Multimodal Emotion Cause Analysis in Conversations

ANURANJAN PANDEY, KRITI JHA, and SWETA SINGH, University of Southern California, USA

This research introduces a computational framework for identifying and extracting the emotion-cause pairs from a conversation. While most research focuses solely on textual data to identify emotion-cause pairs, the complexity of human communication cannot be encapsulated by a single modality. The tone of voice and facial expressions also play a crucial part in conveying the emotion. We make use of the multimodal Emotion-Cause-in-Friednds (ECF) dataset which is derived from the television sitcom "Friends". We use various sophisticated encoding techniques, fusion strategies, and deep learning-based classification models for emotion cause pair prediction. Our results demonstrate the potential of this framework in enhancing emotional intelligence in AI systems and improving human-computer interaction by accurately predicting emotion-cause pairs.

Additional Key Words and Phrases: Multimodal Emotion Analysis, Emotion-Cause Extraction, Conversational AI, Emotional Intelligence, Human-Computer Interaction, Natural Language Processing, Audio-Visual Data Analysis, Deep Learning, Sentiment Analysis, Computational Emotion Recognition

1 PROBLEM DEFINITION

The goal of this project is to create a computational framework that can effectively identify and extract pairs of emotions and their causes from conversations. What makes this approach unique is that it uses a multimodal approach that includes textual, acoustic, and visual data. This approach recognizes the complex nature of human communication, where emotions are expressed and perceived not only through words but also through tone of voice and facial expressions. By combining these three modalities, the proposed framework aims to capture a more nuanced and complete understanding of emotional expressions and their underlying causes in conversations.

The framework takes a conversation as input, which includes the speaker's identity, the textual content, and an accompanying audio-visual clip for each utterance. This input format enables the system to analyze conversations in a comprehensive manner, allowing it to consider the verbal content, vocal characteristics, and visual cues. Each of these elements has a crucial role in conveying and interpreting emotional states and intentions.

The aim of the framework is to produce a set of emotion-cause pairs that are derived from a conversation. Each pair in this set comprises an emotion utterance, which is an utterance that expresses emotion and is classified into one of the predetermined emotion categories, along with single or multiple-cause utterances that are identified as the trigger or reason for the expressed emotion. This format of output shows how conversations unfold and what triggers emotional responses, giving insight into the emotional tone and interactions.

The ultimate goal of this project is to enhance the ability of AI to interact with humans in a more empathetic and contextually aware way. This would aid in improving the comprehension of emotional expression within conversational contexts, which could be utilized to better emotional AI, human-computer interaction, and psychological research.

2 LITERATURE REVIEW

Emotion cause pair prediction has been explored through various techniques, from multi-task learning frameworks to knowledge-enhanced models. This section reviews the relevant literature, highlighting the key contributions and approaches.

Several studies have employed multi-task learning frameworks to leverage the interdependencies between emotion and cause detection. [Bhat and Modi 2023] proposed MuTECCSE, an end-to-end architecture that performs cause span extraction as the main task and emotion prediction as an auxiliary task, using a pre-trained transformer encoder and multi-sample dropout. They also introduced MuTECCEE, a multi-task model that learns contextual representations, models cause-emotion utterance relationships using BiLSTMs, and performs entailment prediction along with emotion prediction. [Li et al. 2023] presented Joint-GCN and Joint-Xatt frameworks that jointly detect emotion and cause utterances using graph convolutional networks and cross-attention as information-sharing mechanisms, respectively. Their Joint-EC model first detects emotional and cause utterances, then extracts emotion-cause pairs using an auxiliary EC-chunk extraction task.

Recognizing the importance of conversational context, [Lee et al. 2023] proposed CPRG-MoE, a context-aware model that incorporates speaker information and conversational context using special tokens. It employs a mixture-of-experts approach, with each expert specializing in a specific pairwise relation category based on speaker and emotion relations. Similarly, [Jeong and Bak 2023] introduced PRG-MoE, a pair-relationship guided mixture-of-experts model that routes pair candidates to experts based on their relationship types, enabling each expert to learn patterns specific to that relationship.

