Multimodal Emotion Recognition in Conversation

Athiya Deviyani, Abuzar Khan, Njall Skarphedinsson, Prasoon Varshney
11-777 Multimodal Machine Learning - Team 11

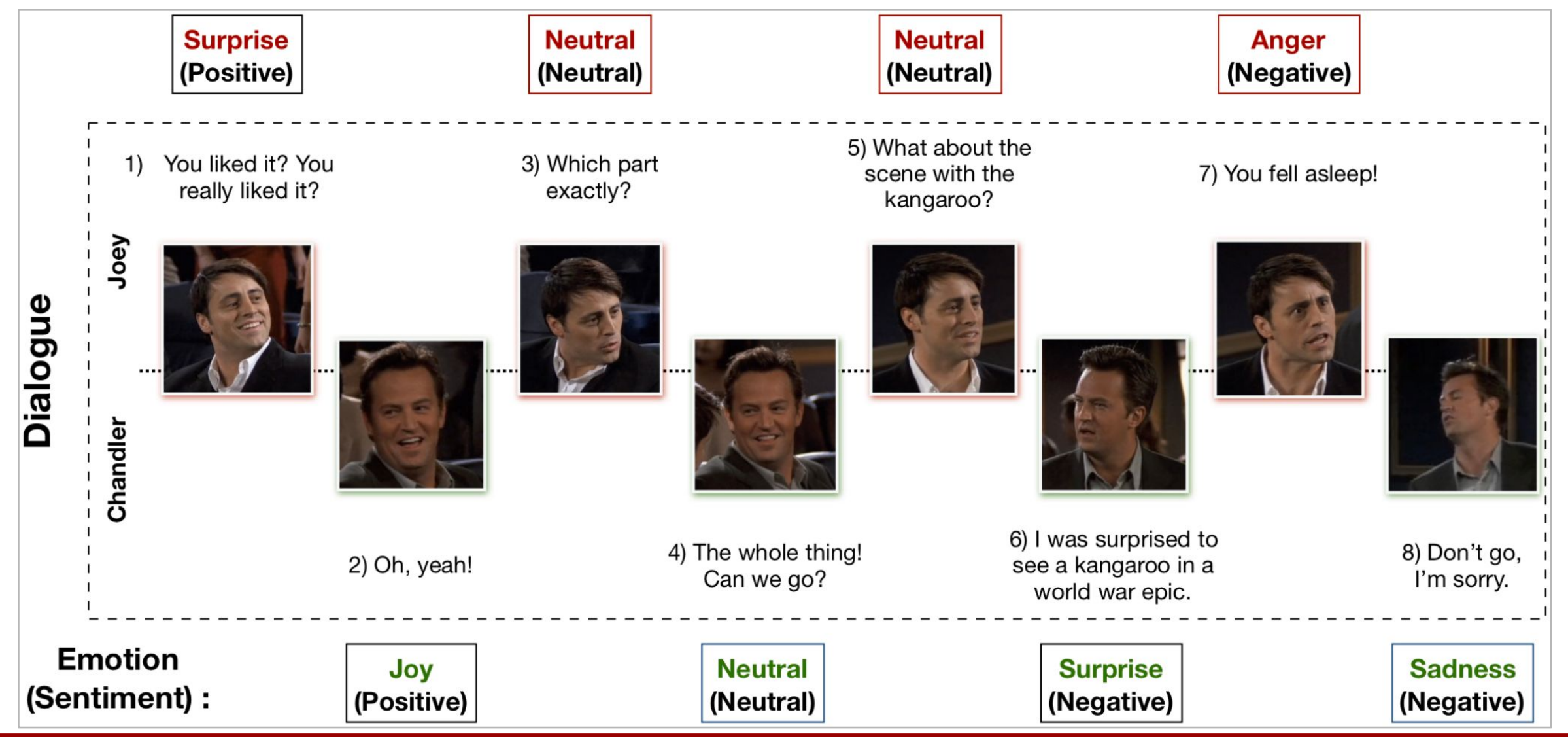
Motivation

- Emotion recognition in conversation is inherently multimodal, as **humans express emotional cues across various modalities** such as language, facial expressions, and speech
- **Multimodal emotion recognition in conversation (mERC):** using more than one modality, identify different emotions at each turn within a conversation, where there is more than one person participating in the conversation (multi-party)
- Large **potential applications in many challenging tasks** such as dialogue generation, behavior understanding, and multimodal interaction in various domains such as healthcare
- **Project goal:** explore techniques to improve cross-modal information sharing for affect recognition

