UIR-ISC at SemEval-2024 Task 3: Textual Emotion-Cause Pair Extraction in Conversations

Hongyu Guo, Xueyao Zhang, Yiyang Chen, Lin Deng*, Binyang Li
Lab of Intelligent Social Computing
University of International Relations, Beijing, China
{chloe_guo,Zhang_X_Y,uiryangyc0114,denglin,byli}@uir.edu.cn

Abstract

The goal of Emotion Cause Pair Extraction (ECPE) is to explore the causes of emotion changes and what causes a certain emotion. This paper proposes a three-step learning approach for the task of Textual Emotion-Cause Pair Extraction in Conversations in SemEval-2024 Task 3, named ECSP. We firstly perform data preprocessing operations on the original dataset to construct negative samples. Secondly, we use a pre-trained model to construct token sequence representations with contextual information to obtain emotion prediction. Thirdly, we regard the textual emotion-cause pair extraction task as a machine reading comprehension task, and fine-tune two pre-trained models, RoBERTa and SpanBERT. Our results have achieved good results in the official rankings, ranking 3rd under the strict match with the *Strict F1-score* of 15.18%, which further shows that our system has a robust performance.

1 Introduction

Emotions are innate to humans and significantly affect people's social interactions, decision-making, and cognition. People are becoming more interested in developing human-like reactions as social media evolves. Therefore, the recognition of emotions in the text is an important topic in natural language processing and its applications (Zhao et al., 2016). In addition to emotion recognition, the research on the cause behind emotions in conversation scenarios is more complex, such as customer support, mental health care, human-computer interaction, etc (Wang et al., 2023b). Thus, it is important to recognize the potential cause behind an individual's emotional state, i.e., Emotion Cause Analysis (ECA). [1]

In recent research, Xia and Ding (2019) proposed the Emotion Cause Pair Extraction (ECPE) task, which is used to automatically predict emotions in documents and recognize the corresponding causes of those emotions. This task has attracted attention from a number of academics (Ding et al., 2020; Wei et al., 2020; Chen et al., 2020). However, the ECPE task studies the emotion-cause relationship of specific events in the document, while in the conversational scene, due to the interaction of multiple speakers, the dialogue contains more diverse and richer emotional expressions, which makes the conversation continue to advance as the conversation progresses. Emotions are also constantly changing, and the emotion of one utterance may be caused by multiple utterances.

In this paper, we propose a three-step learning approach, **E**motion-**C**ause-**S**pan **P**air Extraction in Conversation (**ECSP**), for Subtask 1 of SemEval-2024 Task 3: Textual Emotion-Cause Pair Extraction in Conversations. ECSP consists of three modules: the data preprocessing module, the emotion classification module, and the textual emotion-cause pair extraction module. We first preprocessed the dataset to obtain a large number of negative examples. Then, the pre-trained model BERT is used to construct token sequence representations with contextual information that are fed into a feed-forward neural network layer for emotion prediction. In the textual emotion-cause pair extraction module, in order to obtain causal span, we fine-tuned pre-trained models such as RoBERTa and SpanBERT to make it a machine reading comprehension (MRC) task (Poria et al., 2021).

In the official ranking, our team ranked 3rd under the strict match with the *Strict F1-score* of 15.18%, and ranked 7th under the Proportional match with the *Proportional F1-score* of 19.63%.

*Corresponding author

[1] Description of the task by the organizer of SemEval-2024 Task 3

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 770–776

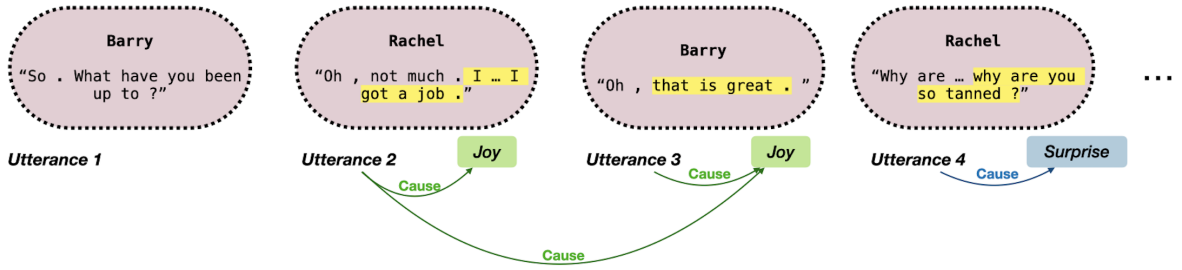

Figure 1: Description of the task of textual emotion-cause pair extraction in conversations.

