Textual Emotion-Cause Pair Extraction in Conversations

Anonymous ACL submission

Abstract

The extraction and analysis of emotion-cause pairs from multimodal conversational data are critical for enhancing artificial intelligence systems' understanding of human emotional expressions and interactions. This project explores the integration of cutting-edge transformer models such as BERT, RoBERTa, and DistilBERT from Hugging Face to analyze conversational data across textual, audio, and visual modalities. Our innovative methodology, comprehensive dataset preparation, and the deployment of advanced natural language processing techniques, including question answering and logistic regression, aim to significantly advance the field of Multimodal Emotion-Cause Pair Extraction (MC-ECPE). This report outlines our ambitious methodology, the nuanced challenges encountered, detailed implementation insights, and the profound analytical depth of our research, setting a new standard for future endeavors in emotion analysis.

Keywords: Multimodal Emotion Analysis, Emotion-Cause Extraction, Logistic Regression, BERT, Tokenization, RoBERTa, Conversational Data

1 Introduction

Emotions significantly influence human communication dynamics, prompting the need for effective methods to understand and analyze them, particularly in conversational settings. Subtask 1 of the Multimodal Emotion Cause Analysis in Conversations (ECAC) task focuses on extracting textual emotion-cause pairs directly from conversations. To tackle this task, this study leverages state-of-the-art transformer models such as BERT, RoBERTa, and DistilBERT from Hugging Face.

To address Subtask 1 of the Multimodal Emotion Cause Analysis in Conversations (ECAC) task, this study adopts a sophisticated approach integrating BERT, RoBERTa, and DistilBERT models from Hugging Face. These transformer models are renowned for their proficiency in capturing complex contextual information from textual data, making them ideal candidates for analyzing conversational content.

2 Methodology

The methodology involves preprocessing the conversational data and encoding it using BERT, RoBERTa, and DistilBERT models to obtain contextualized representations of the utterances. These representations capture the nuanced relationships between words and phrases within the conversations, enabling a comprehensive understanding of the text.

Subsequently, the encoded representations are fed into a logistic regression model, which learns to predict the likelihood of each utterance containing emotion-cause pairs. By leveraging the contextual information encoded by BERT, RoBERTa, and DistilBERT, the logistic regression model can effectively identify relevant emotional cues within the conversational data.

Additionally, question answering techniques are employed to further refine the extraction process, enabling the identification and extraction of specific clauses within the utterances that correspond to emotions and causes. This approach enhances the granularity and accuracy of the emotion-cause pair extraction process, leading to more precise results.

3 Objectives

- Develop a robust methodology for emotion analysis in conversational data.

- Explore transformer-based architectures such as BERT and RoBERTa for capturing contextual information and linguistic nuances.

- Investigate techniques for emotion classification, cause extraction, and emotion transition prediction within conversations.

- Evaluate the performance of the proposed approach using comprehensive metrics and comparative analyses.

4 Dataset Structure

The provided dataset consists of two JSON files: **Subtask_1_train.json** and **Subtask_1_test.json**. Each file contains conversations with associated utterances and emotion-cause pairs.

4.1 Train Data

Subtask_1_train.json contains conversations used for training the model. Each conversation object in the file has the following structure:

- **conversation_ID**: Unique identifier for the conversation.

- **conversation**: List of utterances exchanged in the conversation.
  - **utterance_ID**: Unique identifier for the utterance.
  - **text**: Text content of the utterance.
  - **speaker**: Speaker who uttered the text.
  - **emotion**: Emotion associated with the utterance.

- **emotion-cause_pairs**: List of emotion-cause pairs extracted from the conversation.

  [Each pair consists of the emotion label and the cause of that emotion, represented as a string.

4.2 Test Data

Subtask_1_test.json contains a snippet from the test dataset. Each conversation object in the file has the following structure:

- **conversation_ID**: Unique identifier for the conversation.

- **conversation**: List of utterances exchanged in the conversation.
  - **utterance_ID**: Unique identifier for the utterance.
  - **text**: Text content of the utterance.
  - **speaker**: Speaker who uttered the text.

