# Speech Sentiment Analysis

Katerina Xagorari
ds1200019@di.uoa.gr
National and Kapodistrian University of Athens
Athens, Greece

Christina Borovilou
ds1200008@di.uoa.gr
National and Kapodistrian University of Athens
Athens, Greece

Marios Diamantopoulos
ds2200004@di.uoa.gr
National and Kapodistrian University of Athens
Athens, Greece

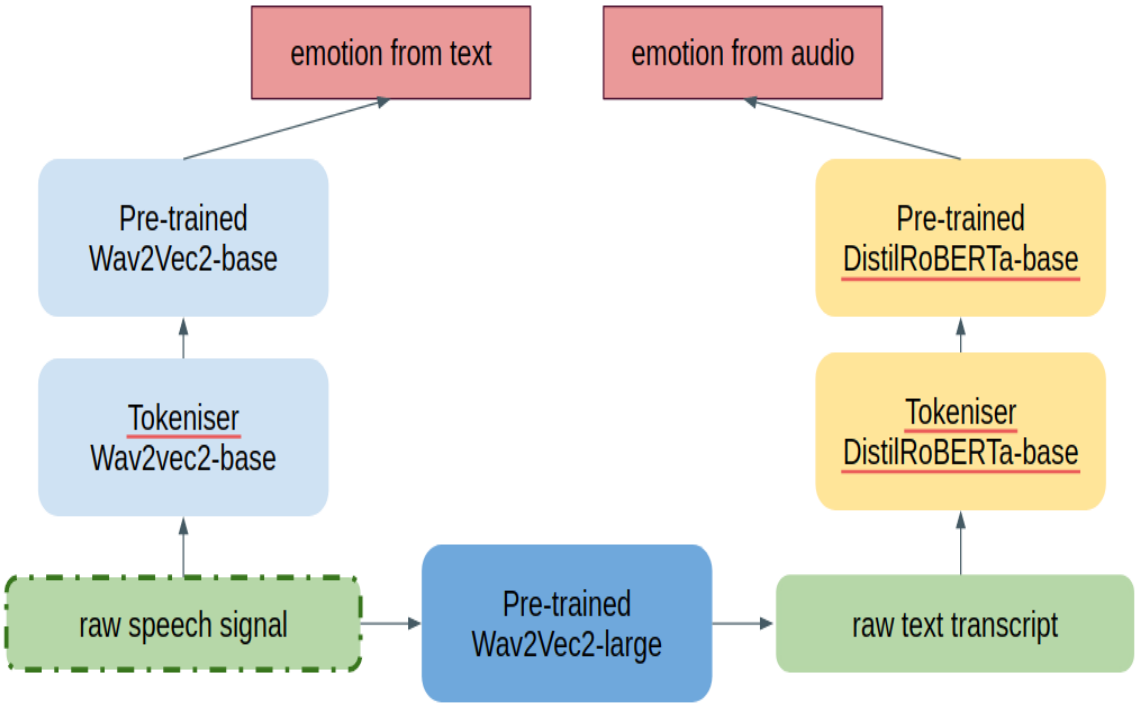


**Figure 1.** Architecture of emoAI

## Abstract

***CCS Concepts:*** • **Computer systems organization** → **Neural networks**.

***Keywords:*** transformers, speech emotion recognition, bert

## 1 Introduction

Emotions are an integral part of social interactions. Emotion recognition is a highly important task in the field of affective computing [10], with great applications in human-computer interactions and medicine. Emotional states can be identified through subjective (speech, facial expressions, body postures) as well as objective (EEG, heart rate) tests [15]. Among these, speech can be very useful in this task, as speech data are easy to be collected, even in non-laboratory settings, and adequate to contribute in recognizing mood states without the need for more information.

