Sirindhord International Institude of Technology, Thammasat University

Multi-Channel Sentiment Analysis with Loosely Coupled Voting Mechanism

Presented by: 6422782399, 6422782423

Advisor: Dr. Cholwich Nattee

CN3-2

1 Purpose

Develope a multi-channel sentiment analysis for human emotion detection using voting algorithm

1 Purpose

The business world isn’t a Kaggle competition where accuracy is everything.

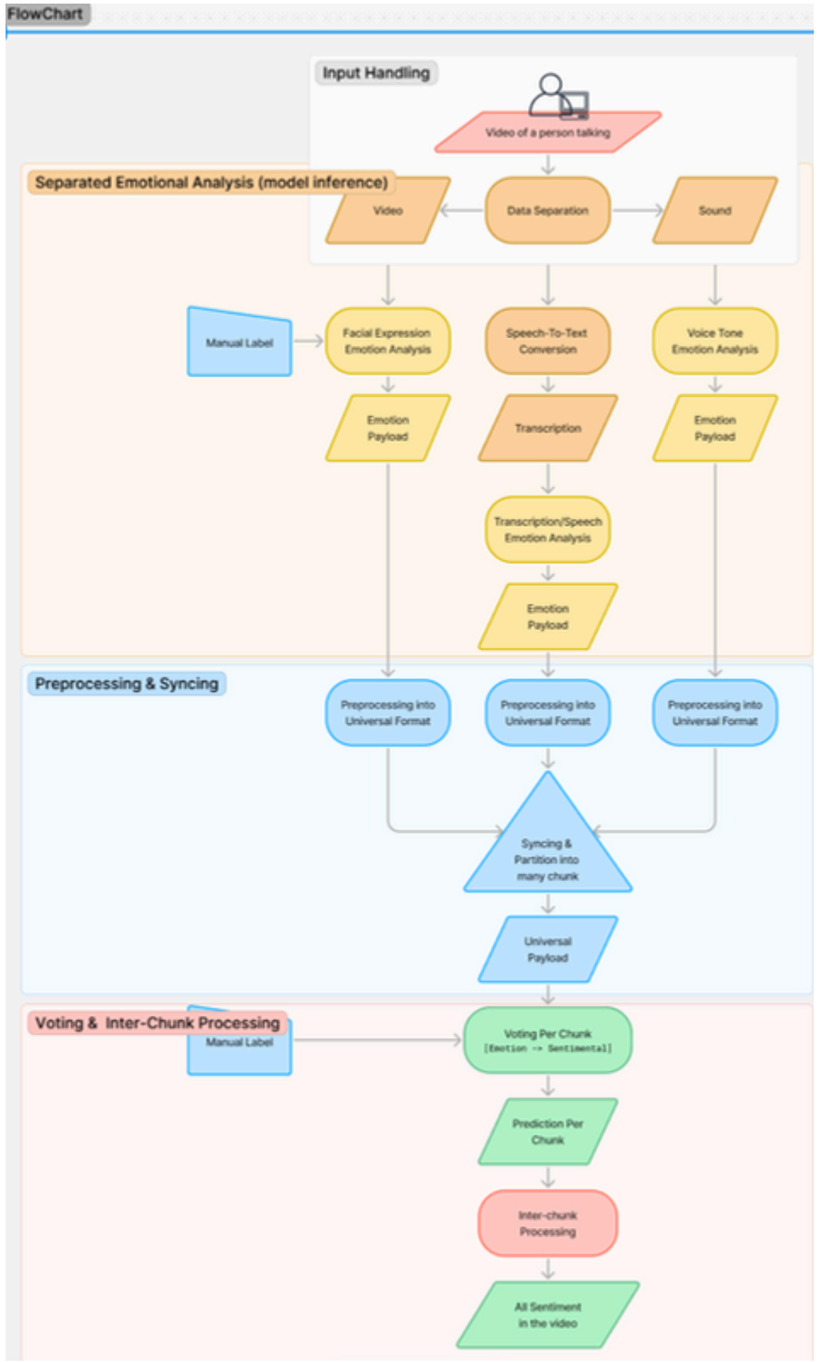
- Real-world challenges like complex emotions, noise, and evolving models.
- Prioritize flexibility and practicality, not just benchmark accuracy.

1 Purpose

Approach	Description	Challenges	Advantages
Single-Channel	Focuses on one input (e.g., facial only).	Can't capture complex emotions.	Simple to implement.
Multi-Channel Voting	Our method: combines multiple inputs post-process.	Conflict resolution is challenging.	Flexible, modular, scalable.
Tightly Coupled Multimodal	Early fusion of inputs into a single model.	Expensive, inflexible, noisy inputs degrade accuracy.	High benchmark accuracy.