Incorporating commonsense knowledge has been explored to enhance emotion cause pair prediction. [Zhao et al. 2022b] proposed KBCIN, a Knowledge-Bridged Causal Interaction Network that leverages commonsense knowledge from ATOMIC-2020 to bridge the reasoning gap between utterances and target emotions. It introduces semantics-level, emotion-level, and action-level knowledge bridges to capture semantic dependencies, emotional interactions, and action tendencies between utterances. [Wang et al. 2023] introduced SHARK, a generative framework for Emotion Cause Triplet Extraction using a BART-based index generation approach with a dual-view gate mechanism to integrate commonsense knowledge for enhancing utterance representations.

Recognizing the multimodal nature of conversations, few studies have explored the fusion of textual, audio, and visual modalities. [Wang et al. 2021] introduced the task of Multimodal Emotion-Cause Pair Extraction (MC-ECPE) and constructed the Emotion-Cause-in-Friends dataset, incorporating multimodal features like BiLSTM for text, 3D-CNN for visuals, and openSMILE for audio. [Chudasama et al. 2022] proposed M2FNet, a multi-modal fusion network for Emotion Recognition in Conversations, using a multi-head attention-based fusion mechanism to combine emotion-relevant features from different modalities.

Several other approaches have been explored for emotion cause pair prediction. [Li et al. 2019] introduced a multi-attention-based neural network model that considers contextual information surrounding emotional words and interactions between emotion expressing clauses and candidate clauses. [Zhao et al. 2022a] proposed CauAIN, a Causal Aware Interaction Network that leverages commonsense knowledge to detect emotion causes by distinguishing between intra-cause and inter-cause utterances. [Poria et al. 2020] introduced RECCON, a novel dataset for recognizing emotion causes in conversations, and employed transformer-based models for identifying these causes.

[Yoo and Jeong 2023] proposed a token classification-based attention model that simultaneously extracts multiple emotion-cause pairs from conversations using the BIO tagging scheme and a pretrained language model. [Singh et al. 2023] introduced DeCoDE, a multi-modal, multi-task framework that leverages textual, acoustic, visual modalities, and external knowledge to detect cognitive distortions and extract emotion causes in clinical conversations.

[Khunteta and Singh 2023] constructed an enhanced English dataset and proposed a multi-task learning framework using BiLSTM and BiLSTM with attention mechanisms for classifying and filtering emotion-cause pairs without requiring prior emotion annotations. [Anand et al. 2023] introduced SeMuL-PCD, a self-supervised multi-label peer collaborative distillation learning framework that utilizes a multimodal transformer network for estimating multiple emotions simultaneously in conversations.

[Li et al. 2022] proposed a model that builds conversations as graphs, introducing social commonsense knowledge to enhance causal reasoning between utterances, especially when causal utterances have different emotions than the targeted utterance. [Li et al. 2021] focused on span-level emotion cause analysis, proposing sequence labeling and pointer network-based approaches to identify precise text spans conveying emotion causes.

3 DATA DESCRIPTION

We utilize the multimodal Emotion-Cause-in-Friends (ECF) dataset for the task of emotion recognition and emotion cause extraction in conversational videos. This dataset is derived from the popular television sitcom "Friends," known for its rich emotional content and diverse conversational scenarios.

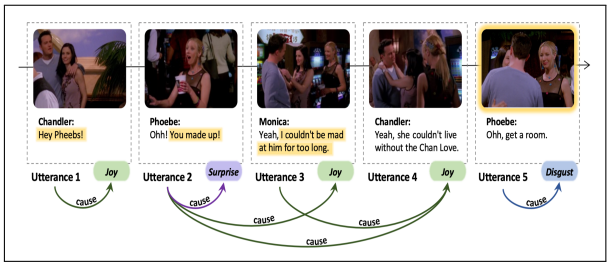

Fig. 1. An example of a conversation from the dataset [Wang et al. 2021]

Figure 1 shows an example of a conversation with multiple utterances, where each utterance has an emotion associated with it. Each arrow points from the causal utterance to the emotion it triggers; an emotion can have many causal utterances. For example, the emotion of Utterance 4 is Joy, and it has multiple cause utterances like Utterance 2 and 3.

