# SemEval-2024 Task 3: The Competition of Multimodal Emotion Cause Analysis in Conversations

**Anup Bhutada**
University of Colorado
Boulder

anbh1796@colorado.edu

**Uday Gadge**
University of Colorado
Boulder

udga1318@colorado.edu

**Shouvik Sengupta**
University of Colorado
Boulder

shse8233@colorado.edu

## Abstract

We present a sub-task from Sem Eval-2023 Task 3, a shared task on extracting Textual emotion-cause pair extractions in conversations. This dataset consists of conversations involving multiple people. Each dialogue in the conversation is annotated with the corresponding emotion - neutral, joy, sadness, anger, surprise, disgust, and fear. Every utterance with a non-neutral emotion has causal spans within the conversation. The task of participating systems was to predict the emotion of each utterance in a set of conversations and also predict the causal spans within the conversation. We are a group of three who took part in this task. We present our findings and results in this paper. In the end, the best-performing model for emotion classification achieved a macro F1 score of 0.47 on the validation set while the proportional span match F1 score for the span detection model was 0.6.

## 1 Introduction

Finding relationships between a pair of texts within a given context requires an understanding of the entire context and the contextual relationships between the pair. Extracting this information is particularly useful for conversational data. The emotion cause pair extraction has potential in areas of automated conflict resolution, therapeutic conversational agents, and marketing research.

The task presented in the SemEval task [1] is broken into two sub-tasks. As a first step, every utterance in the conversation is assigned an emotion based on the probability of each emotion associated with the text. This sub-task is followed by emotion-cause span extraction aimed at extracting the span within the conversation. One of our approaches, however, tried to do both the sub-tasks in a single model. However, the final solution we present is a two-model-based approach. Due to the scarcity of training data, we relied on fine-tuning Bert models pre-trained on similar tasks.

## 2 Data Overview

### 2.1 Annotated data for training

Every instance in training data is a conversation involving multiple speakers. Each utterance is annotated with the speaker and an annotation. Every utterance in the conversation with an emotion other than neutral is annotated with a list of spans from all the utterances in the conversation. These spans are usually from the utterances before the target utterance and the target utterance but there are cases with the span being from a future utterance in cases with the same speaker.

Example:

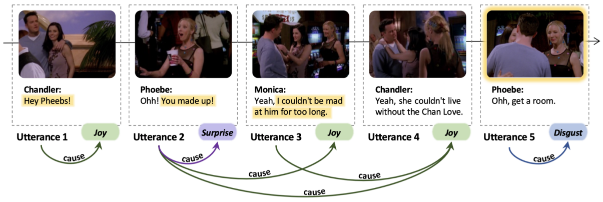
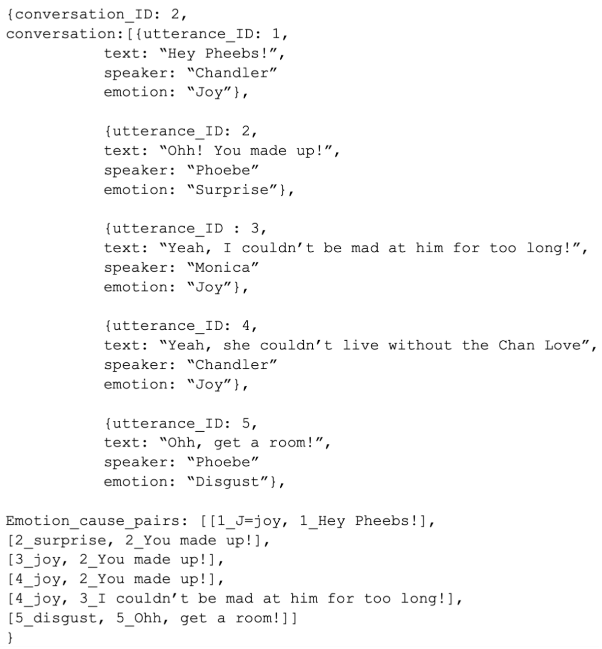


Figure 1: An example of our task and annotated dataset. Each arc points from the cause utterance to the emotion it triggers. The cause spans have been highlighted in yellow. Background: Chandler and his girlfriend Monica walked into the casino (they had a quarrel earlier but made up soon, and then started a conversation with Phoebe.

