

[SLIDE 1]

Hello, everyone. Today we're presenting our project titled multi-channel sentiment analysis. This project explores an alternative approach to emotion detection by integrating facial expression, voice tone, and speech transcription, each analyzed independently before being combined through a custom voting mechanism. Note that by sentiment, we mean positive or negative emotion.

[SLIDE 2]

Historically, sentiment analysis often relies on a single data channel, like text, and often fails to capture the full complexity of human emotions. People express emotions not only through words but also through facial expressions and voice tone. For example, someone might be sad but mask it with a smile or sound angry but is sarcastic joking. Single-channel systems can miss these subtleties.

[SLIDE 3]

This has led to the rise of multimodal approaches, which use neural networks to combine text, voice, and visual data for better emotion recognition. However, current multimodal techniques often face technical challenges like dealing with noise, missing data, the "black box" nature of neural networks, making it hard to interpret, and most importantly, complex emotions create conflicting output from each model, causing misclassifications.

[SLIDE 4]

Our project offers an alternative solution. Instead of solely relying on neural networks, we analyze each data channel independently and use a customized voting mechanism to combine results. This makes our system more transparent, explainable, and customizable, offering greater control in handling edge cases by using custom rules based on human psychology. It's important to note that our system is not meant to replace multimodal systems, but rather to complement them. In situations where multimodal techniques are effective, they should still be used.

[SLIDE 5]

Now, let's talk about the framework of our system.

[SLIDE 6]

First, the input of the system is a short pre-recorded video file of about 30 seconds of a person speaking. The video file contains both video and sound information, which are then separated.

[SLIDE 7]

Next, the separated video is put into the face analysis model. The sound is put into the voice tone analysis model, as well as converted into transcription and put into the transcription analysis model. The models can be downloaded from websites such as huggingface.com and we do not train our own models.

[SLIDE 8]

After getting the output from each model, we apply the voting. This is the core of our system. The voting mechanism will employ a custom algorithm, which is yet to be developed, that weighs the contributions of each model, rather than simply using majority rule or mode. A key aspect of this mechanism will involve timestamp matching across channels, which allows for a more nuanced evaluation of sentiment, as it takes into account when specific words were spoken in relation to vocal tone and facial expressions. The output of this process is the final predicted sentiment.

[SLIDE 9]

Lastly, we simply compare the predicted sentiment to the actual to get the performance metric F1-score.

[SLIDE 10]

Next, let's see the perspective of the user using this system.

[SLIDE 11]

From this diagram, the user can interact with the system by inputting their video. Then, the system will separate data and do a sentiment classification for each data separately. After that, the system will weigh each result and produce the final output.

we will not be talking about requirements because of the time limit

[SLIDE 17]

We have conducted some preliminary testing. For now, we have only tested

single-channel sentiment analysis on two data channels: transcription-based analysis and voice tone analysis. So, this testing is not about our about our main voting mechanism yet, but rather testing separate analysis models in the framework.

[SLIDE 18]

For the transcription analysis, we used Python's Natural Language Toolkit to analyze speech after converting it to text. For voice tone analysis, we employed pre-trained models from HuggingFace's library. Testing involved a set of 5 cleaned audio samples from YouTube, representing various emotional contexts. Let's see some of the clips used.

[SLIDE 19]

The first clip is penguinz0 ranting about U.S. immigration services.

[PLAY VDO]

[SLIDE 20]

You can see that each model captured the strong anger just fine. The value in the brackets is in the range negative 1 to positive one. The value here shows that it was a strong negative sentiment. Now let us show you 2 edge cases.

[SLIDE 21]

The second clip is Markiplier mourning a loss of a friend

[PLAY VDO]

[SLIDE 22]

For transcription: Words like "hope" and "respect" might be why the clip is classified as positive. While for voice tone: His shaking voice tone is somehow misclassified as happy, that could potentially be the model's issue. This is why we also need to account for the facial expressions, since we clearly see he was crying.

[SLIDE 23]

The third clip is TommyInnit starting the stream excitedly.

[PLAY VDO]

[SLIDE 24]

Both are classified correctly, but the transcription sentiment score is weaker than expected. It shows transcription alone is unable to account for vocal tone and energy levels.

