DialogueLLM: Context and Emotion Knowledge-Tuned LLaMA Models for Emotion Recognition in Conversations

Yazhou Zhang[1,2,3], **Mengyao Wang**[2], **Prayag Tiwari**[4], **Qiuchi Li**[5], **Benyou Wang**[6], **Jing Qin**[1*]
[1]The Hong Kong Polytechnic University
[2]Zhengzhou University of Light Industry
[3]China Mobile Communication Group Tianjin Co.
[4]Halmstad University
[5]Copenhagen University
[6]The Chinese University of Hong Kong, Shenzhen

Abstract

Large language models (LLMs) and their variants have shown extraordinary efficacy across numerous downstream natural language processing (NLP) tasks, which has presented a new vision for the development of NLP. Despite their remarkable performance in natural language generating (NLG), LLMs lack a distinct focus on the emotion understanding domain. As a result, using LLMs for emotion recognition may lead to suboptimal and inadequate precision. Another limitation of LLMs is that they are typical trained without leveraging multi-modal information. To overcome these limitations, we propose DialogueLLM, a context and emotion knowledge tuned LLM that is obtained by fine-tuning LLaMA models with 13,638 multi-modal (i.e., texts and videos) emotional dialogues. The visual information is considered as the supplementary knowledge to construct high-quality instructions. We offer a comprehensive evaluation of our proposed model on three benchmarking emotion recognition in conversations (ERC) datasets and compare the results against the SOTA baselines and other SOTA LLMs. Additionally, DialogueLLM-7B can be easily trained using LoRA on a 40GB A100 GPU in 5 hours, facilitating reproducibility for other researchers.

1 Introduction

Scaling up language models has been proved to be an effective way to improve the performance and sample efficiency in downstream NLP tasks. The rise of instruction-following LLMs has garnered considerable attention from academy and industry, due to their outstanding performance in human instruction understanding and responsing. Language modeling has evolved from small language models (SLMs), e.g., GPT (Radford et al., 2018), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), etc., to LLMs, e.g., ChatGPT[1] GPT-4 (OpenAI, 2023), Claude[2], etc.

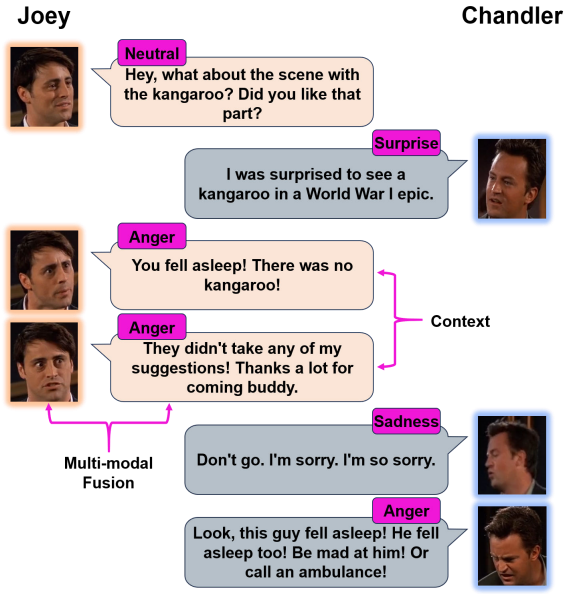

Figure 1: Sample utterances in a multi-modal conversation from the MELD dataset.

Compared with SLMs, LLMs are characterized by their enormous parameter size, typically reaching tens of billions or even more. They often have stronger generalization across various downstream tasks and unique emergent ability to tackle complex tasks. Despite that LLMs possess numerous commendable qualities, they also present a couple of limitations that deserve careful consideration and in-depth exploration: (1) the non-open source status may restrict the development of LLMs community and (2) they are not specifically designed for emotion recognition task. Their broad domain knowledge frequently proves insufficient when tackling such specialized domains. For example, Zhang et al. (Zhang et al., 2023a) showed LLMs' unsatisfactory performance in many emo-

*Corresponding author

[1]https://chat.openai.com/
[2]https://www.anthropic.com/product

tion recognition tasks without fine-tunning on emotional knowledge. Hence, the potential of LLMs in understanding emotional communication has been unexplored.

Human communication is the process of exchanging information, thoughts, ideas, and feelings between individuals, which is naturally filled with the participant's subjective attitudes or emotions. Emotion recognition in conversations (ERC) aims to accurately detect the feelings and emotions expressed in the utterances. It has immense potential in dialogue understanding and intent analysis, and has been an active task in the recent literature (Liu et al., 2023; Zhang et al., 2023c,b). In general, there are two key factors that contribute the classification performance, i.e., multi-modal fusion and context dependency (also known as intra- and inter-speaker dependency) (Ma et al., 2023). Multimodal fusion involves combining information from different sources or modalities, such as text, visual cues, to obtain a more comprehensive and accurate understanding of the emotional utterance. In view that emotions are influenced by the surrounding environment, the relationship between the participants, etc., context is a critical factor in accurately classifying emotions in conversations. The same utterance in different contexts might express different emotions. Fig. 1 illustrates an example to introduce the presence of both challenges.