Dataset

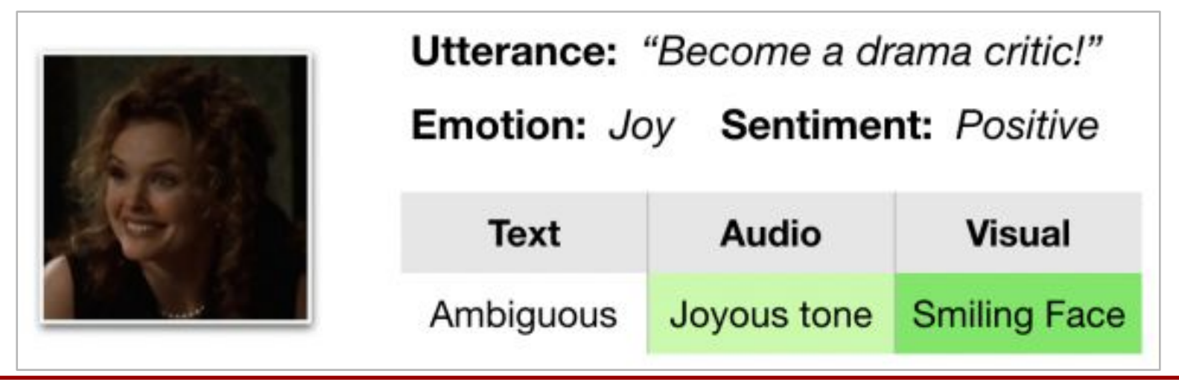
Multimodal EmotionLines Dataset (MELD) [1]

- More than **1400 multi-party dialogues** and **13000 utterances** from the Friends TV series (**modalities:** text, audio, and video)
- Each utterance in a dialogue has **emotion** and **sentiment** labels
  - Emotions: anger, disgust, sadness, joy, neutral, surprise, fear
  - Sentiments: positive, negative, neutral



Challenges

- A conversation in a mERC task often contains different speakers and goes over several turns, therefore **emotions can vary drastically during the conversation**
- Current state-of-the-art architecture utilizes the **Supervised Prototypical Contrastive Learning (SPCL)** loss [2], however it only uses the text modality
- The main challenges in our project are:
  (1) adapting existing SOTA architectures to multiple modalities
  (2) exploring approaches for appropriate alignment and fusion



