Emotion Discovery & Reasoning its Flip in Conversation

Mattia Maranzana, Antonios Pantelis & Nalin Sharma

Natural Language Processing

Master's Degree in Artificial Intelligence

Table of Contents

Table of Contents

Task at Hand

Emotion and trigger classification using a BERT baseline model, for the MELD dataset:

	episode	speakers	emotions	utterances	triggers
0	utterance_0	[Chandler, The Interviewer, Chandler, The Inte...	[neutral, neutral, neutral, neutral, surprise]	[also I was the point person on my company's t...	[0.0, 0.0, 0.0, 1.0, 0.0]
1	utterance_1	[Chandler, The Interviewer, Chandler, The Inte...	[neutral, neutral, neutral, neutral, surprise,...	[also I was the point person on my company's t...	[0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0]
2	utterance_2	[Chandler, The Interviewer, Chandler, The Inte...	[neutral, neutral, neutral, neutral, surprise,...	[also I was the point person on my company's t...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, ...
3	utterance_3	[Chandler, The Interviewer, Chandler, The Inte...	[neutral, neutral, neutral, neutral, surprise,...	[also I was the point person on my company's t...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
4	utterance_4	[Joey, Rachel, Joey, Rachel]	[surprise, sadness, surprise, fear]	[But then who? The waitress I went out with la...	[0.0, 0.0, 1.0, 0.0]
...	...	...	...	...	...
3995	utterance_3995	[Chandler, All, Monica, Chandler, Ross, Chandl...	[neutral, joy, neutral, neutral, surprise, dis...	[Hey., Hey!, So how was Joan?, I broke up with...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
3996	utterance_3996	[Chandler, All, Monica, Chandler, Ross, Chandl...	[neutral, joy, neutral, neutral, surprise, dis...	[Hey., Hey!, So how was Joan?, I broke up with...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
3997	utterance_3997	[Chandler, All, Monica, Chandler, Ross, Chandl...	[neutral, joy, neutral, neutral, surprise, dis...	[Hey., Hey!, So how was Joan?, I broke up with...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
3998	utterance_3998	[Chandler, All, Monica, Chandler, Ross, Chandl...	[neutral, joy, neutral, neutral, surprise, dis...	[Hey., Hey!, So how was Joan?, I broke up with...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
3999	utterance_3999	[Chandler, All, Monica, Chandler, Ross, Chandl...	[neutral, joy, neutral, neutral, surprise, dis...	[Hey., Hey!, So how was Joan?, I broke up with...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...

4000 rows × 5 columns

Figure: Initial look at the MELD dataset.

Some Words about Triggers

- Triggers are utterances that cause an emotion flip.
- They are not in a one-to-one correspondence with emotion flips.

Speakers	Dialogue	Emotions	Triggers	Emotion Flips
Monica	u_1: You wanna hear something that sucks?	neutral	t_1: 0	No
Chandler	u_2: Do I ever.	joy	t_2: 0	No
Monica	u_3: Chris says they're closing down the bar.	sadness	t_3: 1	Yes
Chandler	u_4: No way!	surprise	t_4: 0	Yes

Table: Example of a row in MELD with identified emotion flips.

Challenges of Task

Challenges of Emotion Classification:

- Subjectivity of emotion tagging (influenced by cultural, biased perceptions, or mixed emotions).
- Contextual information is lost between raw script and actual series (e.g. an utterance of anger might not be as clear in written, as it is when one sees the expression and/or hears the tone of the person pronouncing it).

Challenges of Trigger Classification:

- Counter-intuitive way of recording emotion flips (it is the triggering phrase that is singled-out and not the phrase were the emotion flip happens).
- Not every emotion flip is caused by an utterance tagged as a trigger (e.g. the case of self-trigger emotion flips).
- In a way, the intuition for the triggers is based on the preexisting knowledge of emotions that we do not have (think of the aforementioned example).

