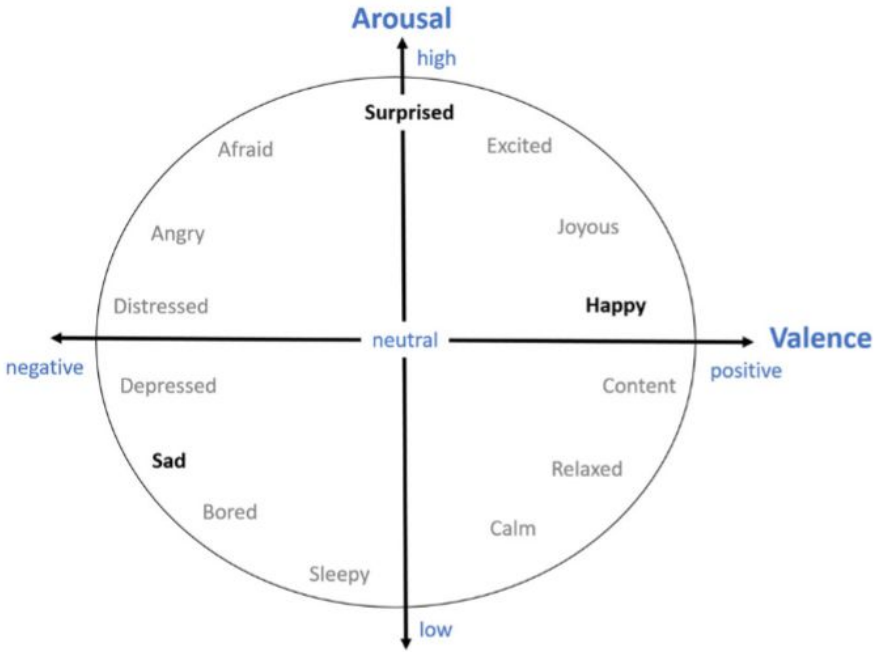
SPEECH EMOTION RECOGNITION

Katerina Xagorari - Big Data & AI
Christina Borovilou - Big Data & AI
Marios Diamantopoulos - Bioinformatics & Biomedical

MSc Data Science and Information Technologies
March 2022

MOTIVATION

- Emotion Recognition audio/text
- Audio characteristic & Sentence Content
- Prognosis/ Follow up of patient's condition
- Diagnosis should not be dependent on the model but used as a complementary component (clinical scales - questionnaire e.g. Beck's Depression Inventory)
- Audio emotion does not necessarily correlate with sentence's semantic context



Happy voice: Everyone is going to war! (Irony)

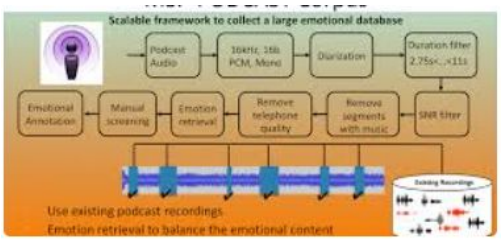
GOALS

1. Find an unbiased real-world dataset to train our model over Sentiment analysis though Speech. The model is annotated with more analyzed voice characteristics like arousal, valence and dominance.