Utterance: *"Become a drama critic!"*
Emotion: *Joy* **Sentiment:** *Positive*

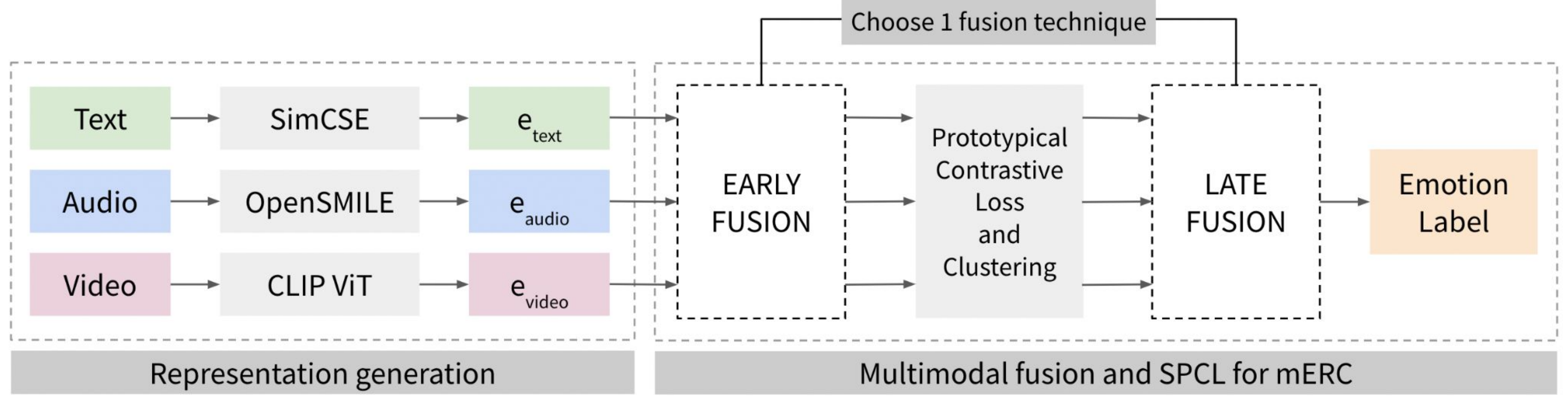
Text	Audio	Visual
Ambiguous	Joyous tone	Smiling Face

Utterance: *"Great, now he is waving back"*
Emotion: *Disgust* **Sentiment:** *Negative*

Text	Audio	Visual
Positive/Joy	Flat tone	Frown

Research ideas

Multimodal Supervised Prototypical Contrastive Learning (MM-SPCL)


Choose 1 fusion technique
Text
SimCSE
e_{text}
Audio
OpenSMILE
e_{audio}
Video
CLIP ViT
e_{video}
EARLY FUSION
Prototypical Contrastive Loss and Clustering
LATE FUSION
Emotion Label
Representation generation
Multimodal fusion and SPCL for mERC

Representation generation

Modality representations were generated through the following:

- **Audio:** extract features using OpenSMILE and L2-based selection
- **Video:** select mid-utterance frame; obtain features using CLIP ViT
- **Text:** prompt-based technique using SimCSE

  `$C_t \oplus P_k$ [ for $u_k$, $s_k$ feels **<mask>** ]`

  Where C_t is context, P_k is prompt, u_k is utterance, and s_k is speaker

Multimodal fusion approaches

- **Early:** Before prototypical contrastive loss and clustering. Experimenting with the following approaches:
  - Concatenation and **L**inear **P**rojection (LP)
  - Pairwise **C**ross-**M**odal Attention (CM) [3]
  - **M**ultimodal **B**ottleneck **T**ransformer (MBT) [4]

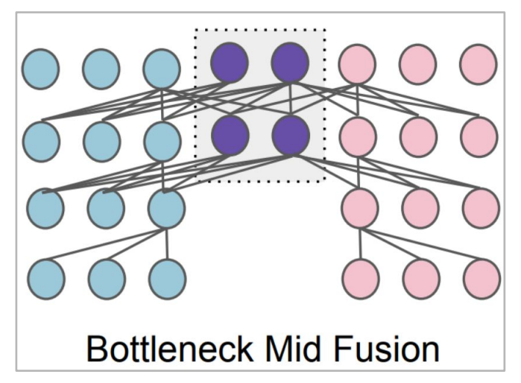


Image Source: [4]