### 2.2 Test data and submission format

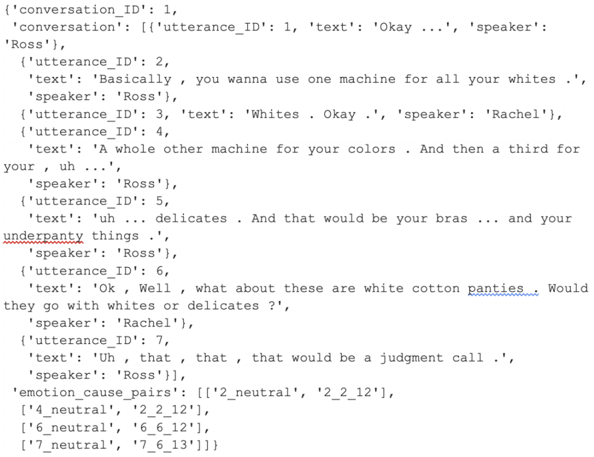
From the training data of around 1300 conversations, 20% of the conversations were randomly sampled and kept out of the training functions for evaluating the models built. The trial data provided by the SemEval is a batch of 40 conversations without annotated emotions and emotion cause spans. We provided the emotions and emotion cause spans for these conversations that had a emotion cause span predicted. Although in the training data only



```
{conversation_ID: 2,
conversation:[{utterance_ID: 1,
           text: “Hey Pheebs!”,
           speaker: “Chandler”
           emotion: “Joy”},

           {utterance_ID: 2,
           text: “Ohh! You made up!”,
           speaker: “Phoebe”
           emotion: “Surprise”},

           {utterance_ID : 3,
           text: “Yeah, I couldn’t be mad at him for too long!”,
           speaker: “Monica”
           emotion: “Joy”},

           {utterance_ID: 4,
           text: “Yeah, she couldn’t live without the Chan Love”,
           speaker: “Chandler”
           emotion: “Joy”},

           {utterance_ID: 5,
           text: “Ohh, get a room!”,
           speaker: “Phoebe”
           emotion: “Disgust”},

Emotion_cause_pairs: [[1_J=joy, 1_Hey Pheebs!],
[2_surprise, 2_You made up!],
[3_joy, 2_You made up!],
[4_joy, 2_You made up!],
[4_joy, 3_I couldn’t be mad at him for too long!],
[5_disgust, 5_Ohh, get a room!]]
}
```

Figure 2: JSON format of the data from Figure 1

provides cause spans for utterances with an emotion that is non neutral, we choose to provide emotions and cause spans based on the span based model. The spans are submitted as per the standards of SemEval with span of indices of tokens rather than the text itself.



```
{'conversation_ID': 1,
 'conversation': [{'utterance_ID': 1, 'text': 'Okay ...', 'speaker':
'Ross'},
  {'utterance_ID': 2,
   'text': 'Basically , you wanna use one machine for all your whites .',
   'speaker': 'Ross'},
  {'utterance_ID': 3, 'text': 'Whites . Okay .', 'speaker': 'Rachel'},
  {'utterance_ID': 4,
   'text': 'A whole other machine for your colors . And then a third for
your , uh ...',
   'speaker': 'Ross'},
  {'utterance_ID': 5,
   'text': 'uh ... delicates . And that would be your bras ... and your
underpanty things .',
   'speaker': 'Ross'},
  {'utterance_ID': 6,
   'text': 'Ok , Well , what about these are white cotton panties . Would
they go with whites or delicates ?',
   'speaker': 'Rachel'},
  {'utterance_ID': 7,
   'text': 'Uh , that , that , that would be a judgment call .',
   'speaker': 'Ross'}],
 'emotion_cause_pairs': [['2_neutral', '2_2_12'],
  ['4_neutral', '2_2_12'],
  ['6_neutral', '6_6_12'],
  ['7_neutral', '7_6_13']]}
```