To overcome the above-mentioned limitations, it's crucial to develop emotion-tailored LLMs that can better understand human-human conversation and take a further step towards emotion intelligence. In this paper, we present DialogueLLM, a collection of emotion and context knowledge enhanced language models, which is specifically designed for ERC based on the open-source base models, namely LLaMA 2 (Touvron et al., 2023b). By collecting diverse instruction conversational data based on emotional knowledge from five open-source benchmarking datasets (i.e., MELD (Poria et al., 2018), MEISD (Firdaus et al., 2020), Dailydialog (Li et al., 2017), IEMOCAP (Busso et al., 2008), EmoryNLP (Zahiri and Choi, 2017).), we obtain 13,638 multi-party dialogues, over 120,000 utterances. Meanwhile, the visual information (i.e., videos) will be forward into GPT-4 API to generate the text descriptions, which is considered as the supplementary knowledge to construct high-quality instructions. We adopt an end-to-end supervised instruction-finetuning approach on the open-source LLaMA 2 (7B and 13B) base models. Additionally, DialogueLLM-7B can be easily trained using LoRA on a 40GB A100 GPU in 5 hours, facilitating reproducibility for other researchers.

We offer a comprehensive evaluation of our proposed DialogueLLM model across three ERC tasks and compare the results against 13 state-of-the-art ERC baselines, including bc-LSTM (Zhang et al., 2023d), MTL (Li et al., 2020), ICON (Hazarika et al., 2018a), DialogXL (Shen et al., 2021a), TODKAT (Zhu et al., 2021), CoG-BART (Li et al., 2022), DialogueGCN (Ghosal et al., 2019), RGAT (Ishiwatari et al., 2020), DAG-ERC (Shen et al., 2021b), DialogueRNN (Majumder et al., 2019), DialogueCRN (Hu et al., 2021), CauAIN (Zhao et al., 2022), COIN (Zhang and Chai, 2021) and three SOTA LLMs, i.e., LLaMA, Alpaca[3] and LLaMA 2. The experimental results show the effectiveness of DialogueLLM with the margin of 1.88%, 5.37% and 0.19% for three ERC tasks. The study reveals that DialogueLLM significantly outperforms the SOTA baselines in ERC tasks requiring deeper understanding or conversational emotion information. A series of sub-experiments underscore how emotion and context knowledge enhanced LLMs deal with ERC tasks. The main innovations of the work are concluded as follows:

- To the best of our knowledge, DialogueLLM is the first open source emotional LLM that is specifically designed for ERC tasks.
- The visual information is proposed to construct high-quality instructions.
- We show a comprehensive dataset of 120K utterances to serve as a training resource, ensuring our model has accurate and domain-specific knowledge.
- Our model achieves state-of-the-art performance on ERC tasks. We show that an open-sourced model finetuned with emotional knowledge has the potential to achieve even higher accuracy than SOTA.

The rest of this paper is organized as follows. Section II briefly outlines the related work. In Section III, we describe the proposed DialogueLLM in detail. In Section IV, we report the empirical experiments and analyze the results. Section V

[3]https://crfm.stanford.edu/2023/03/13/alpaca.html

concludes the paper and points out future research directions.

2 Related Work

We depict two lines of research that form the basis of this work: large language models and emotion recognition in conversations models.

2.1 Large Language Models

In recent years, significant advancements in natural language processing (NLP) have been attributed to the emergence of large language models. These models have showcased remarkable capabilities such as in-context learning, few-shot prompting, instruction following, etc. These dynamic abilities have greatly contributed to boosting the effectiveness of language models, thus enabling AI algorithms to achieve unparalleled levels of effectiveness and productivity. Typically, models like the transformer architecture-based LLMs are first pretrained using extensive datasets comprising diverse languages and domains (Zhao et al., 2023).

OpenAI has achieved significant milestones with the creation of two groundbreaking models: ChatGPT and GPT-4. These models herald a new era in language processing. However, due to their proprietary nature, there has been a proliferation of LLM variants featuring tens or even hundreds of billions of parameters. Our aim is to categorize these LLMs into two distinct groups based on their specialization: general LLMs and specialized LLMs. General LLMs are designed for versatility across a wide spectrum of NLP tasks, including machine translation, language comprehension, and dialogue generation. Prominent examples of these models are GPT-4, Claude, ChatGPT, LLaMA, PanGu-Σ (Ren et al., 2023), Bard [4], Falcon (Penedo et al., 2023), etc. Such LLMs are not specifically optimized for any particular task. While they can perform well across a range of tasks, but their potentials in specific scenarios await further explore.

In contrast, specialized LLMs also known as task-specific LLMs, are fine-tuned for specific tasks via task-specific architectures and knowledge, allowing them to achieve higher or comparable performance against general LLMs with fewer parameters. For example, Wang et al. (Chen et al., 2023) released a large language model 'Phoenix' to meet the needs of multiple languages. Liu and Low (Liu and Low, 2023) fine-tuned a Goat model based on LLaMA model to deal with arithmetic tasks. In view that LLMs have not yet performed optimally in medical domain tasks, a few Chinese and English medical knowledge enhanced LLMs have been proposed, such as HuaTuo (Wang et al., 2023), PMC-LLaMA (Wu et al., 2023), Dr. LLaMA (Guo et al., 2023), ChatDoctor (Li et al., 2023). Different from the above-mentioned works, we aim to explore the potential of LLMs in ERC domain and take a further step towards emotional intelligence.