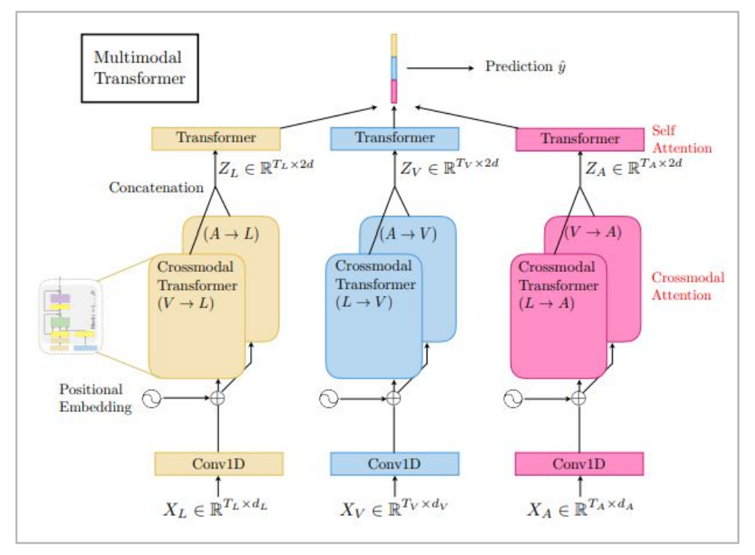


Image Source: [3]