2 Architecture

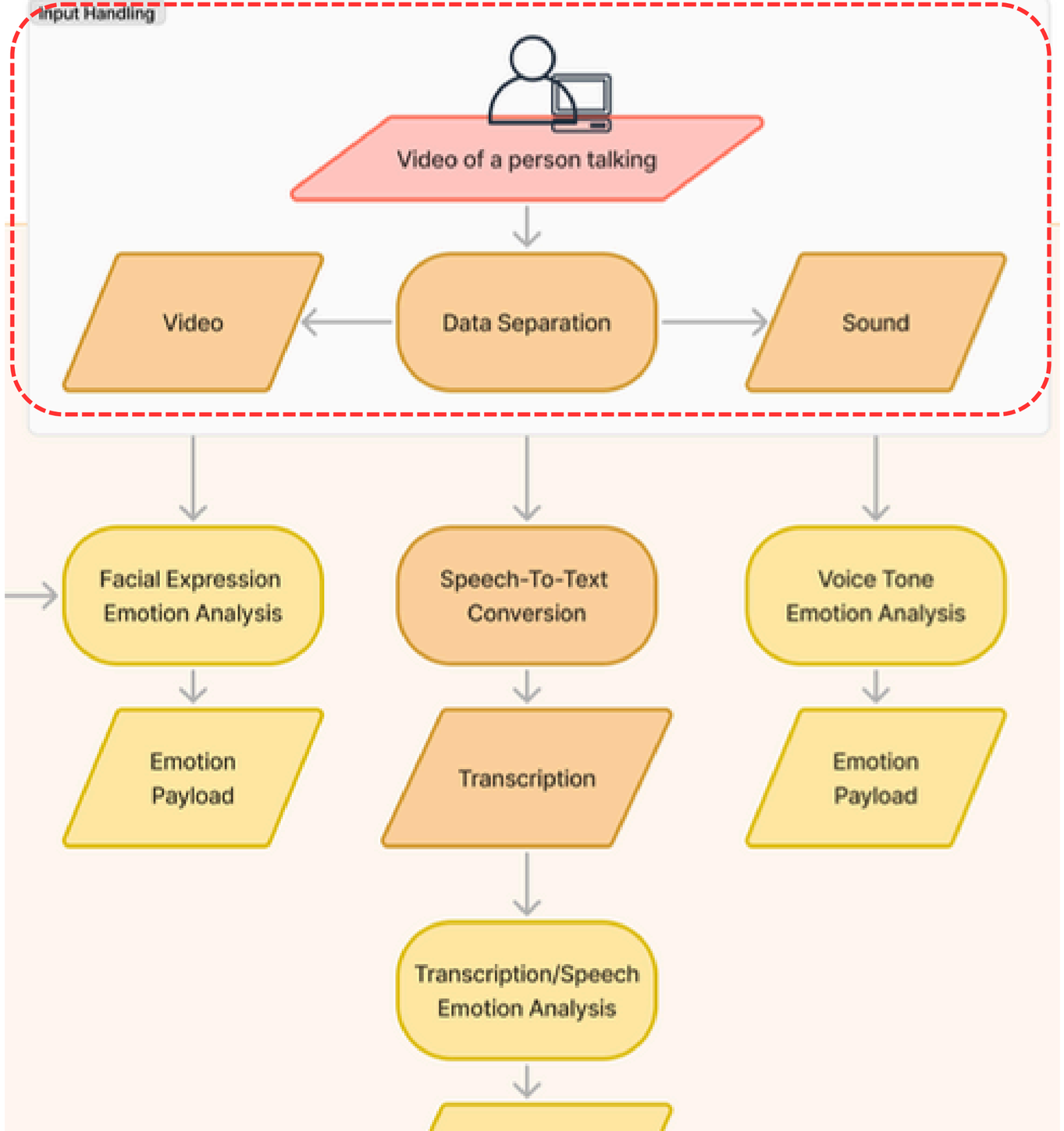
1. Input Handling
2. Separated-Channel Model Inference
3. Pre-processing
4. Syncing
5. Pre-Chunk Voting
6. Inter-Chunk Processing



2.1 Architecture

Input Handling

Short pre-recorded video of a person talking. Separated data into 2 types.

