CentraleSupélec

Projet ST7

# *Deciphering non-verbal behaviors based on speech and text*

***Students:***
Daniel Stulberg Huf
Lawson Oliveira Lima
Luan Rocha do Amaral
Lucas Vitoriano de Queiroz Lira
Rajeh El Saddi
Tomas Gonzalez Villaroel

***Advisor:***
Simon LEGLAIVE

March 31, 2023

# Contents

# 1 Introduction

## 1.1 Context

As social animals, humans are deeply influenced by emotions in their relationships. Deciphering and analyzing emotions, particularly those related to non-verbal behaviors, seemed like an impossible task just a few decades ago. However, advances in technology, psychology, and medicine have allowed us to become more proficient in this area. As a result, we can now communicate more accurately and efficiently with machines, paving the way for further progress in the field.

In this context, we have been tasked with participating in an international research challenge focused on automatic analysis of human behaviors. To aid us in this endeavor, we will be utilizing the Interactive Emotional Motion Capture Database (IEMOCAP), created by the University of Southern California. This comprehensive database contains thousands of labeled data points concerning emotions expressed through audio, text, and motion capture in a variety of communication scenarios performed by ten actors in dyadic sessions. Unlike other conventional databases that only include one speech channel, IEMOCAP can be an extremely useful resource for analyzing non-verbal, multi-modal emotions.

The data from IEMOCAP was extracted and preprocessed before being utilized to train a model that can automatically detect human emotions. This model utilizes Artificial Intelligence techniques and was trained on a subset of the IEMOCAP database. It was then evaluated on another subset to determine its effectiveness in accurately extracting emotions.

## 1.2 Goals

To meet the requirements of our project and address the associated challenges, we chose to focus on audio and text representations to detect emotions in discrete categories. To accomplish this, we developed an algorithm capable of classifying emotions based on the sound and transcription of an individual's speech. The algorithm was tested on a variety of cases, ranging from classifying two to five different emotions.

To perform such task, we have split the process into 3 fronts, namely audio processing, text processing, and AI modeling. All the routines were performed using the Google Colab platform.

## 1.3 Report structure

This document is organized as follows. Section *2* presents a bibliographic research that contemplates some of the previous research and work done with regards to multimodal analysis, using both the IEMOCAP database and other resources. Section *3* will describe, step by step, the pipeline for building the model that is capable of classifying emotions based on audio and text data from IEMOCAP. Then, section *4* presents the results obtained from the conceived algorithm considering both the independent audio an text models, as well as the consolidated model and the visual interface to test it. Finally, section *5* evaluates the performance of the obtained results in comparison with other approaches. Possible room for improvement will also be discussed.

# 2 State of the Art

Sentiment analysis is a major field in academia and it has become increasingly important over time. This is because it allows for the automatic interpretation of human emotions, which in turn helps machines interact with people in a more compassionate manner. Typically, the human brain uses various sources to interpret emotions, such as sound, speech context, facial expressions, posture, and many other forms of information. In this context, common research efforts focus on Multimodal Sentiment Analysis (MSA) to detect emotions using different sources, such as voice, the content of speech, and facial expressions. Other models such as [1] also apply a multimodal approach with physiological signals, such as heart rate.

Sentiment analysis is divided into two main areas: opinion mining and emotion mining, which are usually interrelated subjects (Figure 1). However, our project is focused on emotion mining, particularly on emotion classification. As previously mentioned, the most common approach to emotion classification involves analyzing sound, text, and images of human faces. While images can help detect the current emotion of someone, text can be used not only to detect people's emotions, but also their opinions. Lastly, audio can be used for emotion recognition, but it is limited by issues such as different voice features (accent, speaker, speaking style, language, etc.). The task becomes even more complex when expressions and actions related to emotions may vary from one culture to another. Additionally, human emotions are mainly divided into six basic sentiments (anger, sadness, surprise, fear, disgust, and joy) and other non-basic emotions (such as fatigue, pain, agreement, engagement, curiosity, irritation, and contemplation) [2].

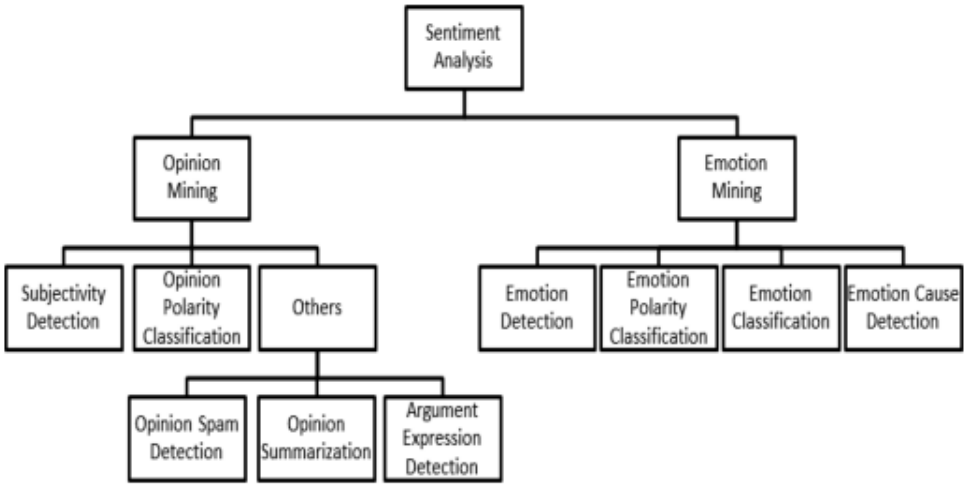


Figure 1: Taxonomy of sentiment analysis tasks [3]

Today, the world is more connected than ever, and social networks are a great tool for people to express themselves and share data. [4] demonstrates how useful data is being collected by social networks such as Instagram and Twitter. With the gathered information, it is possible to detect users' opinions and satisfaction with different publications and advertisements [5].

Due to the significant developments within this area, both the academia and the corporations have seen a need for large datasets with quality content. Big companies like Google are creating and working on huge databases like YouTube-100M, which contains more than 100 million YouTube videos, and also ImageNet, an image database. Another massive dataset is IEMOCAP ([6]), which contains more than 10GB of videos, sound, text, and motion capture of faces in various situations performed by actors and labeled by professionals.

