PetKaz at SemEval-2024 Task 3: Advancing Emotion Classification with an LLM for Emotion-Cause Pair Extraction in Conversations

Roman Kazakov, Kseniia Petukhova, Ekaterina Kochmar
Mohamed bin Zayed University of Artificial Intelligence
{roman.kazakov, kseniia.petukhova, ekaterina.kochmar}@mbzuai.ac.ae

Abstract

In this paper, we present our submission to the SemEval-2023 Task 3 "The Competition of Multimodal Emotion Cause Analysis in Conversations", focusing on extracting emotion-cause pairs from dialogs. Specifically, our approach relies on combining fine-tuned `GPT-3.5` for emotion classification and a BiLSTM-based neural network to detect causes. We score 2nd in the ranking for Subtask 1, demonstrating the effectiveness of our approach through one of the highest weighted-average proportional F_1 scores recorded at 0.264. Our code is available at https://github.com/sachertort/petkaz-semeval-ecac.

1 Introduction

Developing dialog systems is a complex task that has attracted considerable attention from many technology companies and universities over the last 70 years since the introduction of Eliza in 1966 (Weizenbaum, 1966). Modern large language models (LLMs) like GPT-4 (OpenAI, 2023) are trained to avoid causing harm and often assert their lack of personal opinions on intricate matters, which is not at all natural for conversations. They do not respond in a way that is truly human, and they do not understand the range of feelings that words can cause. Recognizing the emotional implications of an utterance provides a deeper understanding of dialog, enabling the development of more human-like dialog systems. These systems could navigate conversations using a comprehensive understanding of emotional dynamics and planning responses based on this understanding rather than just predicting likely outcomes.

To bridge the gap between machine-generated dialogs and rich, complex human communication, we develop models for SemEval-2024 Task 3 "The Competition of Multimodal Emotion Cause Analysis in Conversations"[1] (ECAC) (Wang et al., 2024). This task was previously introduced in Xia and Ding (2019a) and later in Wang et al. (2023), where the authors also described a multimodal dataset called *Emotion-Cause-in-Friends* (ECF) for this task.

We focus only on Subtask 1, "Textual Emotion-Cause Pair Extraction in Conversations" (ECPE),[2] where the goal is to classify emotions and extract the corresponding textual causal spans. To accomplish this, we propose a two-stage pipeline: (1) first, emotions are classified using a fine-tuned LLM, and then (2) causes are extracted with a simple neural network consisting of BiLSTM and linear layers (see Figure 1). Our system achieved a weighted-average proportional F_1 score of 0.264, the primary metric in this competition's evaluation phase on the test set. Consequently, our team ranked 2nd out of 15 participating teams based on this metric. We provide an extensive analysis of the model's performance in Section 5.2.

2 Related Work

Recent research in the field of dialog systems and emotion-cause extraction has seen significant advancements through various innovative approaches, some of which we overview in this section. For instance, Chen et al. (2023) introduce a novel technique that uses graphs to model "causal skeletons" alongside a causal autoencoder (CAE) for refining these models by integrating both implicit and explicit causes.

Following closely, Zhang et al. (2023) present Dual Graph Attention Networks (DualGATs) that leverage discourse structure and speaker context through a combination of Discourse-aware GAT (DisGAT) and Speaker-aware GAT (SpkGAT), enriched with an interaction module for effective information exchange and context capturing.

Moving to earlier work, Kong et al. (2022) pro-

[1] https://nustm.github.io/SemEval-2024_ECAC/

[2] We did not participate in the multimodal track.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1127–1134
June 20-21, 2024

