TECPEC: Textual Emotion-Cause Pair Extraction in Conversations

Shreyas Kabra
CSAI IIIT Delhi
shreyas21563@iiitd.ac.in

Parthiv A Dholaria
CSE IIIT Delhi
parthiv21078@iiitd.ac.in

Kartik Singhal
CSAM IIIT Delhi
kartik21259@iiitd.ac.in

Lakshya IIITD
CSAM IIIT Delhi
lakshya21262@iiitd.ac.in

***Abstract*—In this project, we present our submission to SemEval-2024 Task 3, Subtask 1, "Textual Emotion-Cause Pair Extraction in Conversation", which focuses on extracting emotion-cause pairs from conversations. We introduce a two-step pipeline architecture for identifying these pairs. Initially, using our novel architectures we conducted emotion classification for the utterances. Subsequently, we developed a Question-Answer model to detect the causes for the target utterance given the identified emotions. We ranked 5th in Subtask 1. Our code is available at https://github.com/parthivdholaria/TECPEC**

***Index Terms*—Textual Emotion-Cause Pair Extraction in Conversations, Question-Answer model, two-step pipeline**

I. Introduction

Understanding human emotions in textual conversations is a fundamental aspect of improving human-computer interaction. The ability to extract not only emotions but also their underlying causes from text can significantly enhance the capability of machines to interact more naturally with humans. It all revolves around how well your system can engage the user. Through this Project, we aim to develop novel model architectures and employ novel techniques to improve the detection of textual emotions and their causes in conversations. This is particularly crucial in developing applications such as emotional support bots, sentiment analysis tools, and advanced customer service automation that can empathise and respond appropriately to human needs.

In this project, we present our models for SemEval-2024 Task 3, "The Competition of Multimodal Emotion Cause Analysis in Conversations" (ECAC) (Wang et al., 2024). We focus exclusively on Subtask 1, "Textual Emotion-Cause Pair Extraction in Conversations", where the goal is to classify emotions and extract the corresponding textual causal spans. For this, we present a two-step pipeline architecture where we (1) first predict the emotions using some of our novel architectures, and then (2) extract causes using a Question-Answering model employing simple-transformers. Our team ranked 5th out of 31 participating teams based on their metrics of evaluation, discussed later. We later present a thorough analysis of the performance of our models and their scopes for improvement.

II. Related Work

A. Emotion flip reasoning in multiparty conversations

In 2021, Kumar et al. [1], proposed a novel approach to understanding emotional dynamics within multi-party conversations. It not only focuses on emotion recognition in conversations (ERC) but also proposes a new task called Emotion-Flip Reasoning (EFR), which aims to identify past utterances that have triggered an emotional state flip in a speaker. The authors utilise a combination of masked memory networks and transformers to tackle this task and present a modified version of the MELD dataset, augmented with ground-truth labels for EFR. The development of a transformer-based network that specifically addresses the detection of triggers causing emotional flips in conversational contexts.

B. Recognizing Emotion Cause in Conversations

In 2021, Poria et al. [2] proposed a new dataset called RECCON specially designed for the Emotion cause extraction in conversations. They introduced a variety of emotional causes and established a strong baseline for the given tasks. They define two subtasks. Subtask1: Causal Span Extraction and Subtask2: Causal Emotion Entailment. The first subtask is about identifying the causal span (emotion cause) for a target non-neutral utterance. The other subtask is about giving a target non-neutral utterance (Ut), the goal is to predict which particular utterances in the conversation history are responsible for the non-neutral emotion in the target utterance. They incorporated contextual pre-trained embeddings (like RoBERTa) and outperformed many existing emotion cause extraction approaches.

C. End-to-End Emotion-Cause Pair Extraction based on Sliding Window Multi-Label Learning

In 2020, Ding et al. [3] proposed two joint frameworks: Multi-label learning for extracting cause clauses related to specific emotions (CMLL) and for emotions associated with specific causes (EMLL). Using a sliding window mechanism centred on targeted clauses, their integrated approach outperforms existing systems.