3.1 Dataset Composition

Table 1 presents an overview of the ECF dataset which comprises a significant collection of data, including 1,344 conversations and 13,509 utterances, which span textual, audio, and visual modalities. These conversations reflect a wide range of emotions that occur in natural social interactions, making them an excellent source for analysis. The dataset includes annotations for 7,528 utterances that express emotion, indicative of the high emotive content within the dialogue of "Friends." Among these, 6,876 utterances have been annotated with associated causes, revealing the contextual triggers of emotional responses. Additionally, there are 9,272 emotion-cause pairs. The existence of a greater number of emotion-cause pairs than emotions with causes suggests that emotional triggers are multifaceted and can span across various utterances. To fully understand how conversational cues affect emotional reactions, we need to use sophisticated analytical methods. This is because there are many factors at play and many different emotions involved. It's important to consider all of these factors when trying to decode the relationship between them.

Table 1. Basic Statistics of the Dataset

Items	Number
Conversations	1,344
Utterances	13,509
Emotion (utterances)	7,528
Emotion (utterances) with cause	6,876
Emotion-cause (utterance) pairs	9,272

3.2 Emotion Annotations

The annotation process identified six basic emotions, following the theory of basic emotions: joy, surprise, anger, sadness, disgust, and fear. Each utterance in the dataset has been carefully analyzed and 55.73% of the utterances are annotated with one of the six basic emotions. Further, 91.34% of the annotated utterances are linked with their corresponding causes. This high percentage underscores the depth and complexity of the dataset, making it an excellent resource for training and evaluating computational models for emotion recognition and cause extraction. Figure 2 illustrates the distribution of emotions across the different categories, with joy being the most frequently occurring emotion, aligning with the sitcom's comedic genre.

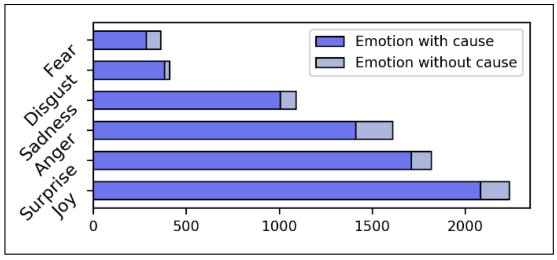

Fig. 2. Emotion Distribution across different categories. [Wang et al. 2021]

4 METHOD

Our approach for extracting emotion-cause pairs from multimodal data incorporates an ensemble of encoding techniques, fusion strategies, and deep learning-based classification models; for a thorough understanding of emotions and their triggers.

4.1 Encoding Techniques

We employ various pre-trained models for distinct modalities to perform feature extraction.

The MARLIN encoder [Cai et al. 2023], pre-trained on a large video dataset, is utilized for video data to capture visual cues crucial for identifying emotional states.

The BERT encoder [Devlin et al. 2019], pre-trained on a vast corpus of textual data for the task of emotion detection, processes textual data to understand semantic and contextual elements. For the textual data, we concatenated the utterance text with the name of the speaker as "speaker: text".

The HuBERT encoder [Hsu et al. 2021], pre-trained on diverse audio samples for the task of emotion detection, is employed to analyze audio for prosodic features and tonal variations that signify emotional states.

4.2 Creation of Emotion-Cause Pairs

We generate emotion-cause pairs by considering all utterances preceding and including the current one within a conversation. Utterances annotated with 'neutral' emotion are purposefully omitted, streamlining the focus on emotionally salient interactions.

4.3 Emotion Detection with Fusion Techniques

The initial stage in our methodology is emotion detection, which is important for the subsequent pairing process. We make use of the following two techniques in order to model the conversation

4.3.1 MLP. To perform this, we employ a multi-layer perceptron (MLP) that leverages both early and late fusion techniques. Early fusion involves the concatenation of encoded video, text, and audio features, providing the MLP with a comprehensive feature set for emotion classification. Conversely, late fusion entails integrating the discrete outputs of modality-specific classifiers, capitalizing on the distinct advantages of each modality. This dual fusion approach in emotion detection ensures that all aspects of the multimodal data are effectively utilized.