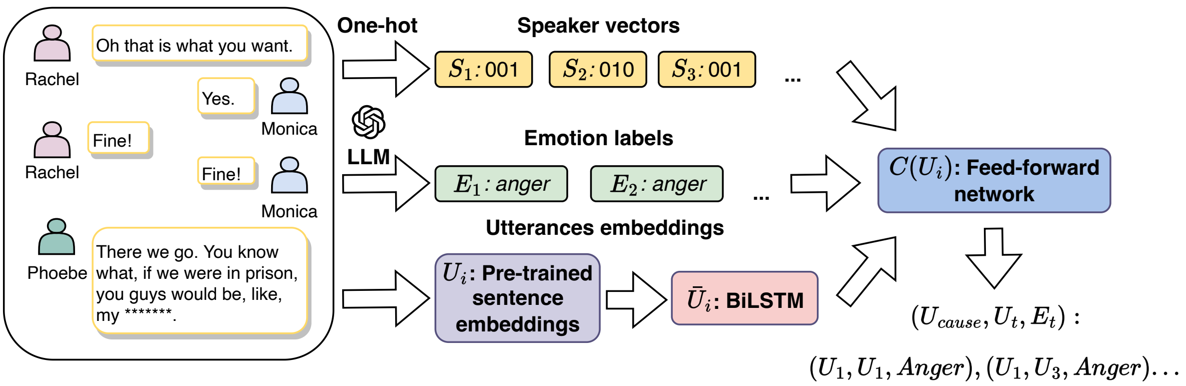


Figure 1: The pipeline for ECPE. Utterances are classified with emotion labels E_i, and speakers are represented with one-hot vectors S_i. Utterances are then encoded with pre-trained sentence embeddings U_i and enriched with context by BiLSTM $\bar{U}_i$. For each target utterance U_t, we detect whether any other utterance from the conversation history $H(U_t)$ is causal using a feed-forward network. $\bar{U}_i$, S_i (of a potential causal utterance), $\bar{U}_t$, S_t, and E_t are concatenated, and then, binary classification is performed. The pipeline outputs labelled emotion-cause pairs (U_i, U_t, E_t).

pose a discourse-aware model (DAM) that integrates emotion cause extraction with discourse parsing, using a Gated Graph Neural Network (GNN) to encode discourse structures and conversation features within a multi-task learning framework, enhancing the understanding of conversational context and structure.

Finally, Gao et al. (2021) focus on improving dialog systems' empathetic response generation by identifying emotion causes. Their framework combines an emotion reasoner for predicting emotion and its cause with a response generator that employs a gated attention mechanism to emphasize important words, exploring both hard and soft gating strategies.

3 System Overview

Our pipeline consists of two stages. Specifically, to identify emotion-cause pairs and emotion types, dialogs are passed through the following modules:

1. classification of utterances with emotion types (including *neutral* for non-emotional utterances) with a supervised fine-tuned LLM; and

2. extraction of cause utterances with a BiLSTM-based network.

The full pipeline is shown in Figure 1. Due to the limitations of the data, we perform the tasks separately, and we elaborate on each of the stages in the following sections.

3.1 Emotion classification

To categorize an utterance with an emotion label E_t, within our pipeline an LLM should consider both the target utterance U_t, which is the t^{th} utterance in a conversation, and the preceding utterance U_{t-1}. It is especially important when we deal with very short turns, such as "Instead of... ?", "No.", "Yeah, maybe...". Indeed, it would be more accurate to utilize causal utterances rather than antecedent ones; however, at the initial stage, these are unknown to us, necessitating the use of a meaningful alternative.

For this stage, we fine-tune `GPT-3.5`.[3] As a `system`'s input, we provide the prompt consisting of an instruction, U_{t-1} (<UTT_1>), and U_t (<UTT_2>) as is shown in Figure 2. This particular prompt was selected during the preliminary prompt engineering stage. The `assistant`'s output consists of one word – the emotion type.

We also note that preliminary experiments showed that the LLM performed poorly in zero-shot and few-shot settings on the emotion detection task, at least on the ECF dataset (see Section 5.1 and Table 2). Therefore, we had to fine-tune it.

3.2 Cause extraction

The second stage is concerned with the detection of the causal utterances for non-emotional utterances in a binary way. Let the whole conversational history of an utterance U_t be $H(U_t) = [U_1, U_t]$, then the set of all causal utterances is $C(U_t) \subseteq H(U_t)$.

[3] `gpt-3.5-turbo-1106`: https://platform.openai.com/docs/models/gpt-3-5-turbo

```
Take a deep breath. Your task: given two
dialog utterances, predict an emotion of
the second utterance. Select the emotion
from the following options: neutral,
anger, disgust, fear, joy, sadness,
surprise. Do not use any other
emotions!!! Respond only with the chosen
emotion, without any additional
explanation. Remember that you can only
use listed emotions!!!

Utterance 1: <UTT_1>
Utterance 2: <UTT_2>

Emotion:
```

Figure 2: The prompt used to perform emotion classification with GPT-3.5.

In addition, speakers are encoded with one-hot vectors $S_1...S_n$ within each dialog.