Speech Emotion Recognition (Speech ER) is the task of recognizing the affective features of speech, without traditionally taking into account its semantic contents. In order to be able to represent emotions schematically in 3-dimensional space, the following major features have been used: valence (which refers to the pleasantness of a stimulus), arousal (the intensity of emotion provoked by a stimulus) and dominance (the degree of control exerted). This dimensional approach has gained grounds over the last years, with researchers claiming that emotions are not independent of one another. In fact, it has been shown that valence and arousal alone can capture the majority of emotional variability [3]. The representation of emotions in 2-D space (regarding valence and arousal) are presented in Figure 1. Therefore, machine

learning models that conduct SER are based on these features, as well as the categorical labels of emotions (anger, happiness, sadness etc).

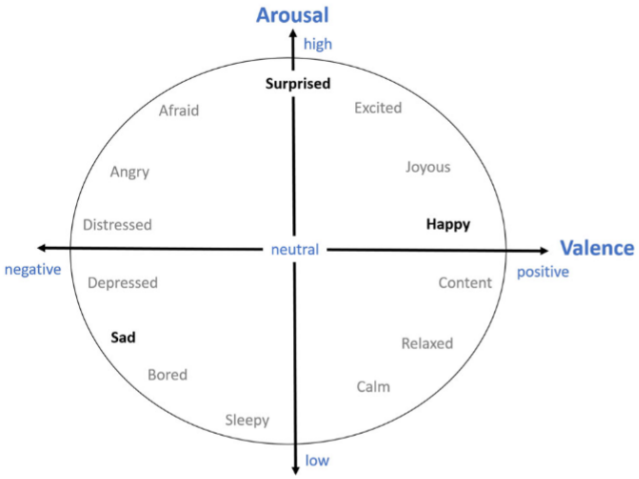


**Figure 2.** The valence-arousal model of emotion representation proposed by Russel [12]

In this work we will use only the main labels of emotions and we can use the same dataset to try multi-label models as a following step.

While humans are socially trained to recognize emotions in daily communication, automatic SER by programmable devices can be challenging. Deep Neural Networks (DNN) have been widely used to tackle this problem over the past decade [7]. An important step in this effort has been the introduction of attention mechanism in DNN models. Attention, both as a cognitive and a computational mechanism, is used to focus on relevant information and is highly significant for learning. Newer attention-based DNN models, that also take into account the semantic content of speech data, tend to perform better in SER tasks [6, 7, 17].

### 1.1 Motivation

The first motivation to work with SER was the overt need of proposing metrics useful to social sciences. Analyzing one problem quantitatively one psychologist gets a more clear image of the case, a researcher has access to arithmetic data other than abstract observations & observations and a teacher can analyze and recognize better specific behaviors of his/her students. A promising idea of the project was to combine emotion recognition simultaneously from audio & text context to frame up the result from two different approaches. A very well-developed, clearly recorded dataset came to complete the idea, providing 3 dimensions on emotion representation, that the majority of scientific community has not yet encountered (it is more common to output emotion recognition predictions to one (summarized) emotion-class labels). Taking the advantage of having access to pretrained models in Speech emotion recognition and in Text emotion recognition, we will examine the possibility to enrich the existing models with our approach.

### 1.2 Goals

1. Find an unbiased real-world dataset to train our model over Sentiment analysis though Speech. The model needs to be annotated with more analyzed voice characteristics like arousal, valence and dominance.
2. Fine-tune a model built for this specific problem [16] using both Speech text content and voice sentiment
3. Evaluate model and its efficacy

## 2 Methods

### 2.1 Dataset

The dataset used in our project was the MSP-Podcast corpus. MSP-Podcast is one of the largest speech emotional datasets, containing naturalistic speech from podcasts that are freely available to the public [8]. Furthermore, the emotional content of the dataset is balanced, so that all included emotions are equally represented. MSP-Podcast corpus is annotated with attribute-based labels (valence, arousal, dominance) with respect to James Russell and Albert Mehrabian [11], as well as categorical ones (anger, happiness, sadness, disgust, surprise, fear, contempt, neutral and other-secondary emotions) [8]. All podcasts are recorded in English. Our model was trained upon the latest version of the MSP-Pod [1] corpus (Version 1.7), which contains 62,140 speaking turns (100hrs).

