NLP PROJECT

Multimodal Cause & Effect Detection (Joy)

Aadit Sharma - 2101AI02

Akash Sinha - 2101CS90

Atul Pande - 2101CS88

Problem Statement

The problem statement was majorly divided into 3 tasks:

Dataset Collection

1000 video conversations that portray “joy” emotion. Deriving the audio and textual features from them.

Annotation

Labelling frames, audio parts and text segments into emotion clause and cause clause

Model

Using a transformer model to predict emotion and its cause

Dataset Research

We researched and found 2 multi-modal datasets (audio, video and text) relevant to our task

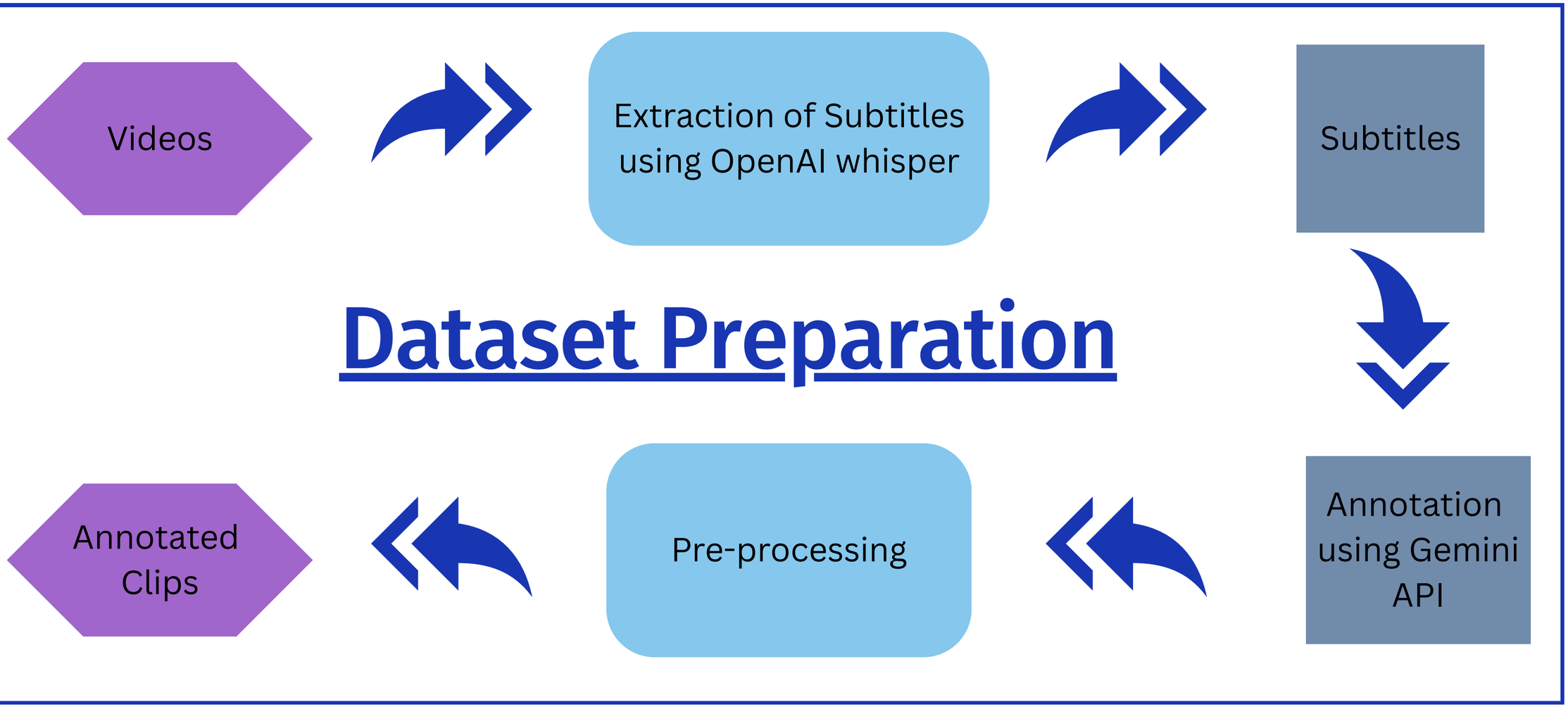
IEMOCAP Dataset:

Each segment is annotated for nine emotions (angry, excited, fear, sad, surprised, frustrated, happy, disappointed, and neutral) along with valence, arousal, and dominance.