[4] https://bard.google.com/

2.2 Emotion Recognition in Conversations

Emotion recognition in conversation (ERC) has become a popular research topic. In this task, the conversational context dependency and multi-modal fusion have been considered thtough deep learning approaches. These efforts can be broadly categorized into methods based on sequences and those based on the Transformer architecture.

Sequence based approaches often use the sequential information in a dialogue to capture the contextual and emotional features. For example, Poria et al. (Poria et al., 2017) introduced an LSTM-based model that effectively captured conversational context from surrounding videos, thereby enhancing the classification process. Building upon this idea, Hazarika et al. (Hazarika et al., 2018b) presented the conversational memory network (CMN), which harnessed contextual information from the conversation history to improve ERC. Another approach by Majumder et al. (Majumder et al., 2019) introduced the DialogueRNN model, which meticulously monitored the states of individual participants throughout the conversation, utilizing this information for ERC. In terms of multimodal advancements, Poria et al. (Poria et al., 2018) played a pivotal role by crafting the first-ever multimodal conversational dataset named the multimodal emotionlines dataset (MELD). This dataset was instrumental in propelling the field of conversational sentiment analysis. Further innovation came from Zhang et al. (Zhang et al., 2019), who devised the quantum-inspired interactive network (QIN) model for conversational emotion recognition, showcasing its effectiveness. Moreover, their research extended to the realm of multi-task learning. Zhang et al. (Zhang et al., 2021) devised a quantum-inspired multi-task learning framework catering to both sarcasm detection and emotion recognition in conversations.

Transformer based approaches often adopt the

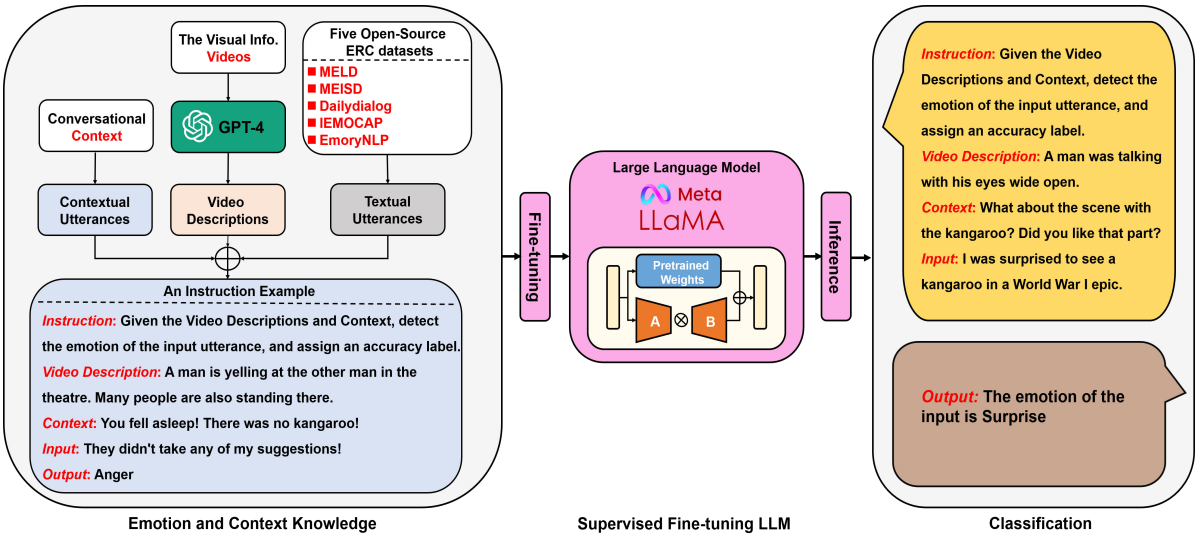

Figure 2: Overview of DialogueLLM fine-tuning and classification pipeline.

"fine-tuning" paradigm. They build the models upon the foundation of Transformer based pre-trained language models. Then, such models are supervised-fine-tuned with labeled samples and are adapted to the specific task. For instance, Zhong et al. (Zhong et al., 2019) presented a Transformer with graph attention to incorporate commonsense knowledge and contextual features. Li et al. (Li et al., 2022) used a supervised contrastive term and a response generation task to enhance BART's ability for ERC. Zhang et al. (Zhang et al., 2023b) proposed a multi-modal multi-task network based on BERT and graph attention network (GAT) to detect sentiment and emotion. They also proposed a quantum inspired multi-task interactive Transformer to model sentiment and emotion (Liu et al., 2023). Chudasama et al. (Chudasama et al., 2022) presented a multi-modal fusion network (M2FNet) to learn emotion-relevant multi-modal features by revising the Transformer encoder. Qiao et al. (Qiao et al., 2023) built a mutual-enhanced incongruity learning network upon RoBERTa and graph convolution networks to identify sarcasm. Pramanick (Pramanick et al., 2022) combined self-attention with BERT to model intra-modal correspondence and optimal transport for cross-modal correspondence, aiming to discover sarcasm and humor.

Compared with them, DialogueLLM possesses the abilty to understand complex emotions. In addition, our model would also benefit the development of task-specific LLMs.