2. Fine-tune a model built for this specific problem using both Speech text content and voice sentiment

3. Evaluate model and its efficacy

DATASET CHARACTERISTICS

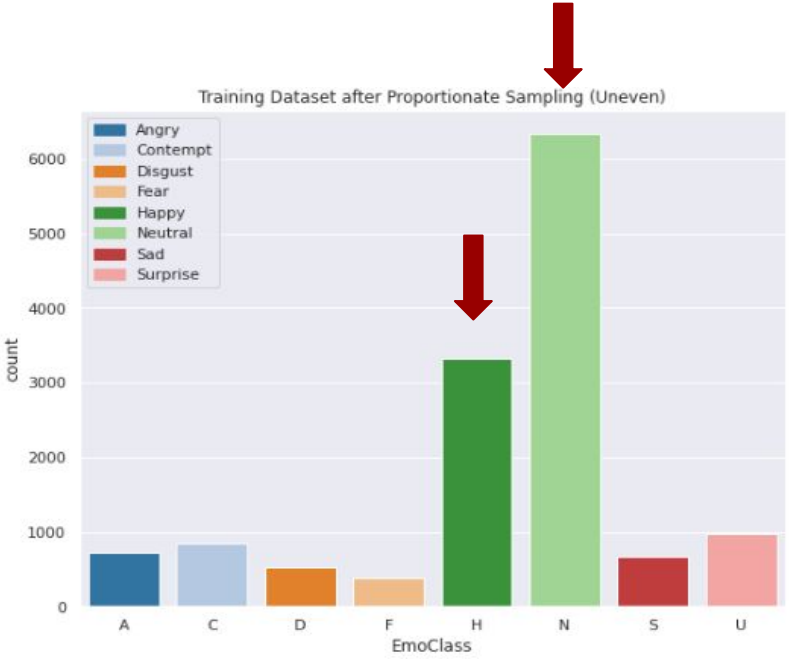


MSP-Podcast (utdallas.edu)

- One of the largest freely available speech emotional datasets (English)
- Real-world data from Podcasts has been collected & annotated from different annotators
- Audio files that have duration in total 113hr 41min (*quite rich dataset*)
- Dataset gives labels for all dimensions of emotions (Activation, Valence, Dominance) as well as single emotion labels, mentioned below:
  - ➔ Angry
  - ➔ Sad
  - ➔ Happy
  - ➔ Surprise
  - ➔ Fear
  - ➔ Disgust
  - ➔ Contempt
  - ➔ Neutral
  - ➔ Other

DATA PREPROCESSING

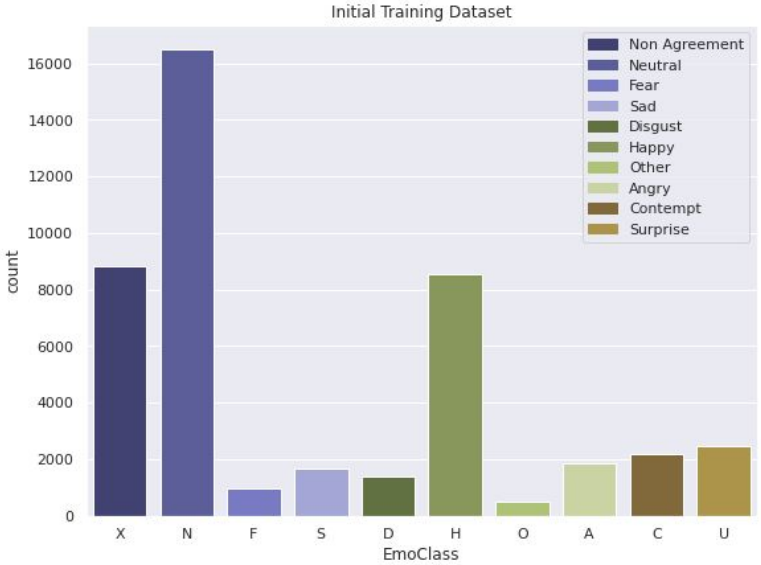
Data Cleaning is vital to get good results.
Remove unnecessary labels and even label distribution

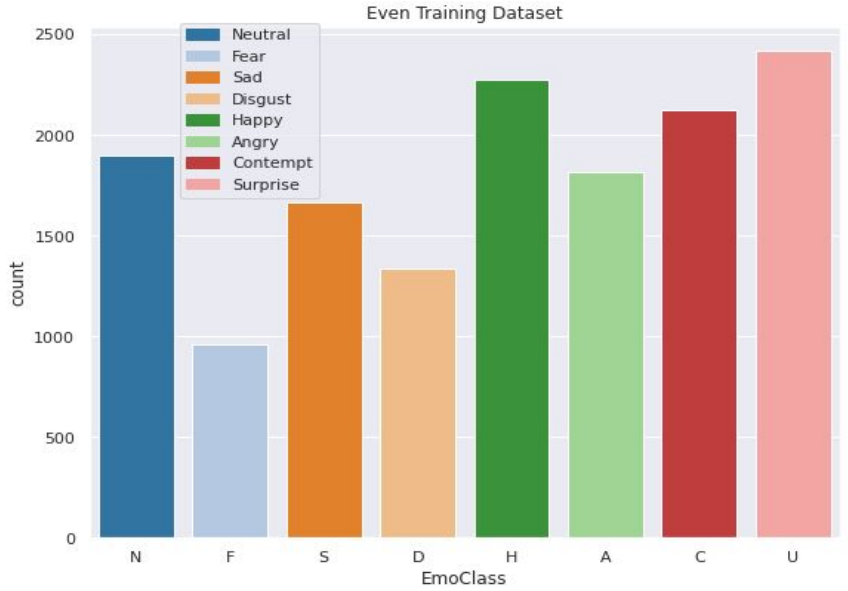


DATA PREPROCESSING

Final Dataset, data from almost all the labels have remained the same,
Happy and Neutral data have been truncated to create uniform distribution