MELD Dataset:

- MELD includes dialogues with multiple speakers.
- Each utterance is annotated with emotion and sentiment labels.

Dataset Collection



Dataset

Video Clip (12.mp4)

OpenAI whisper

```
1
00:00:00,000 --> 00:00:02,000
You had the tooth knocked out this time.

2
00:00:02,000 --> 00:00:03,000
Oh, yes.

3
00:00:03,000 --> 00:00:06,500
I had a concussion, whiplash, stitches in my hand,

4
00:00:06,500 --> 00:00:07,500
dislocated shoulder.
```

SRT file (subtitles12.srt)

Gemini API

Annotated (res12.csv)

timestamp_range	cause_text	emotion_text
00:01:00,000 --> 00:01:03,000	But I have a 10 month old son.	So, it still works.
00:02:21,000 --> 00:02:23,000	I had an injury.	I got a rubber stuck on my nose.
00:02:54,000 --> 00:03:09,000	My brother dared me to put a pencil up my nose.	LAUGHTER
00:06:36,000 --> 00:06:43,000	Did you see how politely though we handled that?	Oh, that's so good.
00:06:46,000 --> 00:06:48,000	Oh, my God.	It's my favorite.
00:09:01,000 --> 00:09:05,000	And I had chosen a day when all of a sudden, randomly,	Oh!

Preprocess

video_name	cause_text	emotion_text	timestamp_range
12-001.mp4	But I have a 10 month old son.	So, it still works.	00:01:00,000 --> 00:01:03,000
12-002.mp4	I had an injury.	I got a rubber stuck on my nose.	00:02:21,000 --> 00:02:23,000
12-007.mp4	Stop kicking me under the table.	And I've got my hair in his foot.	00:04:00,000 --> 00:04:05,000
12-009.mp4	Did you see how politely though we handled that?	Oh, that's so good.	00:06:36,000 --> 00:06:43,000
12-010.mp4	Oh, my God.	It's my favorite.	00:06:46,000 --> 00:06:48,000
12-011.mp4	And I had chosen a day when all of a sudden, randomly,	Oh!	00:09:01,000 --> 00:09:05,000

final.csv (clips)

Intuition

Outline for model training after annotation.

Subtitle generation

To obtain the srt files containing the dialogues, openai/whisper was used, subsequently producing the subtitles separating every sentence.

Annotate utterances

The collected dataset was passed through gemini-2.0-flash to annotate utterances in a particular dialogue as either the emotion clause or as the cause clause.

Classification Problem

Treat this as a classification problem by passing the emotion/cause pair embeddings

Sample Preparation

video_name	cause_text	emotion_text	timestamp_range
12-009.mp4	Did you see how politely though we handled that?	Oh, that's so good.	00:06:36,000 --> 00:06:43,000

Processed annotated clip

12-009-001.mp4	12-009.mp4	12.mp4	Did you see how politely though we handled that?	0:06:36 --> 0:06:39

Utterance 1

12-009-002.mp4	12-009.mp4	12.mp4	Because we thought you would lost your mind, though.	0:06:39 --> 0:06:41