D. Emotion-Cause Pair Extraction: A New Task to Emotion Analysis in Texts

In 2019, Xia et al. [4] proposed a new task in the field of emotion analysis focused on extracting pairs of emotions and their corresponding causes directly from texts. They proposed a 2-step approach to address this new ECPE task, which first performs individual emotion extraction and then cause extraction via multi-task learning, and then conducts emotion-cause pairing and filtering.

III. Methodology

A. Task Definition

Before describing our method, we will first define our task. Our goal is to identify emotion-cause pairs within conversations. These conversations involve multiple speakers and text exchanges describing a scene. An emotion-cause pair is defined and annotated as a textual span. Here's an example illustrating the expected input and output:

- **Input:** a conversation containing the speaker and the text of each utterance.
- **Output:** all emotion-cause pairs, where each pair contains an emotion utterance along with its emotion category (Ekman's six emotions [5]) and the textual cause span in a specific cause utterance, e.g., (U3_Joy, U2_"You made up!").

B. Task Pipeline Overview

Our task pipeline comprises two stages:

- **ERC(Emotion Recognition in Conversation):** Classification of utterances into Ekman's six emotions [5] (anger, joy, surprise, disgust, fear, sadness), with a "neutral" label assigned to non-emotion utterances.
- **CEE(Cause-Emotion Extraction):** Extraction of cause utterance within a conversation corresponding to a target utterance (also in the same conversation) based on its associated emotion.

Next, we will provide a detailed description of our approaches in the subsequent subsection.

C. Emotion Recognition in Conversation (ERC)

To classify an utterance U_t with an emotion label E_t, our architectures should consider both the target utterance which is U_t, which is the t^{th} utterance in a conversation, and the preceding utterances U_1, U_2, U_3, ..., U_{t-1}. We propose two levels for determining these labels:

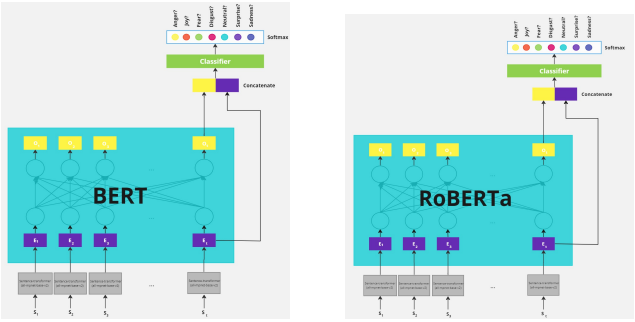

Fig. 1: BERT and RoBERTa For Sentence Classification Given Context

- **Utterance Level**: We called this method as "*Sentence Classification given Context*". Here, we classify only the target utterance U_t considering its preceding utterances U_1, U_2, U_3, ..., U_{t-1} as context. In other words, for a conversation comprising of D utterances, we generate D data points containing target utterances U_1, U_2, ... and U_d respectively, each considering its previous utterances as context. For this task, we have transformer-based encoder architectures like BERT and Roberta both having a classification head for classification (Fig. 1).
- **Conversational Level**: This method is commonly employed by researchers for the ERC Task. Here, the entire conversation serves as a single data point, and each utterance is labelled collectively, taking into account the previous utterances. For this task, we utilized a GPT2-based architecture (fig 2). We opted for GPT2 because it is a decoder-based model which is important because, akin to text generation where a word is predicted based on previous context, our task also requires considering the preceding context for classifying an utterance.