Additionally, there is a significant effort being made to improve technology to better classify human emotions. Google proposed Vggish ([7]), which is a state-of-the-art CNN for detecting emotions in audio. For text classification, one primary algorithm is BERT ([8]), the Bidirecional Encoder Representations from Transformers, which uses a "masked language model" (MLM) for pre-training objectives. Furthermore, other models have shown good results, such as [9] that used L3-Net (Look, Listen, Learn), which is capable of outperforming Vggish and SoundNet while having fewer parameters and less data.

Emotion detection has proven useful in other areas. For example, [10] proposed a model that uses emotions to predict election consequences based mainly on Twitter posts and concluded that these social networks can plausibly reflect the offline political landscape. Also, [11] proposed a multi-modal user-emotion detection using face images and voice to detect emotions. For this purpose, the article uses the Gender and Emotion Voice Analysis (GEVA) algorithm for the voice and the Gender and Emotion Facial Analysis (GEFA) algorithm for face detection, integrated with a Robot Operating System (ROS). [12] uses a Convolutional Neural Network (CNN) that performs significantly well by using transfer learning to reduce the size of the dataset, making

it possible to analyze large-scale data from multimedia content. Figure 2 shows many applications of MSA that can be addressed.

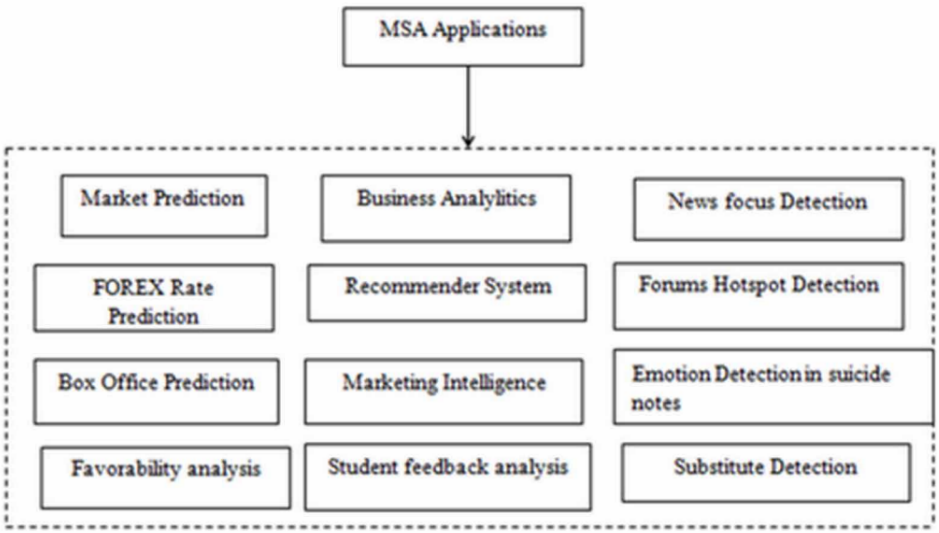


Figure 2: Multimodal sentiment analysis applications [13]

# 3 Methodology

## 3.1 IEMOCAP dataset

The IEMOCAP dataset was selected to tackle the task of identifying emotions. This database was created in 2007 and is composed of five dyadic sessions featuring ten actors, both following scripted scenarios and engaging in spontaneous conversations. The aim was to elicit five specific types of emotions: happiness, anger, sadness, frustration, and neutral states. Over time, the range of emotions expanded to include disgust, fear, excitement, and surprise. The data was collected using three different modalities: audio, text, and motion capture. In total, there is approximately 30 hours of recorded data, resulting in a final database of size 11 GB.

After extracting the IEMOCAP database, we decided to focus on two modalities: text and audio. This choice was made with the intention of improving the performance of our emotion predictive model compared to using only one modality. By considering multiple modalities, we can identify correlations between text and audio, which can enhance the identification of different emotions. However, selecting all three modalities would complicate the task beyond the initial scope of the project.

To ensure the accuracy of our analysis, we conducted an initial preprocessing of the database to filter out emotions with very few occurrences, as well as unbalanced ones. The resulting filtered database included only the following emotions: neutral state, frustration, anger, sadness, and excitement. Figure 3 illustrates the number of occurrences for each of these emotions.

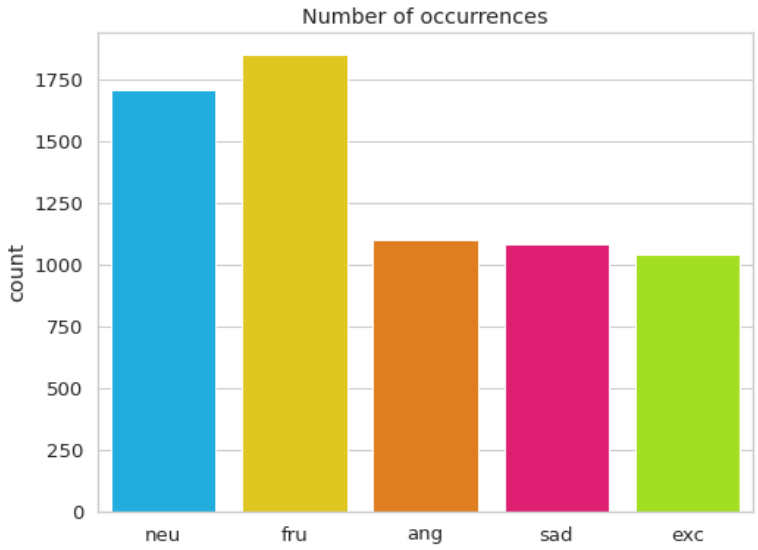


Figure 3: Histogram of the five most balanced emotions of IEMOCAP

Next, the dataset was split into 80%(Train), 10%(Validation) and 10%(Test). To avoid the problem of class imbalance between the emotions in the database we made this processes and used a weighted loss function, more specifically the Weighted Cross Entropy Loss.

## 3.2 Audio processing

The audio files utilized in this research were initially obtained from the IEMOCAP database in the MP3 format. To ensure that only relevant audio files were used for the subsequent analysis, a preliminary filtration process was conducted, which involved removing unlabeled audio files and those labeled to classes that would not be considered for the proposed model. Following the filtration process, the total number of audio files available for analysis amounted to approximately 7500. Each audio file was then segmented into single 3-second samples. To standardize the sample length across all audio files, the audios with a duration shorter than 3 seconds were padded with zeros at the end of each original audio, while those with a duration exceeding 3 seconds were cropped to a 3-second length.

