**Methodology: Multimodal Audio Sentiment Classification**

### 1. **Wav2Vec 2.0 Embeddings + MLP Classifier**

* **Feature extraction**: Used pretrained wav2vec2-base model to extract embeddings from audio files.
* **Model architecture**: 3-layer MLP with 256 → 128 → num\_classes.
* **Training details**:
  + Optimizer: Adam, LR: 5e-4
  + Epochs: 60
  + Batch size: 32
* **Performance**:
  + Accuracy plateaued around **71%**
  + Used confusion matrix to evaluate class-wise performance.

### 2. **BoVW from Spectrograms + MLP**

* **Spectrogram generation**: Converted audio to spectrograms and stored as PNG images.
* **Feature extraction**:
  + ORB keypoints and descriptors were extracted from each spectrogram.
  + Clustered descriptors using **KMeans** (100 clusters) to form a visual vocabulary.
  + Constructed BoVW histograms for each spectrogram.
* **Label encoding**: Alphabetically encoded 7 emotion labels.
* **MLP classifier**: Same 3-layer structure.
* **Performance**:
  + Accuracy plateaued around **40%**
  + T-SNE visualization showed high overlap, indicating BoVW was lossy.

### 3. **MFCC Feature Extraction + MLP**

* **MFCC extraction**:
  + Extracted 13 MFCCs using torchaudio.
  + Averaged over time to get fixed-length vectors.
* **Classifier**: 3-layer MLP.
* **Performance**: Plateaued around **52%** accuracy.

### 4. **CNN on Spectrogram Images**

* **Input**: Raw spectrogram images.
* **CNN architecture**:
  + Conv2D → ReLU → MaxPool → Conv2D → ReLU → Flatten → Linear
* **Performance**: Peaked at ~65% accuracy.

### 5. **Multimodal Feature-Level Fusion (MFCC + BoVW + Wav2Vec)**

* **Fusion method**:
  + Matched filenames across all features.
  + Concatenated aligned vectors to form fused embeddings.
* **Classifier**: 3-layer MLP.
* **Performance**:
  + Final accuracy: **75%**
  + Binary accuracy between emotion pairs ranged from **80–97%**
  + T-SNE plots showed improved class separability.

### 6. **Advanced Evaluation (Top-K Accuracy + Thresholding)**

* **Top-1 accuracy**: ~73%
* **Top-2 accuracy**: ~88%
* **Selective prediction**: Applied softmax temperature scaling (T=0.8) to improve confidence calibration.
* **Threshold-based accuracy**:
  + At threshold 0.4, selective accuracy reached **84.44%**.

### 7. **Planned / Ongoing Enhancements**

* **WavLM**: Integration of WavLM embeddings for further boosting accuracy.
* **Extended Fusion**: Planned fusion: WavLM + Wav2Vec + MFCC + BoVW.

### Summary of Individual Component Accuracies

| Component | Accuracy |
| --- | --- |
| Wav2Vec2 | ~71% |
| BoVW | ~40% |
| MFCC | ~52% |
| CNN | ~65% |
| Fusion | **75%** |