CentraleSupélec

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

Supervisor:
Simon LEGLAIVE

March 28, 2023

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 paper 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. Finally, section 5 evaluates the performance of the obtained results in comparison with other approaches. Possible space for improvement will also be discussed.

2 State of the Art

Sentimental detection is a major field in academy and it become more and more important, since it allow automatic interpretations about humans satisfaction and helps machine to interact with people in a better and kinder way. Usually the human brain uses different sources to interpret emotions like the sound, the context of the speech, the face expressions, the posture and many others information's. Due that, normal researches works with Multimodal Sentiment Analysis (MSA) to perform detection using different sources, normally the voice, the content of the voice and the face, but other models as [1] uses a multimodal approach with physiological signals, like the heart rate.

Sentimental analysis is divided in opinion mining and emotion mining, that are usually correlated subject (figure 1). However, in this work the focus is in emotion mining, especially emotion classification. As previously mentioned the most common approach for emotion classification are the sound, the text and the image of the human's face. Using text is possible to detect people emotion and opinions, images makes possible to detect the current emotion of someone and audio allows emotion recognition but it face problems due to different voice features such as accent, speaker, speaking style, language, etc. Even through, all the expressions and actions related with emotions changes related to the culture. The human's emotions are mainly divided in 6 simple sentiments (i.e. anger, sadness, surprise, fear, disgust, and joy), and non-basic emotions (e.g. fatigue, pain, agreeing, engaged, curiosity, irritation and thinking) [2]

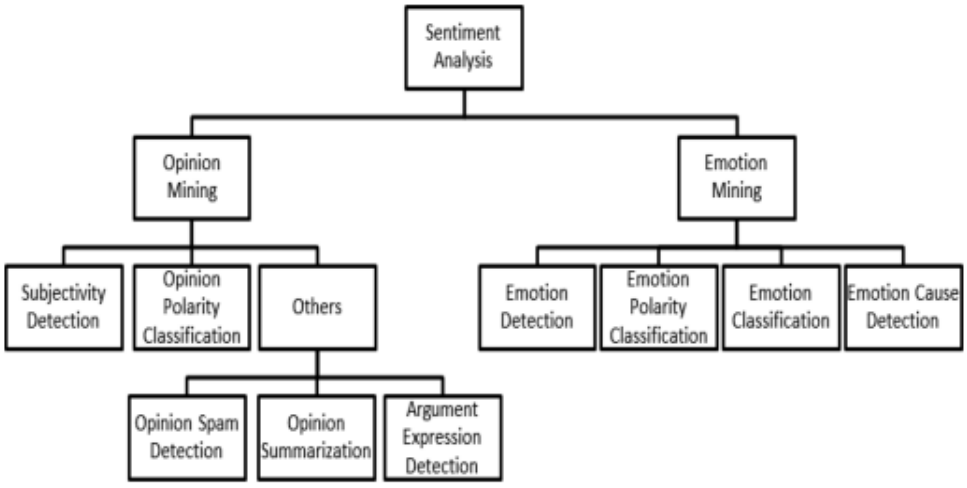

Figure 1: Taxonomy of sentiment analysis tasks [3]

Today, the world is more connected than ever and the social networks are a great tool to expresses themselves and to collect data. [4] shows how useful data are being collected by social networks as Instagram and twitter. With this data, is possible to detect user's opinion and satisfaction about some content and advertisements ([5]). [6] uses a CNN with performs better than the previous state-of-art algorithm using

transfer learning to reduce the size of the dataset to understand users feedback and emotions in general, being able to analyse large scale big data multimedia content.

Due to the great development of this area, it is needed to have large datasets with quality content. Big companies like google are creating and working on huge database like YouTube-100M, which is a dataset with more than 100 milions YouTube videos, and also ImageNet, an image database. Other huge dataset is IEMOCAP ([7]) which counts with more than 10Gb of videos, sound, text, and motion capture of face in many situations performed by actors and labeled by professionals.

Also, a huge effort is being made to improve the tecnology to better classify human emotion. Google proposed Vggis ([8]) which is a state-of-the-art Convolutional Neural Network to detect emotions in audio. For text classification, one main algorithm is bert [9] which uses a "masked language model" (MLM) for pre-training objective. Also, other models has shown good results as [10] that used a L3-net (Look, Listen, Learn) which is capable of out-peform vggish and sound net having less parameters and less data.