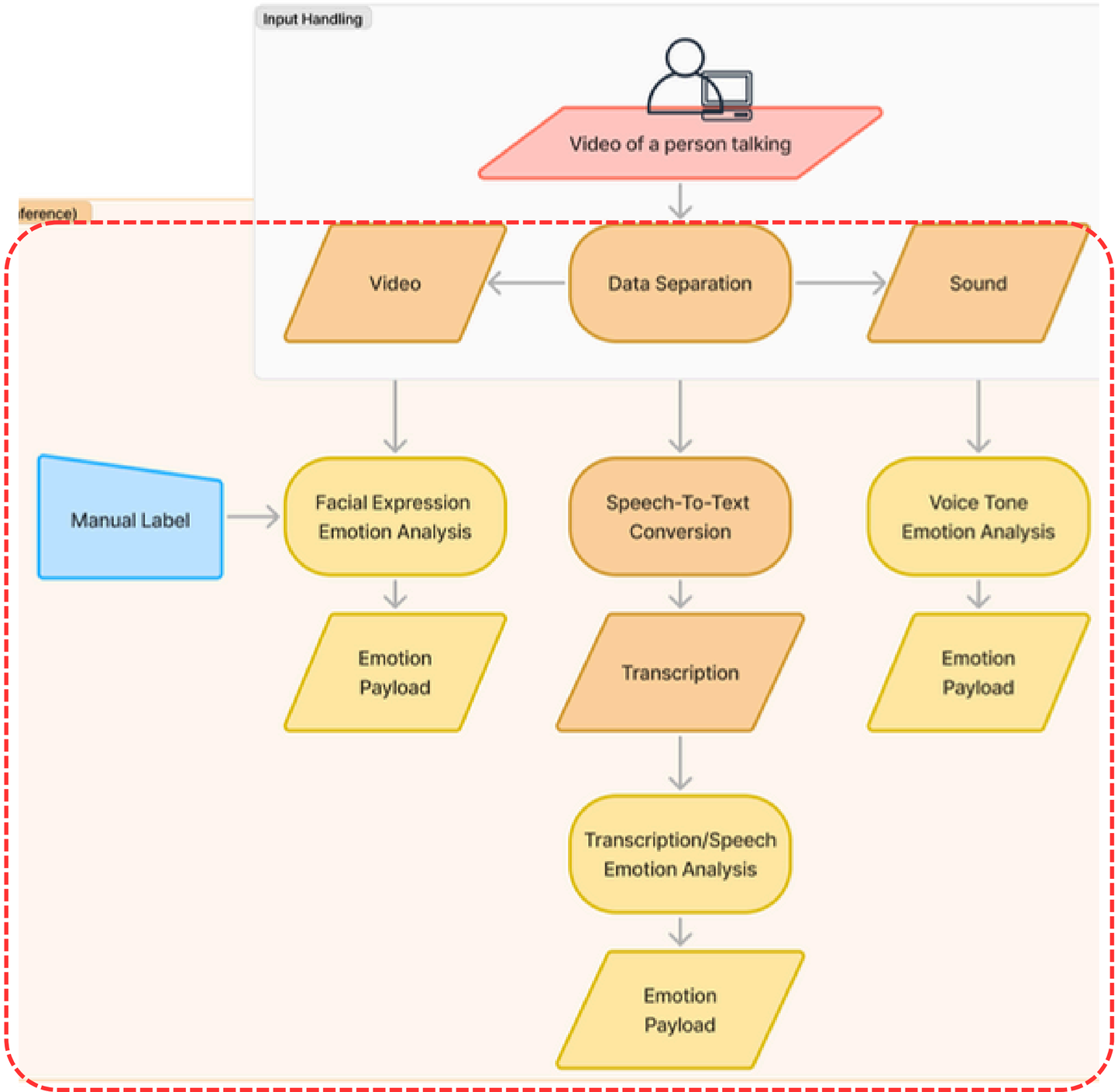


2.2 Architecture

Separated Emotional Analysis (model inference)

Each data type is processed independently.

- Video use for Face sentiment analysis
- Sound use for Voice tone and Speech sentiment analysis



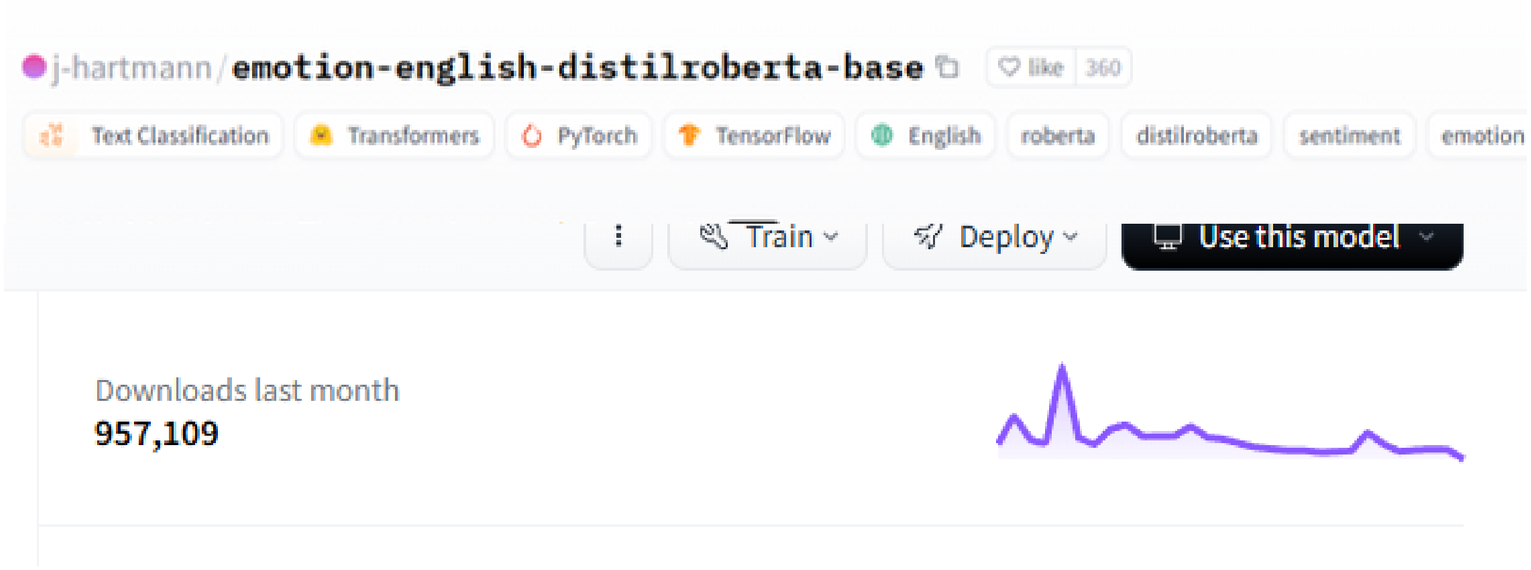
2.2 Separated Emotional Analysis (model inference)

Voice Transcription:

Model:

- OpenAI Whisper → High-accuracy speech-to-text conversion.
- DistilRoBERTa (emotion-english-distilroberta-base) → Fine-tuned for emotion detection.

Output: Emotion classification per sentence(chunk).