3 Methodology

In this section, we detail the comprehensive pipeline for training DialogueLLM models, as shown in Fig. 2.

3.1 Problem Formulation

Assume that there are N conversation instances in the instruction dataset, the i^{th} conversation D_i contains K multi-modal utterances, which is represented as $D_i = \{(C_k, M_k), Y_k\}$, where C_k denotes the contextual utterances, M_k represents the k^{th} target utterance to be classified, Y_k means the emotion labels of the k^{th} utterance, where $i \in [1, 2, ..., N]$, $k \in [1, 2, ..., K]$. The target utterance consist of textual (T) and visual (V) modalities, i.e., $M_k = (T_k, V_k)$, where $T_k \in \mathcal{R}^{l_{T_k} \times d_{T_k}}$, $V_k \in \mathcal{R}^{l_{I_k} \times d_{I_k}}$. Here, l_{T_k} and l_{V_k} denote the sequence length of textual and visual utterances, d_{T_k} and d_{V_k} represents the dimensions of the textual and visual features.

Now, we summarize our research problem as: *Given one multi-speaker conversation including K multi-modal utterances, how to detect their emotions?* It could be written as:

$$\zeta = \prod_{k} p\left(Y_k | C_k, M_k, \Theta\right) \tag{1}$$

where Θ denotes the parameter set.

3.2 Base Model

The first key component is to select open-source and strong foundation language models. LLaMA is a collection of open source foundation language

models ranging from 7B to 65B parameters, which is trained on trillions of tokens using publicly available datasets. It achieves state-of-the-art performance on numerous benchmarks, which has greatly promoted the research progress of LLMs. A considerable number of researchers choose to expand the capabilities of LLaMA models though instruction tuning, due to the lower computational costs.

Furthermore, OpenAI has just developed and released LLaMA 2, which is an updated version of LLaMA 1. Compared with LLaMA 1, the training data used for LLaMA 2 was increased by 40% and the context length was doubled. LLaMA 2 also incorporated grouped query attention mechanisms. LLaMA 2 shows many behaviors similar to ChatGPT, but is also surprisingly small and easy to reproduce. Hence, we adopt LLaMA 2-7B model as our base model. Furthermore, LLaMA-7B, LLaMA-13B and LLaMA 2-13B have also been attempted and evaluated in the experiments. We use low-rank adaptation (LoRA) to finetune them with only 2.1 million trainable parameters.

3.3 Emotion and Context Knowledge Based Instruction Dataset

Human conversation is filled with different emotions, such as neutral, anger, happiness, surprise, etc. To satisfy complex emotion recognition needs, DialogueLLM undergos an instruction-tuning step which involves training on supervised instruction/input/output data. Instruction tuning helps align DialogueLLM with human prompts, enabling precise customization for emotional domains. This allows DialogueLLM to become adaptable and proficient at generating accurate emotional responses. In this paper, we create a high-quality instruction dataset by leveraging five widely used benchmarking ERC datasets. Since many potential shortcomings exist in automatic generation of samples using strong language models (e.g., ChatGPT), such as low quality, repetition, and lack of diversity, etc., different from the existing works (Peng et al., 2023; Liu and Low, 2023), we do not use ChatGPT to generate instances. The benchmarking ERC datasets have provided clean samples with precise annotations, which will be an optimal choice for creating instruction dataset.

The training sets of five benchmarking datasets (i.e., MELD, MEISD, Dailydialog, IEMOCAP and EmoryNLP) are treated as the data source, altogether 13,638 multi-party dialogues, over 120,000 utterances are collected. In view that the labels are from different datasets, we first pre-process the lables. For example, "joy", "happy" and "happiness" will be normalized to be "happiness". The instructions are constructed based on the task definition and label space, e.g., *"Given the Video Description and Context, detect the emotion of the input, and assign an accuracy label from ['happiness', 'anger', 'fear', 'sadness', 'disgust', 'surprise', 'neutral'].".* The textual raw samples and the counterpart labels are normalized to the input/output pairs.

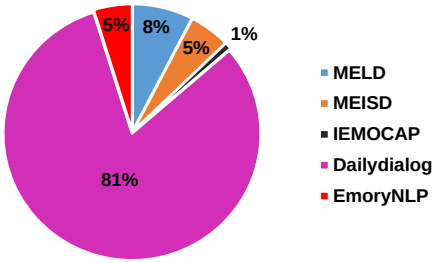

Figure 3: The distribution of five ERC datasets.

In view of the importance of the conversational context and multi-modal knowledge, the contextual utterances and the visual information are incorporated into instruction instances. Assume that there are k contextual utterances before the target utterance, we would list them before the input content. In this work, the default size is $k = 1$. In addition, the corresponding video is split into frames and forward them through GPT-4 API, to generate the descriptions of this video. Then, such descriptions are considered as the supplementary knowledge. More statistics of this dataset are presented in Fig. 3 and Fig. 4. Notably, "Happiness" constitutes the largest proportion, accounting for approximately 31.2% of the total instances. In contrast, "Fear" and "Surprise" are represented at lower proportions. "Fear" comprises around 8.8% of the dataset. Similarly, "Surprise" constitutes approximately 6.3% of the dataset, indicating a notable but still comparatively moderate occurrence of this particular emotion. "Anger", "Sadness", and "Disgust" collectively account for around 23.9% of the dataset. Specifically, "Anger" represents approximately 12.6%, "Sadness" makes up about 11.5%, and "Disgust" comprises around 5.3%.

