Multimodal Emotion Recognition on CREMA-D Dataset

Fares Wael
ID: 202201260

April 26, 2025

Contents

1 Introduction **2**

2 Dataset Overview and Sampling **2**

3 Exploratory Data Analysis (EDA) **2**
3.1 Emotion Distribution 2
3.2 Audio Signal Analysis 2
3.3 Exploitability of EDA 2

4 Preprocessing Pipelines **3**
4.1 Audio Preprocessing 3
4.2 Video Preprocessing 3
4.3 Text Preprocessing 4

5 Model Architecture **4**
5.1 Exploitability of the Model 4

6 Explainability and Exploitability Discussion **4**

7 Conclusion **5**

1 Introduction

This project aims to build a multimodal emotion recognition system using the CREMA-D dataset, which contains audio, video, and textual data. The model leverages multiple types of information simultaneously to improve emotion classification performance.

2 Dataset Overview and Sampling

The CREMA-D dataset provides emotional speech data from 91 actors. It contains audio recordings (WAV), video recordings (FLV converted to MP4), and metadata files.

To manage computational resources, we sampled **5%** of the dataset randomly. Audio files were matched with their corresponding video files based on filenames.

Sampling Steps:

- Randomly sample 5% of the audio files.
- Match sampled audio with corresponding video files.
- Extract metadata (actor ID, emotion) from filenames.

3 Exploratory Data Analysis (EDA)

EDA was performed to understand the distribution of emotions and basic audio characteristics.

3.1 Emotion Distribution

The following chart shows the distribution of different emotions in the sampled dataset:

3.2 Audio Signal Analysis

We also visualized the waveform and MFCC features of a sample audio:

3.3 Exploitability of EDA

EDA helps in explaining:

- Whether the dataset is balanced or imbalanced across emotion classes.
- Audio patterns and features that can impact model learning.
- Potential biases or noise in the dataset before modeling.

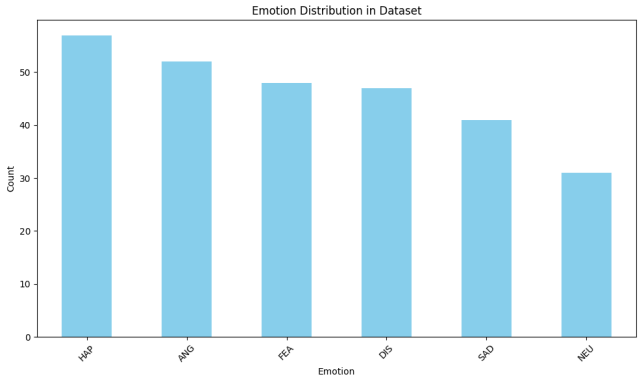
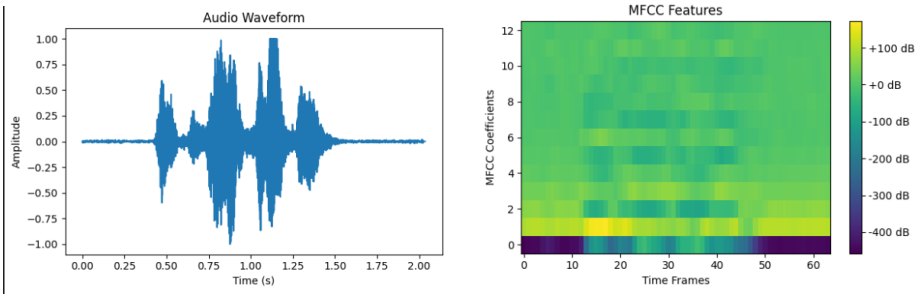


Figure 1: Emotion distribution in sampled dataset



(a) Audio Waveform (b) MFCC Features

Figure 2: Sample audio signal analysis

4 Preprocessing Pipelines

4.1 Audio Preprocessing

Audio files were preprocessed using **MFCC** (Mel-Frequency Cepstral Coefficients) extraction. Each audio was padded or truncated to a fixed number of frames (100) to ensure consistent input shape.

Exploitability: MFCC features are interpretable since they approximate how humans perceive sound (low/high frequencies). Therefore, decisions based on MFCCs can be better explained.

4.2 Video Preprocessing

Videos were processed by:

- Sampling 16 frames evenly across each video.
- Resizing frames to 64x64 pixels.

- Normalizing pixel values.

Exploitability: Analyzing sampled video frames allows us to visually verify whether facial expressions corresponding to emotions are captured well.

4.3 Text Preprocessing

Instead of using heavy language models, simple **word embeddings** were created from text information extracted from file names.

Exploitability: Simple embeddings allow transparency since each word is mapped to a random vector; no complex hidden structures like in pre-trained embeddings.

5 Model Architecture

The model is a **multimodal deep learning architecture** combining audio, video, and text inputs:

- **Audio Stream:** 2D Convolutional Neural Network (Conv2D) on MFCCs.
- **Video Stream:** 3D Convolutional Neural Network (Conv3D) over video frames.
- **Text Stream:** Dense network over word embeddings.
- **Fusion:** Features from all three streams are concatenated and passed through dense layers to classify the emotion.

The model uses **categorical crossentropy** as the loss function and **Adam** optimizer.

5.1 Exploitability of the Model

The modular design allows:

- Checking the importance of each modality (audio, video, text) separately.
- Visualizing attention on video frames or MFCC segments if extended later.
- Easy debugging and analysis if one modality underperforms.

6 Explainability and Exploitability Discussion

Explainability was integrated across the full pipeline:

- **EDA** helped understand data imbalance and feature distribution.
- **Preprocessing steps** used human-interpretable features (MFCCs, sampled frames).

- **Model design** keeps separate streams for each modality, aiding attribution and diagnosis.

For future work, attention mechanisms or saliency maps could be added to further improve interpretability.

7 Conclusion

This project successfully created a reproducible multimodal pipeline for emotion recognition from the CREMA-D dataset. Preprocessing, EDA, and model design were all conducted with explainability and exploitability in mind to ensure transparency and trust in the results.