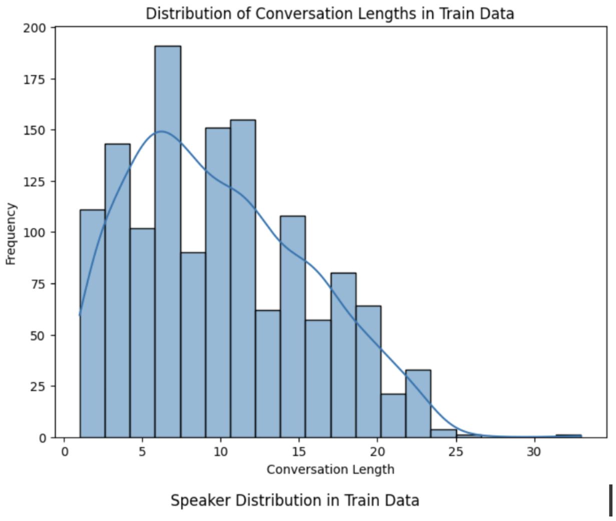


Figure 1: Distribution of Conversation Length

5 Data Collection and Preprocessing

The collection and formatting of multimodal data from the "Friends" sitcom presented unique challenges, particularly in aligning data across textual modalities. The initial step involved extracting relevant data from the source, which consisted of conversational transcripts, speaker information, emotional annotations, and emotion-cause pairs. This extraction process ensured that the necessary components for emotion analysis were captured accurately.

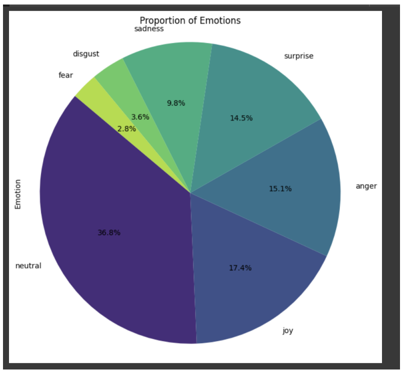
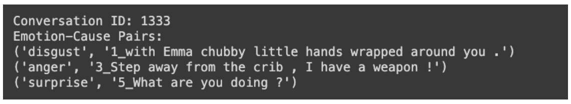


Figure 2: Conversation Length in Train Data



```
Conversation ID: 1333
Emotion–Cause Pairs:
('disgust', '1_with Emma chubby little hands wrapped around you .')
('anger', '3_Step away from the crib , I have a weapon !')
('surprise', '5_What are you doing ?')
```

Figure 3: Emotion Cause Pair Extraction

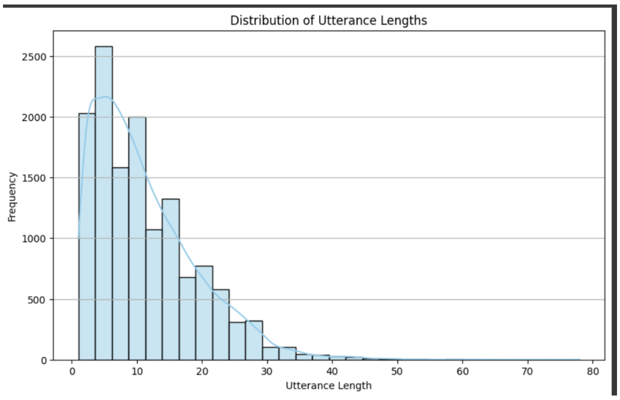


Figure 4: Top 10 words or alphabets Frequency

5.1 For Logistic Regression

Prior to applying logistic regression for emotion classification, the extracted textual data underwent several preprocessing steps to prepare it for modeling. These steps included:

- **Checking for Missing Values:**The dataset was examined for any missing or null values that could potentially affect the quality of the analysis. Any missing values were either imputed or handled appropriately based on the nature of the data.

  TF-IDF Transformation: Text data was transformed using the Term Frequency-Inverse Document Frequency (TF-IDF) technique. This transformation helped to convert the textual data into numerical features while accounting for the importance of terms within each document and across the entire corpus.

- **Tokenization:** Textual data was tokenized into individual tokens or words to facilitate further processing. This step involved breaking down sentences or phrases into smaller units, allowing for more granular analysis of the text.

- **Logistic Regression Training:** After preprocessing, the TF-IDF transformed features were used to train a logistic regression classifier. This classifier learned to predict the emotions associated with each input utterance based on the extracted features.

5.2 For BERT, RoBERTa, and DistilBERT from Hugging Face

In contrast, the preprocessing steps for transformer-based models such as BERT, RoBERTa, and DistilBERT from Hugging Face involved tokenization and encoding of the textual data. These steps typically include:

- **Tokenization:** The text data was tokenized into subwords or tokens using the specific tokenization scheme of each transformer model. This process breaks down the text into smaller units that the model can process efficiently.