1. anger
2. disgust
3. fear
4. joy
5. neutral
6. sadness
7. surprise

https://huggingface.co/j-hartmann/emotion-english-distilroberta-base

2.2 Separated Emotional Analysis (model inference)

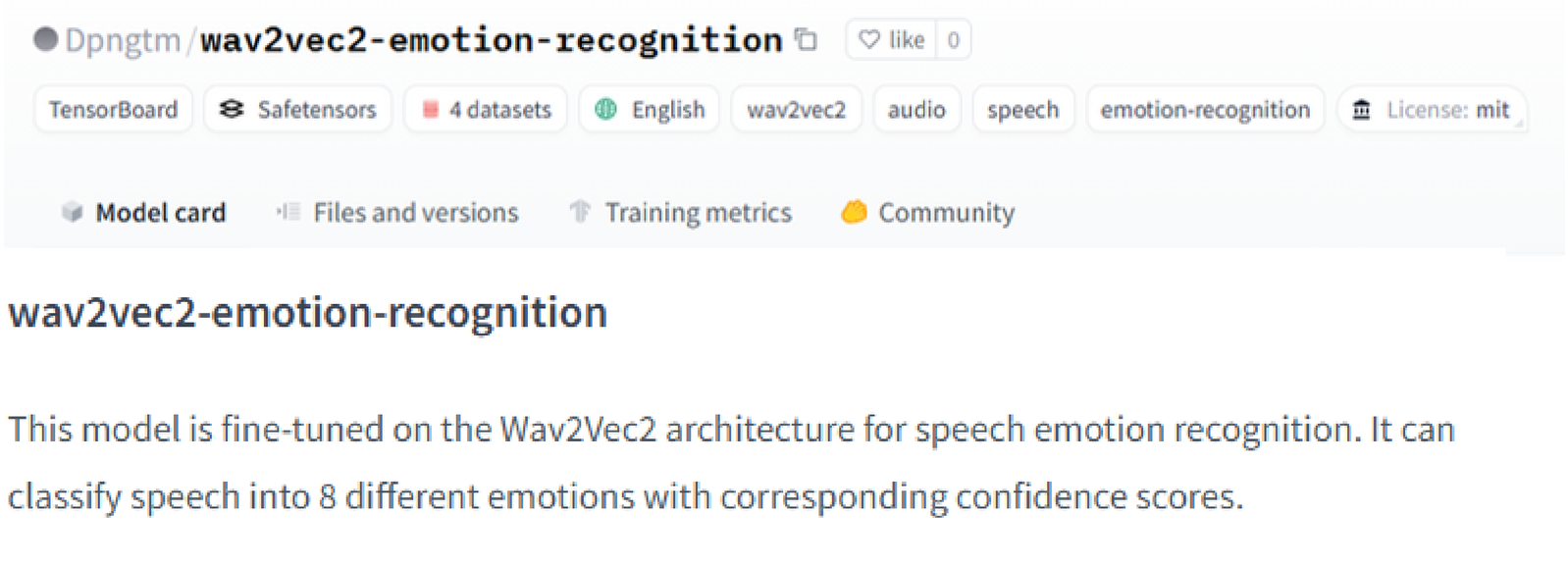
Voice Tone:

Model: Wav2Vec 2.0 Emotion Recognition.

Reasons for Choosing:

1. State-of-the-Art Audio Architecture: Processes raw audio directly.
2. Fine-tuned for Emotion: Trained to detect tone variations.
3. High Accuracy: Reliable in noisy audio environments.

Output: Emotion classification per 5-second interval.



Supported Emotions

- Angry
- Calm
- Disgust
- Fearful
- Happy
- Neutral
- Sad
- Surprised

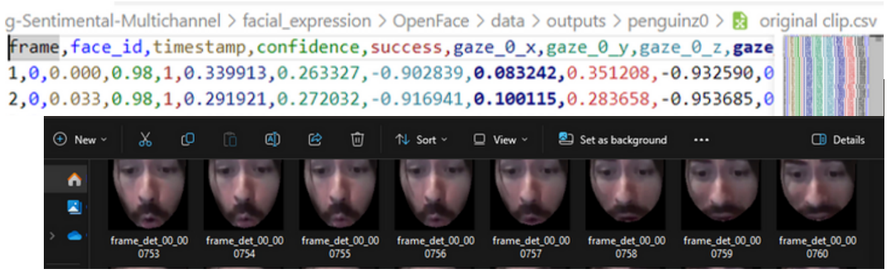
https://huggingface.co/Dpngtm/wav2vec2-emotion-recognition

2.2 Separated Emotional Analysis (model inference)

Facial Expression: (output)

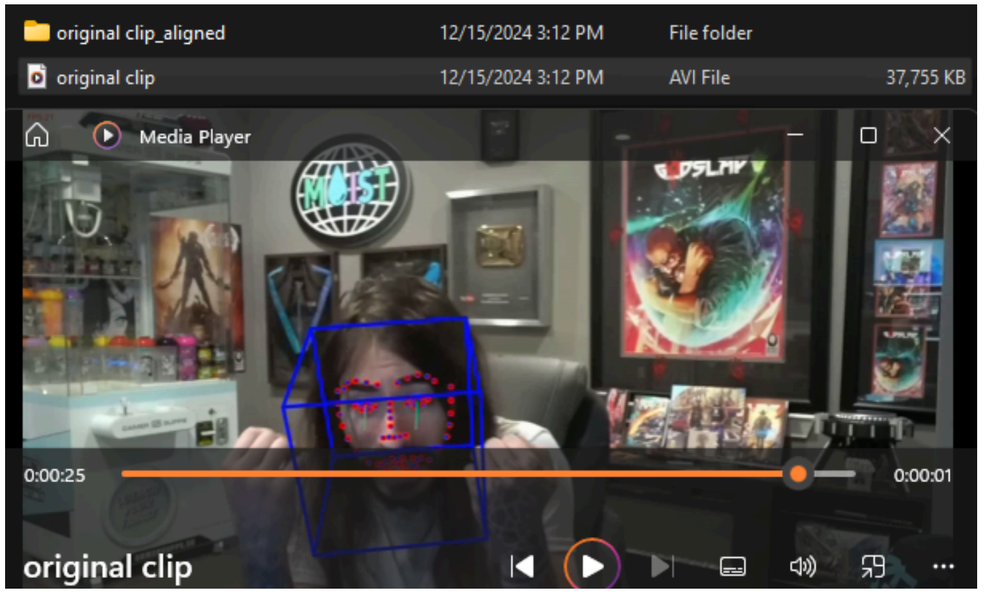
OpenFace: Muscle Movement

less than 1-second intervals



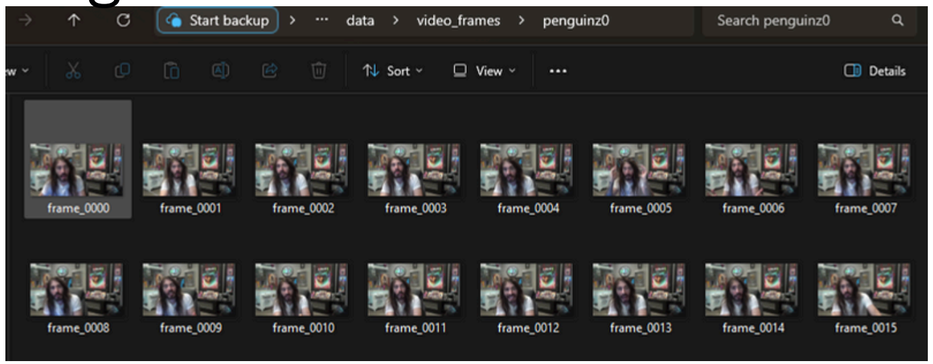

```
Gaze: 1
AUs: 1
Landmarks 2D
Landmarks 3D
Pose: 1
Shape parame
```