First, we need to enrich utterance embeddings $U_1...U_n$[4] obtained from a pre-trained model with the context within the conversation. Bidirectional LSTM (Hochreiter and Schmidhuber, 1997) was chosen because it can preserve context information in sequential settings using the content of the previous hidden state in encoding the current one. This way, we get new utterance representations $\bar{U}_1...\bar{U}_n$.

Then, for each target utterance U_t with $E_t \neq neutral$, we construct t representations:

$$\bar{U}_i \parallel S_i \parallel \bar{U}_t \parallel S_t \parallel E_t, \forall i \in [1, ..., t] \quad (1)$$

containing one of the previous utterances or the target utterance embedding itself $\bar{U}_i$ as a potential cause, its speaker vector S_i, the target utterance embedding $\bar{U}_t$, its speaker vector S_t, and the emotion label E_t. We pass them to a feed-forward neural network and obtain binary predictions $\{0, 1\}$, where 1 means that U_i is a causal utterance and 0 stands for the opposite. All U_i for which 1 is predicted make up $C(U_t)$. Thus, for each U_t with $E_t \neq neutral$ we obtain from 0 to t labelled emotion-cause pairs (U_i, U_t, E_t), where $U_i \in C(U_t)$, consisting of the causal utterance,[5] the emotion utterance, and the emotion label.

We have decided not to extract specific spans from the utterances classified as causes, following a thorough review of the dataset. This decision is based on our observation that these spans often defy straightforward explanations, even from a human annotator perspective. Here are some examples, where the rationale behind the spans remains unclear to us:

- The final punctuation marks are often not included in the cause span: e.g., while the complete utterance is *Instead of [...]?*, the identified cause span is *Instead of [...]*

- For the statement *Me, I ... I went for the watch*, the span is *I went for the watch*

- For the sentence *You know you probably did not know this, but back in high school, I had a, um, major crush on you*, the cause span is defined as *you probably did not know this, but back in high school, I had a, um, major crush on you*

We believe that this part of the task can be more accurately defined as a causal emotion entailment (Poria et al., 2021). Additionally, we note that there is an inconsistency in the dataset's annotation: specifically, the task organizers define emotion causes by identifying specific spans within an utterance, yet the emotional responses themselves are treated as consisting of entire utterances. For these reasons, we have decided that it would be methodologically more appropriate to omit the exact span detection step from our pipeline.

4 Experimental Setup

4.1 Data

The dataset proposed for the shared task contains conversations from the *Friends* series annotated with emotion-cause pairs and emotion labels, including *anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise* from Ekman et al. (1987), or *neutral* for non-emotional utterances.

The shared task organizers highlight that 91% of emotions have corresponding causes and one emotion may be triggered by multiple causes in different utterances. In addition, we have noticed that 16% of them cause several different emotions.

The organizers did not provide a standalone development set, so we had to split the training set ourselves using a ratio of 9:1 relative to the dialogs. The final data splits are shown in Table 1.

[4] We use the same notation for utterances and their embeddings for simplification purposes.

[5] We did not extract causal spans and used the whole causal utterance in the evaluation.

Set	# dialogs	# utterances	# EC
Training	1,236	12,346	8,565
Development	138	1,273	799
Total	1,374	13,619	9,364

Table 1: Distribution of dialogs, utterances, and emotion-cause pairs ("EC") across the split sets.

4.2 Training and architecture details

We fine-tune `GPT-3.5` with the default hyperparameters recommended by OpenAI[6] using two epochs, which is the number automatically chosen by the platform.

The cause extractor model is initialized with mean pooling from the penultimate layer's hidden state of the pre-trained `bert-base-uncased`.[7] Our neural network consists of three BiLSTM layers, one hidden linear layer accompanied by batch normalization, and a ReLU activation function.

For training, we employ the Adam optimizer with the learning rate of $1e$-4, weight decay (L_2-norm regularization) of $1e$-5, and cross-entropy as the loss function. We train the model for 200 epochs using a batch size of 32.

As a framework for training and evaluation, we use `PyTorch`[8] (Paszke et al., 2019).

4.3 Evaulation measures

As proposed in the shared task, we apply the weighted average (w-avg.) F_1 score by emotion type for evaluation. The specific implementation of F_1 score for the ECPE task can be found in Xia and Ding (2019b). In this setting, an emotion-cause pair is considered as correctly predicted if the index of an emotion utterance, an emotion type, and the index of the cause utterance match the entry in the gold dataset. There are two strategies related to causal span detection: *strict* F_1 (the same span) and *proportional* F_1 (overlap).[9]

5 Results

Our final submission was evaluated on the test set and achieved the following results:

- w-avg. proportional F_1: 0.264;
- w-avg. strict F_1: 0.104.