The decision for making the duration of every audio the same relies on the fact that the processed audios will be the inputs of a neural network (explained with more details below), which requires a constant shape for its input. In addition, the choice for a 3-second standard was due to the fact that the mean duration of all the audios present in the database stands around such value.

Following the standardization, each audio file was translated into its corresponding waveform, from which it was computed its corresponding Mel spectrogram. A spectrogram is a suitable representation for analysing audio such as music and speech. It is computed by performing the short-time Fourier transform, which briefly consists on breaking the audio signal into its component frequencies and corresponding amplitudes

over overlapping windows of time [14]. A spectrogram allows us to get the energy at various frequency bins at each time step. However, instead of a vanilla linear spectrogram, the representation was computed using the Mel logarithmic scale, which better corresponds to human perception [15].

Finally, the Mel spectrogram representation was fed to a neural network. The input shape for the network was 192 frames by 128 Mel bands. Such process will be better explained in the subsection following the text processing part.

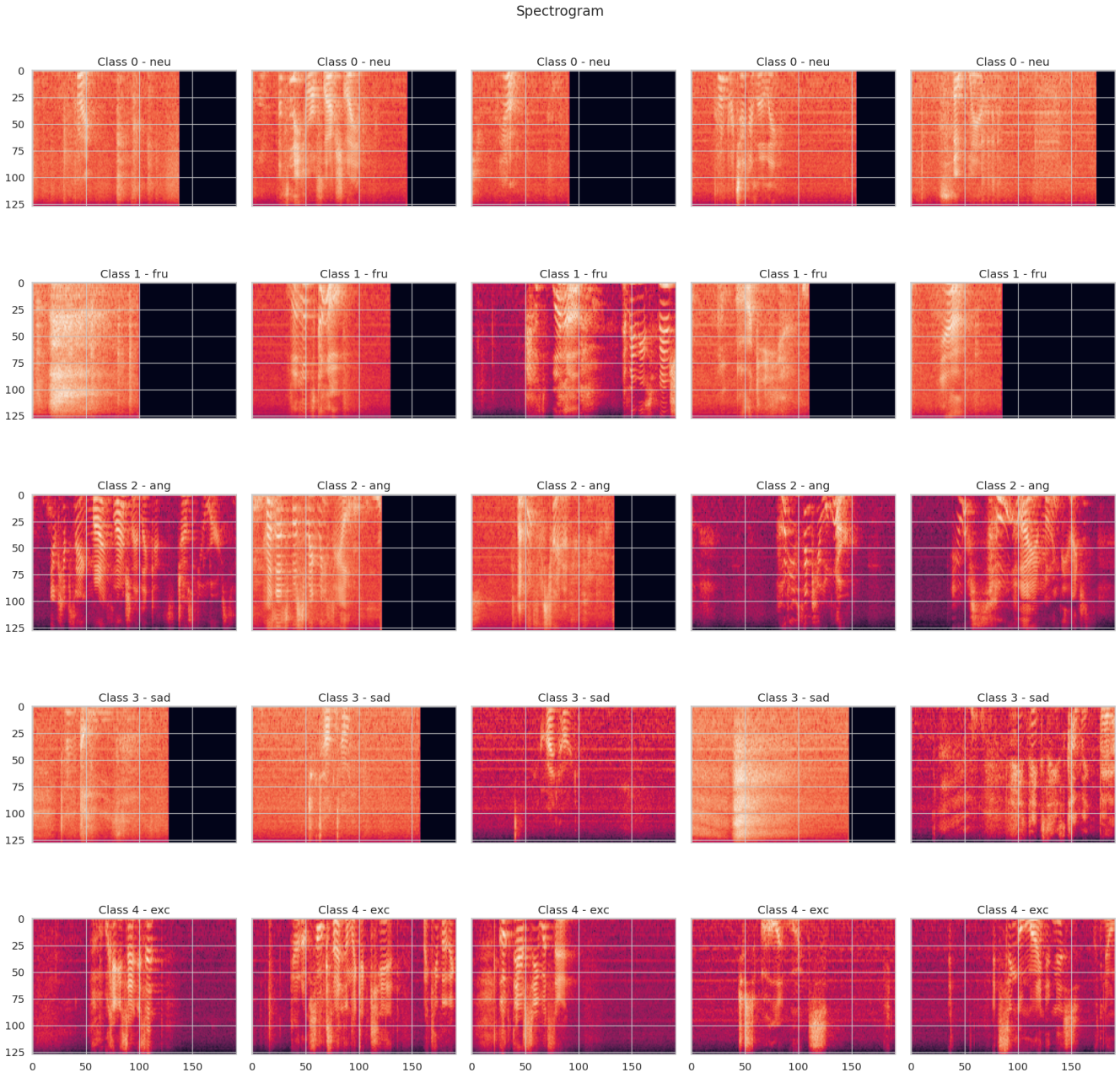


Figure 4: 5 Mel spectrograms for 5 different classes (time x frequency x intensity)

## 3.3 Text processing

The goal in this part was building a text detection model that can predict the emotion of a given text on 5 different classes of emotions. The model was built using the BERT tokenizer and the BERT model from Hugging Face.

The first step in building the model is to collect the data from the IEMOCAP database and then preprocess it. To ensure that only relevant sequences of text were used for the subsequent analysis, a preliminary filtration process was conducted, which involved removing unlabeled text data and those labeled to classes that would not be considered for the proposed model. Following the filtration process, the total number of text data available for analysis amounted to approximately 7500 sequences of characters, labeled to one of the 5 corresponding emotions (Neutral, Frustrated, Angry, Sad, Excited). After such step, the data was preprocessed by removing any unwanted characters and converting all text to lowercase to ensure consistency and uniformity within the transcriptions.

The following step was passing this data to the BERT tokenizer, wich is a tool that is used to break the input text into tokens that can be processed by the BERT model. The input text is tokenized into a sequence of subwords, and each subword is assigned to an index in the tokenizer's vocabulary. For each sequence of text, the BERT tokenizer generates *input_ids* and *attention_mask*, which are two important parameters used to prepare the input data for the model.

*Input_ids* is a sequence of token IDs representing the input text after tokenization. Each token in the text is replaced with a unique ID from the BERT tokenizer's vocabulary. The sequence is then padded or truncated to a fixed length to ensure that all inputs have the same size. *Attention_mask* is another sequence that is used to indicate which tokens in the *input_ids* sequence should be attended to by the model and which should be ignored. It is a binary mask that has the same length as the *input_ids* sequence, where each value is either 0 or 1. The value 1 indicates that the corresponding token in *input_ids* is valid and should be attended to by the model, while the value 0 indicates that the corresponding token is a padding token and should be therefore ignored.