Figure 3: Sample of submission format

# 3 Methods and analysis

As mentioned earlier, the scarcity of the data prevented us from exploring the possibility of building BERT models on the data. We instead relied on fine tuning models trained for similar tasks on other large datasets. We present the different model architectures we explored for different sub-tasks along with the specifics of fine tuning for the model.

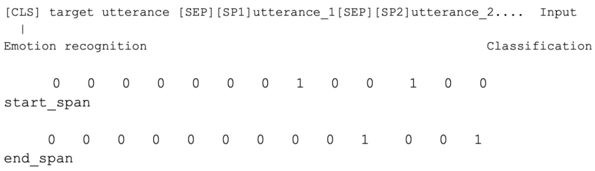
## 3.1 SGNLP

SGNLP is a collection of models developed by Singapore's NLP researchers for various NLP tasks [2]. These tasks include emotion recognition and emotion entailment included in the Recognizing Emotion Cause in Conversation (RECCON) package. These tasks proved to be very useful for the task we were trying to achieve. However, the pre trained model failed to extract spans from the context utterance. The complex architecture of SGNLP was difficult to fine tune but we endorse the packages in SGNLP as the available pre trained models are very similar to the provided tasks.

## 3.2 BERT with 3 heads

We used a small-BERT based pre-trained model to build a three headed classification model that would give the classification output for emotion detection, start spans detection and end spans detection. The model architecture used the last hidden layer of the pre-trained transformer model to build a three way classification model. The losses for all three tasks are added together and backpropagated for fine tuning. The input batch for the model is formatted by treating each utterance as the target utterance for which we need to predict the emotion and extract the cause. Each utterance is appended with the entire conversation as context for span extraction. The target utterance and the context conversation are separated by a [SEP] token. Moreover, each utterance in the context conversation is preceded by a special speaker token ([SP1], [SP2] .., [SP8]) to indicate the speaker of the utterance in the conversation. For the classification task, the last hidden state for the [CLS] token is used to generate the softmax probabilities of the seven emotions using two fully connected layers.

The other two heads (start token and end token) use the last hidden layer representations corresponding to all the tokens in the input, to generate logits for span start and end probabilities using a single fully connected layer. This model was trained to extract all possible causal spans for the emotion exhibited in the target utterance by setting the start and end tokens corresponding to all causal spans in the context conversation to 1. The model achieved acceptable results on classification of emotion in the target utterance (sub-task 1), but could not achieve good results on cause extraction (sub-task 2).



```
[CLS] target utterance [SEP][SP1]utterance_1[SEP][SP2]utterance_2....  Input
  |
Emotion recognition                                                Classification

     0   0   0   0   0   0   0   1   0   0    1   0   0
start_span

     0   0   0   0   0   0   0   0   0   1    0   0   1
end_span
```

Figure 4: Bert with 3 classification heads

### 3.3 Emotion recognition

We experimented with different transformer based models to predict the emotion of a target utterance in a conversation. Specifically, 4 different models were tried: small-BERT model [7], distilroberta_base model , BERT uncased model pre trained for emotion classification in text and a RoBERTa model trained on emotion cause extraction using a Question Answering architecture [3]. All the parameters in the small-BERT model were fine-tuned while training, whereas only the top dense layer was fine-tuned in case of the distilroberta_base and BERT uncased emotion models. This was because the latter models were pre-trained on twitter data to predict six emotions. In both the cases the loss was a weighted cross entropy loss to take into account the imbalance in the distribution of emotions in the data. The input batch for all the models used the same format as the inputs in the three-classification heads BERT model described in the previous section.