Table of Contents

Preprocessing

- **Data cleaning:** Set some problematic values under the column "triggers" that were initially set to "*None*[1]" in the `pandas` dataframe to 0.0.
- **One-Hot Encoding:** Using the `LabelBinarizer`, we perform some one-hot encoding to the set of emotions, converting each emotion into a list of 7 entries, where each entry corresponds to one of the seven emotions that we consider in MELD. All the entries of the list in question are set to 0, but one that is set to 1 corresponding to the emotion of the respective utterance.

	episode	speakers	emotions	utterances	triggers
0	utterance_0	[Chandler, The Interviewer, Chandler, The Inte...	[[0, 0, 0, 0, 1, 0, 0], [0, 0, 0, 0, 1, 0, 0],...	[also I was the point person on my company's t...	[0.0, 0.0, 0.0, 1.0, 0.0]
1	utterance_1	[Chandler, The Interviewer, Chandler, The Inte...	[[0, 0, 0, 0, 1, 0, 0], [0, 0, 0, 0, 1, 0, 0],...	[also I was the point person on my company's t...	[0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0]
2	utterance_2	[Chandler, The Interviewer, Chandler, The Inte...	[[0, 0, 0, 0, 1, 0, 0], [0, 0, 0, 0, 1, 0, 0],...	[also I was the point person on my company's t...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, ...
3	utterance_3	[Chandler, The Interviewer, Chandler, The Inte...	[[0, 0, 0, 0, 1, 0, 0], [0, 0, 0, 0, 1, 0, 0],...	[also I was the point person on my company's t...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
4	utterance_4	[Joey, Rachel, Joey, Rachel]	[[0, 0, 0, 0, 0, 0, 1], [0, 0, 0, 0, 0, 1, 0],...	[But then who? The waitress I went out with la...	[0.0, 0.0, 1.0, 0.0]
...	...	...	...	...	...
3995	utterance_3995	[Chandler, All, Monica, Chandler, Ross, Chandl...	[[0, 0, 0, 0, 1, 0, 0], [0, 0, 0, 1, 0, 0, 0],...	[Hey., Hey!, So how was Joan?, I broke up with...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
3996	utterance_3996	[Chandler, All, Monica, Chandler, Ross, Chandl...	[[0, 0, 0, 0, 1, 0, 0], [0, 0, 0, 1, 0, 0, 0],...	[Hey., Hey!, So how was Joan?, I broke up with...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
3997	utterance_3997	[Chandler, All, Monica, Chandler, Ross, Chandl...	[[0, 0, 0, 0, 1, 0, 0], [0, 0, 0, 1, 0, 0, 0],...	[Hey., Hey!, So how was Joan?, I broke up with...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
3998	utterance_3998	[Chandler, All, Monica, Chandler, Ross, Chandl...	[[0, 0, 0, 0, 1, 0, 0], [0, 0, 0, 1, 0, 0, 0],...	[Hey., Hey!, So how was Joan?, I broke up with...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
3999	utterance_3999	[Chandler, All, Monica, Chandler, Ross, Chandl...	[[0, 0, 0, 0, 1, 0, 0], [0, 0, 0, 1, 0, 0, 0],...	[Hey., Hey!, So how was Joan?, I broke up with...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...

4000 rows × 5 columns

Figure: The MELD dataset after some preprocessing.

[1] "*NaN*" in the .json file.

Preprocessing

- **Data splitting:** We performed the recommended 80/10/10 split.
- **Tokenization:** Chose `BertTokenizer` as our *tokenizer* and 'bert-base-uncased' as our *model card*. The tokenization is performed using the function `tokenize_padding`. Firstly, it appends each speaker to their corresponding utterance. Secondly, it performs the tokenization and pads the tokenized dialogues into a maximum length of 128. Finally, it converts the tokenized dialogues into `Pytorch` tensors and returns the `input_ids`, as well as the `attention_mask`.
- Custom `datasets` class, called `MyDataset`, that receives as inputs the previously mentioned `input_ids`, and `attention_mask`, as well as our to label classes `emotions` and `triggers`.