Finally, this instruction dataset is used for supervised fine-tuning. Notably, all the instances in the dataset are normalize as "instruction/video descriptions/context/input/output" pairs (see Fig. 2).

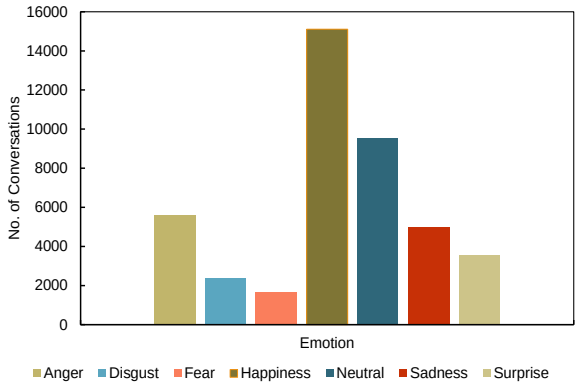

Figure 4: The distribution of seven basic emotions across five datasets.

3.4 Training and Implementation

The DialogueLLM-7B model is fine-tuned LLaMA 2-7B with the emotional knowledge based instruction data to acquire emotion recognition skills. LLaMA-7B/13B and LLaMA 2-13B are also fine-tuned through the same procedure. Training a Alpaca-7B model will cost about 5 hours on a 40GB A100 GPU. The total approximate tokens seen during pre-training is approximately 22 billion tokens. We optimize our model with the AdamW optimizer with the following hyper-parameters: $\beta_1 = 0.9$, $\beta_2 = 0.95$. We use a cosine learning rate schedule, such that the final learning rate is equal to 10% of the maximal learning rate. The activation function is set to SwiGLU to improve performance. The target utterances are forward through DialogueLLM models to generate the emotion labels.

4 Experiments

4.1 Research Question

RQ1: Is it effective to propose an emotion-tailored LLM?

RQ2: Does modeling of the contextual dependency and multi-modal information help improve performance?

RQ3: Does DialogueLLM has powerful in-context learning abilities?

To answer RQ1, we compare the proposed DialogueLLM with a wide range of state-of-the-art baselines and other LLMs on three benchmark datasets in Sec. 4.4. To answer RQ2, we conduct a ablation test by removing one component at one time in Sec. 4.5. To answer RQ3, we consider zero-shot and few-shot prompting setups, and report their results in Sec. **??**.

4.2 Experimental Settings

Datasets. Three benchmark ERC datasets which include the textual and visual utterances with high quality emotion annotations, are selected as the experimental beds, $viz.$ MELD[5] (Poria et al., 2018), IEMOCAP[6], and EmoryNLP[7].

MELD. It consists of 13,708 multi-modal utterances from 1,433 multi-party dialogues of Friends TV series. The utterances in each dialogue are annotated with one of three sentiments (positive, negative or neutral) and one of seven emotions (anger, disgust, fear, joy, neutral, sadness or surprise). The overall Fleiss' kappa score reaches 0.43. In this work, we only use textual and visual information.

IEMOCAP. It is comprised of 151 recorded dialogue videos, encompassing a total of 302 videos across the entire dataset, each involving two speakers per session. The annotations for this dataset encompass 9 distinct emotions (anger, excitement, fear, sadness, surprise, frustration, happiness, disappointment, and neutrality). The recordings are distributed across five sessions, with each session featuring five pairs of speakers.

EmoryNLP. consists of 97 episodes, 897 scenes, and 12,606 utterances, which is a textual corpus that comprises multi-party dialogue transcripts of the Friends TV show. Each utterance is annotated with one of seven emotions, i.e., sad, mad, scared, powerful, peaceful, joyful, and neutral. The detailed statistics are shown in Table 1.

Evaluation metrics. In line with the previous approaches, *accuracy* (Acc) and *weighted-F1* (w-F1) are used as evaluation metrics. For each method, we run five random seeds and report the average result of the test sets.

Hyper-parameter. We report the detailed hyperparameter settings of DialogueLLM on three datasets in Table 2. The maximum context length is set to 4,096. We use a weight decay of 0.1 and gradient clipping of 1.0. The batch size is set to 128.

4.3 Compared Baselines

A wide range of SOTA baselines are included for comparison including pre-trained language model (PLM) based and LLM based approaches. They are:

- ***PLMs based approaches:***

[5]https://github.com/declare-lab/MELD.
[6]https://sail.usc.edu/iemocap/.
[7]https://github.com/emorynlp

Table 1: Testing dataset statistics.

Type	Dataset	Dialogue			Utterance			Class	Metric
		Train	Validation	Test	Train	Validation	Test		
Main Datasets	MELD	1,039	114	280	9,989	1,109	2,610	7	Weighted-F1
	IEMOCAP	120		31	5,810		1,623	9	Weighted-F1
	EmoryNLP	659	89	79	7,551	954	984	7	Weighted-F1
Auxiliary Datasets	Dailydialog	11,118	1000	1000	87,832	7912	7863	6	BLEU
	MEISD	702	93	205	14040	1860	4100	8	Micro-F1

Table 2: Hyperparameters for fine-tuning DialogueLLM.