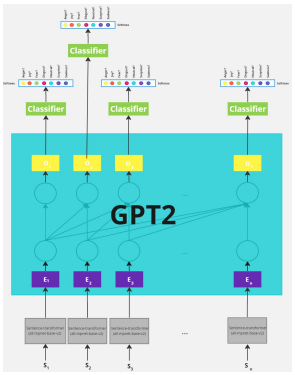

Fig. 2: GPT2 Conversational Level ERC Model

D. Cause-Emotion Extraction (CEE)

For this task, we utilized the method proposed by Poria et al. [2], referred to as "Causal Span Extraction," which entails identifying the causal span (emotion cause) for a target non-neutral utterance using a Question-Answering model. The framework suggested by Poria et al. [2] is as follows: For a target utterance U_t, the causal utterance $U_i \notin C(Ut)$, and a causal span $S \notin CS(U_t)$ from U_i, the context, question, and answer are as follows:

- **Context:** The context of a target utterance U_t, is a concatenation of all the previous utterances and U_t.
- **Question:** The question is framed as follows: "The target utterance is $< U_t >$ and the evidence utterance is $< U_i >$ then what is the causal span from evidence in the context that is relevant to the target utterance's $< U_t' semotion >$?".
- **Answer:** The causal span $S \in CS(U_t)$ appearing in U_i if $U_i \in C(U_t)$. Otherwise, if $U_i \notin C(U_t)$, then the answer returned will be "empty".

IV. Dataset Analysis

The dataset proposed for the task contains conversations from the famous American Sitcom F.R.I.E.N.D.S., annotated with emotion-cause pairs and emotion labels which include the six Ekman's emotions [5] i.e. anger, disgust, fear, joy, sadness

and surprise for emotional utterances and the non-emotional utterances are labelled as "neutral".

Set	# Conversations	# Utterances
Training	1236	12144
Validation	138	1475
Testing	665	6301

TABLE I: Data distribution

Since the organizers do not provide a separate validation set and the test set does not contain the emotions and emotion-cause pairs, therefore we have to split the provided training data into a 9:1 ratio as training and validation dataset. The final dataset split is shown in Table 1.

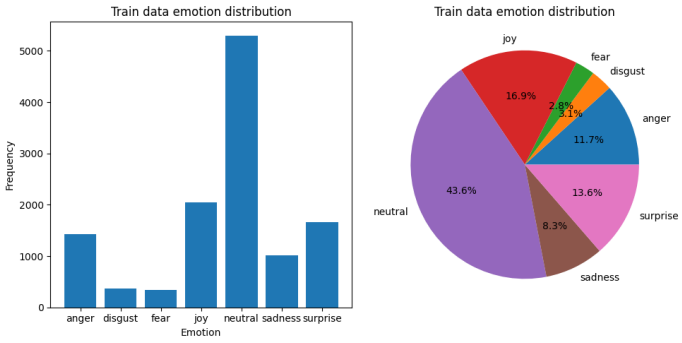

Fig. 3: Emotion Label Distribution on Training Data

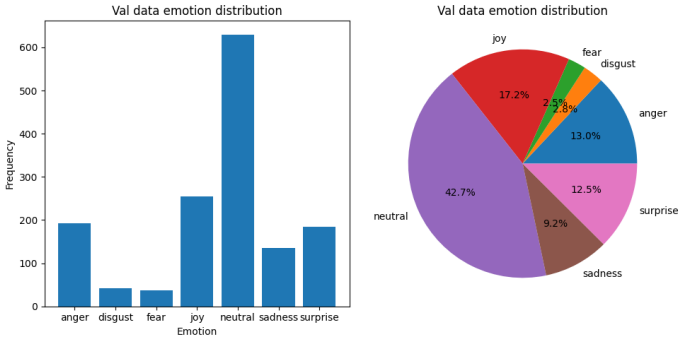

Fig. 4: Emotion Label Distribution on Validation Data

While performing visualisation on the training and validation set, we found that the non-emotional label "neutral" is the most occurring label and emotions like fear and disgust are least occurring, hence the provided dataset is very skewed. Figures 3 and 4 show further analysis of the emotion label distribution on training and validation data.

V. Experimental Setup

A. Sentence Embedding

Since computers can't work directly on textual data, we need a way to express it numerically. In our task, we need a way to express each utterance as a 768-dimensional vector (Input dimensions of transformer-based models). For this, we used the *all-mpnet-base-v2* model from sentence-transformer. We used sentence-transformer because sentence-transformer models are trained to encode the semantic meaning of a sentence into the embeddings they produce.