The best results for the classification sub-task, however, were obtained in case of the RoBERTa model trained on emotion cause extraction. The last layers in this model were replaced with a 2 layer fully connected network on the [CLS] token for classification. The motivation behind using this setup was that the previous setup with the 3 classification heads that is trying to predict multiple spans in the context conversation may be backpropagating too many losses that are not weighed accurately. And therefore separating these tasks and training two simpler models may be a better idea to train these models to perform better.

The loss used to train this model was the categorical cross-entropy loss with weighted class contributions. Since the dataset had a relatively large number of utterances with neutral emotions, the weight of this class was reduced and some classes that had very few data points (like disgust and fear) were weighted heavily.

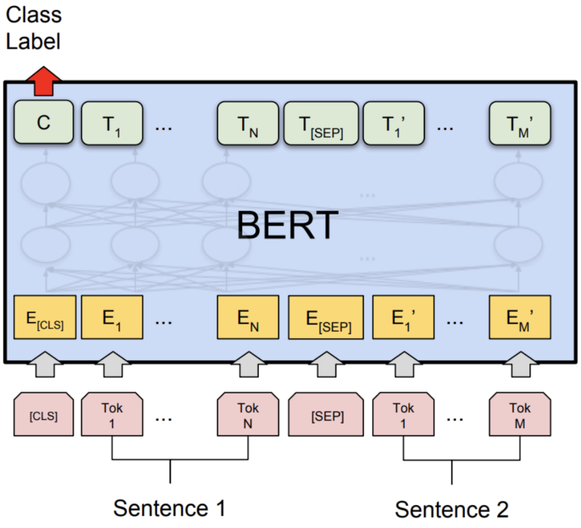


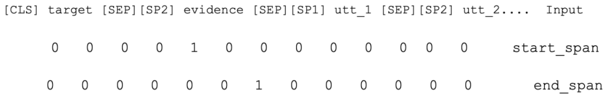
Figure 5: Bert for classification

### 3.4 Emotion Span Extractor

To detect cause spans in conversation we used pre-trained BERT models for span extraction. We experimented with some Question Answering models and Emotion span extraction models. The pre-trained model that performed best on fine-tuning was the same one from the last section: RoBERTa model trained on emotion cause extraction using a Question Answering architecture [nakul cite]. In both the models the context from which span or answer is to be extracted is appended to the target utterance separated by a [SEP] token. This is called the evidence utterance or the utterance that may contain the cause for the emotion exhibited in the target utterance.

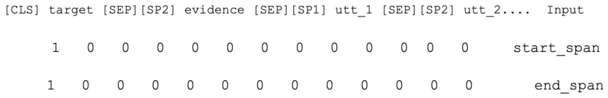
Since every utterance in conversation can contain the causal span for each target utterance, the input batch is formatted to contain all possible pairs of target utterance and evidence utterance. At the end of every input, the entire conversation, in the same format as described above, is again appended as context to provide the model with all the information in that conversation. In case of no relation between the target and evidence utterance the start and end span labels are set to the [CLS] token. The model, in this way, also learns to predict whether a relationship exists between the target and evidence utterance along with the start and end tokens for causal spans.

If the evidence has no cause for the target utterance:



```
[CLS] target [SEP][SP2] evidence [SEP][SP1] utt_1 [SEP][SP2] utt_2....  Input
    0   0   0   0   1   0   0   0   0   0   0   0   0     start_span
   0   0   0   0   0   0   1   0   0   0   0   0   0       end_span
```

Figure 6: Tokenized input with cause



```
[CLS] target [SEP][SP2] evidence [SEP][SP1] utt_1 [SEP][SP2] utt_2....  Input
    1   0   0   0   0   0   0   0   0   0   0   0   0     start_span
   1   0   0   0   0   0   0   0   0   0   0   0   0       end_span
```

Figure 7: Tokenized input with no cause

## 4 Results

In this section, results for the best performing model on the validation dataset are summarized. The best performing model for the classification sub-task was the fine-tuned RoBERTa base model pre-trained on emotion cause extraction. The same pre-trained model fine-tuned on the span extraction sub-test achived the best results in this experiment.