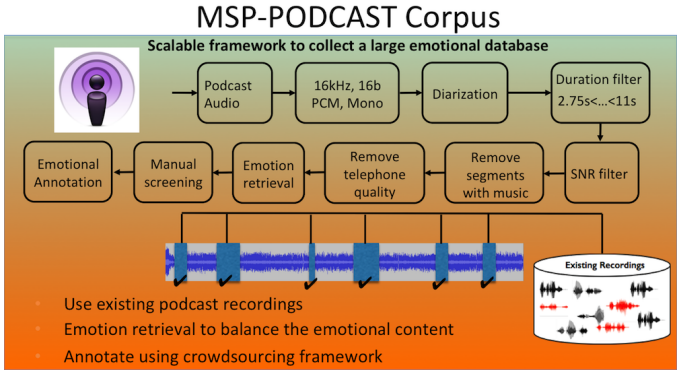


**Figure 3.** The MSP-Podcast corpus curation [9]

### 2.2 Data Meta-Info

The MSP-PODSCAST Dataset provides the following material:

[1] https://ecs.utdallas.edu/research/researchlabs/msp-lab/MSP-Pod cast.html

1. A file with suggested partition for train, test and validation sets (which we also adopted in our implementation) using as ID, the name of the audio file combined with the name of the set that is proposed to belong to (e.g. Train)
2. A folder with the audio files (73,042 speaking turns amounting to 113hr 41min in total)
3. A folder of the labels of the corpus (one detailed file giving emotional labels in all dimensions (Arousal, Valence, Dominance) and the emotion-class the audio belongs(e.g. Neutral) and one only with the consensus result) both in JSON & CSV format.

**Labeling & metrics**

1. Emotion-class takes the following values:
   - Angry (A)
   - Sad (S)
   - Happy (H)
   - Surprise (U)
   - Fear (F)
   - Disgust (D)
   - Contempt (C)
   - Neutral (N)
   - Other (O)
2. Arousal, valence and dominance are given using the Self-Assessment Manikin (SAM) where rankings go from 1 to 7
   - valence (1-very negative; 7-very positive)
   - arousal (1-very calm; 7-very active)
   - dominance (1-very weak; 7-very strong)

The proposed split is such that speaker-independent sets of dataset are created for train, test and development. More details are given below:

1. Test set 1: We use segments from 60 speakers (30 female, 30 male) - 15,326 segments
2. Test set 2: We randomly select 5,037 segments from 100 podcasts. Segments from these podcasts are not included in any other partition.
3. Validation set: We use segments from 44 speakers (22 female, 22 male) - 7,800 segments
4. Train set: We use the remaining speech samples - 44,879 segments

Due to limitation of our resources, we used 60% of given data.

### 2.3 Dataset generation

Our approach is to use both text and audio to predict emotion. In order to do so, we generated text from the given audio files using pretrained model facebook/wav2vec2-large-960h[1] This model transforms audio to text, so we were able to create files containing the audio files' names and the corresponding text that will be used for further training with text pre-trained models.

The above model is a state-of-the-art tool developed by Facebook, trained around 53K hours of unlabeled speech & 10 minutes of transcribed speech (on LibriSpeech benchmark) scoring at a word error rate (WER) of 8.6 percent on noisy speech and 5.2 percent on clean speech. The model rely less on labeled data thanks to self-supervised learning, which liberates it from the boundaries that limited labeled data impose.

### 2.4 Data Pre-processing

As described in Data meta-info, the separation of data to train, test  validation is done by referring to a partition file. Using this file we create 3 different files keeping the names of audios for train, validation & test respectively.
For training set, two pretrained models were used:

1. facebook/wav2vec2-large-960h[2]
   Same model will be used this time to execute audio emotion task, using the given labels of the dataset. The model learns contextualized speech representations by randomly masking feature vectors before passing them to a transformer network
2. distilroberta-base [13] This model is a distilled version of the RoBERTa-base model. RoBERTa is a transformers model pretrained on a large corpus of English data in a self-supervised fashion, meaning it was pretrained on the raw texts only, with no humans labelling them; automatic process generates inputs and labels for given text.