&



FaceTorch: Static frame analysis.

extracting frames at 1-second intervals



model inference on each frame

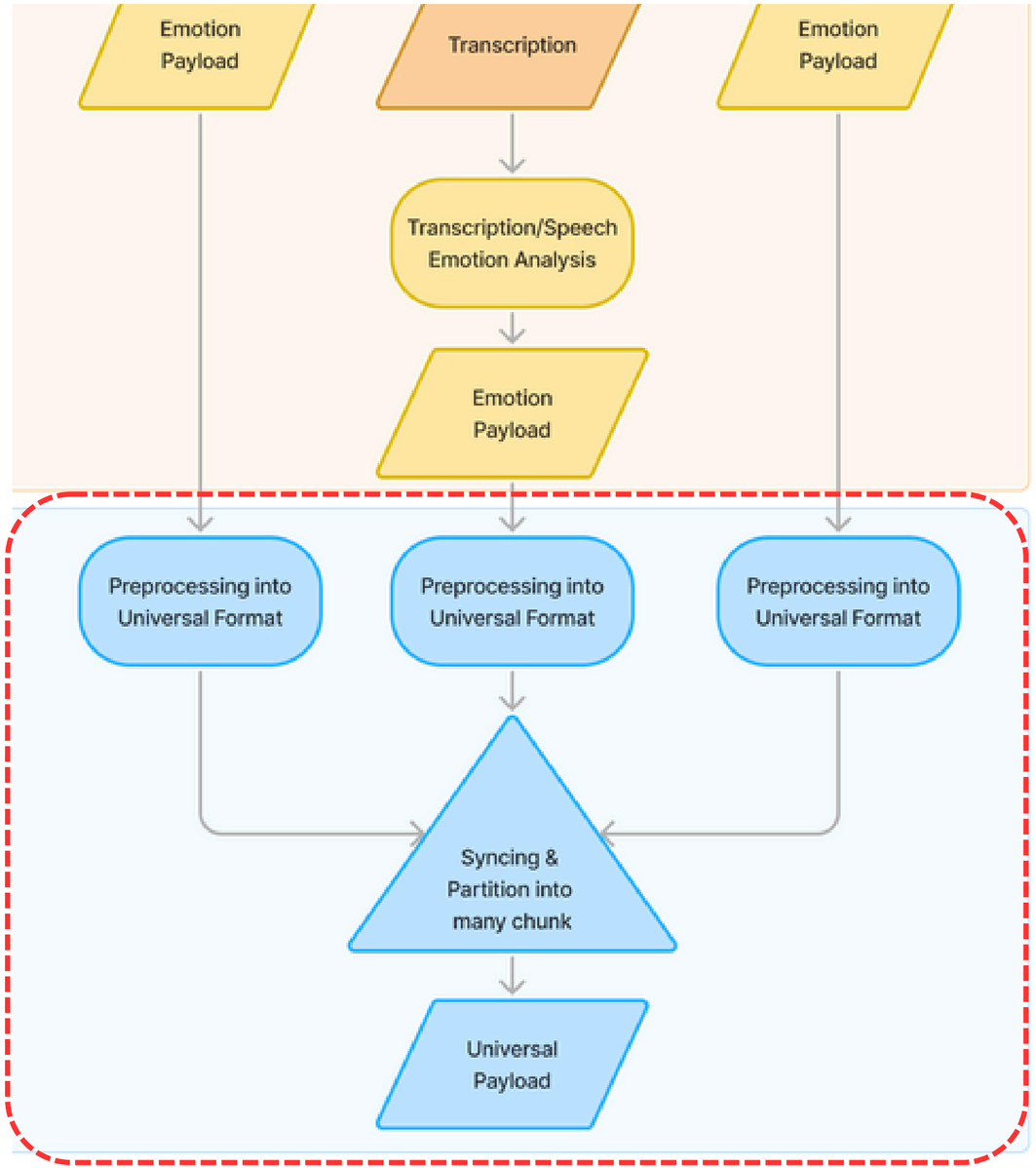
https://github.com/tomas-gajarsky/facetorch
https://github.com/TadasBaltrusaitis/OpenFace

2.3 & 2.4 Architecture

Preprocessing & Syncing

Preprocessing: Transform raw outputs into a universal format.

Syncing: Align outputs into fixed intervals and partition them by transcript chunks.



2.3 Preprocessing

Voice Tone & Voice Transcription

Voice Tone: use video directly

Emotion standardization

- Angry -> Angry
- Calm -> Neutral
- Disgust -> Disgust
- Fearful -> Fear
- Happy -> Happy
- Neutral -> Neutral
- Sad -> Sad
- Surprised -> Surprise


```
{
    "time": 5,
    "emotion": "Angry",
    "confidence": 0.993854820728302
},
{
    "time": 10,
    "emotion": "Happy",
    "confidence": 0.9797731637954712
},
```


Voice Transcript: 1-second intervals

Emotion standardization

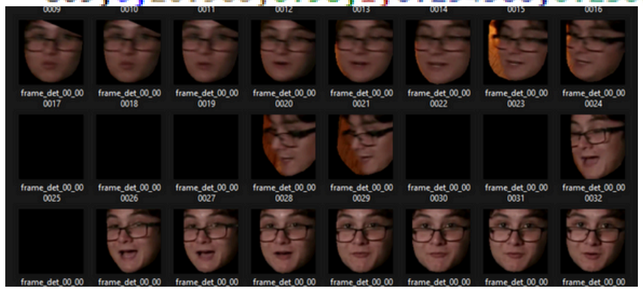
- anger -> Angry
- disgust -> Disgust
- fear -> Fear
- joy -> Happy
- neutral -> Neutral
- sadness -> Sad
- surprise -> Surprise


```
{
    "time": 7.0,
    "emotion": "Neutral",
    "confidence": 0.6356979608535767,
    "transcript": " and I just wanted to reassure you guys that Markiplier as a channel will continue."
},
{
    "time": 13.0,
    "emotion": "Neutral",
    "confidence": 0.6744359731674194,
    "transcript": " I really hope that Ryan and Matt will want to continue moving forward with me,"
},
```