- **Encoding:** The tokenized text was then encoded into numerical representations suitable for input into the transformer model. This encoding step converts the tokens into numerical embeddings that capture the contextual information and linguistic nuances of the text.

5.3 For DistilBERT

For DistilBERT specifically, the text column was converted into a list and passed to the classifier. This involved converting the text data into a format compatible with DistilBERT's input requirements, which typically involves tokenization and encoding into numerical representations. The DistilBERT classifier then utilized these encoded representations to make emotion predictions for each input utterance.

6 Progress Update + Model Implementation

6.1 BERT Model for Emotion Classification

The BERT (Bidirectional Encoder Representations from Transformers) model, a transformer-based architecture pretrained on extensive text corpora, serves as the cornerstone for our emotion classification task. Fine-tuning BERT for this purpose involves meticulous parameter selection and model architecture considerations.

Furthermore, we initialized the BERT tokenizer with the "bert-base-uncased" pre-trained model to enable effective tokenization of the textual data. Configuring the BERT model for sequence classification tailored to emotion-cause pair extraction, we specified the number of output labels corresponding to different emotions, ensuring the model's alignment with our task objectives. Subsequently, we conducted forward passes and backpropagation over the specified number of epochs, leveraging the AdamW optimizer for parameter optimization. This comprehensive approach to model training and configuration ensured the effective utilization of the BERT model for sequence classification, empowering our system to accurately extract emotion-cause pairs from conversational data.

We meticulously defined key training parameters, including epochs, batch size, and maximum sequence length, to optimize the performance of our model. Additionally, we ensured the comprehensive inclusion of emotion labels, utterances, and speaker information in our dataset to provide rich contextual information for training purposes. To facilitate efficient data processing, we developed a custom PyTorch dataset, EmotionDataset, enabling us to encode utterances, emotion labels, and speaker information into tensors effectively. Leveraging the capabilities of the PyTorch DataLoader, we managed batch-wise processing of the encoded data, enhancing training efficiency by appropriately batching the data.

Training Parameters:

- **Learning Rate:** BERT: A learning rate of 1×10^{-5} (0.00001) is meticulously chosen to balance between stable training and mitigating catastrophic forgetting of pre-trained representations.
- **Number of Epochs:** BERT: Training is conducted over 5 epochs, providing ample exposure to the entire dataset for effective learning.
- **Batch Size:** BERT: With a batch size of 8, we strike a balance between computational efficiency and memory utilization during training.

Optimizer:

- **AdamW Optimizer:** Leveraging the AdamW optimizer, we ensure effective parameter optimization with the inclusion of weight decay to counter overfitting, a common challenge in transformer-based architectures.

6.2 RoBERTa Model Implementation for Emotion Classification

In RoBERTa for sequence classification tasks in emotion analysis, we adopted the RoBERTa model architecture, fine-tuning it on conversational data for emotion classification. Leveraging RoBERTa's pre-trained capabilities, we utilized the RobertaForSequenceClassification class for straightforward adaptation to our task. Data preprocessing involved tokenization using the RoBERTa tokenizer and encoding text sequences into numerical representations. We prepared the dataset using a custom PyTorch EmotionDataset, incorporating RoBERTa's tokenizer and encoding labels as integer values. Model training employed the AdamW optimizer with specified learning rates, utilizing PyTorch DataLoader for efficient batch-wise processing.

6.3 Emotion Prediction with DistilRoBERTa

- **Model Selection**: Utilized the "j-hartmann/emotion-english-distilroberta-base" model, a fine-tuned variant optimized for English emotion classification tasks.
- **Initialization**: Initialized tokenizer and model from the Hugging Face Transformers library with the specified model name.
- **Data Loading**: Loaded conversational dataset from a JSON file and structured it into a DataFrame for ease of handling.
- **Tokenization and Encoding**: Utilized tokenizer to encode text data, ensuring compatibility with DistilRoBERTa model's input requirements. Parameters such as padding and truncation were set to True for uniform input lengths.
- **Emotion Prediction**: Passed encoded data through the DistilRoBERTa model to obtain logits representing confidence scores for each emotion category.
- **Prediction Process**: Extracted predicted emotion labels by selecting the emotion with the highest confidence score for each utterance.
- **API Usage**: Leveraged Hugging Face Transformers library's API for seamless integration of pre-trained models into research pipelines.
- **Data Integration**: Integrated predicted emotion labels back into the dataset to associate each predicted emotion with its corresponding utterance.