B. BERT/RoBERTa For Sentence Classification Given Context

These models are a variant of BERT/RoBERTa. It takes into account the target utterance and the contextual information from preceding utterances in the conversation. The meaning of the representations are the following

- S_i (Utterance): represents each padded utterance of a conversation.
- E_i (Sentence Embedding): Each sentence is converted into an embedding using the *all-mpnet-basev2* sentence transformer
- O_i (Output State): represents the final hidden state for each input embedding.

In the architecture (Fig 1.) we extract the output embedding corresponding to the last token of the target utterance from BERT/RoBERTa's output sequence. The extracted target embedding is concatenated with the original target utterance embedding creating a fused representation that contains both the contextual information and the inherent features of the target utterance. This concatenated embedding is passed through a dropout layer to prevent overfitting during training. A linear layer (classifier) projects the final output embedding. The number of output neurons in this layer equals the number of emotion labels which is 7 in our task.

We used Cross-entropy loss as it is a multi-class classification problem. We found optimal results on the following hyperparameters: Number of epochs = 30, Learning Rate = 1e-6, BatchSize = 16.

C. GPT2 Conversational Level

For this model, the input is the whole conversation. First, we compute the sentence embedding of all the utterances in the conversation which are later fed into the GPT2 model. We opted for GPT2 because it is a decoder-based model which is important because, akin to text generation where a word is predicted based on previous context, our task also requires considering the preceding context for classifying an utterance. The output from the GPT2 model are O_1, O_d, ..., O_d which are later fed into a classifier for labels. The label from the output O_i will be the label for the utterance S_i in the conversation.

We used Cross-entropy loss as it is a multi-class classification problem. We found optimal results on the following hyperparameters: Number of epochs = 30, Learning Rate = 1e-7, BatchSize = 16.

D. Question Answering Model (simple-transformer)

This model is used for finding the emotion-cause pair. Since it requires data in a different format, therefore we first convert the original data into the required format and then use it for training. The format can be found on their official website.

We have used the pre-trained *mrm8488/spanbert-finetuned-squadv2* and fine-tuned it for our task. We train the model only on 3 epochs because of hardware constraints.

VI. Evaluation Metric

A. Emotion Recognition in Conversation (ERC)

- **Accuracy:** Number of Correct Predictions over Total Number of Predictions.
- **Weighted F1:** Average of class-wise F1 Score considering the proportion for each class in the dataset
- **Macro F1:** Compute class-wise F1 Score then average it without considering the proportion for each class in the dataset.

B. Cause-Emotion Extraction (CEE)

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

VII. Results and Analysis

A. ERC Performance on Validation Set

Model	Accuracy	Macro F1	Weighted F1
BERT	0.32	0.27	0.32
RoBERTa	0.31	0.27	0.30
GPT2	0.36	0.30	0.37

TABLE II: Scores on the Validation Set (ERC Task)