The results on the classification sub-task are presented through a confusion matrix and a classification report containing the precision, recall and f1 scores for all classes. This report also contains the overall accuracy and averaged F1 scores.

For span extraction, metrics described in the evaluations seciton of the SemEval Task 3 description were used. These metrics are strict match F1 score, proportional match F1 score, weighted strict match F1 score and weighted proportional match F1 score.

These scores are described below.

**Strict F1**:

$$P = \frac{\sum CorrectParis}{\sum PredictedPairs}$$

$$P = \frac{\sum CorrectParis}{\sum PredictedPairs}$$

$$F_1 = \frac{2 \times P \times R}{P + R}$$

**Proportional F1**:

$$P = \frac{\sum Overlapinspans_i}{\sum Len(PredictedSpans_i)}$$

$$R = \frac{\sum Overlapinspans_i}{\sum Len(AnnotatedSpans_i)}$$

$$F_1 = \frac{2 \times P \times R}{P + R}$$

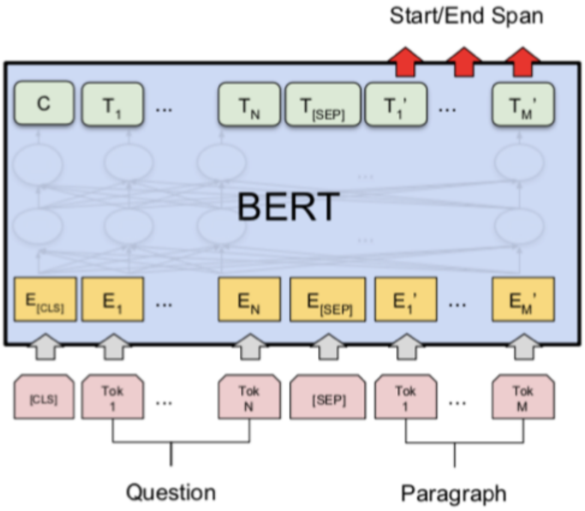
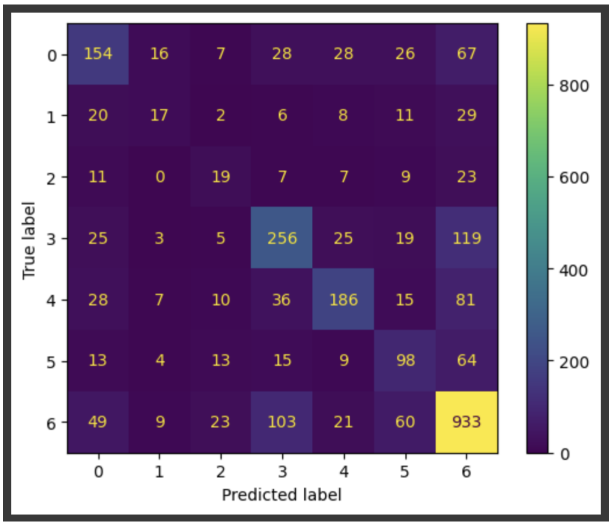


Figure 8: Bert on SQuAD [5]

The weighted versions of these scores compute these metrics for each class and compute the weighted mean of these scores.