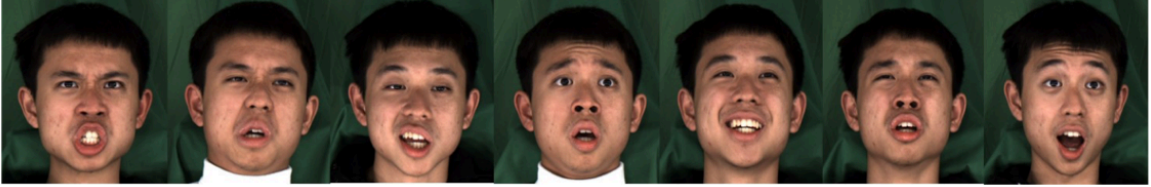
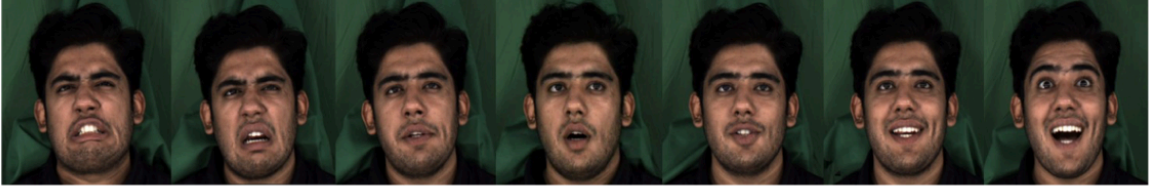
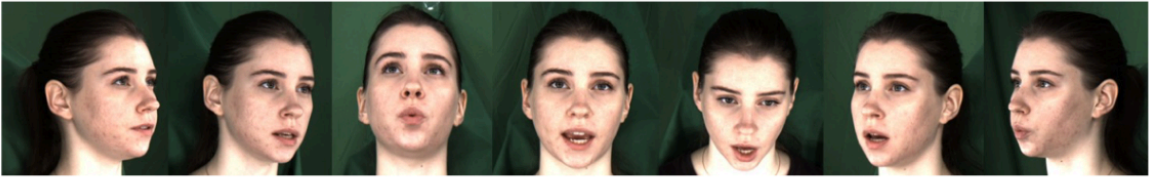
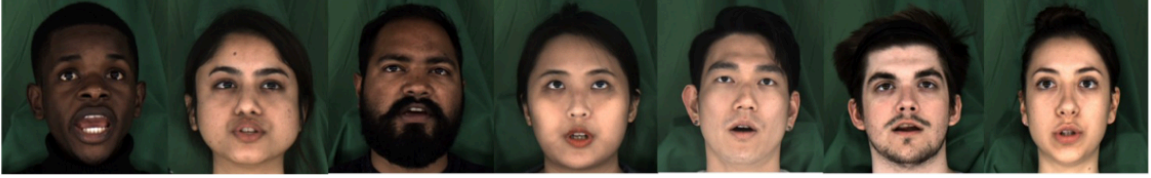
[SLIDE 25]

The tests highlight the limitations of using a single data source for sentiment detection, which is why multimodal approaches are standard. Facial data provides non-verbal cues like smiles or tears, transcription provides the emotional content of the words, and voice tone provides intensity through pitch and rhythm.

But even then, they may not be able to handle edge cases that well. Thus, our project.

[SLIDE 26]

That's the end of the presentation. Thank you for listening.

Emotion Category							
	Angry	Disgust	Contempt	Fear	Happy	Sad	Surprise
Emotion Intensity							
	Strong (disgust)	Medium (disgust)	Weak (disgust)	Neutral	Weak (happy)	Medium (happy)	Strong (happy)
Multi-view							
	Left-60	Left-30	Down-30	Front	Top-30	Right-30	Right-60
Actor							
	ID-0	ID-1	ID-2	ID-3	ID-4	ID-5	ID-6

←↻🏠🔒https://paperswithcode.com/dataset/iemocap🔍🔖🌟⚙️📄🔖🔍🔄⋮👤

🔊Audio

IEMOCAP (The Interactive Emotional Dyadic Motion Capture (IEMOCAP) Database)

Edit

Introduced by Carlos Busso et al. in [IEMOCAP: interactive emotional dyadic motion capture database](#)

Multimodal Emotion Recognition **IEMOCAP** The IEMOCAP dataset consists of 151 videos of recorded dialogues, with 2 speakers per session for a total of 302 videos across the dataset. Each segment is annotated for the presence of 9 emotions (angry, excited, fear, sad, surprised, frustrated, happy, disappointed and neutral) as well as valence, arousal and dominance. The dataset is recorded across 5 sessions with 5 pairs of speakers.

Source: 📄 Multi-attention Recurrent Network for Human Communication Comprehension

Homepage

Benchmarks

Edit

Trend	Task	Dataset Variant	Best Model	Paper	Code
	Emotion Recognition in Conversation	IEMOCAP	SDT		
	Multimodal Emotion Recognition	IEMOCAP	CORECT		
	Speech Emotion Recognition	IEMOCAP	SER with MTL		

Papers

Search for a paper or author

Paper	Code	Results	Date	Stars ↑
Open Implementation and Study of BEST-RQ for Speech Processing		—	7 May 2024	8,673

Usage

License

Edit

🔗 Custom (non-commercial)

Modalities

Edit

🔊Videos

🔊Audio