Emotion detection has proven useful in other areas, as example [11] proposed a model which uses the emotion to predict elections consequences based mainly on twitter's post and concluded that this social networks can plausibly reflects the offline political landscape. Also, [12] proposed a multi-modal user-emotion detection using face images and the voice to detect emotions. For that the article uses Gender and Emotion Voice Analysis (GEVA) algorithm for the voice and Gender and Emotion Facial Analysis (GEFA) algorithm for the face detection, integrating with a Robot Operating System (ROS) . [6] uses a CNN with performs better than the previous state-of-art algorithm using transfer learning to reduce the size of the dataset to understand users feedback and emotions in general, being able to analyse large scale big data multimedia content. Figure 2 shows many applications MSA can be used.

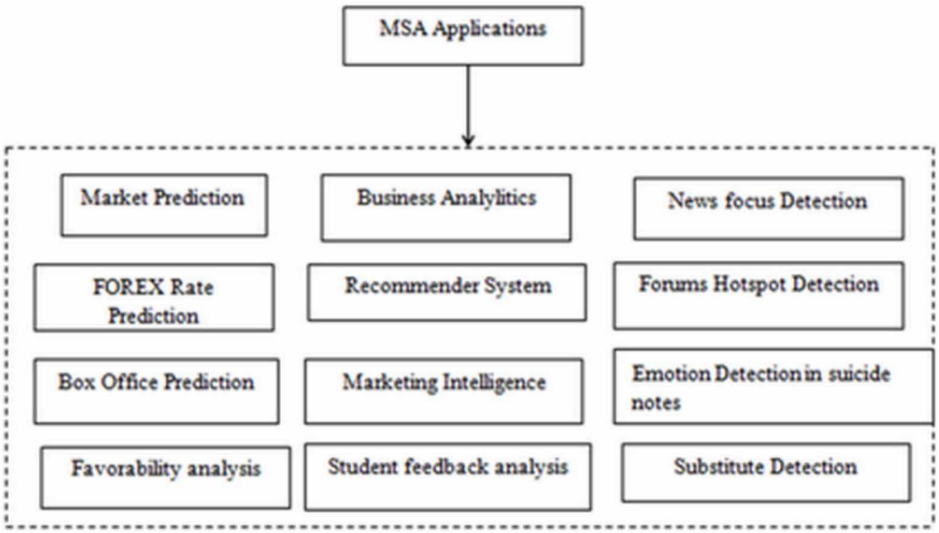

Figure 2: Multimodal sentiment analysis applications [13]

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, speech, and motion capture. In total, there is approximately 30 hours of recorded data, resulting in a final database size of 11 GB.

After extracting the IEMOCAP database, our time has decided to focus on two modalities: speech 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 speech 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 databases included only the following emotions: neutral state, frustration, anger, sadness, and excitement. The figure below illustrates the number of occurrences for each of these emotions.

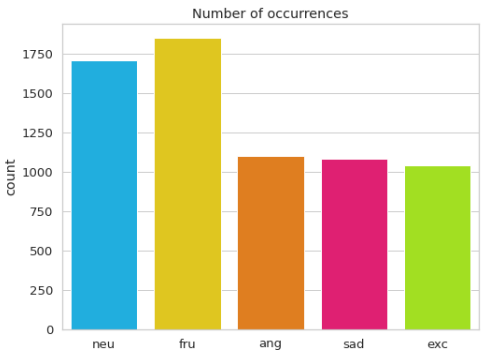

Figure 3: 5 Mel spectrograms for 5 different classes (time x frequency x intensity)

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 analyses, 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 288 frames by 193 Mel bands. Such process will be better explained in the subsection following the text processing part.