- **Late:** Residual learning for non-additive bimodal and trimodal interactions

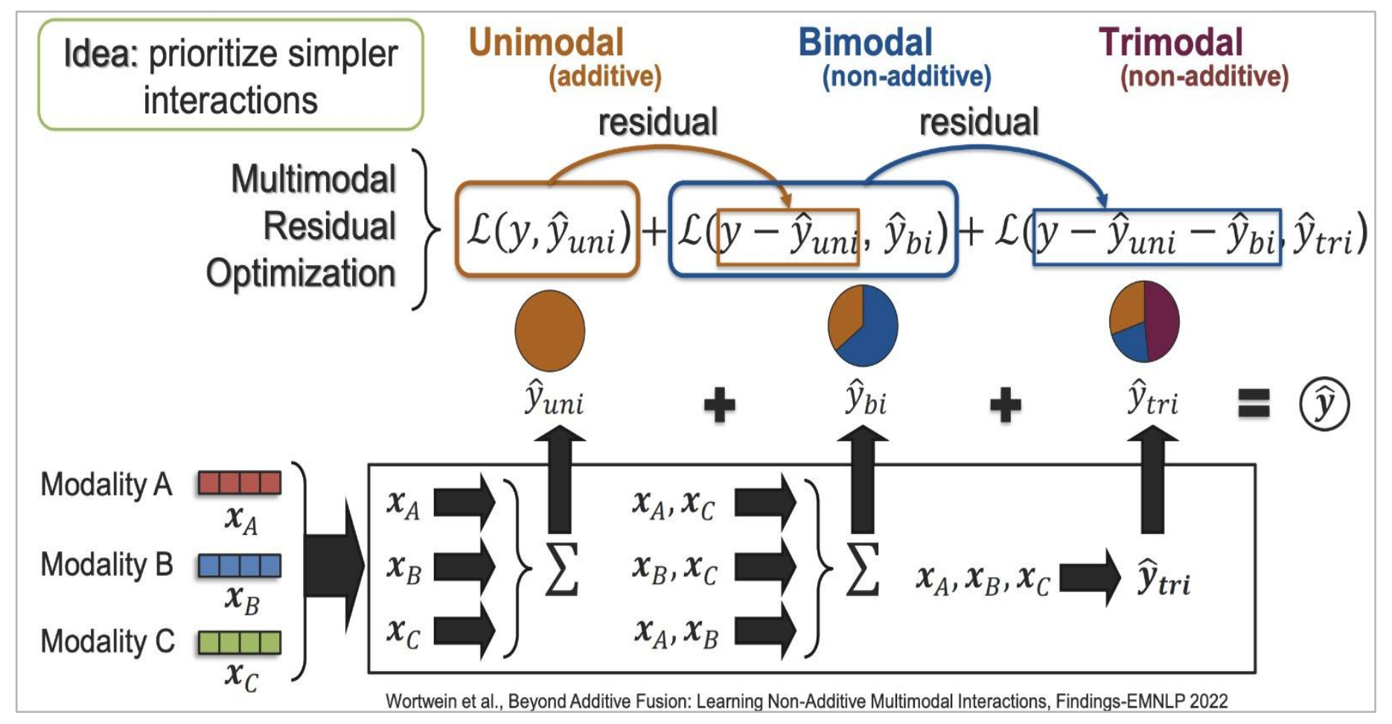

Idea: prioritize simpler interactions
Unimodal (additive)
Bimodal (non-additive)
Trimodal (non-additive)
residual
residual
Multimodal Residual Optimization
$\mathcal{L}(y,\hat{y}_{uni}) + \mathcal{L}(y-\hat{y}_{uni},\hat{y}_{bi}) + \mathcal{L}(y-\hat{y}_{uni}-\hat{y}_{bi},\hat{y}_{tri})$
$\hat{y}_{uni}$ + $\hat{y}_{bi}$ + $\hat{y}_{tri}$ = $\hat{y}$
Modality A x_A
Modality B x_B
Modality C x_C
x_A x_B x_C $\sum$
x_A, x_C x_B, x_C x_A, x_B $\sum$
x_A, x_B, x_C $\hat{y}_{tri}$
Wortwein et al., Beyond Additive Fusion: Learning Non-Additive Multimodal Interactions, Findings-EMNLP 2022

Image Source: Lecture 12.2

Supervised Prototypical Contrastive Learning (SPCL)

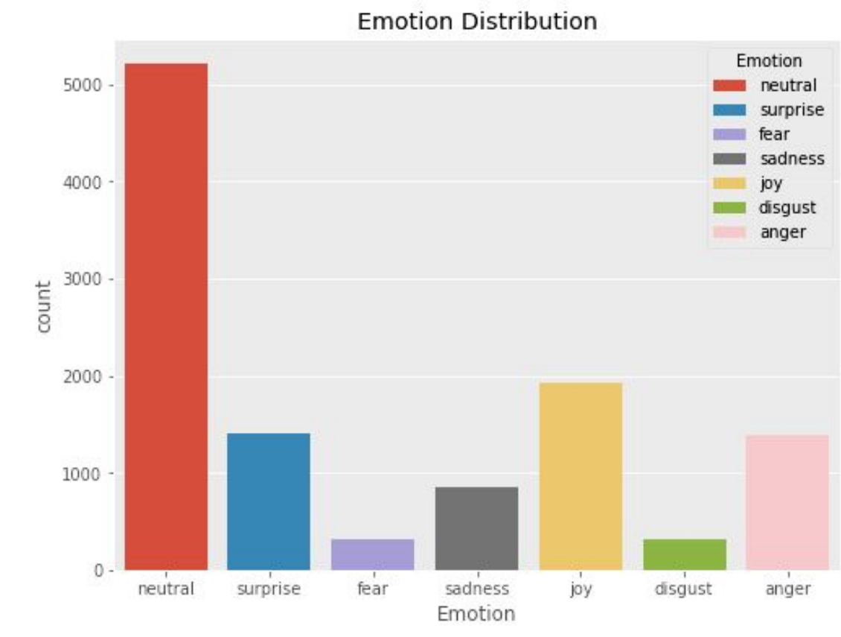
- The SPCL framework uses a **contrastive loss**; it treats same label examples within batch as positive, the rest negative
- The MELD dataset exhibits a heavy **class imbalance**, which can be particularly detrimental for contrastive loss objectives
- **Prototypical learning** is used to introduce vectors from each class in each batch by sampling from a support set
- Further, **curriculum learning** is also used to order the train set in order from easiest to hardest instances
- Difficulty is measured as distance of an instance cluster centers