Together, *input_ids* and *attention_mask* form the input representation of the text that is fed into the BERT model. By using *attention_mask*, the BERT model can focus only on the relevant tokens and ignore the padded ones, which can help improve the model's performance and efficiency.

## 3.4 Model

The modelling step was divided into two parallel sub-steps, each one representing one modal approach (audio and text), which are later combined into a consolidated model that uses both modalities as input.

That being sad, the sub-model developed for audio classification was implemented using a CNN. This was deemed an appropriate choice due to the extensive range of architectures and parameters available for transfer learning in CNNs. Furthermore, as substantiated by prior research, the combination of CNN architecture and Log-Mel spectrogram has yielded promising results. Specifically, the CNN architecture selected for this study was the VGGish model developed by Google, which has been shown to be effective for audio classification tasks. An overview of the VGGish architecture is provided below.

| **Block** | **Components** | **Output shape** |
|---|---|---|
| Input | Mel-spectogram | 1 x 192 x 128 |
| Block-1 | Conv (3 x 3, 64) → Relu → MaxPool | 64 x 96 x 64 |
| Block-2 | Conv (3 x 3, 128) → Relu → MaxPool | 128 x 48 x 32 |
| Block-3 | Conv (3 x 3, 256) → Relu | 256 x 48 x 32 |
| Block-4 | Conv (3 x 3, 256) → Relu → MaxPool | 256 x 24 x 16 |
| Block-5 | Conv (3 x 3, 512) → Relu | 512 x 24 x 16 |
| Block-6 | Conv (3 x 3, 512) → Relu → MeanPool(temporal axis) | 512 x 16 |

Table 1: Implemented VGGish architecture

The chosen sub-model to deal with text classification was BERT, a state-of-the-art framework based on transformers developed by Google in 2018 and trained on large amounts of text data. BERT deals with Natural Language Processing (NLP) problems with high precision and speed. The preprocessed data (*input_ids* and *attention_mask*) was fed to the pre-trained BERT model. By doing so, BERT is able to extract meaningful features from the input text and use those features to predict the emotion contained in the text.

After building and deploying both sub-models, a more general architecture was built in order to support both audio and texts as inputs. The consolidated model processes each input independently and then concatenates the sets of features from each modality. Such pipeline is called late fusion and it is described by pre-processing and extracting the features of each modality independently to later consolidate all contemplated modalities into a single model. Late fusion allows the use of different models on different modalities, thus allowing more flexibility [16].

Once the final model was established, fully connected layers were incorporated into the network to enable classification of input data. The output from this layer provides probabilities for each of the pre-defined emotions and its structure can be seen below.

| Block | Components | Output shape |
|---|---|---|
| Input | Concatenated features | 4096(VGGish) + 768(Bert) |
| Block-1 | Dense (4864, 512) → Relu | 512 |
| Block-2 | Dense (512, 512) → Relu | 512 |
| Block-3 | Dense (512, 5) → Softmax | 5 |

Table 2: Implemented classifier architecture

To ensure optimal performance, the parameters of the two last layers of the VGGish and the entire BERT were fine-tuned. Moreover, a hyperparameter optimization was performed as to find the best parameters for improving the accuracy and F1-score. Finally, the network was trained in 50 epochs using a learning rate of $2 \cdot 10^{-5}$ with a scheduler (warm-up during 20% of the training and a linear decaying). The final set of chosen hyperparameters is shown below.

```
# Hyperparameters used in feature and example generation.
NUM_FRAMES = 96*2 # Number of frames in each band
NUM_MELS = 64*2 # Number of mel bands
EXAMPLE_SIZE = 3 # Max duration of the audio to crop
READ_OR_GEN = True # True -> read data. False -> generate data from IEMOCAP

# Hyperparameters used in training.
LEARNING_RATE = 2e-5 # Learning rate for the Adam optimizer.
BATCH_SIZE = 32
NUM_EPOCHS = 50
OPTIONS = 1 # 1 -> train the model. 2 -> use pre-trained model on IEMOCAP
SEED = 71
```

Table 3: Set of chosen hyper parameters

In order to better understand the whole process, a pipeline for the consolidated algorithm is shown in figure 5 in the form of a diagram.