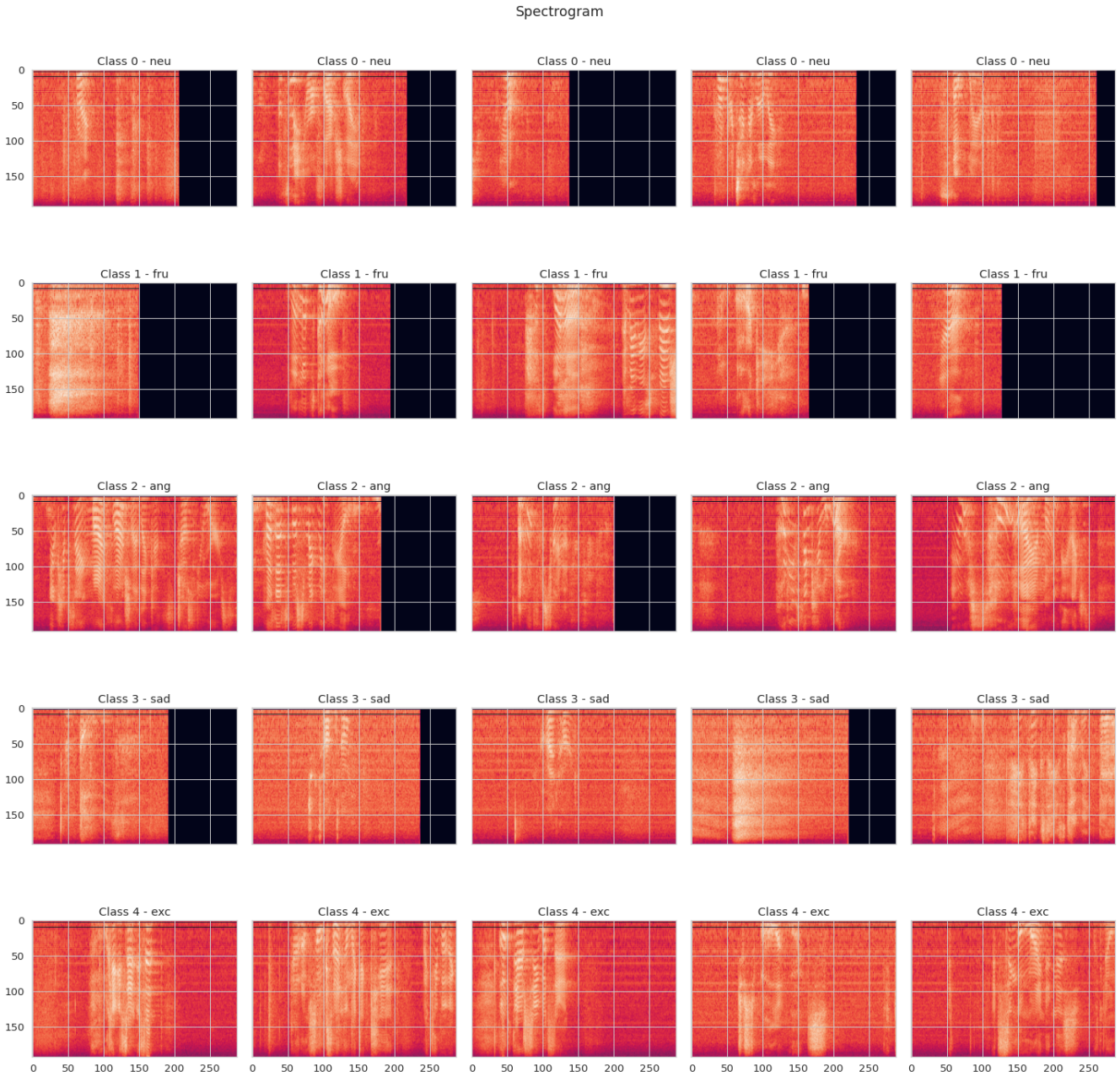

Figure 4: 5 Mel spectrograms for 5 different classes (time x frequency x intensity)

3.3 Text processing

tokenizer maxtoken size padding

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 initial sub-model developed pertained to audio classification, which was implemented using a Convolutional Neural Network (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.
COLOCAR IMAGEM DA ARQUITETURA VGGISH !!!!!!!!!!

The chosen sub-model to deal with text classification was BERT (Bidirecional Encoder Representations from Transformers), a state-of-the-art framework based on transformers developed by Google in 2018. BERT deals with Natural Language Processing (NLP) problems with high precision and speed.

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 models on different modalities, thus allowing more flexibility [16].