Audio to text was done with wav2vec2 Large 960h [14]

## 3 Findings and Analysis

### 3.1 Observing the Dataset

Before we start any training process, we need to "feel the data". The first idea is to check if our dataset shares audio samples for all labels uniformly. As we see at the pictures below for training & validation Dataset, there are important differences on the number of audio files among the emotional labels. This uneven distribution of data concerning their labels, will result in a model that correlates many features to the dominant labels and quite fewer features to the less famous ones. Thus, we will end up with a model that its predictions are depended on the density of data of its labels.

The accuracy calculated with the above shape of dataset is very good; it manages to reach 50%. But getting deeper to the predictions, we can see that this is not necessarily a good result. The labels with the most numerous samples dominate the predictions as well as they manage to match with more and more features in the vector space (we got mainly 3 out of 8 labels). Consequently, the need for reshaping the dataset is obvious. One thought to handle the problem is to use Proportionate Sampling and the results are shown in the next figures 6, 7:

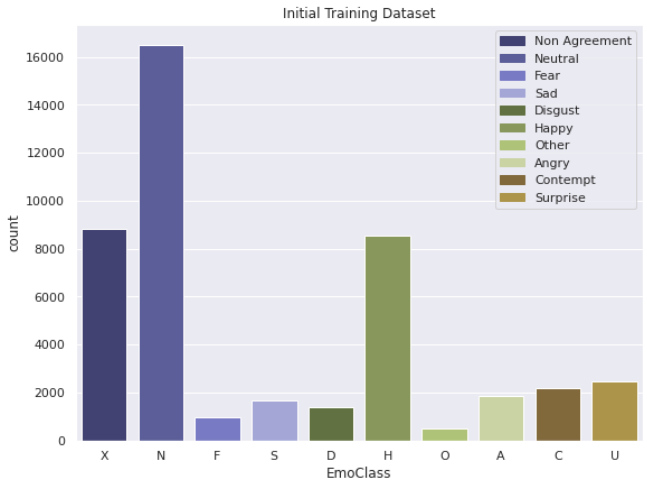


**Figure 4.** Initial Training Dataset

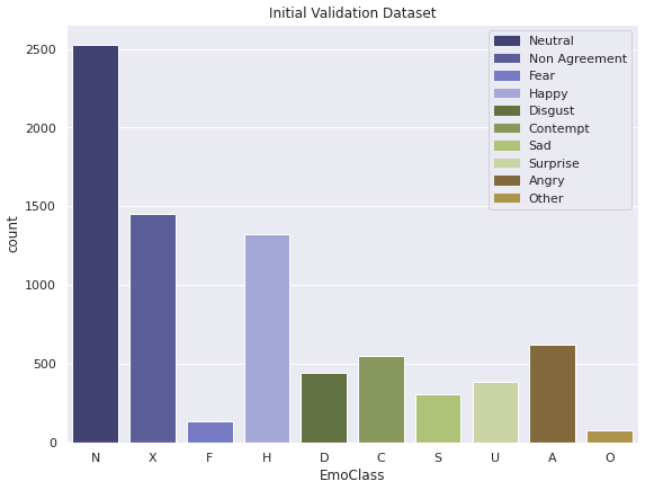


**Figure 5.** Initial Valuation Dataset

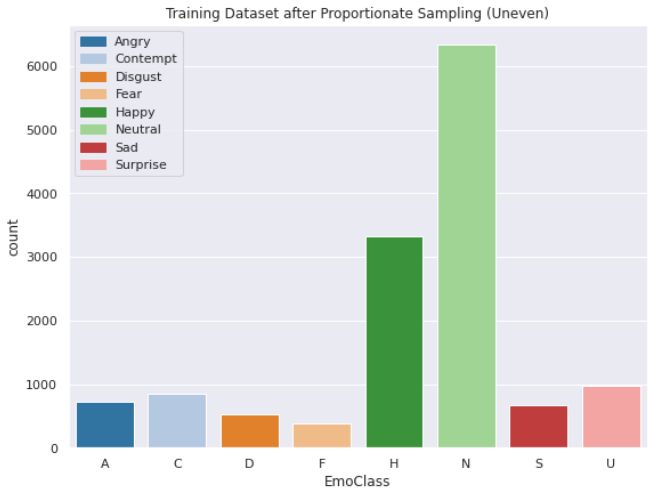


**Figure 6.** Initial Training Dataset

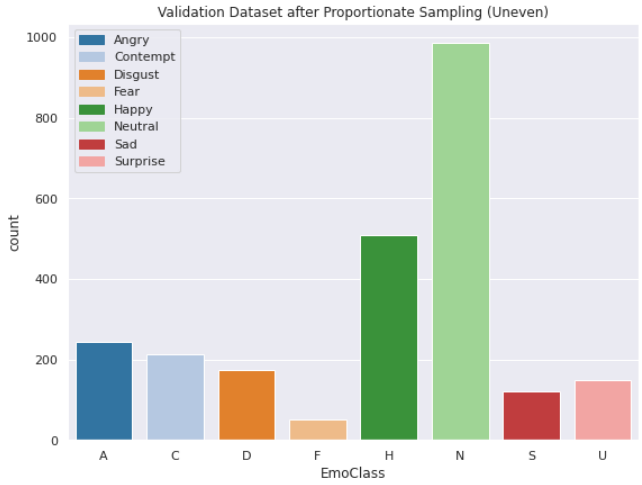


**Figure 7.** Initial Valuation Dataset

The results are not good enough even at this try. In order to surpass this difficulty, we remove the amount of data that exceed such that the population of samples for each and every label is comparable with the others. The new, even distribution of our data among the labels is shown at figures 8, 9.