Table of Contents

Custom BERT Model

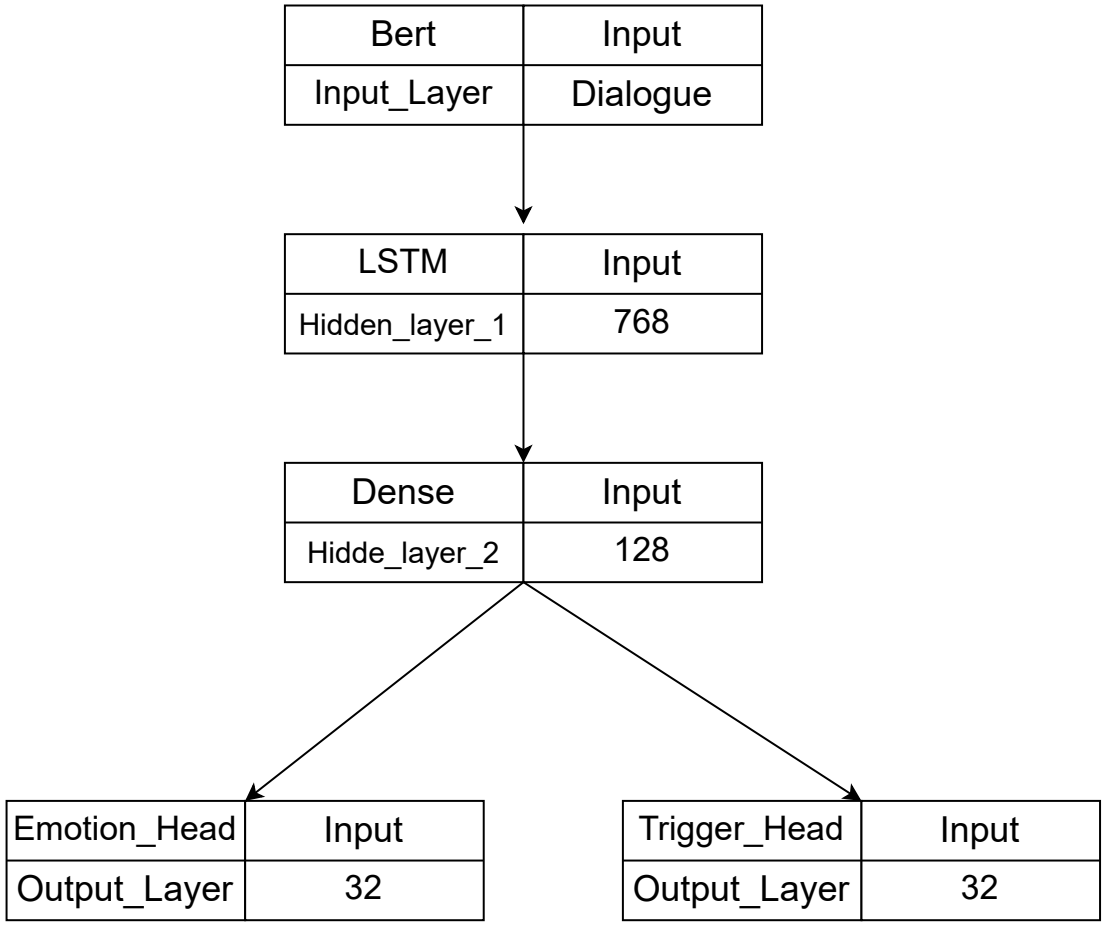

Figure: Architecture of our Custom BERT Model.

Custom BERT Model

Intution for the Additional Layers

We added:

- The Bidirectional LSTM layer to capture more context.
- The Linear layer on top to highlight any sequential dependencies, critical in understanding emotional nuances and triggering events.