Model should learn all emotions, not only the ones in abundance





APPROACH

Emotion recognition with general purpose pre-trained models:

1. Audio
2. Text

We will fine-tune the models with our data and compare their efficiency

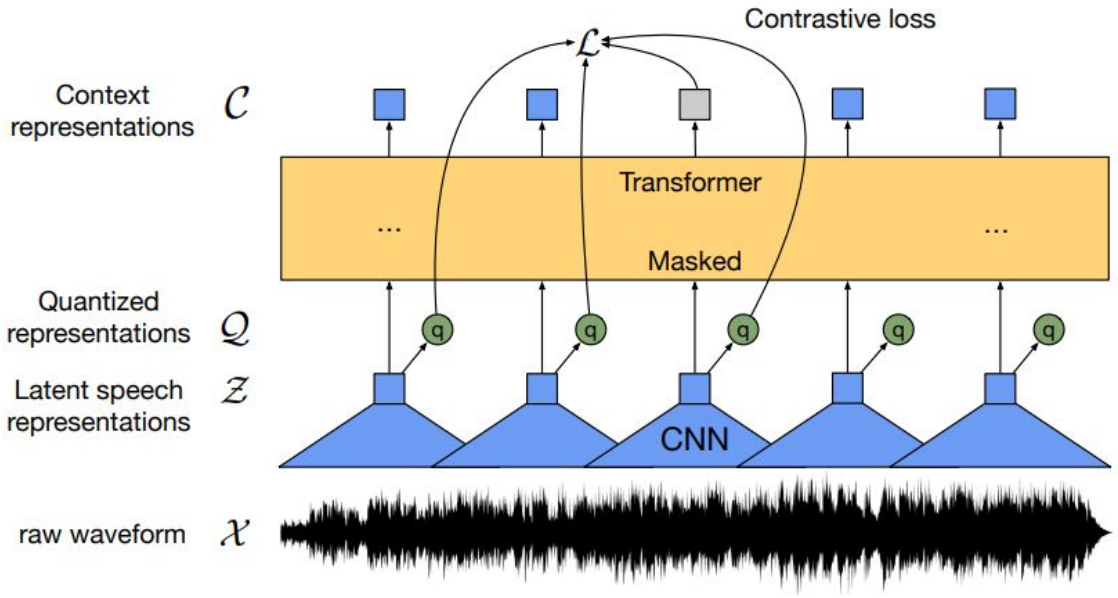
1. Audio : Wav2vec2.0
2. Text: DistilRoBERTa

Why non-conditional Deep Neural Networks? Pre-trained models use transformers which are considered state-of-the art and generalise really well on various, even unseen, tasks