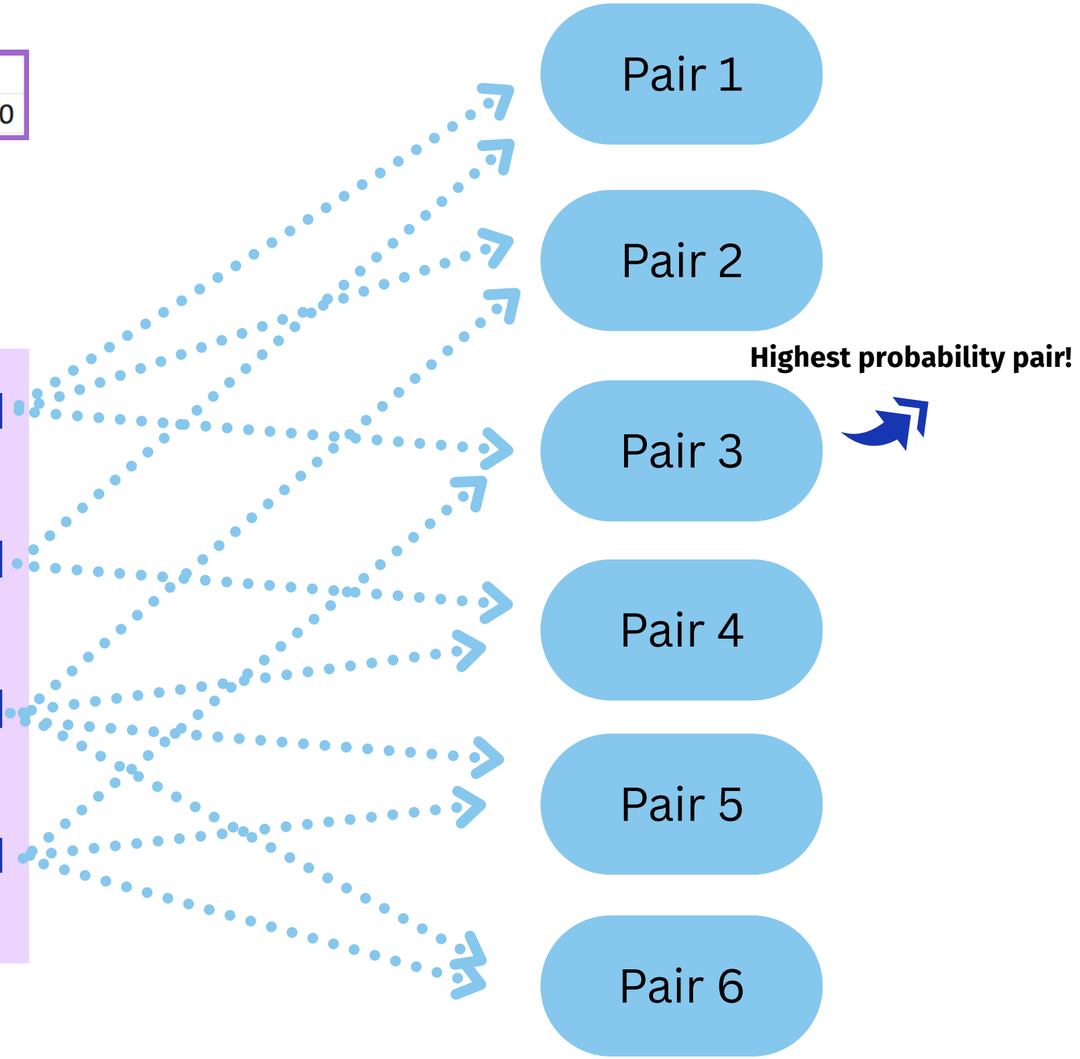
Utterance 2

12-009-003.mp4	12-009.mp4	12.mp4	Yeah.	0:06:41 --> 0:06:42

Utterance 3

12-009-004.mp4	12-009.mp4	12.mp4	Oh, that's so good.	0:06:42 --> 0:06:43

Utterance 4



Training Process

A brief sequence on how the model will be trained

Transformer + Classification layer was used to make predictions

Classification Problem

cause-effect detection task as a binary classification problem → classifying a pair of utterances as joy cause/effect

Embeddings

Extracted multimodal embeddings for each utterance using the models mentioned before and then concatenated them.

Classification Input

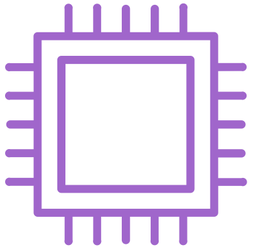
Created pairs of utterances as input samples for classification

Model Training

Trained our model to classify whether each pair exhibits a cause-effect relationship.