### 3.2 Training Process

During training, evaluation process was taking place every 10 steps during an epoch for the whole evaluation dataset. The total number of steps in each epoch is 100. The progress of validation accuracy & validation loss for the first 15 epochs is shown below (figures 10, 11).

### 3.3 Results

**3.3.1 Results from Audio source.** The decrease in our data to achieve homogeneity, resulted to have number of samples with order of magnitude 1,000 for each label. One of the first results we took is shown in figure 12 and the things we find is that:

- The best prediction (and almost the unique) is the *happy* label
- Accuracy is 29%
- The model does not recognize any other label

We need to train our model over more and more data; the number of samples is not big enough to serve the needs of the task. After continuous training tasks, we manage to get maximum score for accuracy 29%. The new results are shown in Picture 13, where now the prediction equals '1' implies that we do not have any false positive result. The total increase of accuracy that we manged to reach varied from 10% at epoch 0 to 17% as a final achievement that we considered as promising if improvements in training & preprocessing can be applied further.

**3.3.2 Results from text source.**


**Figure 8.** Even Training Dataset

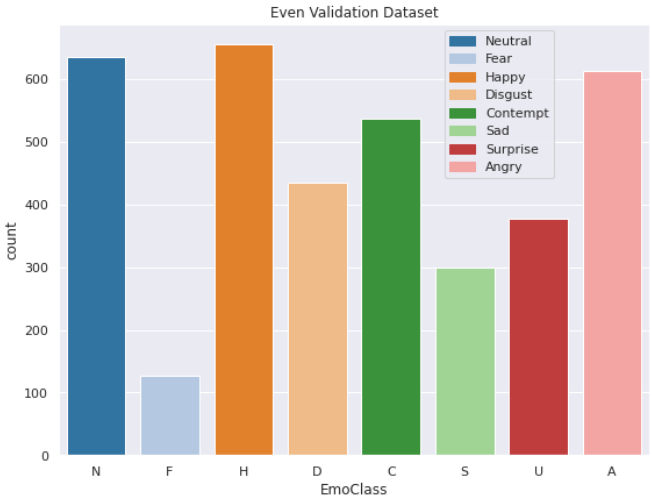


**Figure 9.** Initial Valuation Dataset

**3.3.3 Compare Pretrained model with fine-tuned.** We calculated the accuracy of pretrained model (distilroberta-base) and the accuracy of our fine-tuned model and the results are shown below:

| Model | Accuracy |
|---|---|
| Pretrained | 44% |
| fine-tuned | 35% |

This was an expected result as we know already that *distilroberta-base model* was fine-tuned with quite more data and with fewer labels than our model, which justifies the difference in accuracy