Seed	Avg F1 Emotion	Avg F1 Trigger	U-F1 Emotion	U-F1 Trigger
42	0.528025	0.494125	0.539914	0.517256
45	0.505046	0.514643	0.457864	0.536982
156	0.541744	0.469595	0.522171	0.503125
550	0.537196	0.533296	0.503804	0.520581
777	0.551572	0.483456	0.540413	0.482436

Table: BERT Custom model performance across different seeds.

Table of Contents

Initial Experimentation: Emotion Classification

- **Phase 1 - Emotion Classification:** Our initial focus was on classifying emotions in dialogues. As an experiment, we used this simpler classification to inform/help the trigger classification task looking for better results, but this did not go to plan. We also used this simpler configuration with one classifier head to assist us in developing a full working model.
- **Phase 2 - Emotion and Trigger Detection:** Subsequently, we expanded our scope to include trigger detection alongside emotion classification, which yielded better results and was the approach that we used for testing all our models.

Tokenization, Padding, and Truncation

- **Tokenization Strategies:** Various tokenization setups were tested to determine the optimal approach for our model inputs.
- **Padding and Truncation:** After experimentation, we opted for enabling padding (padding = true) to ensure consistent sequence lengths across inputs, crucial for batch processing. Truncation was not applied (truncation = false), to preserve the integrity of our dialogue data.

XLNet in Emotion Discovery and Reasoning

- XLNet's theoretical advantage: permutation-based training, bidirectional context, theoretically more powerful than BERT.
- **CustomXLNetModel:** Merges XLNet's contextual awareness with LSTM's temporal dynamics, similar to our CustomBERTModel
- **Challenges Encountered:** In practice, XLNet's performance was not superior to BERT for detecting emotion flips in conversations.
- **Insights:** The complexity of conversational emotion recognition may require more than XLNet's advanced pre-training techniques.

XLNet Experiment Results

Below, we report the result of our XLNet implementation:

```
Average Sequence F1 (Emotion):                0.300681
Average Sequence F1 (Trigger):                0.263727
Unrolled Sequence F1 (Emotion):               0.188067
Unrolled Sequence F1 (Trigger):               0.262602
```

RoBERTa in Emotion Discovery and Reasoning

- RoBERTa's refinement: Reimplementation of BERT with modifications to key hyperparameters and tiny embedding tweaks.
- **Model and Adaptation:** Employed the 'roberta-base' model card to leverage RoBERTa's enhanced pre-training. This particular model had 125M parameters, 'bert-base-uncased' has 110M parameters for reference. Note: 'roberta-large' could not be utilized.
- **Tokenizer:** Utilized AutoTokenizer for the 'roberta-base' card, appropriate to the task at hand.
- **Comparative Performance:** Despite improvements, RoBERTa did not surpass our tailored BERT model in capturing emotion transitions.
- **Insights:** Both 'roberta-base' and our custom BERT model were fine-tuned on the same dataset, yet BERT performed better. This outcome suggests that our BERT's specific adjustments and pre-training alignment with our task might have provided an edge over 'roberta-base'. Despite 'roberta-base's' advancements, it shows that newer models do not automatically ensure superior results for all tasks, highlighting the significance of model customization to the task at hand.

RoBERTa Experiment Results

Below, we report the result of our RoBERTa implementation:

```
Average Sequence F1 (Emotion):          0.443786
Average Sequence F1 (Trigger):          0.497131
Unrolled Sequence F1 (Emotion):         0.403676
Unrolled Sequence F1 (Trigger):         0.525890
```

Optimizer and Loss Function Selection

- **Optimizer - AdamW:** Chosen for its advanced regularization, improving training stability over Adam by adjusting weight decay. We also used the LR Scheduler 'ReduceLROnPlateau' to dynamically adjust Learning Rate based on the loss values for each epoch.
- **Emotion Loss:** CrossEntropyLoss is ideal for the 7-class emotion task, efficiently managing class probability distribution.
- **Trigger Loss:** CrossEntropyLoss with weights to counter-act the imbalance of the trigger distribution.
- **Rationale:** These loss functions were selected after testing various options, aligning with the multi-class and binary classification needs of our tasks.