Model Components

The models used for the multimodal embeddings

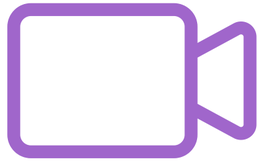


Text

Sentence-Transformer
“all-mpnet-base-v2”

Audio

AST
Audio Spectrogram Transformer

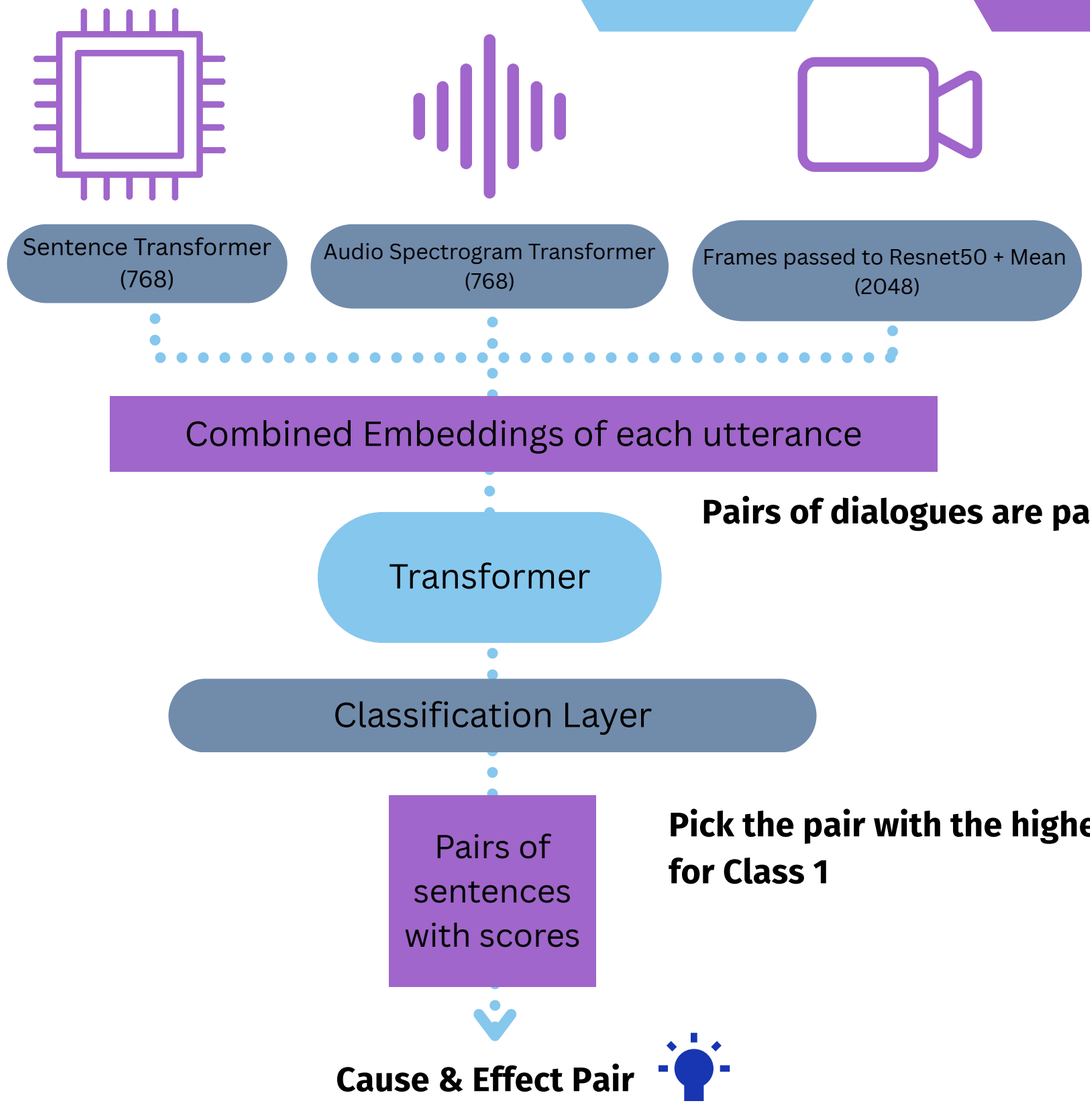


Video

ResNet50
Took mean of
features of frame

Methodology

Video Clip



CLASSIFICATION DETAILS

During the train-test split, we conducted our experimentation on 2 levels - segregating on the basis of video level and on clip level.