[6]https://platform.openai.com/docs/api-reference/fine-tuning/create
[7]https://huggingface.co/bert-base-uncased
[8]https://pytorch.org
[9]For the details on the metrics, refer to https://github.com/NUSTM/SemEval-2024_ECAC/tree/main/CodaLab/evaluation.

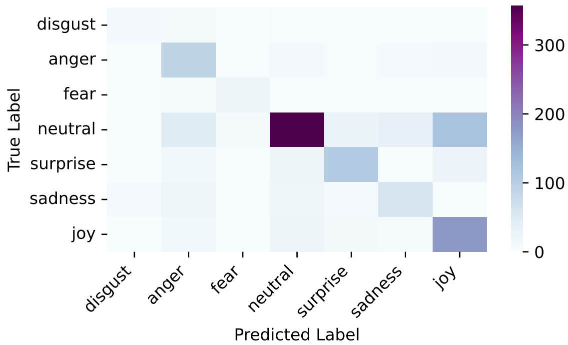


Figure 3: Performance of our emotion classifier on our development set.

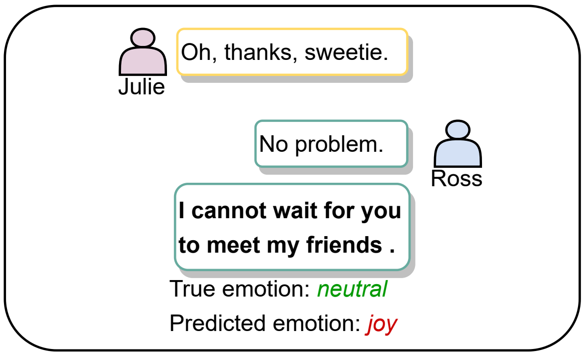


Figure 4: An example of a dialog where our model classified neutral utterance as *joy*.

As a result, we score second out of fifteen teams participating in Subtask 1 according to the main shared task metric – w-avg. proportional F_1.

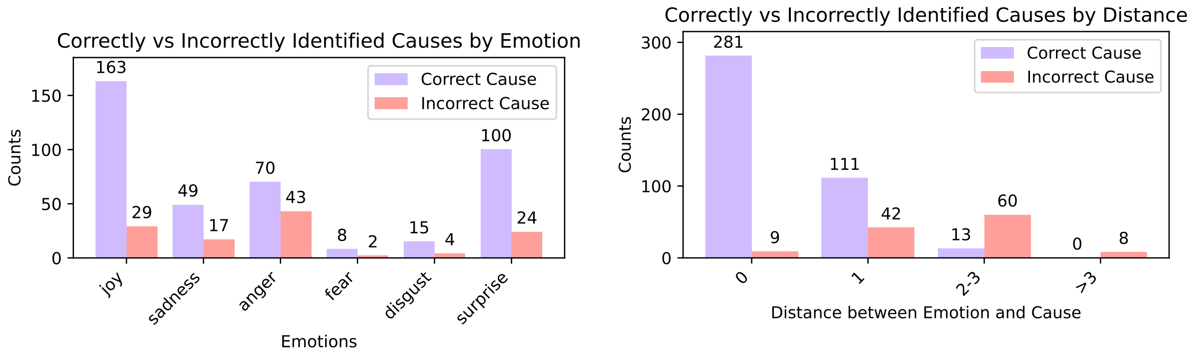
5.1 Emotion classification performance

Table 2 overviews the performance of emotion classification using `GPT-3.5` across different paradigms: zero-shot, few-shot, and fine-tuning. We note that zero- and few-shot settings use the same prompt (see Figure 2), with the few-shot setting including one handpicked example per each emotion type (see Appendix A). As expected, fine-tuning yields the best results on all emotion types and overall. Interestingly, few-shot prompting performs worse than zero-shot, which suggests that examples hamper the model's understanding of emotion types instead of improving it.

Utterances of *disgust* type turn out to be the most difficult to predict correctly: one of the possible reasons is that they are insufficiently represented in the training set (amounting to only about 6% of emotional utterances). However, the zero-shot and few-shot settings also show the poorest performance on *disgust*.