Hyperparameter Tuning via Grid Search

- **Objective:** Optimize model performance on four key F1 metrics—average and unrolled sequence F1 for both emotion and trigger detection—using grid search.
- **Tested Parameters:**
  - *Learning Rate* and *Weight Decay* in the optimizer for balancing learning speed and regularization.
  - *Batch Size* to find the optimal number of samples that yields the most stable and accurate learning.
  - *Embedding Layer Freezing* to determine the impact of keeping pre-trained embeddings static vs. allowing further training.
  - *Dropout Rate Values* to identify the best rate for preventing overfitting while maintaining model learning capacity.
- **Method:** Parameters varied in a grid search to evaluate their impact on the validation set F1 scores, guiding the selection of the best setup for final testing across 5 seeds for reliability.
- **Outcome:** This exhaustive approach allowed us to fine-tune our model with a configuration that maximizes performance, evidenced by improved F1 scores across metrics.

Table of Contents

Model Comparison

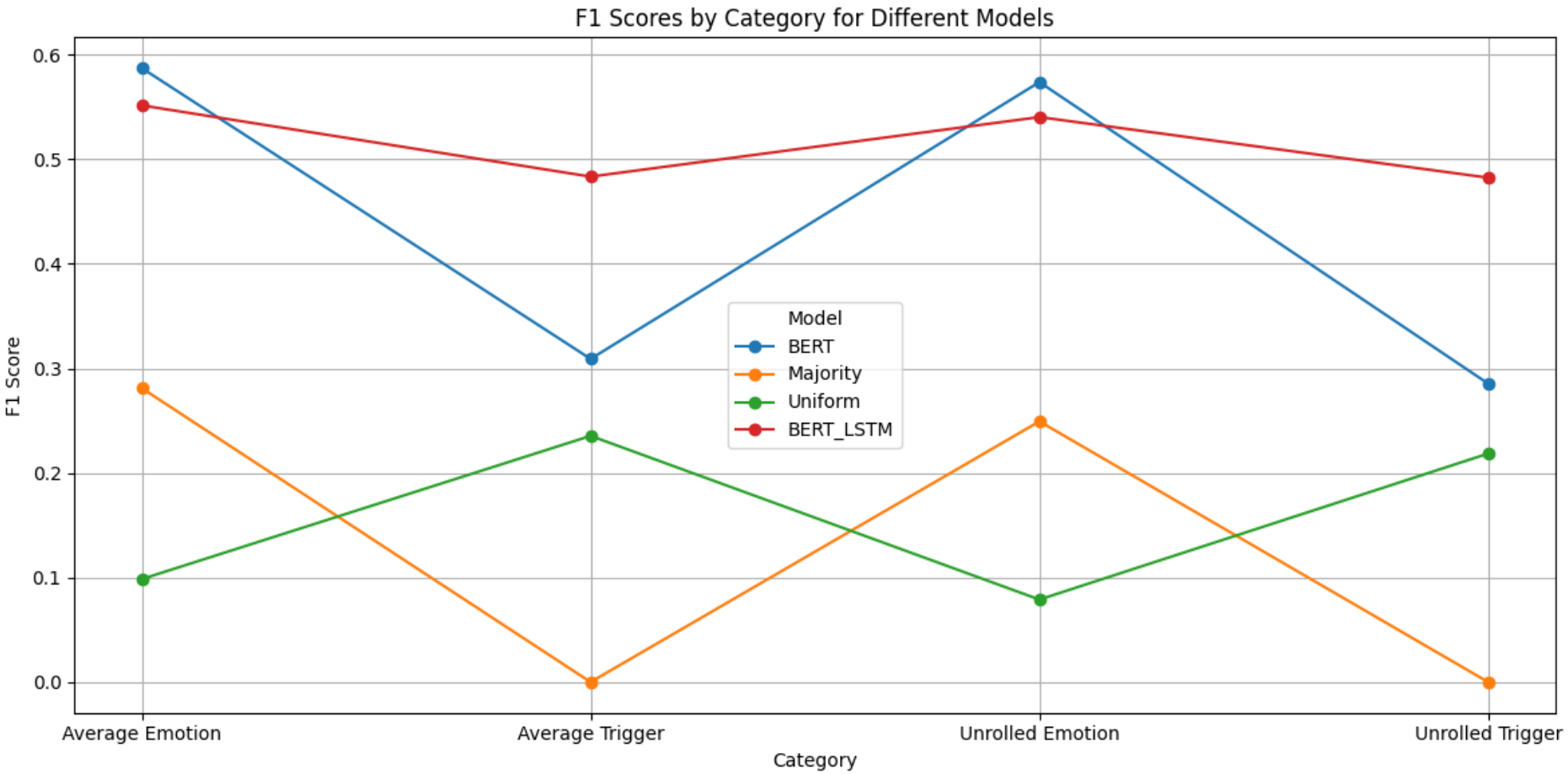

Figure: Comparison of different F_1 scores for our different models.

Table of Contents

Error Analysis

Speaker	Dialogue	Ground Truth	Prediction
Rachel	(D_{3612}, u_1): Oh God, I hate my job, I hate it, I hate my job, I hate it.	disgust	anger
Rachel	(D_{3612}, u_{10}): Oh well then, so I'm just going to go back...	joy	neutral
Stanley	(D_{3989}, u_{10}): I don't know.	neutral	neutral
Stanley	(D_{3989}, u_{11}): A week?	anger	neutral
Stanley	(D_{3989}, u_{12}): Maybe two?	neutral	neutral

Table: Reasoning of misclassifications of emotions for two dialogues of the testing set.

Error Analysis

Speakers	Dialogue	Ground Truth	Prediction
Monica	u_1: Hi! Umm, I'm Monica Geller, I'm the chef at Alessandro's.	neutral, 0	neutral, 0
Food Critic	u_2: Still?	surprise, 0	surprise, 0
Monica	u_3: I think the things that you said about me are really unfair...	neutral, 0	neutral, 1