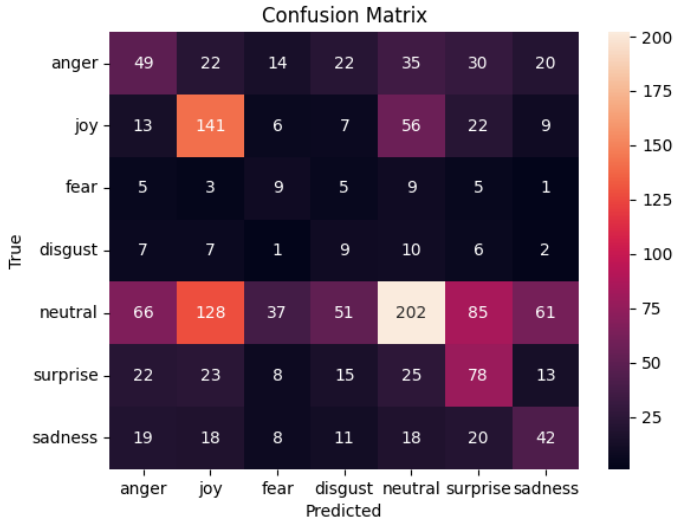

Fig. 5: Validation Confusion Matrix (ERC) by GPT2 Model

Table 2 shows class-wise F1 along with Accuracy, Weighted F1 and Macro F1 Scores on the Validation Set. The models we implemented for utterance level do not perform well enough where which shows why the researchers do not use that approach. The GPT2 model based on Conversational level outperforms the above two models showing significant improvement in all the scores. Therefore GPT2 is our best model. Figure 5 shows the confusion matrix on the Validation by GPT2 Model. It was observed that emotions with low presence in the dataset like "fear" and "disgust", our models show low scores for these emotions.

B. Testing Scores on Leaderboard

TABLE III: Testing Scores on Leaderboard

ERC Model	Cause Model	Wt. Strict F1	Wt. Prop. F1	Strict F1	Prop. F1
GPT2	QA	0.1345	0.1767	0.1283	0.1626
BERT	QA	0.1318	0.1704	0.1283	0.1581
RoBERTa	QA	0.1314	0.1697	0.1301	0.1629

We submitted all the predicting output from the three ERC models—GPT2, BERT, RoBERTa and the Simple Transformer QA cause model. The scores are shown in Table 3. The predictions from GPT2 secure us the 5th position on the leaderboard. The scores from the BERT and RoBERTa exhibit comparable results with marginal differences in the scores.

C. Ablation Study on Validation Set

Table 4 presents an ablation study on a validation set, comparing the performance of three ERC models—GPT2, BERT, and RoBERTa—and the ground truth of prediction. If we use the ground truth emotion of utterances, the QA model presents really good scores, whereas all three models demonstrate significantly lower scores in comparison to the ground truth, highlighting a substantial gap in model efficacy. These results show that we need to work more on our ERC task for predicting emotion; the cause-finding model is already performing well.

TABLE IV: Ablation Study on Validation Set

ERC Model	Cause Model	Wt. Strict F1	Wt. Prop. F1	Strict F1	Prop. F1
Ground truth	QA	0.3430	0.4612	0.0.3441	0.4594
GPT2	QA	0.1153	0.1543	0.1135	0.1443
BERT	QA	0.1181	0.1673	0.1148	0.1568
RoBERTa	QA	0.1132	0.1623	0.1123	0.1557

VIII. Conclusion

In conclusion, our exploration into Textual Emotion-Cause Pair Extraction in Conversations has provided a two-step pipeline that first provides emotion to utterances in a conversation and the QA model will find cause for that emotion. Despite our system's commendable 5th place finish among 31 contenders, the ablation study suggests working more on the emotion recognition task to bridge the gap to ground truth performance.

IX. Future Work

- Using some state-of-the-art models for finding the ERC task.
- Develop an end-to-end architecture that directly identifies the emotion-cause pair because, as observed in our

models, errors can stem from both the emotion and cause models, resulting in greater inaccuracies.

- Using present-day LLMs like ChatGPT for both tasks.

REFERENCES

[1] S. Kumar, S. Dudeja, M. S. Akhtar, and T. Chakraborty, "Emotion flip reasoning in multiparty conversations," *IEEE Transactions on Artificial Intelligence*, 2023.

[2] S. Poria, N. Majumder, D. Hazarika, D. Ghosal, R. Bhardwaj, S. Y. B. Jian, P. Hong, R. Ghosh, A. Roy, N. Chhaya *et al.*, "Recognizing emotion cause in conversations," *Cognitive Computation*, vol. 13, pp. 1317–1332, 2021.

[3] Z. Ding, R. Xia, and J. Yu, "End-to-end emotion-cause pair extraction based on sliding window multi-label learning," in *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, 2020, pp. 3574–3583.

[4] R. Xia and Z. Ding, "Emotion-cause pair extraction: A new task to emotion analysis in texts," *arXiv preprint arXiv:1906.01267*, 2019.

[5] P. Ekman, "Expression and the nature of emotion," *Approaches to emotion*, vol. 3, no. 19, p. 344, 1984.