Once the final model was established, a fully connected layer was incorporated into the network to enable classification of input data. The output from this layer was discrete, providing probabilities for each of the pre-defined emotions. To ensure optimal performance, the parameters of the two previous sub-models were fine-tuned to train the weights of the network. This involved refining the model's hyperparameters to improve its predictive accuracy and reduce the risk of overfitting. The final set of chosen hyperparameters is shown below.

```
# Hyperparameters used in feature and example generation.
NUM_FRAMES = 96*2
NUM_MELS = 64*2
EXAMPLE_SIZE = 3
READ_OR_GEN = True # [True] -> Read data. [False] -> Generate data

# Hyperparameters used in training.
LEARNING_RATE = 2e-5 # Learning rate for the Adam optimizer.
BATCH_SIZE = 32
NUM_EPOCHS = 50
OPTIONS = 1
SEED = 71
```

Table 1: Set of chosen hyper parameters

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

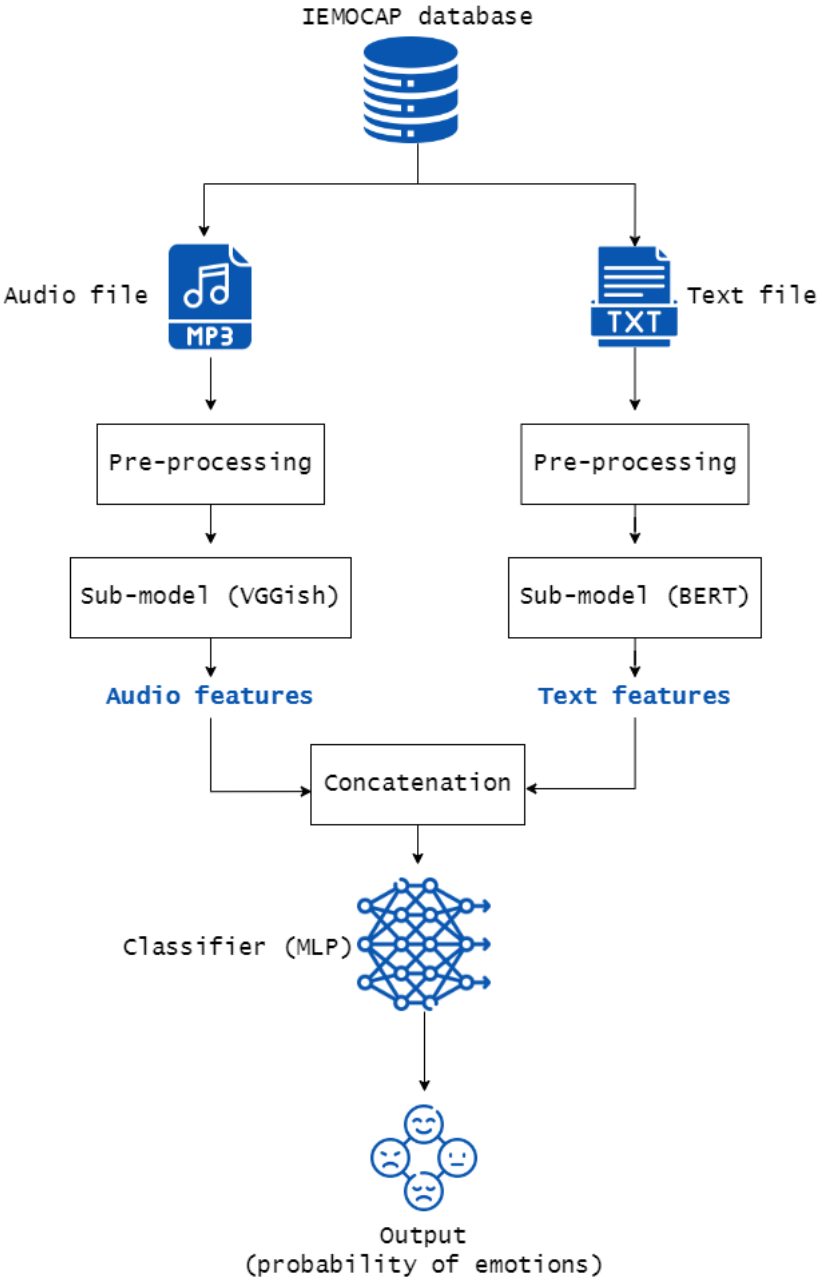

Figure 5: Model pipeline

4 Results

5 Conclusion

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[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**.