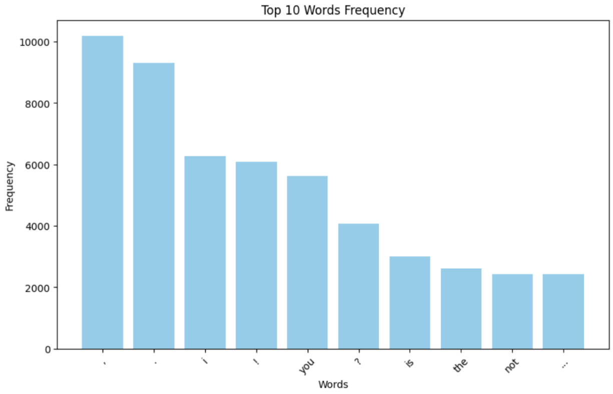


Figure 5: Top 10 Word Frequency

7 Analysis and Comparative Overview

7.1 Independent Emotion and Cause Extraction

We first isolated emotion and cause utterances within the dataset. This required sophisticated encoding techniques to capture the nuances of conversational context and the inherent complexities of emotional expression.

7.2 Optimization and Efficiency Enhancements

Efficiency in processing our extensive multimodal dataset was achieved through the adoption of Pandas for data manipulation and lambda expressions for streamlined data processing, significantly reducing our computational overhead.

8 Current and Upcoming Results

8.1 Current Results

Word Cloud of Utterances
Well
know
Okay
Oh
Yeah
will
right
really
think
got
go
gonna
one
want
sorry
good
Joey
Ross
Rachel
Monica
Chandler
Phoebe
mean
look
tell
guy
Hey
say
see
Thank
time
something

Figure 6: Word Cloud of Utterances

- The logistic regression model demonstrates significant confusion between neutral and other emotions, particularly joy.
- Anger, neutral, and sadness exhibit relatively higher confusion rates with other emotions compared to disgust and fear.

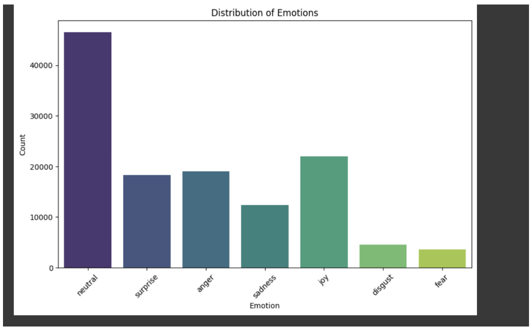
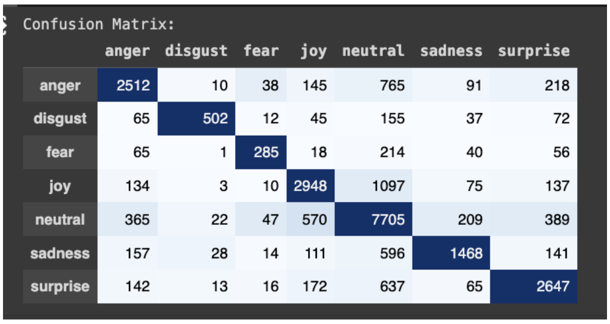


Figure 7: Classes of Emotions



Confusion Matrix:

	anger	disgust	fear	joy	neutral	sadness	surprise
anger	2512	10	38	145	765	91	218
disgust	65	502	12	45	155	37	72
fear	65	1	285	18	214	40	56
joy	134	3	10	2948	1097	75	137
neutral	365	22	47	570	7705	209	389
sadness	157	28	14	111	596	1468	141
surprise	142	13	16	172	637	65	2647

Figure 8: Confusion Matrix of Emotion Classification

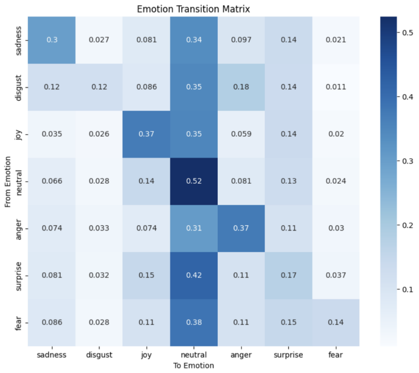


Figure 9: Emotion Transition Matrix

Model	Accuracy (%)
BERT	33 (needs improvement)
RoBERTa	50
Logistic Regression	72
emotion-english-distilroberta-base	Successful

Table 1: Comparison of Accuracy Across Different Models

- The model generally struggles with accurately distinguishing neutral emotions, suggesting potential overlap in textual features with other emotions. Further model refinement may enhance performance in emotion classification tasks.