Hyperparameter	Value
Batch size	128
Micro batch size	8
Epoch	5
Learning rate	3e-4
Lora r	4
Lora alpha	16
Lora dropout	0.05
Cutoff length	256

(1) bc-LSTM (Zhang et al., 2023d) implements an utterance-level LSTM to capture contextual features.

(2) ICON (Hazarika et al., 2018a) hierarchically models the self- and inter-speaker emotional influences into global memories, and generates contextual summaries.

(3) MTL (Li et al., 2020) exploits speaker identification (SI) as an auxiliary task to enhance the utterance representation in conversations.

(4) DialogXL (Shen et al., 2021a) modifies the recurrence mechanism of XLNet to store longer historical context and dialog-aware self-attention to deal with the multi-party structures.

(5) TODKAT (Zhu et al., 2021) designs a transformer-based encoder-decoder architecture fuses the topical and commonsense information, and performs the emotion label sequence prediction.

(6) CoG-BART (Li et al., 2022) uses the pre-trained encoder-decoder model BART as the backbone model and utilizes an auxiliary response generation task to enhance the model's ability of handling context information.

(7) DialogueRNN (Majumder et al., 2019) designs a method based on recurrent neural networks (RNN) that keeps track of the individual party states throughout the conversation and uses this information for emotion classification.

(8) DialogueGCN (Ghosal et al., 2019) leverages self and inter-speaker dependency of the interlocutors to model conversational context for emotion recognition.

(9) DialogueCRN (Hu et al., 2021) designs multi-turn reasoning modules to extract and integrate emotional clues.

(10) RGAT (Ishiwatari et al., 2020) proposes relational position encodings to capture both the speaker dependency and the sequential information.

(11) DAG-ERC (Ghosal et al., 2019) regards each conversation as a directed acyclic graph to model the conversation context.

(12) CauAIN (Zhao et al., 2022) retrieves causal clues provided by commonsense knowledge to guide the process of causal utterance traceback.

(13) COIN (Zhang and Chai, 2021) is a conversational interactive model to mitigate the problem of overlooking the immediate mutual interaction between different speakers by applying state mutual interaction within history contexts.

(13) SACL-LSTM (Hu et al., 2023) applies contrast-aware adversarial training to generate worst-case samples and uses a joint class-spread contrastive learning objective on both original and adversarial samples.

- ***LLMs based approaches:***

(1) LLaMA (Touvron et al., 2023a) takes a sequence of words as an input and predicts a next word to recursively generate text.

(2) Alpaca is a state-of-the-art finedtuning version of LLaMA, by using supervised learning

from a LLaMA 7B model on 52K instruction-following demonstrations.

(3) LLaMA 2 (Touvron et al., 2023b) is trained on 2 trillion tokens, and have double the context length than Llama 1, and outperforms other open source language models on many external benchmarks.

4.4 Results and Anlysis

The experimental performance of all baselines is shown in Table 3. We divide these baselines into two categories, i.e., standard deep learning architectures and large language models without fine-tuning. We will conduct a detailed analysis of their classification performance.

Table 3 shows that when LLM is fine-tuned without using a knowledge base, Llama7B, Alpaca, and Llama2-7B all perform very poorly in emotion recognition tasks. The performance of the large language model DialogueLLM through corpus fine-tuning has been greatly improved on the three datasets, which shows that fine-tuning of LLM is indispensable when performing specific tasks. We can notice that MTL performs very poorly compared to other baseline models, with the worst classification performance in Meld and Emorynlp datasets. One main reason is that MTL ignores the modeling of conversation-level interaction information, and the model cannot learn contextual information, leading to inaccuracy in classification results. The modeling of DialogueCNN, DialogueGCN, and DialogueCRN focuses on the contextual information perceived by the speaker. Compared with MTL, the performance has been significantly improved. This further confirms that modeling of contextual information is crucial for conversation-level emotion recognition.

MELD. In traditional models like bc-LSTM and MTL exhibit lower performance with F1 scores of 65.87% and 62.45% and accuracies of 64.87% and 61.9%, respectively. EmoLLM achieves an impressive F1 score of 71.91% and an accuracy of 71.81% on the MELD dataset. It significantly outperforms the baseline models in terms of F1 score and accuracy. This demonstrates the effectiveness of EmoLLM in modeling emotion in human language on this dataset.

IEMOCAP. On the IEMOCAP dataset, EmoLLM continues to excel with an F1 score of 70.08% and an accuracy of 69.04%. It remains one of the top-performing models in the evaluation.DialogueCRN also delivers strong results on this dataset with an F1 score of 67.39% and an accuracy of 67.53%, showcasing its robust performance.While other models like CauAIN and SACL-LSTM perform well, achieving F1 scores above 65%, they fall short of EmoLLM's performance.

Emorynlp. EmoLLM maintains competitive performance on the Emorynlp dataset with an F1 score of 41.76% and an accuracy of 38.47%.In this dataset, DialogueRNN stands out as one of the top-performing baseline models with an F1 score of 43.66% and an accuracy of 37.54%.While the overall performance on this dataset is lower compared to the others, EmoLLM consistently outperforms the majority of the baseline models.