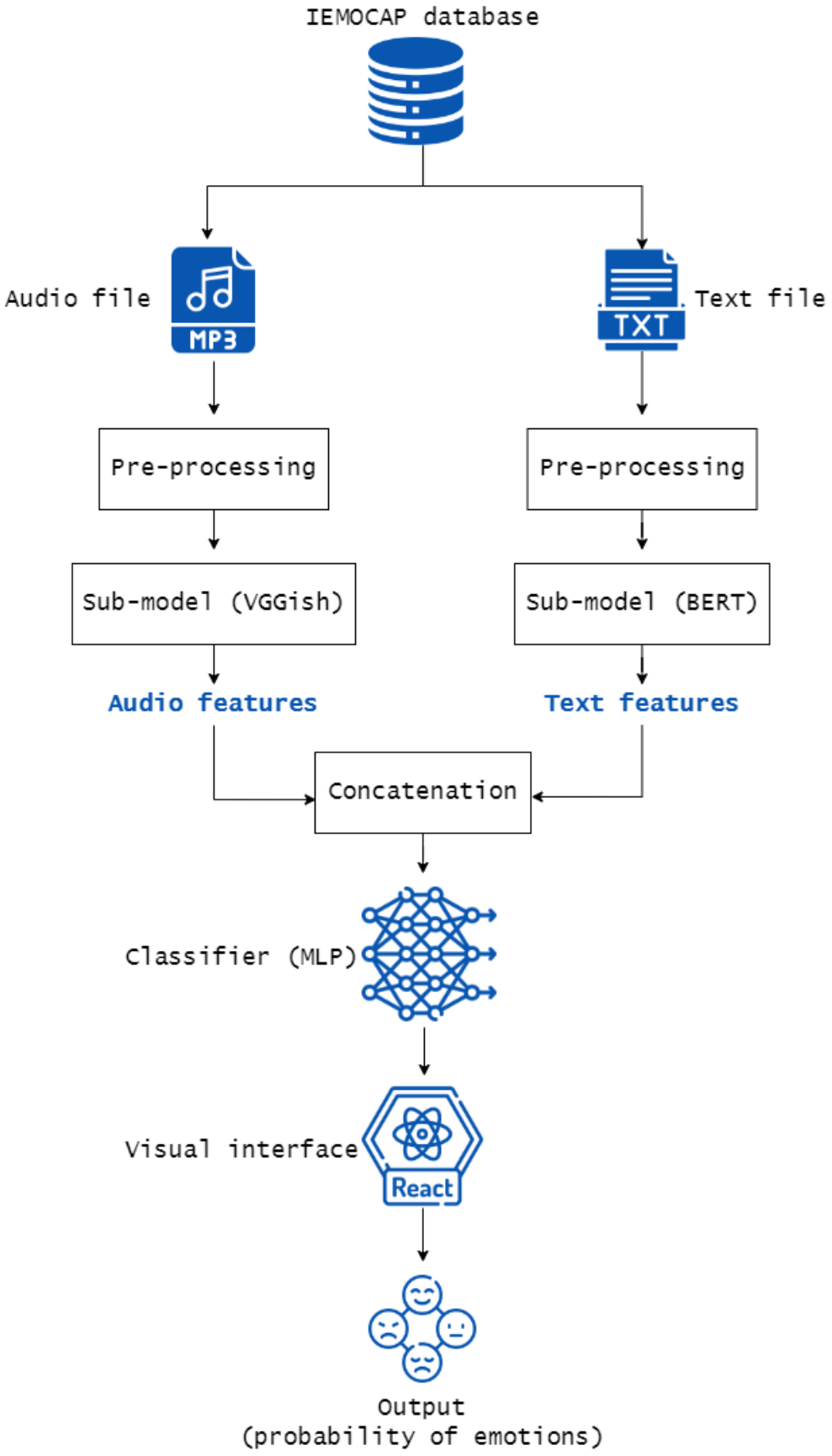


Figure 5: Model pipeline

# 4 Results

## 4.1 Performance metrics

The obtained results involved 5 classes of emotions: Neutral, Frustration, Angry, Sad, and Excited. Figure 6 shows the training history with the loss and accuracy for each epoch. When using only audio and the VGGish model, our system achieved an accuracy of 0.601 and a F1-score of 0.607.

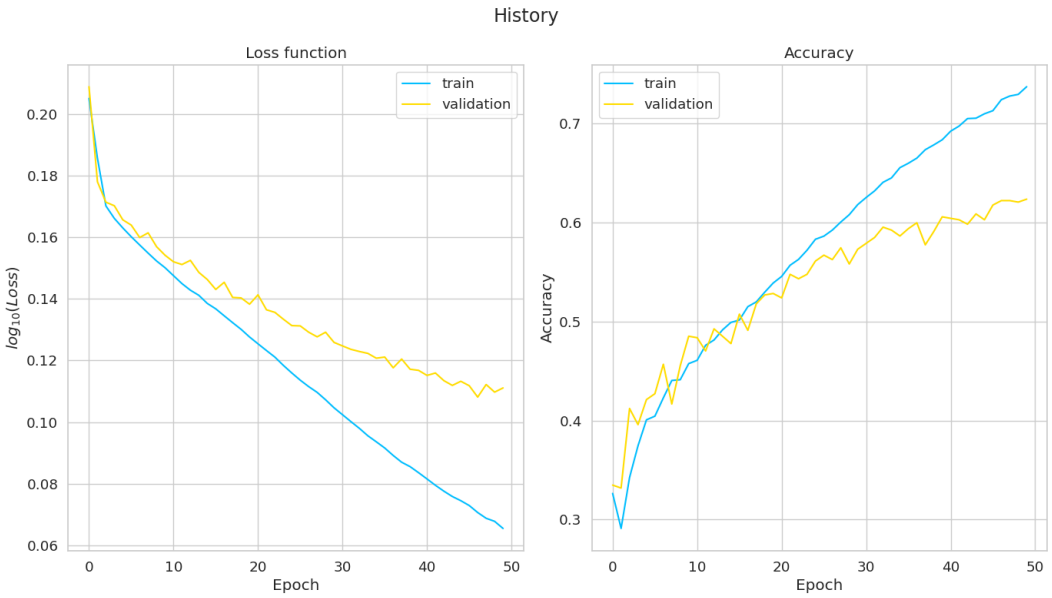


Figure 6: Loss functions and accuracy of the audio model

The analysis of the confusion matrix (Figure 7) revealed that while the system performed well in detecting the Sad and Excited emotions, it struggled to distinguish Neutral, Frustration, and Angry.

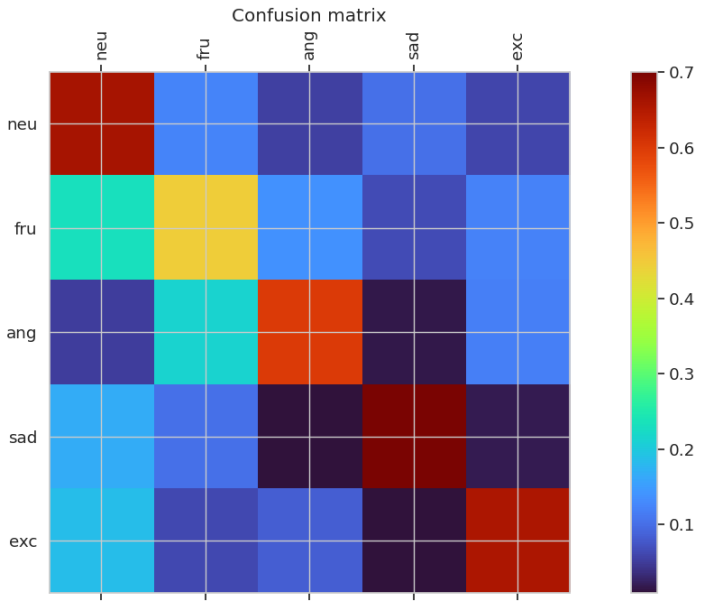


Figure 7: Confusion matrix for the audio model

In contrast, by solely leveraging textual data and applying the BERT model, we achieved a higher accuracy of 0.655 and a F1-score of 0.661, as depicted in Figure 8.