The emotion class mappings are for the confusion matrix shown in figure 9 are:-
{'anger':0, 'disgust':1, 'fear':2, 'joy':3, 'surprise':4, 'sadness':5, 'neutral':6 }



| | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| 0 | 154 | 16 | 7 | 28 | 28 | 26 | 67 |
| 1 | 20 | 17 | 2 | 6 | 8 | 11 | 29 |
| 2 | 11 | 0 | 19 | 7 | 7 | 9 | 23 |
| 3 | 25 | 3 | 5 | 256 | 25 | 19 | 119 |
| 4 | 28 | 7 | 10 | 36 | 186 | 15 | 81 |
| 5 | 13 | 4 | 13 | 15 | 9 | 98 | 64 |
| 6 | 49 | 9 | 23 | 103 | 21 | 60 | 933 |

Figure 9: Confusion Matrix for Classification

The model performs reasonably well for some classes (e.g., class 6 - neutral) as seen all the metrics are above 70% but less effectively for others (e.g., class 1 - disgust) with lowest f1 score of 0.23. The class imbalance is a major factor for inconsistency of metrics among classes. The macro and weighted averages provide an overall assessment of the model's performance, with the weighted av-

| classes | Precision | Recall | F1 score | support |
|---|---|---|---|---|
| 0 | 0.51 | 0.47 | 0.49 | 326 |
| 1 | 0.30 | 0.18 | 0.23 | 93 |
| 2 | 0.24 | 0.25 | 0.25 | 76 |
| 3 | 0.57 | 0.57 | 0.57 | 452 |
| 4 | 0.65 | 0.51 | 0.57 | 363 |
| 5 | 0.41 | 0.45 | 0.43 | 216 |
| 6 | 0.71 | 0.78 | 0.74 | 1198 |
| accuracy | | | 0.61 | 2724 |
| macro avg | 0.49 | 0.46 | 0.47 | 2724 |
| weighted avg | 0.60 | 0.61 | 0.60 | 2724 |

Table 1: Classification report

| Dev Metrics | Precision | Recall | F1 score |
|---|---|---|---|
| Strict | 0.469409 | 0.468902 | 0.469918 |
| Proportional | 0.598752 | 0.576989 | 0.622221 |
| Weighted Strict | | | 0.469506 |
| Weighted Prop | | | 0.600978 |

Table 2: Span Match Results

erage being more influenced by the larger classes. The results for span match (Table 2) show that scores of about 0.62 on proportional span match and about 0.47 on strict span matches were acieved on the validation set.

## 5 Conclusion and Future Scope

In conclusion, we found the task very challenging due to the scarcity of training data available. Emotional recognition was more difficult to train as the samples were lower compared to emotion cause span. The other problem with emotion recognition was that the labels were imbalanced. The sparsity of labels other than neutral was another source of challenge in the task. The final models we present are both BERT-based but there are some alternatives we would like to propose at the end of this paper. One alternative is SGNLP, as their packages were trained on conversation data rather than Twitter data. This might serve the purpose of this data very well. The other alternative is prompt engineering. We briefly attempted GPT prompt engineering but couldn't pursue further due to resource constraints.

## 6 References

[1] F. Wang, Z. Ding, R. Xia, Z. Li and J. Yu, "Multimodal Emotion-Cause Pair Extraction in Conversations," in IEEE Transactions on Affective Computing, vol. 14, no. 3, pp. 1832-1844, 1 July-Sept. 2023, doi: 10.1109/TAFFC.2022.3226559.

[2] Poria, S., Majumder, N., Hazarika, D., Ghosal, D., Bhardwaj, R., Jian, S.Y.B., Hong, P., Ghosh, R., Roy, A., Chhaya, N., Gelbukh, A. and Mihalcea, R. (2020). Recognizing emotion cause in conversations. arXiv preprint arXiv:2012.11820., Dec 2020.

[3] https://huggingface.co/Nakul24/Spanbert-emotion-extraction/tree/main

[4] https://github.com/angelosps/Question-Answering/blob/main/Question_Answering.ipynb

[5] Schwager, Sam, and John Solitario. "Question and answering on SQuAD 2.0: BERT is all you need." ArXiv e-prints of (2019).

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000 6010 .

[7] https://huggingface.co/prajjwal1/bert-small