it is advisable not to operate each component independently, as the preceding component forms the basis for the subsequent one. Central to our approach is the capture of ET features, which is at the cost of additional computation. In addition, our approach is more suitable for predicting the late utterances in the conversation, as their ET information is more complete and sufficient.

9. Acknowledgements

This work is supported by the public technology service platform project of Xiamen City (No.3502Z20231043).

10. Bibliographical References

Yinan Bao, Qianwen Ma, Lingwei Wei, Wei Zhou, and Songlin Hu. 2022. Speaker-guided encoder-decoder framework for emotion recognition in conversation. In *IJCAI*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Paul Ekman. 1993. Facial expression and emotion. *American psychologist*, 48(4):384.

Zhiyi Fu, Wangchunshu Zhou, Jingjing Xu, Hao Zhou, and Lei Li. 2022. Contextual representation learning beyond masked language modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 2701–2714.

Qingqing Gao, Biwei Cao, Xin Guan, Tianyun Gu, Xing Bao, Junyan Wu, Bo Liu, and Jiuxin Cao. 2022. Emotion recognition in conversations with

emotion shift detection based on multi-task learning. *Knowledge-Based Systems*, 248:108861.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. In *Findings of the Association for Computational Linguistics*, pages 2470–2481.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 154–164.

Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2023. Supervised adversarial contrastive learning for emotion recognition in conversations. *Proceedings of the 29th International Conference on Computational Linguistics*.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 7042–7052.

Tatsuya Ide and Daisuke Kawahara. 2022. Building a dialogue corpus annotated with expressed and experienced emotions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 21–30.

Wenxiang Jiao, Haiqin Yang, Irwin King, and Michael R. Lyu. 2019. HiGRU: Hierarchical gated recurrent units for utterance-level emotion recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 397–406.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

Bongseok Lee and Yong Suk Choi. 2021. Graph based network with contextualized representations of turns in dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 443–455.

Joosung Lee and Wooin Lee. 2022. Compm: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *arXiv preprint arXiv:2309.11911*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang. 2021. Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In *Findings of the Association for Computational Linguistics*, pages 1204–1214.

Shimin Li, Hang Yan, and Xipeng Qiu. 2022. Contrast and generation make bart a good dialogue emotion recognizer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11002–11010.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 986–995.

Xiao Liu, Jian Zhang, Heng Zhang, Fuzhao Xue, and Yang You. 2023. Hierarchical dialogue understanding with special tokens and turn-level attention. *International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yukun Ma, Khanh Linh Nguyen, Frank Z Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64:50–70.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6818–6825.

Donovan Ong, Jian Su, Bin Chen, Anh Tuan Luu, Ashok Narendranath, Yue Li, Shuqi Sun, Yingzhan Lin, and Haifeng Wang. 2022. Is discourse role important for emotion recognition in conversation? *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11121–11129.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021a. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13789–13797.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021b. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 1551–1560.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2024. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36.

Dongming Sheng, Dong Wang, Ying Shen, Haitao Zheng, and Haozhuang Liu. 2020. Summarize before aggregate: A global-to-local heterogeneous graph inference network for conversational emotion recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4153–4163.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022a. Supervised prototypical contrastive learning for emotion recognition in conversation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

Xiaohui Song, Liangjun Zang, Rong Zhang, Songlin Hu, and Longtao Huang. 2022b. Emotionflow: Capture the dialogue level emotion transitions. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8542–8546. IEEE.

Livia Tomova, Kimberly L Wang, Todd Thompson, Gillian A Matthews, Atsushi Takahashi, Kay M Tye, and Rebecca Saxe. 2020. Acute social isolation evokes midbrain craving responses similar to hunger. *Nature Neuroscience*, 23(12):1597–1605.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 2023. Should you mask 15% in masked language modeling? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2985–3000.

Gloria Willcox. 1982. The feeling wheel: A tool for expanding awareness of emotions and increasing spontaneity and intimacy. *Transactional Analysis Journal*, 12(4):274–276.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32.

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 44–52.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.

Ling.Yu Zhu, Zhengkun Zhang, Jun Wang, Hongbin Wang, Haiying Wu, and Zhenglu Yang. 2022. Multi-party empathetic dialogue generation: A new task for dialog systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 298–307.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 1571–1582.