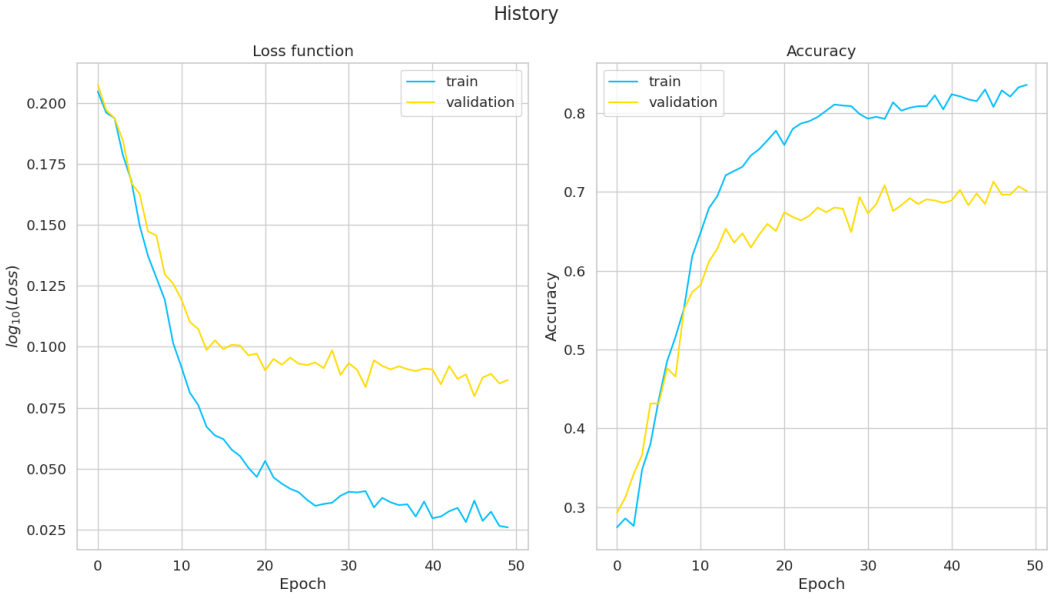


Figure 8: Loss functions and accuracy of the text model

However, analyzing the confusion matrix, as presented in Figure 9, we observed that the model encountered some difficulty in detecting Neutral, often misclassifying it as Frustration, despite the matrix being approximately an identity.

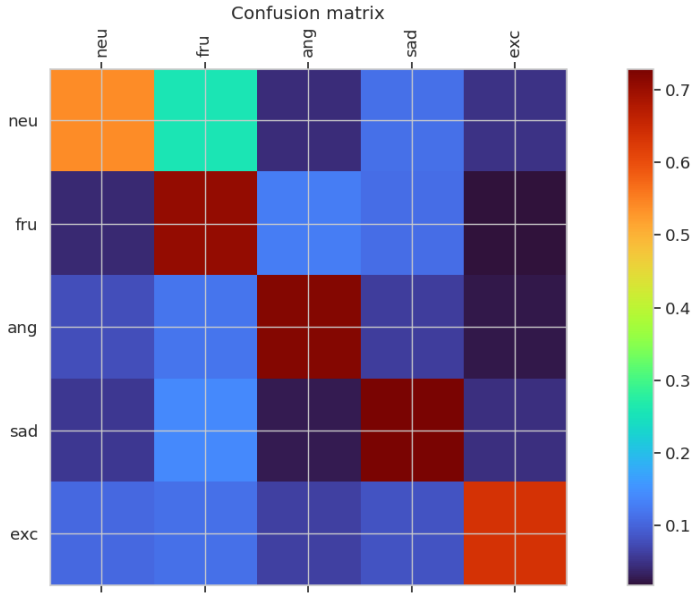


Figure 9: Confusion matrix for the text model

When incorporating both text and audio for multi-modal classification, we significantly improved our results, with an accuracy of 0.766 and a F1-score of 0.772, surpassing the performance of the text and audio individual models (Figure 13).

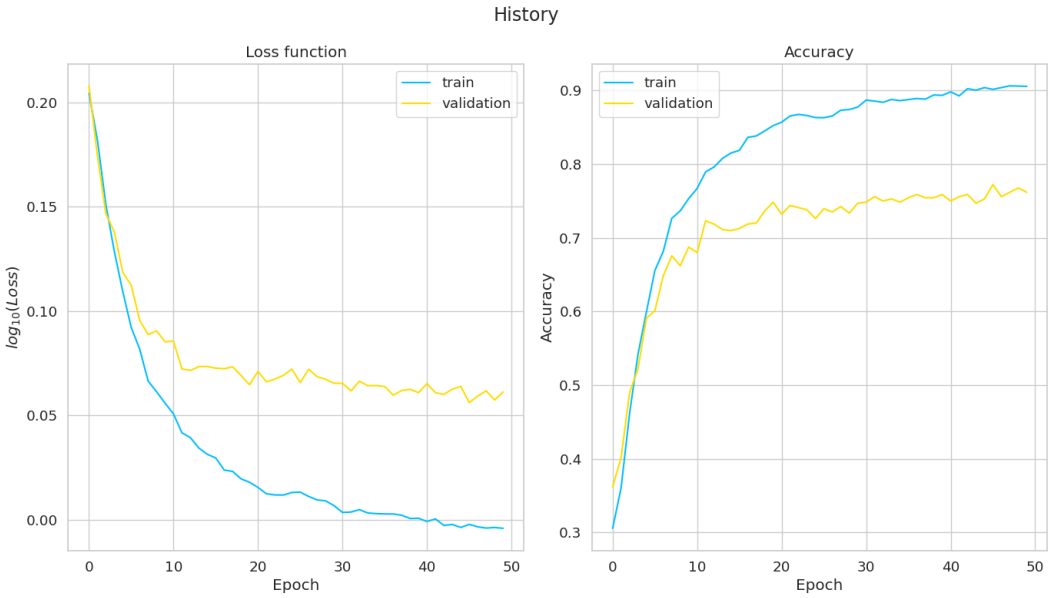


Figure 10: Loss functions and accuracy of the multi-modal model

Our analysis of the confusion matrix (Figure 11) revealed that the system's detection improved, as the matrix was much closer to the identity.