2 Background

2.1 Task Definition

As shown in Figure 1, the task of textual emotion-cause pair extraction in conversations aims to extract all emotion-cause pairs in a given conversation based entirely on text and mark the specific causal span of the emotion cause (Wang et al., 2024).

Input: A conversation containing the speaker and the text of each utterance. Represented as the content in the pink rectangular box in Figure 1.

Output: All predicted emotion-cause pairs, where each pair contains an emotion utterance along with its emotion category and the textual cause span in a specific cause utterance. The utterance pointed by the curve to the emotion in the Figure 1 is the cause utterance of the emotion, and the yellow background text fragment is a specific textual cause span.

2.2 Related Work

Emotions always play a vital role in information exchange, from the communication between human individuals in the real world to the human-computer interaction in the virtual world. Recognizing emotion categories in text is an essential task in NLP and its applications (Zhao et al., 2016). In addition, the causes of emotions play a key role in human-computer interaction and customer service systems, which can provide important information on the reason for any emotion changes.

The aim of Emotion Cause Extraction (ECE) is to explore the causes of emotion changes and what causes a certain emotion (Chen et al., 2010). Xia and Ding (2019) reformed ECE into ECPE (Emotion-Cause Pair Extraction), aiming to extract potential emotions and corresponding causes from documents simultaneously.

Since ECPE does not fully consider the correlation between emotional utterances and causal utterances and the limited availability of background, Shan and Zhu (2020) proposed an Inter-EC model with self-attention, which optimized the interactive multi-task network model. Cheng et al. (2021) reconstructed the emotion-cause pair extraction task into the classification problem of candidate sentence pairs and proposed a goal-oriented, unified sequence-to-sequence model. Poria et al. (2021) constructed a dialogue-level dataset RECCON and introduced a task highly relevant for (explainable) emotion-aware to address causal span extraction and causal emotion entailment.

3 System Overview

In order to implement the task of textual emotion-cause pair extraction in conversations, we have designed the ECSP approach, which contains three main modules, namely data preprocessing, emotion classification, and textual emotion-cause pair extraction.

Firstly, in the data preprocessing module, the dataset is preprocessed to obtain a large number of negative samples. Then the pre-trained model BERT is used to convert token sequences with contextual information in the conversation into semantic representations and predict emotions in the emotion classification module. Finally, textual emotion-cause pairs are extracted based on the predicted emotions in the textual emotion-cause pair extraction module.

The overall architecture of ECSP system is shown in Figure 2, and the detailed description for each part is presented as follows.

3.1 Data Preprocessing Module

Since the original dataset only contains positive examples, i.e., utterances containing emotions, which are annotated using causal spans extracted from the historical context of the conversation, we designed the data preprocessing module to provide a large