Results

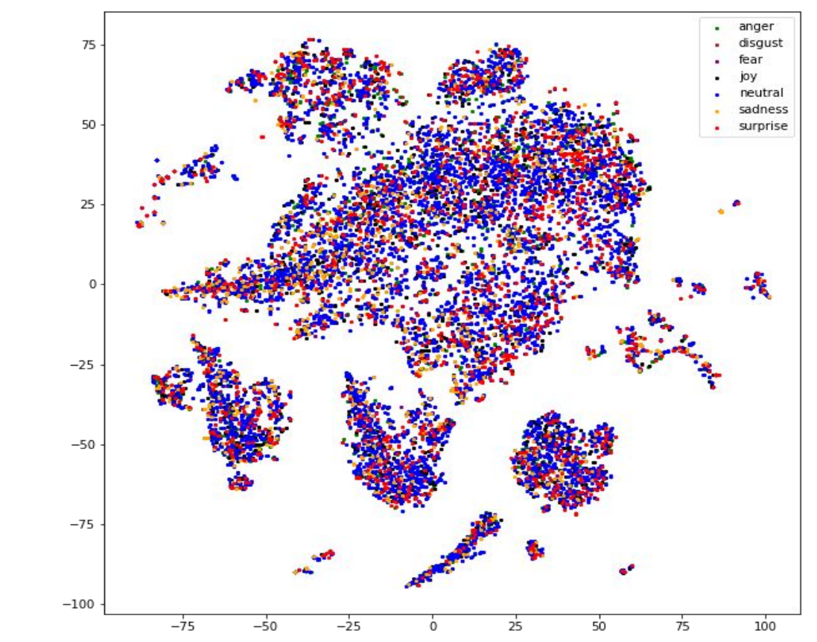
Model {Modalities} (Fusion approach* in parenthesis)	F1 Score (per approach)
SPCL {t}	0.6627
MM-SPCL {t, a} (LP, CM)	(0.6587, 0.6551)
MM-SPCL {t, v} (LP, CM)	(0.6622, 0.6518)
MM-SPCL {t, a, v} (LP, CM)	(0.6543, 0.6514)