Approach	*neutral*	*anger*	*disgust*	*fear*	*joy*	*sadness*	*surprise*	macro	w-avg.
Zero-shot	0.61	0.43	0.30	0.32	0.54	0.47	0.50	0.45	0.54
Few-shot	0.57	0.49	0.31	0.34	0.54	0.37	0.41	0.43	0.51
Fine-tuning	**0.70**	**0.57**	**0.42**	**0.51**	**0.63**	**0.52**	**0.66**	**0.57**	**0.64**

Table 2: F_1 scores on emotion classification with GPT-3.5 across different approaches.



(a) Analysis across emotions on our development set (on correctly identified emotions only).

(b) Break-down of distance between emotion and cause on our development set (on correctly identified emotions only).

Figure 5: Performance of the cause extractor.

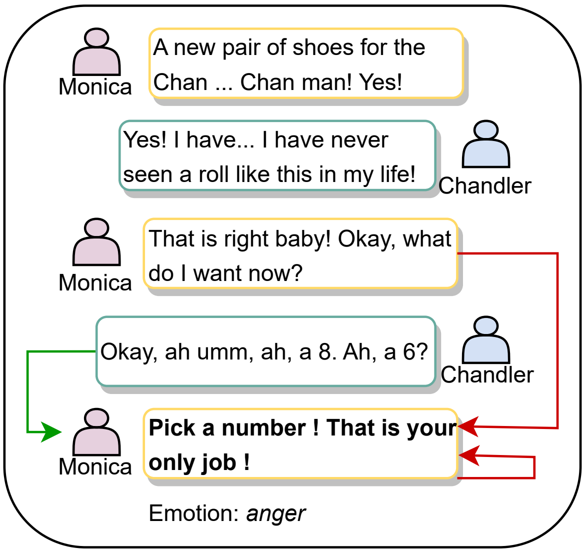


Figure 6: An example of a dialog with annotated causes for *anger* (green for causes correctly identified by our model, and red for causes that our model failed to recognize).

Our analysis of the emotion classifier's performance across different emotion types shows that the model often incorrectly classifies *neutral* utterances as indicative of *joy* (see Figure 3). After further investigation, we have found that a large number of these incorrectly categorized cases contain greetings ("*Hi!*") and expressions of gratitude ("*Thank you!*", "*You're welcome!*), which, according to our dataset, should be *neutral*, yet our classifier interprets them as *joy*. This implies that text alone may not be enough to identify an emotion, given that such utterances can express joy or remain emotionally neutral. There are other controversial cases, such as a conversation between two lovers shown in Figure 4, where the statement "*I cannot wait for you to meet my friends*" is actually more likely to express joy rather than neutrality.

5.2 Analysis

We also evaluate our model on its ability to identify the causes of utterances expressing different emotions, as shown in Figure 5a. Based on this analysis, the greatest challenge for our model is determining causes of *anger*. Similarly, manual analysis shows that this task is difficult for humans as well. As an example, Figure 6 highlights a scenario where the source of anger in Monica's utterance is not only attributed to the preceding utterance from Chandler but is also caused by the utterance that came before Chandler's, as well as the context of Monica's own statement. Intricacies like this one highlight the controversies present in the dataset.

Additionally, we have looked into how well our model performs in determining the emotion's cause based on how close it is to the emotional utterance, as we show in Figure 5b. First of all, it transpires that most emotional utterances are self-caused. Furthermore, our analysis shows that there is a clear correlation between the cause's distance from the emotional utterance and our model's identification accuracy: the further away the cause, the lower the model's performance.

In the course of our analysis, we have discovered instances where emotions appear before their causes. This observation suggests that the organizers' definition of a cause in dialog contexts is non-trivial, as, typically, we would expect that something happens and triggers an emotion. However, in the case of the preceding emotion, the cause is fundamentally different: it is a reason in terms of linguistics and it explains the emotion, but it does not trigger it (for the difference between CAUSE and REASON, please refer to Ruppenhofer et al., 2006).

Overall, accurate identification of emotions and their causes within utterances proves to be a complex challenge, not only for models but also for humans. All issues mentioned above point to important problems in the dataset that need to be carefully thought through and fixed to enhance both the accuracy and reliability of ECPE efforts.