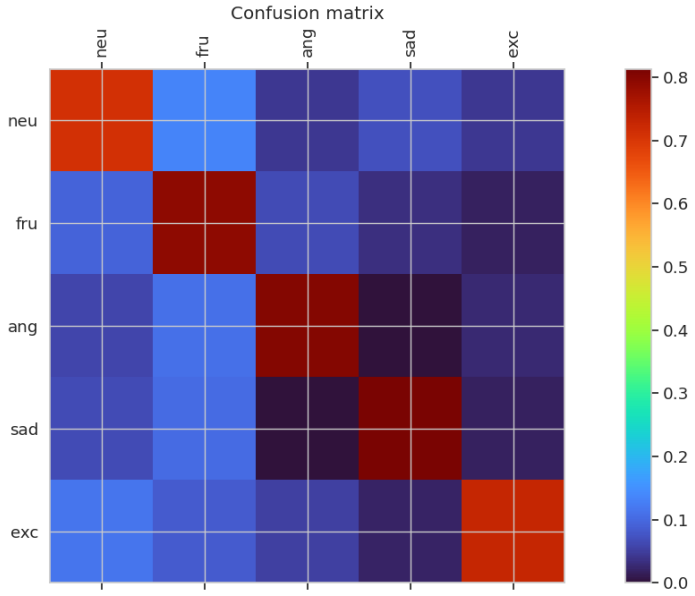


Figure 11: Confusion matrix for the multi modal model

## 4.2 Visual interface

An interface was created with the objective of facilitating the use of the model with custom data. We chose to develop a simple web interface using the React.js library to handle changes using state logic, which allowed to updating and creating a visualization of the predicted emotions easily.

The interface interacted with the AI model through HTTP requests, which were initially received by a Python Flask server, to then initiate the AI pipeline.

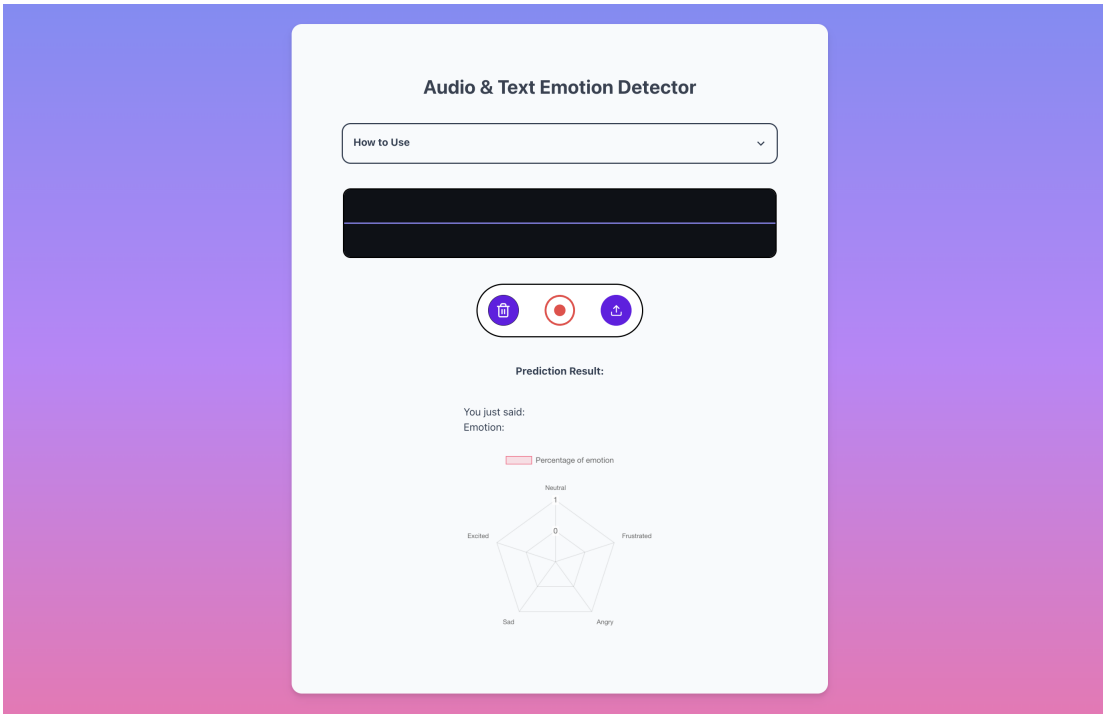


Figure 12: Visual interface

After having received the audio data from the user, the AI uses SpeechRecognitionModel from HuggingSound to get the text from the speech and then predicts the emotions based on both sources. Finally, the returned prediction is shown in the interface.

The visualization shows all of the emotions that the AI can predict along with the percentages of each one, in the format of a radar chart.

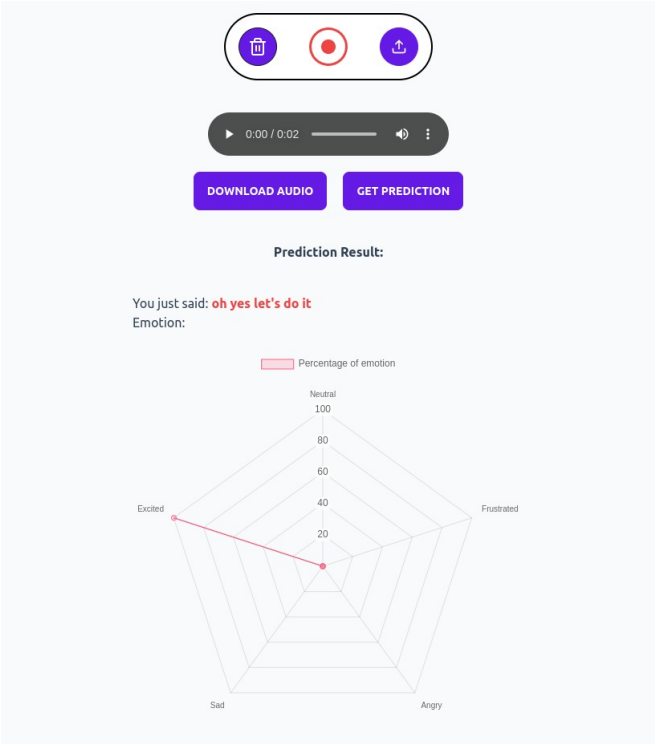


Figure 13: An example of prediction shown in the interface

An advantage of this approach is that the independence between the interface and the model allows us to test several models using the same interface, as well as the possibility of scaling the model to its usage in the cloud.

On the other hand, a potential drawback of this approach is that the use of the speech recognition model introduces a dependency on the accuracy of the predicted text from the audio.