2.3 Preprocessing Facial Expression:

OpenFace: use video directly

```
1    frame,face_id,timestamp,confidence,success
809  808,0,26.927,0.98,1,0.275761,0.222539,-0.9
810  809,0,26.960,0.98,1,0.294560,0.230058,-0.9
```



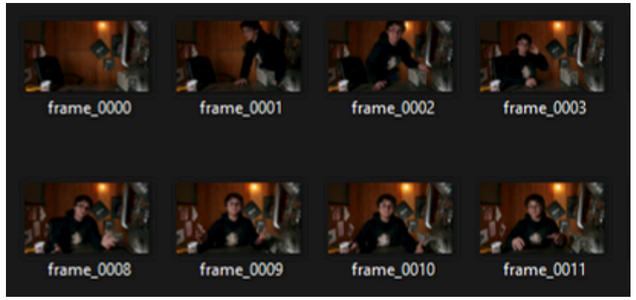
swap! any moment

(grouping)
Interval-based Aggregation

FaceTorch: 1-second intervals

ıce-Syncing-Sentimental-Multichannel > facial_expression > FERPlus > da

```
27  frame_0024.jpg, Sad, Neutral, 0.29, False
28  frame_0025.jpg, Sad, Neutral, 0.30, False
29
```



Temporal Alignment

Universal Format for syncing


```
{
    "frame": "0000.jpg",
    "emotion": "Sad",
    "confidence": 0.19
},
{
    "frame": "0001.jpg",
    "emotion": "Sad",
    "confidence": 0.16
},
{
    "frame": "0002.jpg",
    "emotion": "Sad",
    "confidence": 0.18
},
{
    "frame": "0003.jpg",
    "emotion": "Sad",
    "confidence": 0.2
},
{
    "frame": "0004.jpg",
    "emotion": "Sad",
    "confidence": 0.17
},
{
    "frame": "0005.jpg",
    "emotion": "Sad",
    "confidence": 0.18
},
{
    "frame": "0006.jpg",
    "emotion": "Sad",
    "confidence": 0.16
```

2.4 Architecture

Syncing

Steps:

1. Use transcript timestamps to partition data into chunks.
2. Align:
    - Voice Tone (fixed intervals).
    - Facial Expression (aligned/aggregated).
    - Transcription (already chunked).

Result: Synchronized, chunk-based outputs ready for voting.


```
[
    {
        "partition": "0-6s",
        "transcript": {
            "text": " So what we came across that day in
            "emotion": "Neutral",
            "confidence": 0.40619438886642456
        },
        "facial_expression": [
            {
                "frame": "0000.jpg",
                "emotion": "Sad",
                "confidence": 0.19
            },
            {
                "frame": "0001.jpg",
                "emotion": "Sad",
                "confidence": 0.16
            },
            {
                "frame": "0002.jpg",
                "emotion": "Sad",
                "confidence": 0.18
            },
            {
                "frame": "0003.jpg",
                "emotion": "Sad",
                "confidence": 0.2
            },
            {
                "frame": "0004.jpg",
                "emotion": "Sad",
                "confidence": 0.17
            },
            {
                "frame": "0005.jpg",
                "emotion": "Sad",
                "confidence": 0.18
            },
            {
                "frame": "0006.jpg",
                "emotion": "Sad",
                "confidence": 0.16
            }
        ],
        "voice_tone": [
            {
                "time": 5,
                "emotion": "Fear",
                "confidence": 0.5330254435539246
            }
        ]
    },
    {
        "partition": "6-10s",
        "transcript": {
            "text": " None of us knew how to react",
            "emotion": "Neutral",
            "confidence": 0.853630542755127
        },
        "facial_expression": [
            {
                "frame": "0007.jpg",
                "emotion": "Sad",
                "confidence": 0.18
            },
            {
                "frame": "0008.jpg",
                "emotion": "Sad",
                "confidence": 0.18
            },
            {
                "frame": "0009.jpg",
                "emotion": "Sad",
                "confidence": 0.16
            },
            {
                "frame": "0010.jpg",
                "emotion": "Sad",
                "confidence": 0.17
            }
        ],
        "voice_tone": [
            {
                "time": 10,
                "emotion": "Fear",
                "confidence": 0.851924479007721
            }
        ]
    },
    {
        "partition": "10-12s",
        "transcript": {
            "text": " or how to feel",
            "emotion": "Sad",
            "confidence": 0.5896208882331848
        },
        "facial_expression": [
            {
                "frame": "0011.jpg",
                "emotion": "Sad",
                "confidence": 0.17
            },
            {
                "frame": "0012.jpg",
                "emotion": "Sad",
                "confidence": 0.17
            }
        ],
        "voice_tone": []
    },
```