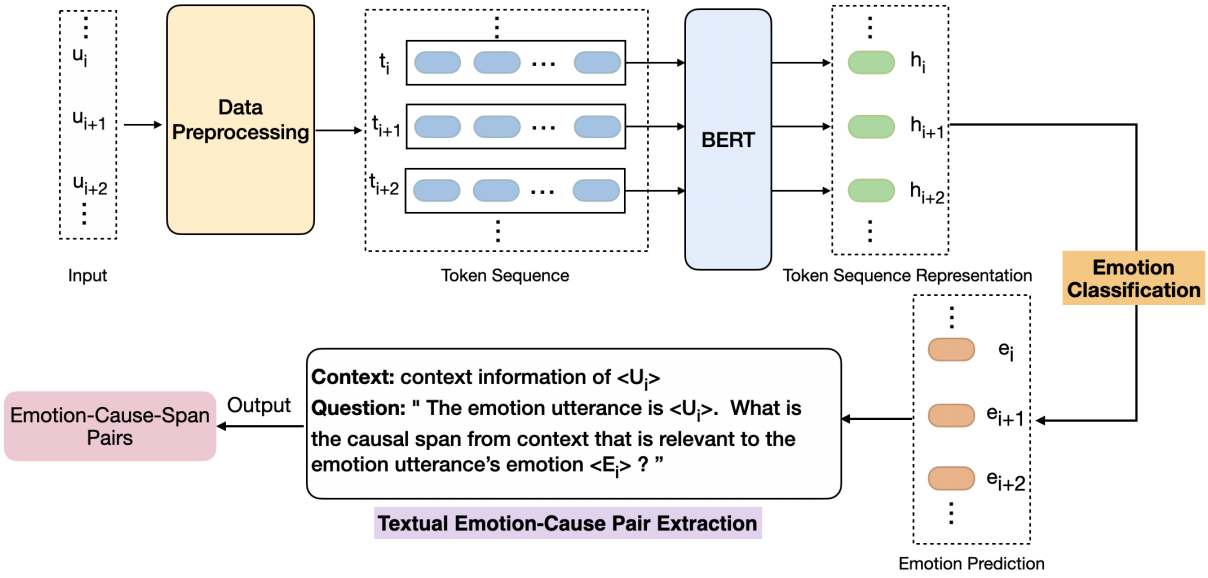

Figure 2: The overall architecture of ECSP consists of three parts: data preprocessing, emotion classification, and textual emotion-cause pair extraction. After preprocessing the origin dataset, BERT is utilized to transform the token sequence with contextual information in the conversation into a semantic representation and predict emotions. Then, extract textual emotion-cause pairs.

Dataset	Train	Val	Test
Positive Samples	7093	900	900
Negative Samples	36778	4247	4247

Table 1: Statistics of the preprocessed dataset, including positive and negative samples.

number of negative examples in which the cause is not expressed in order to better train the model to recognize emotional causes in conversation tasks.

Considering dialogue D and an emotion utterance U_i in D, we construct the complete set of negative examples as $\{U_{Neg}|U_{Neg} \in H(U_i) \setminus C(U_i)\}$, where $H(U_i)$ is the conversational history and $C(U_i)$ is the set of cause utterances for U_i.

Table 1 shows the statistics of the preprocessed dataset.

3.2 Emotion Classification Module

Without the loss of generality, the input can be represented by several utterances, $D = \{U_1, ..., U_i, ...U_n\}$. In our system, BERT is used to build the token sequence representations. Each token sequence is enveloped by predefined special tokens ($[CLS]$, $[SEP]$), $t_i^{'} = \{[CLS], w_{i1}, ..., w_{ik}, [SEP]\}$, where w_{ik} is the k-th token in the i-th utterance's token sequence. The [CLS] token is used for generating representations for classification tasks. The [SEP] token is used to denote the end of a sentence. The utterance's representation h_i is acquired through BERT, which is the final hidden state of [CLS].

$$h_i = BERT(t_i^{'}) \tag{1}$$

The token sequence representation h_i is fed into the Feed-Forward Neural Network (FFNN) layer to obtain the emotion prediction E_i.

$$E_i = Softmax(W^e h_i + b^e) \tag{2}$$

where W^e is a weight and b^e is a bias of the emotion classification layer, respectively.

3.3 Textual Emotion-Cause Pair Extraction Module

In order to implement the extraction of textual span in the ECPE task, we regard this module as a machine reading comprehension (MRC) task. The specific task is defined as follows:

Context: *Context* is the context information $U_j (j \in [1, i])$ of emotion utterance U_i, which is the traversal of all utterances in U_i's conversation history $H(U_i)$.