Food Critic	u_4: I don't see any reason why I would do that to myself again.	joy, 0	neutral, 0
Joey	u_5: Either eat it, or be in it.	anger, 0	neutral, 0
Monica	u_6: Spoon? So, what do you think?	joy, 0	neutral, 0
Food Critic	u_7: I'm torn... But I must be honest, your soap is abysmal.	disgust, 0	neutral, 1
Joey	u_8: Thata girl! Huh? We should get out of here...	joy, 0	surprise, 0
Cooking Teacher	u_9: Welcome to introduction to cooking...	neutral, 1	neutral, 0
Monica	u_{10}: I can.	neutral, 0	neutral, 0

Table: Reasoning of misclassifications of triggers for dialogue D_{3624} of the testing set.

Table of Contents

Limitations

- **Misclassification Challenges:** Contradictory labeling for the triggers, subjectivity of emotion classification.
- **Compute Budget:** Google Colab, Kaggle with one or two Nvidia T4 GPUs respectively. Limited use of model cards.
- **Data Volume and Quality:** Our model faced limitations due to a dataset of only 4000 labeled data points, insufficient for capturing the full spectrum of human emotion and dialogue complexity.

Table of Contents

Conclusion

- **Multimodal Analysis:** Since the dialogues are taken from a series, we also have other information that is not exploited. A good idea would be to perform a multimodal analysis, including *sound* and/or *image*.
- **Alternative Tagging of Triggers:** For a more significant emotion flip analysis and reasoning with clearer patters for the model to identify, we could think of an alternative tagging system. For example, one where in the last utterance of each speaker we *record*[2] the presence or absence of an emotion flip.
- **Alternative Architectures:** We could experiment with more interesting architectures that do not necessarily perform the classification of emotions and triggers at the same time, but prioritize one and appending the predictions to the utterances, pass an updated input with more contextual information for the remaining classification task.

[2] Put 1 if there has been an emotion flip for the speaker in question and 0 if not.