# 5 Conclusion

As to conclude, it is safe to say that the generated model presented reasonable results, given the choices taken and the assumptions made. In the image below, one can observe a comparison between the performances of a set of published models targeted at multimodal emotion recognition on IEMOCAP [17].

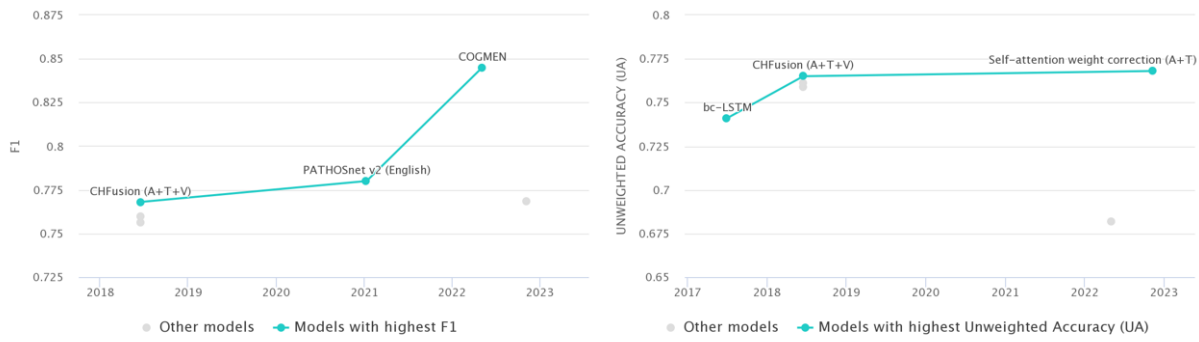


Figure 14: Benchmarks regarding the the F1-score and Unweighted Accuracy (WA) for different models.

As previously discussed, we got the followings F1-score and Unweighted Accuracy:

| **Model** | **F1-score** | **Accuracy** |
|---|---|---|
| VGGish | 0.607 | 0.601 |
| BERT | 0.655 | 0.661 |
| VGGish+BERT | 0.766 | 0.772 |

Table 4: Comparison between our models

Our last model is highly comparable to the results achieved by similar approaches in the literature. Moreover, such better attempts can be explained by a variety of factors, for example the use of the MOSAIC database as a supporting input, the consideration of a smaller subset of emotions, and, of course, the incorporation of all three modalities rather than just two.

As for its limitations, our model can only classify up to five different emotions with reasonable accuracy and execution time. However, this limitation can be overcome by employing a more powerful machine, and also by revising the architecture in order to support a broader range of emotions, even if they are less balanced. Another constraint to consider is that our model is person-dependent, meaning that both the training and testing sets comprise the voices of the same actors featured in the original audio data set. An alternative person-independent approach would involve training the model on the voices of one set of actors and then testing it on a separate set with audio recordings from different actors. While this approach may deliver poorer results for the same architecture, it could offer superior generalization capabilities for detecting emotions in future applications.

Finally, there is room for improvement by incorporating all modalities, which would likely enhance overall performance by allowing for further cross-modality fusion. Additionally, the utilization of cutting-edge algorithms, such as transformers for speech recognition, could also improve feature extraction, although requiring more powerful processing tools. Such avenues for improvement demonstrate opportunities for future research in the topic, ultimately leading to more advanced and effective models for emotion recognition.

# References

[1] LISETTI C., NASOZ F., LEROUGE C., and et al. **Developing multimodal intelligent affective interfaces for tele-home health care**. *International Journal of Human-Computer Studies*, 59(1):245–255, 2003. Applications of Affective Computing in Human-Computer Interaction.

[2] SIKANDAR M.A. **A Survey for Multimodal Sentiment Analysis Methods**. 2014.

[3] YADOLLAHI A., SHAHRAKI A., A.G., and ZAIANE O.R. **Current State of Text Sentiment Analysis from Opinion to Emotion Mining**. *ACM Computing Surveys (CSUR)*, 50:1 – 33, 2017.

[4] VERMA G.K. and TIWARY U.S. **Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals**. *NeuroImage*, 102:162–172, 2014. Multimodal Data Fusion.

[5] TRUTA T.M., CAMPAN A., and BECKERICH M. **Efficient Approximation Algorithms for Minimum Dominating Sets in Social Networks**. *International Journal of Service Science, Management, Engineering, and Technology*, 9:1–32, 04 2018.

[6] BUSSO C., BULUT M., LEE CC., and et al. **IEMOCAP: interactive emotional dyadic motion capture database**. *Language Resources and Evaluation*, 42(4):335–359, 12 2008.

[7] SHAWN S., CHAUDHURI S., ELLIS D., and et al. **CNN Architectures for Large-Scale Audio Classification**. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017.

[8] DEVLIN J., CHANG M.W., LEE K., and et al. **Bert: Pre-training of deep bidirectional transformers for language understanding**. *arXiv preprint arXiv:1810.04805*, 2018.

[9] CRAMER A.L., WU H.H., SALAMON Salamon J., and et al. **Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings**. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3852–3856, 2019.

[10] TUMASJAN A., SPRENGER T., SANDNER P., and et al. **Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment**. volume 10, 01 2010.

[11] ALONSO-MARTÍN F., MALFAZ M., SEQUEIRA J., and et. al. **A Multimodal Emotion Detection System during Human–Robot Interaction**. *Sensors*, 13(11):15549–15581, 2013.

[12] ISLAM J. and ZHANG Y. **Visual Sentiment Analysis for Social Images Using Transfer Learning Approach**. pages 124–130, 10 2016.

[13] ULLAH M.A., ISLAM M.M., NORHIDAYAH B.A., and et al. **An overview of Multimodal Sentiment Analysis research: Opportunities and Difficulties**. In *2017 IEEE International Conference on Imaging, Vision Pattern Recognition (icIVPR)*, pages 1–6, 2017.

[14] ROBERTS L. **Understanding the Mel Spectrogram**. *Medium*, 03 2020.

[15] CHEN K., SHEN M., YIN K., and et. al. **NeuroMV: A Neural Music Visualizer**. *Kayos's Blog*, 05 2022.

[16] SADOK S. **Multimodal Deep Generative Models**. `https://centralesupelec.edunao.com/mod/url/view.php?id=116961`.

[17] **Multimodal Emotion Recognition on IEMOCAP**. `https://paperswithcode.com/sota/multimodal-emotion-recognition-on-iemocap`. Accessed: 2023-03-29.