Question: The *Question* is framed as follows: *" The emotion utterance is $< U_i >$. What is the causal span from the context that causes the emotion $< E_i >$ of the emotion utterance? "*

Answer: The causal span $S \in CS(U_i)$ appearing in U_j if $U_j \in C(U_i)$. For negative examples, S is assigned an empty string.

Among them, emotion utterance U_i is the i-th utterance in dialogue D. $H(Ui)$ is the conversation history set of U_i, a set of all utterances from the beginning of the conversation till the utterance U_i, including U_i. $U_j \in H(U_i)$ is the context of U_i. $C(U_i)$ is the set of cause utterances of U_i, $C(U_i) \in H(U_i)$. $CS(U_i)$ is the cause span set of U_i.

3.4 Loss Function

Loss function is used to evaluate the extent to which the predicted and true values of the model are not the same. For different models and different tasks, the choice of loss function has a great impact on the performance of the model. In this task, the focal loss function is used to better alleviate the problem of unbalanced number of sample categories.

The goal of Focal Loss (Lin et al., 2017) is to address the issue where traditional cross-entropy loss contributes less to the loss of positive samples when there are a large number of easily classified negative samples. The adoption of the focal loss alleviates this issue by balancing the weight assigned to minority classes, facilitating the learning process (Wang et al., 2022).

$$BCEloss(o,t) = -\frac{1}{n}\sum_{i}\Big(t[i]\log(o[i]) + (1-t[i])\log(1-o[i])\Big) \quad (3)$$

As shown in formula 3, we use balance factor to deal with data imbalance in Balance Cross Entropy loss(BCEloss).

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \quad (4)$$

Focal loss reduces the loss weight of easily distinguishable negative samples and increases the dynamic adjustment factor based on BCEloss to achieve the effect of mining difficult samples. We make the model more focused on hard-to-learn samples by setting γ value as 2 in the formula 4, thus the network will not be biased by too many negative examples.

4 Experiments

4.1 Dataset

The SemEval-2024 Task 3 dataset is ECF (Wang et al., 2023a), which contains 1,344 conversations and 13,509 utterances. As shown in Table 2, 55.73% of utterances are labeled with emotion categories, 91.34% of emotions are labeled with corresponding cause, and the same emotion may be caused by multiple cause utterances (the number of emotion-cause pairs is greater than the number of emotion with cause).

Filed	Number
No. of conversations	1,344
No. of utterances	13,509
No. of emotion (utterances)	7,528
No. of emotion (utterances) with cause	6,876
No. of emotion-cause (utterance) pairs	9,272

Table 2: Statistics of ECF dataset.

For each emotion category, the proportion of emotion utterances with reason annotations is shown in Figure 3.

We split the original dataset into 80% train set, 10% valid set, and 10% test set.

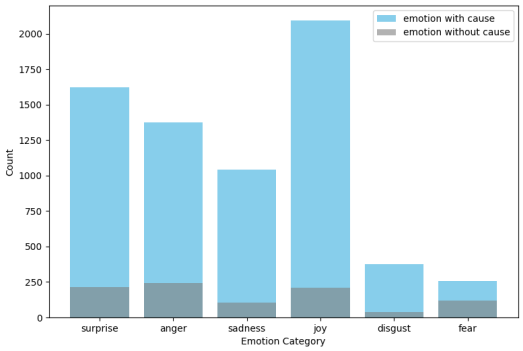

Figure 3: The distribution of emotions (with/ without cause) in different categories.

4.2 Baselines

In our experimental setup, we assume that emotion-cause pairs have two settings:

- Only non-neutral emotional utterances are recognized.
- The cause of emotion only exists in previous or current utterances because speakers cannot predict future utterances in conversational scenarios.

As to the emotion classification module, we used the pre-trained model BERT to obtain the semantic embedding of the input utterance.