4.3.2 Transformer. While the MLP focuses on only classifying the current utterance into an emotion class, we used transformer [Vaswani et al. 2023] to use all the previous utterances in a conversation to make a prediction for the current utterance. This transformer-based approach enables us to effectively model the long-term dependencies and joint reasoning across different modalities. With transformers also, we use early fusion and late fusion. In early fusion, we train the model on the concatenated embeddings of the three modalities; in case of late fusion, we combine the predictions made by these modality-specific models by taking the average of the output probabilities.

Our EmotionDetector model employs a transformer encoder, which processes multi-modal embeddings enhanced with positional encodings to maintain sequential context. This setup allows the encoder to effectively capture the dynamics within the conversation. The processed embeddings are then used to generate logits for various emotion classes through a linear transformation layer.

During the training phase, we focus on mitigating the effects of class imbalance by employing a weighted loss approach, which prioritizes learning from underrepresented emotion classes. This strategy ensures that our model performs robustly across diverse emotional states, enhancing its applicability in real-world scenarios where emotional cues are crucial.

4.4 Binary Classifier for Emotion Pair Prediction

We paired all non-neutral utterances with the previous utterances of in the conversation and used them for binary classification. We discovered that only utterances with non-neutral emotions had a causal pair. Since an utterance can have multiple causes, we decided to use binary classification for our analysis.

4.4.1 MLP. Following the emotion detection phase, we employ a binary classifier using a Multilayer Perceptron (MLP) for predicting emotion-cause pairs, incorporating both early and late fusion techniques. In early fusion, we create a feature vector from the concatenated embeddings of the emotion utterance, the prospective cause utterance, and the one-hot encoded emotion vector. Here the embeddings for the emotion utterance as well as the prospective cause utterance are the concatenated audio, text, and visual embeddings. This comprehensive feature vector is then input to the MLP for classification. For late fusion, we deploy three separate MLPs, each dedicated to one of the modalities (audio, text, and visual). Each MLP processes its modality-specific data to predict emotion-cause pairs independently.

4.4.2 Transformer. For emotion-cause pair prediction, our approach utilizes both early and late fusion techniques, akin to methodologies employed in MLP architectures. We developed the PairDetector model to determine the validity of emotion-cause pairs from given pairs of utterances using a transformer encoder architecture enhanced with attention mechanisms.

Importantly, our model combines multi-modal input embeddings with one-hot encoded emotion predictions from the preliminary emotion classification model. This concatenated input is then processed through the transformer encoder, enabling the model to

leverage both contextual cues from the multi-modal embeddings and specific emotion data for more accurate prediction.

The transformer encoder outputs are subsequently passed through a sigmoid function to generate logits. These logits are compared against a threshold, which is optimized using ROC curve analysis, to decide the binary classification—whether the utterance pair represents a valid emotion-cause relationship.

During the training phase of the PairDetector model, we focused on addressing the class imbalance issue—where non-emotion-cause pairs (negative samples) are more prevalent—by applying a weighted binary cross-entropy loss. Positive samples (valid pairs) receive a higher weight of 4.2, ensuring that the model adequately learns from these less frequent but critical instances.

4.5 Combined Accuracy

Our approach involves a two-step process of determining the emotion in a given utterance and then predicting the cause behind that emotion. To evaluate the overall effectiveness, we calculate the combined accuracy metric. This metric assesses if our framework can accurately predict both the emotion and its corresponding cause. The combined accuracy evaluates the performance of the entire emotion-cause pair prediction pipeline, which consists of a multi-modal embedding model, an emotion prediction attention model, and a pair prediction attention model. It considers the emotion prediction from the emotion classification model as well as the subsequent cause utterance prediction from the pair prediction model.

To calculate the combined accuracy, we first identify the samples or utterance pairs where the predicted emotion for the effect utterance matches the ground truth emotion label. From this subset, we then evaluate whether the predicted cause utterance also matches the ground truth cause utterance label. Only the samples with exact matches for both the emotion and cause utterance predictions are considered correct for the combined accuracy metric.