The experimental results demonstrate the effectiveness of the EmoLLM model in emotion recognition tasks across different datasets. It consistently achieves high F1 scores and accuracy, outperforming the majority of the baseline models. Notably, EmoLLM's performance is robust across various datasets, which underscores its versatility and reliability in handling different data sources and domains.

4.5 Ablation Test

Since contextual information is added to the language model fine-tuned in our experiment and video description is removed, we try to discuss their contribution to performance improvement. To this end, we propose two ablation experimental fine-tuning models: (1) *Remove contextual information* from DialogueLLM knowledge base (2) *Add video description* to the DialogueLLM knowledge base for model training, and perform emotion recognition on three datasets respectively.

The results of different ablation experimental models are shown in Table 4. It's worth noting that in the three data sets, the No-context model shows a slight decrease in F1 score and accuracy relative to DialogueLLM, which means that removing the contextual information of the dialogue is The performance of emotion classification has a certain adverse impact. Compared with DialogueLLM, the Add video description model shows significant performance degradation on meld and emorynlp. This is because our description of the information generated by the intercepted pictures of specific nodes in the conversation video is not accurate enough, and

Table 3: Comparison results (%) on different methods. The best scores are in bold.

Methods	# Param.	MELD		IEMOCAP		EmoryNLP		Avgerage	
		Acc	w-F1	Acc	w-F1	Acc	w-F1	Acc	w-F1
bc-LSTM	1.2M	65.87	64.87	63.08	62.84	40.85	36.84	56.60	54.85
ICON	0.5M	-	-	64.00	63.50	-	-	-	-
MTL	1.2M	62.45	61.90	-	-	36.36	35.92	49.40	48.91
DialogXL	510M	-	62.41	-	65.94	-	34.73	-	54.36
TODKAT	330M	67.24	65.47	61.11	61.33	42.38	38.69	56.91	55.16
CoG-BART	415.1M	64.95	63.82	65.02	64.87	40.94	37.33	56.97	55.34
DialogueRNN	9.9M	65.96	65.30	64.85	64.65	43.66	37.54	58.16	55.83
DialogueGCN	2.1M	63.62	62.68	62.49	62.11	36.87	36.43	54.33	53.14
DialogueCRN	3.3M	66.93	65.77	67.39	67.53	41.04	38.79	58.45	57.36
RGAT	13M	-	60.91	-	65.22	-	34.42	-	53.52
DAG-ERC	9.5M	63.75	63.36	66.54	66.53	39.64	38.29	56.64	56.06
CauAIN	6.1M	65.85	64.89	65.08	65.01	43.13	37.87	58.02	55.92
COIN	0.6M	-	-	66.05	65.37	-	-	-	-
SACL-LSTM	2.6M	**67.51**	**66.45**	**69.08**	**69.22**	**42.21**	**39.65**	**59.60**	**58.44**
LLaMA	7B	15.09	16.02	19.32	18.24	17.78	17.40	17.40	17.22
Alpaca	7B	19.22	18.37	20.35	19.16	17.95	17.33	19.17	18.29
LLaMA 2	7B	23.71	24.12	26.73	24.35	25.50	17.27	25.31	21.91
DialogueLLM	**7B**	**71.91**	**71.81**	**70.48**	**69.40**	**41.76**	**38.47**	**61.25**	**59.77**
Improve	**7B**	↑ 4.46%	↑ 5.36%	↑ 1.40%	↑ 0.18%	↓ 1.90%	↓ 1.18%	↑ 1.65%	↑ 1.33%

Table 4: Ablation experiment results.

Dataset	Models	w-F1	Acc
MELD	No Context	70.91	67.94
	Add Video Description	60.80	59.75
	DialogueLLM	**71.91**	**71.81**
IEMOCAP	No Context	68.14	68.01
	Add Video Description	-	-
	DialogueLLM	**70.48**	**69.40**
Emorynlp	No Context	39.41	36.25
	Add Video Description	38.00	28.57
	DialogueLLM	**41.76**	**38.47**

some noise is introduced to affect the classification of the model. We did not conduct the add video Description experiment on IEMOCAP because this corpus hired actors to sit in chairs for face-to-face conversations. The description of the picture information was too stereotyped, such as a man and a woman sitting on a chair to communicate face-to-face. Therefore, in order to avoid the experiment Impact on results We did not test on this dataset.

Here, we can give the answer to RQ3:Dialogue context information has a positive impact on the performance of sentiment analysis tasks, while increasing video description information may require more accurate processing to avoid adverse effects on sentiment classification results. These results provide valuable guidance for further improving the EmoLLM model.

5 Conclusion and Future Work

In this study, our emotional large language model DialogueLLM, specially designed for ERC, achieves advanced performance, which shows that an open source model incorporating emotion knowledge can achieve higher accuracy than SOTA. In addition, we also found that adding video description will affect the performance of emotional LLM, so adding an accurate picture description to emotional LLM may be particularly important. In the future, we will design and generate more accurate video descriptions and add multi-modal information to continue to explore the potential of LLM in the nlp field.

References

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.