BERT: BERT is a deep pre-trained language model based on the Transformer architecture. Devlin et al. (2018) used the Masked Language Model (MLM) to learn rich language representations and achieve SOTA performance in various downstream

Model		Strict			Proportional		
		P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
w/o context	RoBERTa	16.30	12.19	13.57	21.17	17.49	18.42
	SpanBERT	15.03	13.92	13.72	19.33	20.33	18.71
with context	RoBERTa	**18.35**	12.63	14.63	**22.34**	17.51	19.06
	SpanBERT	17.56	**14.41**	**15.18** (3/16)	20.94	**20.20**	**19.63** (7/16)

Table 3: Experimental results of textual emotion-cause pair extraction task. Shown in ( ) is the official ranking.

tasks. In the emotion classification task, we added contextual information to each utterance such that each utterance contains all its previous utterances as context, then used the BERT tokenizer to generate the input tensor of the emotion classification model, encoded it by BERT, and used a linear layer to predict emotions.

As to the textual emotion-cause pair extraction module, we fine-tuned two pre-trained models: RoBERTa and SpanBERT.

RoBERTa: RoBERTa (Liu et al., 2019) is an improved version of the BERT model, adopting more model parameters, more training data, and larger batch sizes. We used a Roberta-base model and added a linear layer on top of the hidden state to calculate the start and end logic of the span.

SpanBERT: SpanBERT (Joshi et al., 2020) is based on BERT, has made specific optimizations in the pre-training stage for the task of predicting spans of text, and has excellent performance in question and answer tasks. We used the SpanBERT-base model fine-tuned on the SQuAD 2.0 dataset as the second baseline model for the textual emotion-cause pair extraction task.

We utilized the PyTorch library (Paszke et al., 2019) and the HuggingFace library (Wolf et al., 2020) on our models and trained and tested them on the Nvidia A800-40G.

4.3 Evaluation Metrics

Since the task of textual emotion-cause pair extraction involves the textual cause span, the organizers of SemEval-2024 Task 3[2] adopted two strategies to determine whether the span is extracted correctly (Wang et al., 2024):

- **Strict Match**: The predicted span should be exactly the same as the annotated span.
- **Proportional Match**: Considering the overlap proportion between the predict span and the annotated one.

[2]https://github.com/NUSTM/SemEval-2024_ECAC

For the **Strict Match**, we firstly evaluate the emotion-cause pairs of each emotion category separately and then further calculate a weighted average of Strict F1-scores across the six emotion categories.

$$\text{Strict}F1 = \sum_{j=1}^{6} w^j \text{Strict}F1^j \tag{5}$$

Where w_j denotes the proportion of the annotated pairs with emotion category j, $j \in \{anger, disgust, fear, joy, sadness, surprise\}$.

For the **Proportional Match**, match each predicted pair with one of the annotated pairs that has the maximum overlap proportion in terms of the cause span (if the predicted span overlaps with multiple annotated spans):

$$overlap_i = \begin{cases} len(ps_i \cap as_k) & [eu_i, ec_i, cu_i] \\ & \text{are correct and} \\ & ps_i \cap as_k \neq \phi, \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

$$k = \arg\max_t \frac{len(ps_i \cap as_t)}{len(as_t)} \tag{7}$$

where $len(*)$ denotes the number of textual tokens, ps_i and as_k represent the cause span in the predicted pair pp_i and the annotated pair ap_k respectively. Then the proportional F1-score is calculated based on the overlap length between the predicted span and the annotated span, and a weighted average of the six emotion categories is also calculated.

$$\text{Proportional}F1 = \sum_{j=1}^{6} w^j \text{Proportional}F1^j \tag{8}$$

In the SemEval-2024 Task 3, the organizers initially selected the *Strict F1-score* as the main ranking metric. Due to poor overall results, they eventually switched to using the *Proportional F1-score* as

the main ranking indicator. This also shows that it is very difficult to extract the accurate textual cause span of emotion utterances.

4.4 Results

The experimental results of our work are given in Table 3. As shown in the table, we conducted experiments based on whether to add contextual information to the emotion classification module and gave the performance of two baseline models for the textual emotion-cause extraction task under strict match and proportional match, respectively. Among them, the SpanBERT model using the contextual information emotion prediction module achieved the best performance, with the *Strict F1-score* of 15.18% and the *Proportional F1-score* of 19.63%.