Formally, let $\mathcal{D} = \{(x_i, y_i^e, y_i^c)\}_{i=1}^N$ be our dataset of N samples, where x_i is the input (e.g., multi-modal embeddings), y_i^e is the ground truth emotion label for the effect utterance, and y_i^c is the ground truth cause utterance label. Let $\hat{y}_i^e$ and $\hat{y}_i^c$ denote the predicted emotion and cause utterance labels respectively from our pipeline models.

The combined accuracy is then calculated as:

$$\text{Combined Accuracy} = \frac{1}{N}\sum_{i=1}^{N} F\{\hat{y}_i^e = y_i^e\} F\{\hat{y}_i^c = y_i^c\}$$

Where $F\{\cdot\}$ is the indicator function that evaluates to 1 when the condition inside the braces is true, and 0 otherwise.

In our experiments, we found the combined accuracy to be a robust metric for evaluating the end-to-end performance of our emotion-cause pair prediction pipeline. It allowed us to identify failure modes and areas for improvement in both the emotion prediction and pair prediction components of our system.

5 RESULTS

Our investigation focused on the effectiveness of early and late fusion techniques in emotion detection and cause pair prediction using the Multi-Layer Perceptron (MLP) and Transformer classifier. For emotion detection, the transformer architecture only trained on the text modality performed the best with 58.14% accuracy and 55.65% F1 Score. In the case of cause pair prediction, we found that MLP with late fusion performed the best with 88.23% accuracy, where as the F1 score was highest for early fusion on the MLP classifier with 87.79% F1 score.

We found that the transformer models performed poorly on the video and audio embeddings despite training it on complex architectures with up to 8 Attention heads, 7 layers, and Bi-directional transformers. Further investigation is required to find the root cause of this issue.

Task	Modality	Accuracy (%)	F1 Score (%)
Emotion Detection	Text	53.74	52.56
	Video	31.61	31.59
	Audio	45.95	43.62
	Late Fusion	54.29	50.37
	Early Fusion	**54.52**	**53.83**
Cause Pair Prediction	Text	85.30	84.69
	Video	85.89	83.86
	Audio	87.00	85.08
	Late Fusion	**88.23**	86.67
	Early Fusion	88.17	**87.79**

Table 2. Summary of Results for MLP

Task	Modality	Accuracy (%)	F1 Score (%)
Emotion Detection	Text	**58.14**	**55.65**
	Video	12.31	11.80
	Audio	12.47	11.00
	Late Fusion	54.09	45.99
	Early Fusion	45.83	49.54
Cause Pair Prediction	Text	**72.64**	**48.52**
	Video	10.64	39.43
	Audio	10.87	45.17
	Late Fusion	37.69	42.64
	Early Fusion	66.85	42.64
Combined Accuracy	Text	**28.89**	**47.69**
	Video	09.01	36.97
	Audio	10.93	41.96
	Late Fusion	14.31	34.60
	Early Fusion	27.81	42.64

Table 3. Summary of Results for Transformer Model

Text modality performed the best on the combined accuracy metric, this high score is due to transformers' abliltiy to model the sequences effectively and make predictions which are used later on combined with the cause pair classification.

These findings highlight the challenges and complexities inherent in multimodal emotion analysis, demonstrating the importance

of continued exploration into model architectures and fusion techniques to improve accuracy and understanding in emotion and cause pair prediction tasks.

6 CONCLUSION AND LESSONS LEARNED

We have developed a strong framework that uses multi-modal data analysis and advanced deep learning techniques to predict emotion-cause pairs in conversational settings. Our fusion approaches were able to capture the nuanced cues that convey emotional expressions and their underlying causes by integrating encoding strategies for text, audio, and video modalities.

We have developed transformer-based models, such as the EmotionDetector and PairDetector, which can predict emotions and identify valid emotion-cause pairs, respectively. The attention mechanisms used in these models allowed us to effectively model long-range dependencies and contextual information, which is crucial for decoding the relationships between utterances and their emotional triggers.