Zhihong Chen, Feng Jiang, Junying Chen, Tiannan Wang, Fei Yu, Guiming Chen, Hongbo Zhang, Juhao Liang, Chen Zhang, Zhiyi Zhang, et al. 2023. Phoenix: Democratizing chatgpt across languages. *arXiv preprint arXiv:2304.10453*.

Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. 2022. M2fnet: Multi-modal fusion network for emotion recognition in conversation. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4652–4661.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Meisd: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations. In *Proceedings of the 28th international conference on computational linguistics*, pages 4441–4453.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*.

Zhen Guo, Peiqi Wang, Yanwei Wang, and Shangdi Yu. 2023. Dr. llama: Improving small language models in domain-specific qa via generative data augmentation. *arXiv preprint arXiv:2305.07804*.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018a. Icon: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2594–2604.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018b. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2018, page 2122. NIH Public Access.

Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2023. Supervised adversarial contrastive learning for emotion recognition in conversations. *arXiv preprint arXiv:2306.01505*.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. *arXiv preprint arXiv:2106.01978*.

Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto. 2020. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7360–7370.

Jingye Li, Meishan Zhang, Donghong Ji, and Yijiang Liu. 2020. Multi-task learning with auxiliary speaker identification for conversational emotion recognition. *ArXiv*, abs/2003.01478.

Shimin Li, Hang Yan, and Xipeng Qiu. 2022. Contrast and generation make bart a good dialogue emotion recognizer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11002–11010.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Tiedong Liu and Bryan Kian Hsiang Low. 2023. Goat: Fine-tuned llama outperforms gpt-4 on arithmetic tasks. *arXiv preprint arXiv:2305.14201*.

Yaochen Liu, Yazhou Zhang, and Dawei Song. 2023. A quantum probability driven framework for joint multi-modal sarcasm, sentiment and emotion analysis. *IEEE Transactions on Affective Computing*, pages 1–15.

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*.

Junxia Ma, Lu Rong, Yazhou Zhang, and Prayag Tiwari. 2023. Moving from narrative to interactive multi-modal sentiment analysis: A survey. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6818–6825.

OpenAI. 2023. Gpt-4 technical report.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Shraman Pramanick, Aniket Roy, and Vishal M Patel. 2022. Multimodal learning using optimal transport for sarcasm and humor detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3930–3940.

Yang Qiao, Liqiang Jing, Xuemeng Song, Xiaolin Chen, Lei Zhu, and Liqiang Nie. 2023. Mutual-enhanced incongruity learning network for multi-modal sarcasm detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9507–9515.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Xiaozhe Ren, Pingyi Zhou, Xinfan Meng, Xinjing Huang, Yadao Wang, Weichao Wang, Pengfei Li, Xiaoda Zhang, Alexander Podolskiy, Grigory Arshinov, et al. 2023. Pangu-{\Sigma}: Towards trillion parameter language model with sparse heterogeneous computing. *arXiv preprint arXiv:2303.10845*.

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

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021b. Directed acyclic graph network for conversational emotion recognition. *arXiv preprint arXiv:2105.12907*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*.

Sayyed M Zahiri and Jinho D Choi. 2017. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. *arXiv preprint arXiv:1708.04299*.

Haidong Zhang and Yekun Chai. 2021. Coin: Conversational interactive networks for emotion recognition in conversation. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 12–18.

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

Yazhou Zhang, Ao Jia, Bo Wang, Peng Zhang, Dongming Zhao, Pu Li, Yuexian Hou, Xiaojia Jin, Dawei Song, and Jing Qin. 2023b. M3gat: A multi-modal multi-task interactive graph attention network for conversational sentiment analysis and emotion recognition. *ACM Transactions on Information Systems*.

Yazhou Zhang, Qiuchi Li, Dawei Song, Peng Zhang, and Panpan Wang. 2019. Quantum-inspired interactive networks for conversational sentiment analysis. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence IJCAI-19*, pages 5436–5442. International Joint Conferences on Artificial Intelligence Organization.

Yazhou Zhang, Yaochen Liu, Qiuchi Li, Prayag Tiwari, Benyou Wang, Yuhua Li, Hari Mohan Pandey, Peng Zhang, and Dawei Song. 2021. Cfn: a complex-valued fuzzy network for sarcasm detection in conversations. *IEEE Transactions on Fuzzy Systems*, 29(12):3696–3710.

Yazhou Zhang, Jinglin Wang, Yaochen Liu, Lu Rong, Qian Zheng, Dawei Song, Prayag Tiwari, and Jing Qin. 2023c. A multitask learning model for multi-modal sarcasm, sentiment and emotion recognition in conversations. *Information Fusion*, 93:282–301.

Yazhou Zhang, Yang Yu, Dongming Zhao, Zuhe Li, Bo Wang, Yuexian Hou, Prayag Tiwari, and Jing Qin. 2023d. Learning multi-task commonness and uniqueness for multi-modal sarcasm detection and sentiment analysis in conversation. *IEEE Transactions on Artificial Intelligence*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Weixiang Zhao, Yanyan Zhao, and Xin Lu. 2022. Cauain: Causal aware interaction network for emotion recognition in conversations. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4524–4530.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 165–176, Hong Kong, China. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 1571–1582, Online. Association for Computational Linguistics.