Based on the classification results we pair up all the utterances present in a single clip and then predict the probability that a particular utterance pair would be our actual Cause/Effect pair!
We mark the pair giving the highest probability and use it to find the accuracy for all clips given.

Results (Video-level)

```
Accuracy: 55.43%

Weighted Metrics:
  Precision: 0.5687
  Recall: 0.5543
  F1 Score: 0.5299

Per-Class Metrics:
  Class 0 - Precision: 0.6000, Recall: 0.3261, F1: 0.4225
  Class 1 - Precision: 0.5373, Recall: 0.7826, F1: 0.6372

Classification Report:
              precision    recall  f1-score   support

           0       0.60      0.33      0.42        92
           1       0.54      0.78      0.64        92

    accuracy                           0.55       184
   macro avg       0.57      0.55      0.53       184
weighted avg       0.57      0.55      0.53       184


Confusion Matrix:
[[30 62]
 [20 72]]
```

```
Confusion Matrix Explanation:
  True Negatives (Class 0 correctly predicted): 30
  False Positives (Class 0 incorrectly predicted as Class 1): 62
  False Negatives (Class 1 incorrectly predicted as Class 0): 20
  True Positives (Class 1 correctly predicted): 72

Test Results Summary:
  Loss: 0.7391
  Accuracy: 55.43%
  Class 0 - Precision: 0.6000, Recall: 0.3261, F1: 0.4225
  Class 1 - Precision: 0.5373, Recall: 0.7826, F1: 0.6372
```

Classification Results

```
Clip-level Accuracy: 0.8272
Total clips evaluated: 81
```

Final Prediction Results

Results (Clip-level)

```
Accuracy: 66.36%

Weighted Metrics:
  Precision: 0.6688
  Recall: 0.6636
  F1 Score: 0.6610

Per-Class Metrics:
  Class 0 - Precision: 0.6392, Recall: 0.7512, F1: 0.6907
  Class 1 - Precision: 0.6983, Recall: 0.5760, F1: 0.6313

Classification Report:
              precision    recall  f1-score   support

           0       0.64      0.75      0.69       217
           1       0.70      0.58      0.63       217

    accuracy                           0.66       434
   macro avg       0.67      0.66      0.66       434
weighted avg       0.67      0.66      0.66       434


Confusion Matrix:
[[163  54]
 [ 92 125]]
```

```
Confusion Matrix Explanation:
  True Negatives (Class 0 correctly predicted): 163
  False Positives (Class 0 incorrectly predicted as Class 1): 54
  False Negatives (Class 1 incorrectly predicted as Class 0): 92
  True Positives (Class 1 correctly predicted): 125

Test Results Summary:
  Loss: 1.0640
  Accuracy: 66.36%
  Class 0 - Precision: 0.6392, Recall: 0.7512, F1: 0.6907
  Class 1 - Precision: 0.6983, Recall: 0.5760, F1: 0.6313
```

Classification Results

```
Clip-level Accuracy: 0.8632
Total clips evaluated: 190
```

Final Prediction Results

Contributions

Aadit Sharma - 2101AI02

- Annotated clips using Gemini to identify and separate cause-effect pairs
- Extracted multimodal embeddings from each segment
- Contributed to model architecture design and implementation

Akash Sinha - 2101CS90

- Collected half of the video dataset from YouTube.
- Generated the srt files for the video clips using whisper.
- Prepared classification pairs for cause-effect detection in training.

Atul Pande - 2101CS88

- Collected half of the video dataset using YouTube.
- Segmented videos into utterance-wise clips
- Collaborated on model architecture development and implementation

Thank You