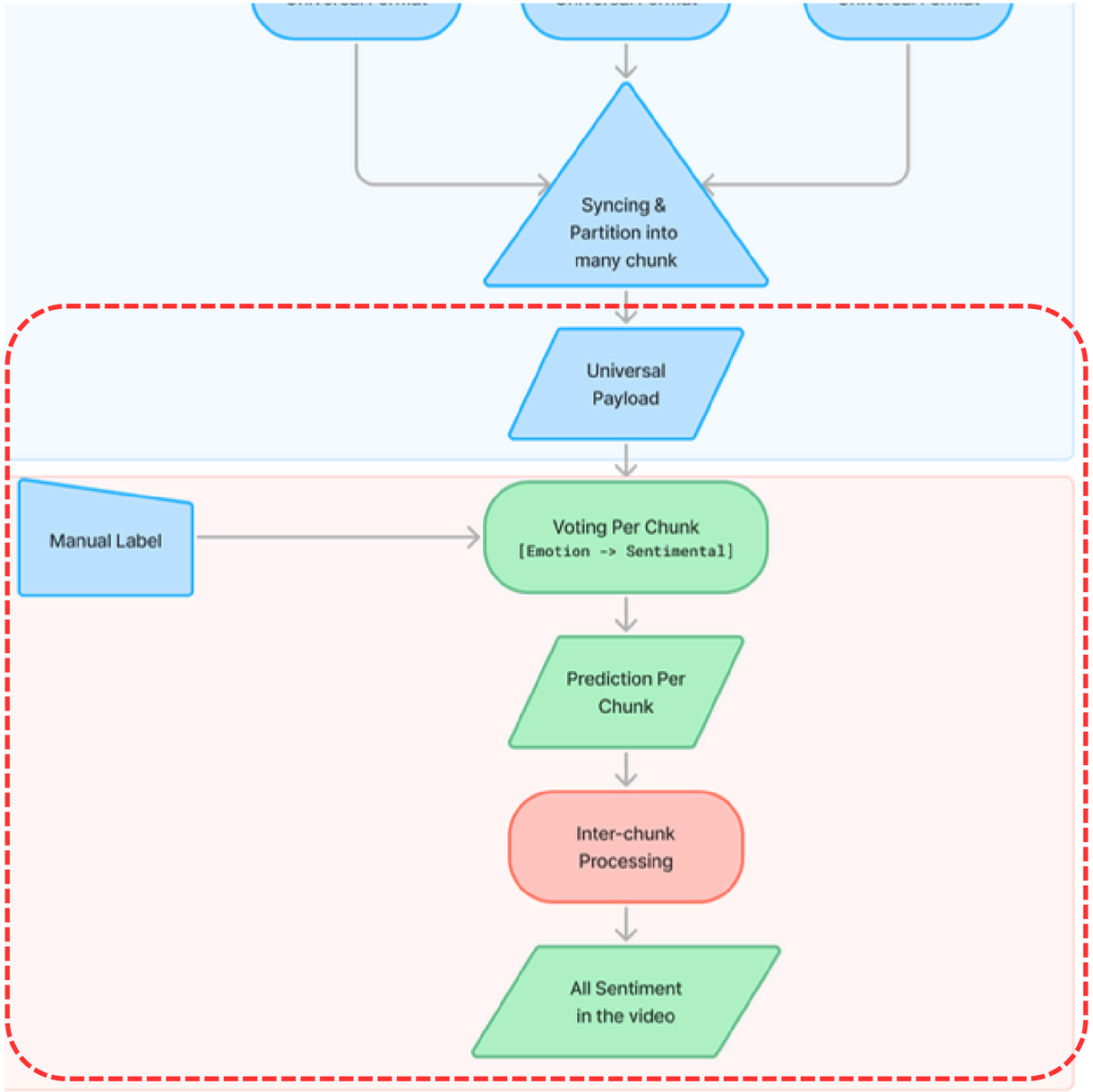
2.5 & 2.6 Architecture: Voting & Inter-Chunk Processing

Voting Mechanism:

- Assign weights to each modality:
  - Transcription: 50%
  - Facial Expression: 30%
  - Voice Tone: 20%.
- Adjust weights dynamically based on emotion conflict.

Inter-Chunk Processing:

- Smooth transitions between emotions.
- Replace short transient states with Neutral for realistic results.



2.5 Per-Chunk Voting

Initial Weight on every Chunk

- Transcript (text or speech content): 50% weight (0.5)
- Facial Expression: 30% weight (0.3)
- Voice Tone: 20% weight (0.2)

Adjust Weight by Conflict

Negative Emotions: ["Angry", "Disgust", "Fear", "Sad"]

Positive Emotions: ["Happy", "Surprise"]

Neutral Emotion: "Neutral" (does not contribute to conflict)

$$conflict\_ratio = \frac{min(positive\_count, negative\_count)}{positive\_count + negative\_count}$$

$$adjusted\_weight = 1 - \text{conflict\_ratio}$$

Multiply adjusted weight with the initial weight.

2.5 Per-Chunk Voting

3.5.3 Example for Adjusting Weight

Emotion List for Voice Tone:

[{"emotion": "Happy", "confidence": 0.7},

{"emotion": "Sad", "confidence": 0.6},

{"emotion": "Fear", "confidence": 0.8}]

Modality Weight: 0.2

Step-by-Step:

1. **Count Emotions:**
   - Positive emotions: 1 (Happy)
   - Negative emotions: 2 (Sad, Fear)
2. **Check Conflict:**
   - Both positive and negative counts > 0, so a conflict exists.
3. **Compute Conflict Ratio:**

$$conflict\_ratio = \frac{min(1,2)}{1+2} = \frac{1}{3} \approx 0.33$$

4. **Adjust Weight:**

$$adjusted\_weight = 1 - 0.33 \approx 0.67$$

5. **Apply Adjust Weight:**
   - Original weight: 0.2
   - Adjusted weight: 0.2 * 0.67 ≈ 0.134

2.6 Inter-Chunk Processing

This function smooths out transitions between different emotions in a list of partitioned results. It ensures that short, transient emotional states are replaced by a "Neutral" state, providing more realistic emotion in the sequence, since human emotion does not normally change from positive to negative and back to positive immediately.