## 4 Discussion and Future Work

To sum up, the performance of the model was not quite efficient for 2 main reasons: the technical resources that were available had important limitations and secondly the homogenous data distribution was quite poor. A possible solution to this could be Data Augmentation in order to have a decent amount of data for our training. Overall in this work, we created an emotion recognition model based on a *pre-trained "BERT-like" architectures* on the one hand and *Wav2Vec* model on the other hand. "BERT-like" architectures are taking speech data as input, it can extract their semantic content into text form and combine them with sound features of speech to predict one's emotions. The first aim of our

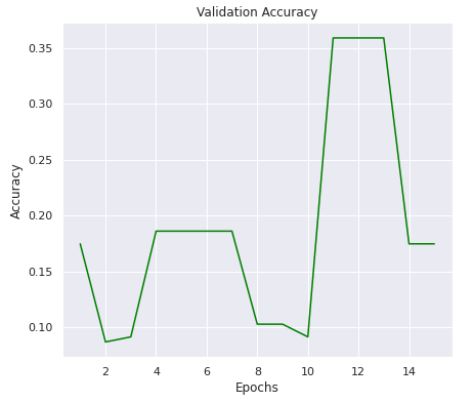


**Figure 10.** Validation Accuracy

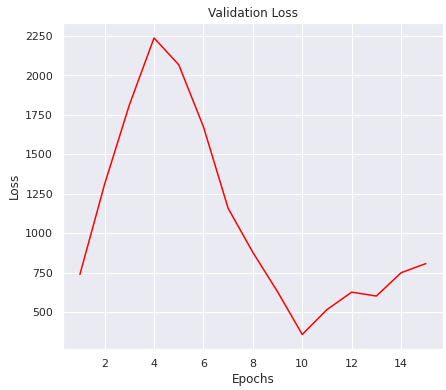


**Figure 11.** Validation Loss

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| A | 0.00 | 0.00 | 0.00 | 61 |
| C | 0.00 | 0.00 | 0.00 | 73 |
| D | 0.00 | 0.00 | 0.00 | 71 |
| F | 0.00 | 0.00 | 0.00 | 36 |
| H | 0.29 | 1.00 | 0.45 | 342 |
| N | 0.00 | 0.00 | 0.00 | 440 |
| S | 0.00 | 0.00 | 0.00 | 49 |
| U | 0.00 | 0.00 | 0.00 | 102 |
| accuracy | | | 0.29 | 1174 |
| macro avg | 0.04 | 0.12 | 0.06 | 1174 |
| weighted avg | 0.08 | 0.29 | 0.13 | 1174 |

**Figure 12.** Audio Score - initial try

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| A | 1.00 | 0.00 | 0.00 | 31 |
| C | 1.00 | 0.00 | 0.00 | 37 |
| D | 1.00 | 0.00 | 0.00 | 34 |
| F | 1.00 | 0.00 | 0.00 | 16 |
| H | 0.29 | 1.00 | 0.45 | 172 |
| N | 1.00 | 0.00 | 0.00 | 222 |
| S | 1.00 | 0.00 | 0.00 | 23 |
| U | 1.00 | 0.00 | 0.00 | 51 |
| accuracy | | | 0.29 | 586 |
| macro avg | 0.91 | 0.12 | 0.06 | 586 |
| weighted avg | 0.79 | 0.29 | 0.13 | 586 |