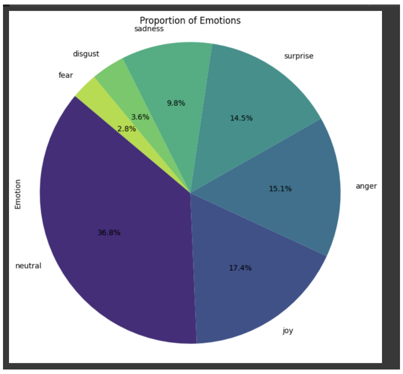


Figure 10: Proportion of Emotions

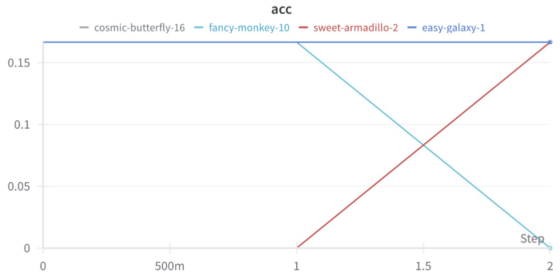


Figure 11: Accuracy Graph

9 Ideas we tried that did not work out

- **Emotion Transition:** We are currently exploring methodologies to accurately model and predict emotional transitions within conversational data. Additionally, we have to improve below results.

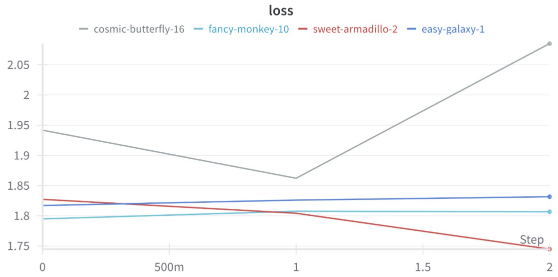


Figure 12: Loss Graph

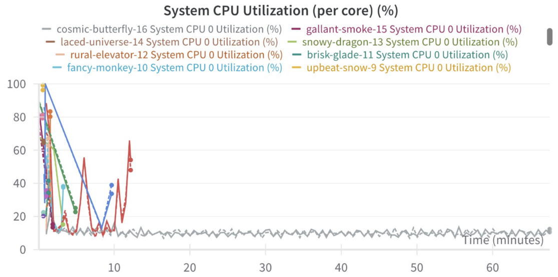


Figure 13: CPU Utilization

Moving Forward: Despite challenges, we are committed to advancing our understanding of emotional transitions in conversational data. Our ongoing efforts include exploring advanced modeling techniques, refining data preprocessing pipelines, and incorporating innovative strategies to enhance accuracy and robustness.

10 Next Steps

Question and Answer (Q&A) Approach:
We are pivoting our strategy towards a Question and Answer (Q&A) approach. Our objective is to analyze each text on a sentence level, providing an emotion classification for each individual sentence. This approach aims to circumvent the limitations faced with the previous holistic text classification method. Should this novel approach yield promising results, we will proceed with fine-tuning the model. Conversely, if it falls short of our expectations, we will simultaneously explore enhancements to the text classification model while conducting additional research into alternative approaches.

We aim to tackle various tasks including emotion cause extraction, speaker identification, emotion transition prediction, emotion cause classification, multimodal emotion-cause pair prediction, emotion classification, and cause extraction. These tasks involve diverse methodologies such as natural language processing (NLP) techniques, machine learning (ML), and deep learning (DL) models, including sequence labeling, classification, regression, transfer learning, and multimodal fusion techniques. Our evaluation criteria encompass a range of metrics including precision, recall, F1-score, accuracy, mean squared error (MSE), and mean absolute error (MAE), tailored to the specific requirements of each task.

11 Conclusion

This project represents a significant leap forward in the field of emotion analysis, offering new insights and methodologies for extracting emotion-cause pairs from multimodal conversational data. Our work lays the groundwork for future research and development in creating more emotionally intelligent AI systems.

12 Acknowledgements

We acknowledge the *SemEval 2024* and the organizers of the Knowledge Representation and Reasoning Task 3: The Competition of Multimodal Emotion Cause Analysis in Conversations. The organizers Rui Xia, Jianfei Yu, Fanfan Wang, Erik Cambria who provided the training data are highly appreciated. We are also indebted to our professor, Jonathan Rusert for his guidance.