6 Conclusions

Our work presents a novel approach to emotion-cause pair extraction in conversations, using the capabilities of an LLM (specifically, `GPT-3.5`) for emotion classification. This methodology is further enhanced by the use of a BiLSTM-based neural network for extracting causes. Our system outperforms most of the submissions to the shared task, scoring 2nd in the overall ranking according to the main metric of weighted-average proportional F_1. For future enhancements to our pipeline, we consider the following improvements:

- Firstly, data annotation itself can be expanded and improved, potentially via the use of an LLM for annotation.
- Secondly, speaker representations can be improved to enhance the understanding and processing of the dialogs.
- Finally, more accurate methods of LLM-based cause extraction can be developed further.

Limitations

Due to OpenAI's policy, we are unable to share our fine-tuned model. Therefore, those wishing to reproduce our experiments will need to do the fine-tuning independently. Overall, the usage of an open-source solution instead of a proprietary LLM can be one of the future directions. Also, it may be applied using a more specific framework like InstructERC (Lei et al., 2024).

Furthermore, our research is limited to the emotions present in the provided task data. Consequently, adding new emotions would require further fine-tuning. Due to the shared task rules, we have to develop our system based only on the presented dataset that is limited to a single concrete domain (*Friends* series) and the English language.

Acknowledgements

We are grateful to Mohamed bin Zayed University of Artificial Intelligence (MBZUAI) for supporting this work. We also thank the anonymous reviewers for their valuable feedback.

References

Hang Chen, Xinyu Yang, Jing Luo, and Wenjing Zhu. 2023. How to enhance causal discrimination of utterances: A case on affective reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 494–512, Singapore. Association for Computational Linguistics.

Paul Ekman, Wallace Friesen, Maureen O'Sullivan, A. Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William LeCompte, Tom Pitcairn, and Pio Ricci Bitti. 1987. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53:712–7.

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

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–80.

Dexin Kong, Nan Yu, Yun Yuan, Guohong Fu, and Chen Gong. 2022. Discourse-aware emotion cause extraction in conversations. *arXiv preprint arXiv:2210.14419*.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2024. InstructERC: Reforming emotion recognition in conversation with a retrieval multi-task LLMs framework.

OpenAI. 2023. GPT-4 technical report.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor

Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13:1317–1332.

Josef Ruppenhofer, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, California. Distributed with the FrameNet data.

Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. 2023. Multimodal emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing*, 14(3):1832–1844.

Fanfan Wang, Heqing Ma, Rui Xia, Jianfei Yu, and Erik Cambria. 2024. Semeval-2024 task 3: Multimodal emotion cause analysis in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2022–2033, Mexico City, Mexico. Association for Computational Linguistics.

Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Rui Xia and Zixiang Ding. 2019a. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012.

Rui Xia and Zixiang Ding. 2019b. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012, Florence, Italy. Association for Computational Linguistics.

Duzhen Zhang, Feilong Chen, and Xiuyi Chen. 2023. DualGATs: Dual graph attention networks for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7395–7408, Toronto, Canada. Association for Computational Linguistics.

A Prompt for Few-shot

```
Take a deep breath. Your task: given two dialog utterances, predict an emotion of the second
utterance. Select the emotion from the following options: neutral, anger, disgust, fear, joy,
sadness, surprise. Do not use any other emotions!!! Respond only with the chosen emotion,
without any additional explanation. Remember that you can only use listed emotions!!!

Examples:

Utterance 1: Alright , so I am back in high school , I am standing in the middle of the
cafeteria , and I realize I am totally naked .Utterance 2: Oh , yeah . Had that dream .
Emotion: neutral

Utterance 1: Do not you realise what you are ... you are doing to yourself ?
Utterance 2: Hey , you know , I have had it with you guys and your cancer and your emphysema
and your heart disease .
Emotion: anger

Utterance 1: Oh , hey , do not do that ! Cut it out !
Utterance 2: It is worse than the thumb !
Emotion: disgust

Utterance 1: I am not moving out .
Utterance 2: You would tell me if you were moving out right
Emotion: fear

Utterance 1: So , what do you think ?
Utterance 2: I think It is the most beautiful table I have ever seen .
Emotion: joy

Utterance 1: No , wait , oh , what are we sorry about ?
Utterance 2: I do not know ... right , he is the pig !
Emotion: sadness

Utterance 1: No , wait , oh , what are we sorry about ?
Utterance 2: How did I not see this ?
Emotion: surprise


Utterance 1: UTT_1
Utterance 2: UTT_2
Emotion:
```

Figure 7: The prompt used to perform emotion classification with GPT-3.5 in the few-shot setting.