Although we encountered several challenges throughout our research, we gained valuable insights that will guide future explorations in this domain. For instance, we learned that the choice of modality and its corresponding encoding technique played a pivotal role in the performance of our models and that addressing class imbalance was important.

7 FUTURE WORK

We can explore several avenues to improve the models. First, we can make an attempt to fix the issues with the current transformed models as they are not able to learn anything for the Video and Audio Modality.

The integration of a Common Sense Knowledge Graph, specifically utilizing ATOMIC [Sap et al. 2019], can help improve the model's understanding of complex emotional cues and their causes within conversations. Knowledge graphs could provide contextual insights, helping in predicting emotional nuances.

Advancements in model architecture also offer promising directions for exploration. Experimenting with Cross-Attention Across Modalities and Graph Convolutional Networks (GCN) could lead to significant improvements in how different types of data—textual, acoustic, and visual—are integrated, allowing for a more nuanced analysis of emotion-cause relationships. Moreover, the exploration of Multimodal Models such as MuIT (Multimodal Transformer) [Tsai et al. 2019] and VATT (Video and Text Transformer) [Akbari et al. 2021] can further enhance the model's capability to integrate disparate modalities seamlessly, providing a more holistic understanding of the conversational context.

8 TEAM MEMBERS' CONTRIBUTION

In our project on emotion-cause pair extraction, each team member's unique contributions have collectively advanced our research.

Anuranjan's main focus was on generating video embeddings using MARLIN, concatenating embeddings, and preprocessing data. He developed a custom dataloader for emotion classification, where previous utterances were modeled as context for current utterances, using positional encoding. Anuranjan also implemented the transformer model architecture for emotion classification for individual modalities and late fusion. Lastly, he was in charge of writing the final report.

Kriti played a significant role in our project by working on the creation of textual embeddings using BERT, generating the pair embeddings, and developing the MLP-based late fusion model. She also implemented a custom dataloader, which included the get_item function and Dataset class for pair prediction. In this process, the predicted emotional context was concatenated with the utterance pair embeddings. Moreover, Kriti worked on implementing the transformer model architecture for individual modalities and late fusion for pair prediction. Additionally, she carried out hyperparameter tuning for the models on CARC.

Sweta has made significant contributions to our project by generating audio embeddings using HuBERT and developing an MLP-based early fusion model. Additionally, she has created an evaluation script for both emotion classification and pair prediction, implemented a custom collate method as part of the dataloader for pair prediction, and devised a combined accuracy metric for the final model evaluation.

ACKNOWLEDGMENTS

We would like to express our gratitude to Professor Mohammad Soleymani for his invaluable guidance on using feature extractors such as MARLIN, BERT, and HuBERT for video, text, and speech analysis, respectively. His insights on dataset utilization and problem-solving approaches have been extremely helpful. Additionally, TA Di Chang has been instrumental in helping us make the right modeling decisions.

REFERENCES

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text. arXiv:2104.11178 [cs.CV]

Sidharth Anand, Naresh Kumar Devulapally, Sreyasee Das Bhattacharjee, and Junsong Yuan. 2023. Multi-label Emotion Analysis in Conversation via Multimodal Knowledge Distillation. In *Proceedings of the 31st ACM International Conference on Multimedia*. ACM, Ottawa ON Canada, 6090–6100. https://doi.org/10.1145/3581783.3612517

Ashwani Bhat and Ashutosh Modi. 2023. Multi-Task Learning Framework for Extracting Emotion Cause Span and Entailment in Conversations. In *Proceedings of The 1st Transfer Learning for Natural Language Processing Workshop*. PMLR, 33–51. https://proceedings.mlr.press/v203/bhat23a.html ISSN: 2640-3498.