**Figure 13.** Audio Score - best try

project was to build a model that provides quantitative information about emotions, regarding the valence and arousal features of speech. Emotional attributes (valence, arousal, dominance) can be associated with various neuropsychiatric disorders. For example, a study showed that higher depression scores in Depression Anxiety Stress Scales (DASS) test were correlated with more negative valence of speech[5]. Furthermore, higher anxiety scores at the same test were associated with higher arousal and lower valence [5]. More studies have shown a link between depression and lower valence, arousal and dominance in speech [17]. Another robust feature of our model is the inclusion of textual data. Studies have shown that semantic features offer better results in valence estimation, whereas acoustic features are adequate to capture arousal [6]. Our model, based on "BERT-like" architectures, is able to focus on textual data and detect words that have a strong correlation with specific emotions and/or valence-arousal values. This characteristic may be able to detect brooding rumination; in other words, the tendency to focus on negative aspects of one's life. People with depression or suicidal ideation tend to ruminate in negative thoughts and repeat words of negative valence in their speech (about their low self-esteem or even death) [4]. A multimodal ER model, like the one presented here, could be further trained to detect suicidality in high-risk populations, taking advantage of semantic contents of speech. The contribution f resources is appearing at this point as audio files are very time-consuming in processing and training tasks where conventional computers may not be enough.

## 5 Conclusion

Our vision regarding this project was to create a model that could be beneficial in the neuropsychiatric field. Speech data, in combination with other modalities (facial expressions, EEG, heart rate etc), can be very useful in emotion recognition and, therefore, in patient diagnosis, follow-up and prognosis. Emotion recognition is a very challenging task.

## References


[1] Michael Auli Alexei Baevski, Alexis Conneau. [n.d.]. Wav2vec 2.0 Learning the structure of speech from raw audio. https://ai.facebook.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio/.

[2] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *CoRR* abs/2006.11477 (2020). arXiv:2006.11477 https://arxiv.org/abs/2006.11477

[3] Hatice Gunes and Maja Pantic. 2010. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions (IJSE)* 1, 1 (2010), 68–99.

[4] Alex S Holdaway, Aaron M Luebbe, and Stephen P Becker. 2018. Rumination in relation to suicide risk, ideation, and attempts: Exacerbation by poor sleep quality? *Journal of affective disorders* 236 (2018), 6–13.

[5] Philipp Kanske and Sonja A Kotz. 2012. Auditory affective norms for German: testing the influence of depression and anxiety on valence and arousal ratings. *PLoS One* 7, 1 (2012), e30086.

[6] Seliz Gülsen Karadoğan and Jan Larsen. 2012. Combining semantic and acoustic features for valence and arousal recognition in speech. In *2012 3rd International Workshop on Cognitive Information Processing (CIP)*. IEEE, 1–6.

[7] Eva Lieskovská, Maroš Jakubec, Roman Jarina, and Michal Chmulík. 2021. A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism. *Electronics* 10, 10 (2021), 1163.

[8] Reza Lotfian and Carlos Busso. 2017. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing* 10, 4 (2017), 471–483.

[9] Reza Lotfian and Carlos Busso. 2017. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing* 10, 4 (2017), 471–483.

[10] Rosalind W Picard. 2000. *Affective computing*. MIT press.

[11] James Russell and Albert Mehrabian. 1977. Evidence for a Three-Factor Theory of Emotions. *Journal of Research in Personality* 11 (09 1977), 273–294. https://doi.org/10.1016/0092-6566(77)90037-X

[12] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[13] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv* abs/1910.01108 (2019).

[14] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862* (2019).

[15] Nicu Sebe, Ira Cohen, Theo Gevers, and Thomas S Huang. 2005. Multimodal approaches for emotion recognition: a survey. In *Internet Imaging VI*, Vol. 5670. International Society for Optics and Photonics, 56–67.

[16] Shamane Siriwardhana, Andrew Reis, Rivindu Weerasekera, and Suranga Nanayakkara. 2020. Jointly Fine-Tuning "BERT-like" Self Supervised Models to Improve Multimodal Speech Emotion Recognition. arXiv:2008.06682 [eess.AS]

[17] Brian Stasak, Julien Epps, Nicholas Cummins, and Roland Goecke. 2016. An Investigation of Emotional Speech in Depression Classification.. In *Interspeech*. 485–489.