*Results for MBT Fusion and Residual Optimization approaches in progress

Analysis

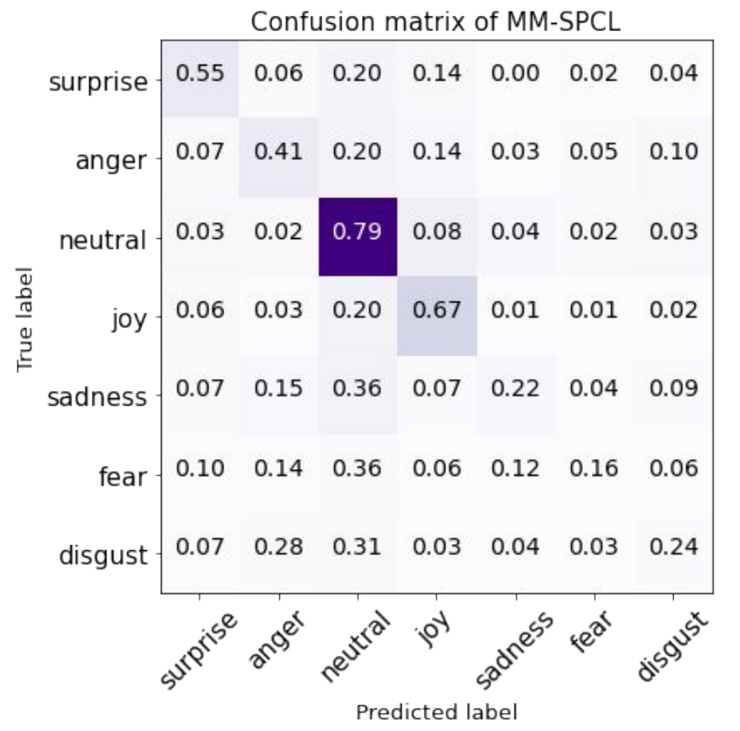


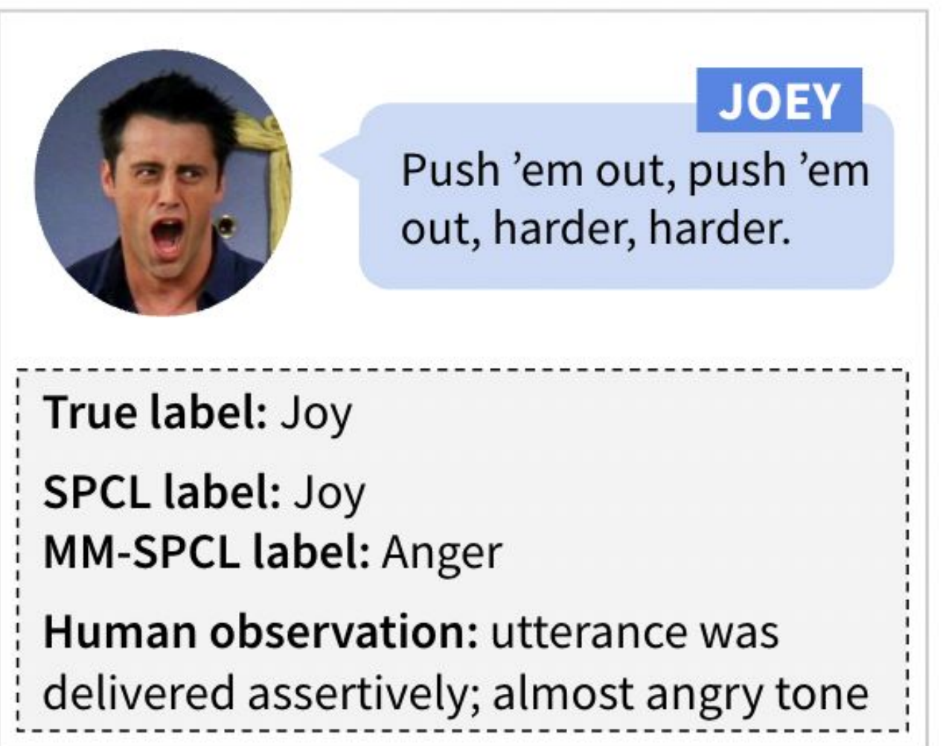
Class Imbalance in MELD



Inseparable Audio Features

When is someone *really* happy?
We found that the Multimodal SPCL models joy and disgust much better than its unimodal counterpart, yet it models surprise, fear and sadness more poorly.





True label: Joy
SPCL label: Joy
MM-SPCL label: Anger
Human observation: utterance was delivered assertively; almost angry tone

MM-SPCL correctly disagrees with SPCL!
This dialogue exemplifies the lack of context in audio, and how this results in erroneous predictions when its effect dominates. Upon listening to the audio of each utterance in isolation, the tone of the speaker does indeed sound assertive (almost angry), but from context it is clear that the scene depicts a joyous moment.

Future work

- Obtain video and audio representations using a **speaker-aware** and **context-aware** methodology, similar to how the text representations are obtained in the original SPCL paper [2]
- **Identify biases** in emotion recognition performance that may stem from the stereotypical portrayals of characters, e.g. female characters being portrayed as being more emotional/dramatic
- Use **explainability and interpretability techniques** for multimodal models such as SHAP and LIME to further **understand the contribution made from each modality** towards the final emotion classification

[1] Poria, Soujanya, et al. "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversation." (2018)
[2] Song, Xiaohui, et al. "Supervised Prototypical Contrastive Learning for Emotion Recognition in Conversation." (2022)
[3] Tsai, Yao-Hung, et al. "Multimodal Transformer for Unaligned Multimodal Language Sequences" (2019)
[4] Nagrani, Arsha, et al. "Attention bottlenecks for multimodal fusion." (2021)