3.5.5 Example for Inter-chunk Outlier Handling

Output:

- **Partition: 0-6s, Final Emotion: Sad, Confidence: 0.37**
- **Partition: 6-10s, Final Emotion: Sad, Confidence: 0.63**
- **Partition: 10-12s, Final Emotion: Happy, Confidence: 0.40**
- **Partition: 12-17s, Final Emotion: Fear, Confidence: 0.46**
- **Partition: 17-21s, Final Emotion: Neutral, Confidence: 0.36**
- **Partition: 21-24s, Final Emotion: Neutral, Confidence: 0.23**

After Adjust:

- Partition: 0-6s, Final Emotion: Sad, Confidence: 0.37
- Partition: 6-10s, Final Emotion: Sad, Confidence: 0.63
- Partition: 10-12s, Final Emotion: Neutral, Confidence: 0.40
- Partition: 12-17s, Final Emotion: Fear, Confidence: 0.46
- Partition: 17-21s, Final Emotion: Neutral, Confidence: 0.36
- Partition: 21-24s, Final Emotion: Neutral, Confidence: 0.23

Multi-Channel Sentiment Analysis System 🎭🎤📝

This project implements a multi-channel sentiment analysis system that integrates **facial expressions**, **voice tone**, and **speech transcription** to analyze human emotions. A **voting mechanism** is employed to combine results, offering transparency, modularity, and robustness compared to traditional multimodal systems.

Abstract 📄

This system independently analyzes sentiment across three modalities:

- **Facial Expressions** (FERPlus, OpenFace)
- **Voice Tone** (Wav2Vec2 fine-tuned for RAVDESS)
- **Speech Transcriptions** (DistilRoBERTa for emotion detection)

A **post-processing voting mechanism** is applied to combine outputs. Unlike black-box neural networks, this approach is interpretable and adaptable, allowing conflict resolution between emotional signals for a more accurate and explainable sentiment classification.

System Workflow 🚀

The system processes a video file (30-second segments) and performs the following steps:

1. **Frame Extraction:**
    - Convert video into frames at **1-second intervals** for facial expression analysis.
      📜 Refer: FacialExpression.md
2. **Speech and Voice Analysis:**
    - Extract audio for **voice tone** and **speech transcription**.
3. **Modality-Specific Analysis:**
    - **Facial Expression:** Detect frame-level emotions using **Facetorch** or **OpenFace**.
    - **Voice Tone:** Analyze audio tone with **Wav2Vec2**.
    - **Speech Transcription:** Sentiment detection using **DistilRoBERTa**.
4. **Syncing and Voting:**

3

Evaluation & Advantage

3 Evaluation & Advantage

1. Feasibility and Practical Applicability

"Current Single-Channel Models Are Already Multimodal."

Evidence:

Voice Tone Multimodality:

- Wav2Vec 2.0 incorporates multiple audio features, including frequency, waveform, and phonetic patterns (Baevski et al., 2020, "Wav2Vec 2.0: A Framework for Self-Supervised Learning of Speech Representations", NeurIPS).
- Wav2Vec proves that voice analysis models are already complex and deeply multimodal.

Text Multimodality:

- Transformers like BERT and DistilRoBERTa analyze syntactic (structure) and semantic (meaning) features, combining multiple linguistic dimensions into one decision (Devlin et al., 2018, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", NAACL).

Facial Expression Multimodality:

- OpenFace utilizes temporal (time-based) and spatial (frame-based) analysis of action units (AUs), making it inherently multimodal for facial sentiment detection (Baltrusaitis et al., 2016, "OpenFace: An Open Source Facial Behavior Analysis Toolkit", IEEE Winter Conference on Applications of Computer Vision).

3 Evaluation & Advantage

2. Accessibility of Black-Box Models

Black-Box Model → Universal Format → Voting Mechanism.

Application: Organizations can leverage high-performing models for real-world tasks without violating licensing restrictions or developing their own systems."

Evidence:

- Proprietary limitations of models cited in Google NLP API (2023). Documentation: https://cloud.google.com/natural-language
- Similar black-box discussions in Chen et al. (2021). "Explainability in AI Applications."

3 Evaluation & Advantage

3. Flexibility and Modularity

Switching between OpenFace ⇔ Facetorch (Facial Expression Models)

Application: Businesses can stay up-to-date with rapidly advancing models, saving both time and resources."

Evidence:

- Flexibility as a core principle of late fusion systems (Ramachandram et al., 2017). "Deep Multimodal Learning", Pattern Recognition Letters.

NOTE: this research is not related to emotional or sentimental analysis (multimodal-related)

3 Evaluation & Advantage

4. Enhanced Robustness Through Multi-Channel Fusion

better than single channel:

- Text models misinterpret sarcasm.
- Voice tone models struggle with noise.
- Facial expression models may fail under poor lighting.

Evidence:

- Studies confirm that multi-channel fusion enhances robustness (Ngiam et al., 2011). "Multimodal Deep Learning", ICML Proceedings.

3 Evaluation & Advantage

5. Transparency and Maintenance

Early Fusion (High Complexity) vs. Late Fusion (Low Complexity).

Voting Mechanism Table: Channel Outputs → Final Decision (visible weights).

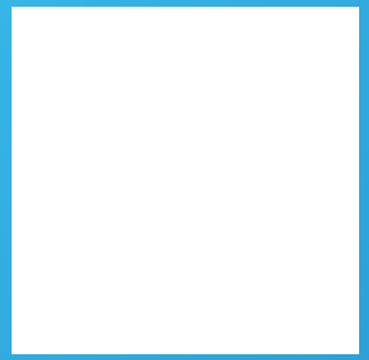
Application: This makes our system practical for small research labs and companies with limited computational resources."
Evidence:

- Late fusion reduces training overhead (Atrey et al., 2010). "Multimodal Fusion Techniques", Multimedia Systems.
- Transparency issues in black-box AI are well-documented (Lipton, 2016). "Mythos of Model Interpretability".

3 Future Work

we aim to make condition-based adjsut weight further as emotion or detect sarcasm and want to make it better to keep up with multi-modal

Application: This is particularly useful for detecting sarcasm or ambiguous emotions in customer interactions."



THANK YOU