Zhixi Cai, Shreya Ghosh, Kalin Stefanov, Abhinav Dhall, Jianfei Cai, Hamid Rezatofighi, Reza Haffari, and Munawar Hayat. 2023. MARLIN: Masked Autoencoder for facial video Representation LearnINg. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 1493–1504. https://doi.org/10.1109/CVPR52729.2023.00150

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, New Orleans, LA, USA, 4651–4660. https://doi.org/10.1109/CVPRW56347.2022.00511

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. arXiv:2106.07447 [cs.CL]

DongJin Jeong and JinYeong Bak. 2023. Conversational Emotion-Cause Pair Extraction with Guided Mixture of Experts. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Andreas Vlachos and

Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 3288–3298. https://doi.org/10.18653/v1/2023.eacl-main.240

Arunima Khunteta and Pardeep Singh. 2023. Emotion Cause Pair Extraction By Multi Task Learning on Enhanced English Dataset. *Procedia Computer Science* 218, 766–777. https://doi.org/10.1016/j.procs.2023.01.057

Jaehyeok Lee, DongJin Jeong, and JinYeong Bak. 2023. Enhancing Emotion–Cause Pair Extraction in Conversation With Contextual Information. (Nov. 2023). https://doi.org/10.36227/techrxiv.24548494.v1

Jiangnan Li, Fandong Meng, Zheng Lin, Rui Liu, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. 2022. Neutral Utterances are Also Causes: Enhancing Conversational Causal Emotion Entailment with Social Commonsense Knowledge. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Vienna, Austria, 4209–4215. https://doi.org/10.24963/ijcai.2022/584

Wei Li, Yang Li, Vlad Pandelea, Mengshi Ge, Luyao Zhu, and Erik Cambria. 2023. ECPEC: Emotion-Cause Pair Extraction in Conversations. *IEEE Transactions on Affective Computing* 14, 3 (July 2023), 1754–1765. https://doi.org/10.1109/TAFFC.2022.3216551

Xiangju Li, Shi Feng, Daling Wang, and Yifei Zhang. 2019. Context-aware emotion cause analysis with multi-attention-based neural network. *Knowledge-Based Systems* 174, 205–218. https://doi.org/10.1016/j.knosys.2019.03.008

Xiangju Li, Wei Gao, Shi Feng, Yifei Zhang, and Daling Wang. 2021. Boundary Detection with BERT for Span-level Emotion Cause Analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, Online, 676–682. https://doi.org/10.18653/v1/2021.findings-acl.60

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. 2020. Recognizing Emotion Cause in Conversations. https://doi.org/10.48550/ARXIV.2012.11820 Publisher: arXiv Version Number: 4.

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. arXiv:1811.00146 [cs.CL]

Gopendra Vikram Singh, Soumitra Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. 2023. DeCoDE: Detection of Cognitive Distortion and Emotion Cause Extraction in Clinical Conversations. In *Advances in Information Retrieval*, Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo (Eds.), Vol. 13981. Springer Nature Switzerland, Cham, 156–171. https://doi.org/10.1007/978-3-031-28238-6_11 Series Title: Lecture Notes in Computer Science.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 6558–6569. https://doi.org/10.18653/v1/P19-1656

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. arXiv:1706.03762 [cs.CL]

Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. 2021. Multimodal Emotion-Cause Pair Extraction in Conversations. arXiv. http://arxiv.org/abs/2110.08020 arXiv:2110.08020 [cs].

Fanfan Wang, Jianfei Yu, and Rui Xia. 2023. Generative Emotion Cause Triplet Extraction in Conversations with Commonsense Knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 3952–3963. https://doi.org/10.18653/v1/2023.findings-emnlp.260

Soyeop Yoo and Okran Jeong. 2023. A Token Classification-Based Attention Model for Extracting Multiple Emotion–Cause Pairs in Conversations. *Sensors* 23, 6, 2983. https://doi.org/10.3390/s23062983

Weixiang Zhao, Yanyan Zhao, Zhuojun Li, and Bing Qin. 2022b. Knowledge-Bridged Causal Interaction Network for Causal Emotion Entailment. http://arxiv.org/abs/2212.02995 arXiv:2212.02995 [cs].

Weixiang Zhao, Yanyan Zhao, and Xin Lu. 2022a. CauAIN: Causal Aware Interaction Network for Emotion Recognition in Conversations. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Vienna, Austria, 4524–4530. https://doi.org/10.24963/ijcai.2022/628