In addition, we draw the following observations:

- Firstly, the context of whole dialogue is curcial for the prediction of causal spans. When contextual information is added to the input utterances in the emotion classification module, the overall performance of the model will be improved to a certain extent. In the RoBERTa model, after adding contextual information, the *Proportional F1-score* increased by 1.36%, and the *Strict F1-score* increased by 1.06%. In the SpanBERT model, the *Proportional F1-score* increased by 0.92%, and the *Strict F1-score* increased by 1.46%.

- Secondly, it can be seen from the experimental results that the SpanBERT model always achieved good performance compared with the RoBERTa model in the textual emotion-cause pair extraction task. When there is no context information in the emotion extraction module, the *Strict F1-score* of the SpanBERT model is 0.12% higher than the RoBERTa model, and the *Proportional F1-score* is 0.29% higher. When there is context information in the emotion extraction module, the *Strict F1-score* of the SpanBERT model is 0.55% higher than the RoBERTa model, and the *Proportional F1-score* is 0.57% higher.

In the official ranking, our team used the three-step learning approach ECSP, which consists of an emotion classification module with contextual information and a textual emotion-cause pair extraction module with SpanBERT as the baseline. The ranking obtained is shown in Table 3. Among them, our team ranked 3rd under the strict match with the *Strict F1-score* of 15.18%, and ranked 7th under the Proportional match with the *Proportional F1-score* of 19.63%.

5 Conclusion

In this paper, we introduce the system implementation of SemEval-2024 Task 3: Textual Emotion-Cause Pair Extraction in Conversations. We propose an integrated system named Emotion-Cause-Span Pair Extraction in Conversation (ECSP), which was implemented in three modules: preprocessing data, emotion classification with contextual information input, and textual emotion-cause pair extraction, and it performed well in the official rankings. In the future, we will utilize this dataset to investigate if the *Speaker* attribute affects the extraction task of emotion-cause pairs, as well as to implement methods such as external knowledge bases to improve our system's recognition performance on ECPE tasks.

Acknowledgment

This paper was partially supported by National Natural Science Foundation of China (Grant number: 61976066), Beijing Natural Science Foundation (Grant number: 4212031), and Research Funds for NSD Construction, University of International Relations (Grant numbers: 2021GA07).

References

Ying Chen, Wenjun Hou, Shoushan Li, Caicong Wu, and Xiaoqiang Zhang. 2020. End-to-end emotion-cause pair extraction with graph convolutional network. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 198–207.

Ying Chen, Sophia Yat Mei Lee, Shoushan Li, and Chu-Ren Huang. 2010. Emotion cause detection with linguistic constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 179–187.

Zifeng Cheng, Zhiwei Jiang, Yafeng Yin, Na Li, and Qing Gu. 2021. A unified target-oriented sequence-to-sequence model for emotion-cause pair extraction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2779–2791.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020. Ecpe-2d: Emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3161–3170.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the association for computational linguistics*, 8:64–77.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

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*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, Alexander Gelbukh, and Rada Mihalcea. 2021. Recognizing emotion cause in conversations.

Jingzhe Shan and Min Zhu. 2020. A new component of interactive multi-task network model for emotion-cause pair extraction. In *Journal of Physics: Conference Series*, volume 1693, page 012022. IOP Publishing.

Cheng Wang, Jorge Balazs, György Szarvas, Patrick Ernst, Lahari Poddar, and Pavel Danchenko. 2022. Calibrating imbalanced classifiers with focal loss: An empirical study. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 145–153.

Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. 2023a. 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.

Fanfan Wang, Jianfei Yu, and Rui Xia. 2023b. Generative emotion cause triplet extraction in conversations with commonsense knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3952–3963.

Penghui Wei, Jiahao Zhao, and Wenji Mao. 2020. Effective inter-clause modeling for end-to-end emotion-cause pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3171–3181.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Rui Xia and Zixiang Ding. 2019. 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.

Jun Zhao, Kang Liu, and Liheng Xu. 2016. Sentiment analysis: mining opinions, sentiments, and emotions.