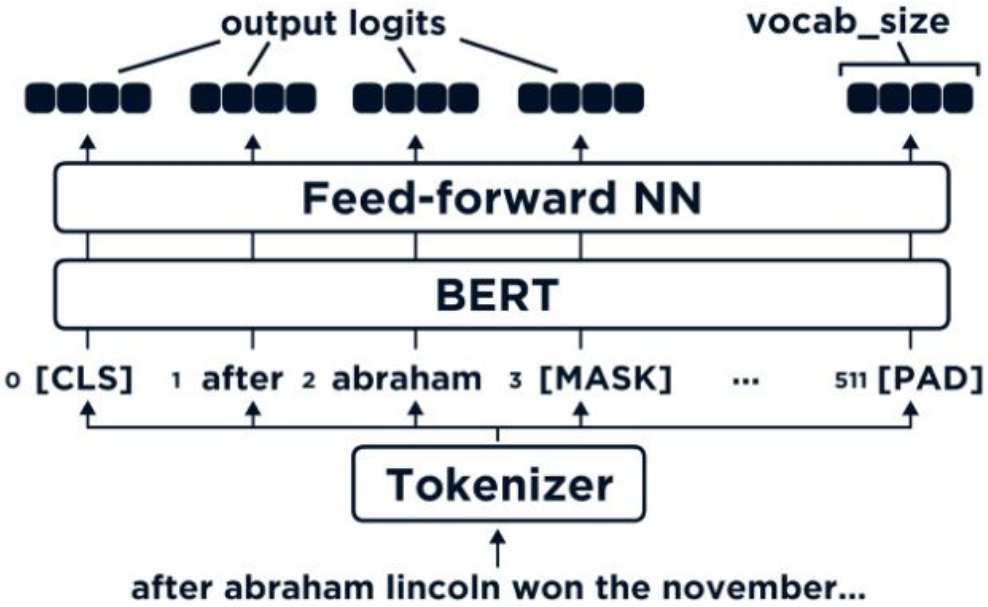
WAV2VEC 2.0 - SELF-SUPERVISED LEARNING OF SPEECH REPRESENTATIONS

Learns contextualized speech representations by randomly masking feature vectors before passing them to a transformer network

The model builds context representations over continuous speech representations and self-attention captures dependencies over the entire sequence of latent representations end-to-end



WHAT IS MASKING IN BERT?



DISTILROBERTA

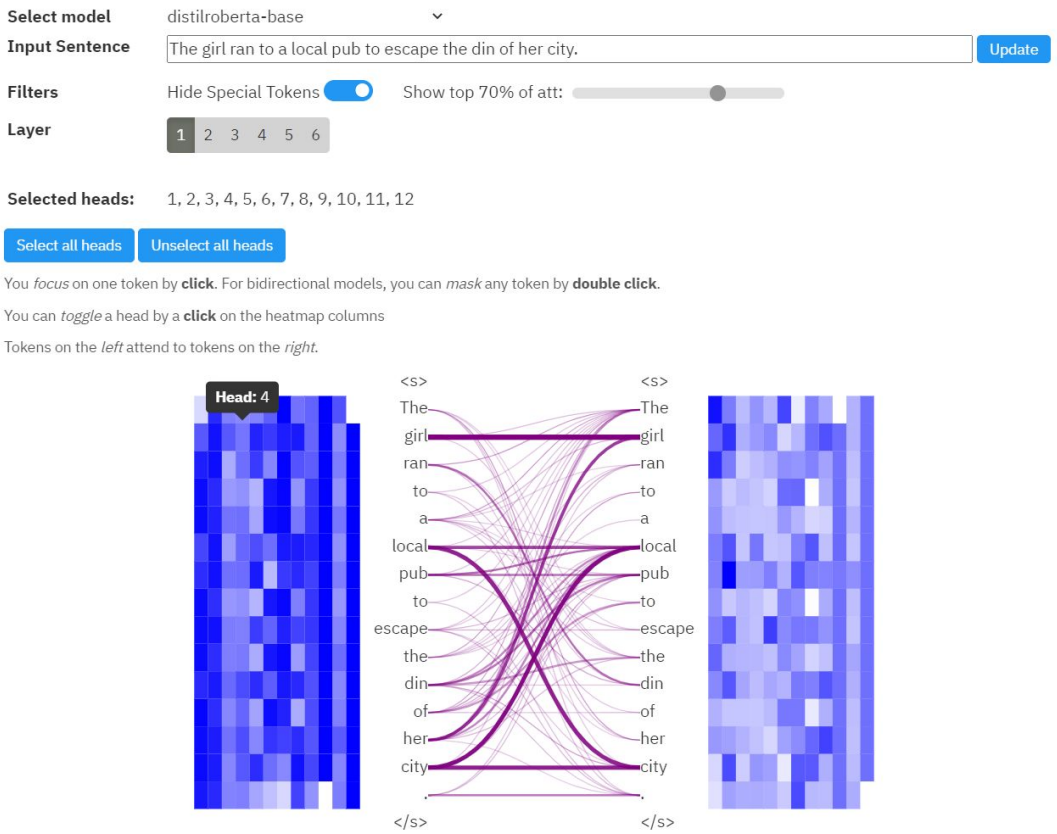
RoBERTa (**R**obustly **o**ptimized **BERT** **a**pproach) :

- dynamic masking
- FULL-SENTENCES without NSP loss
- large mini-batches
- larger byte-level BPE

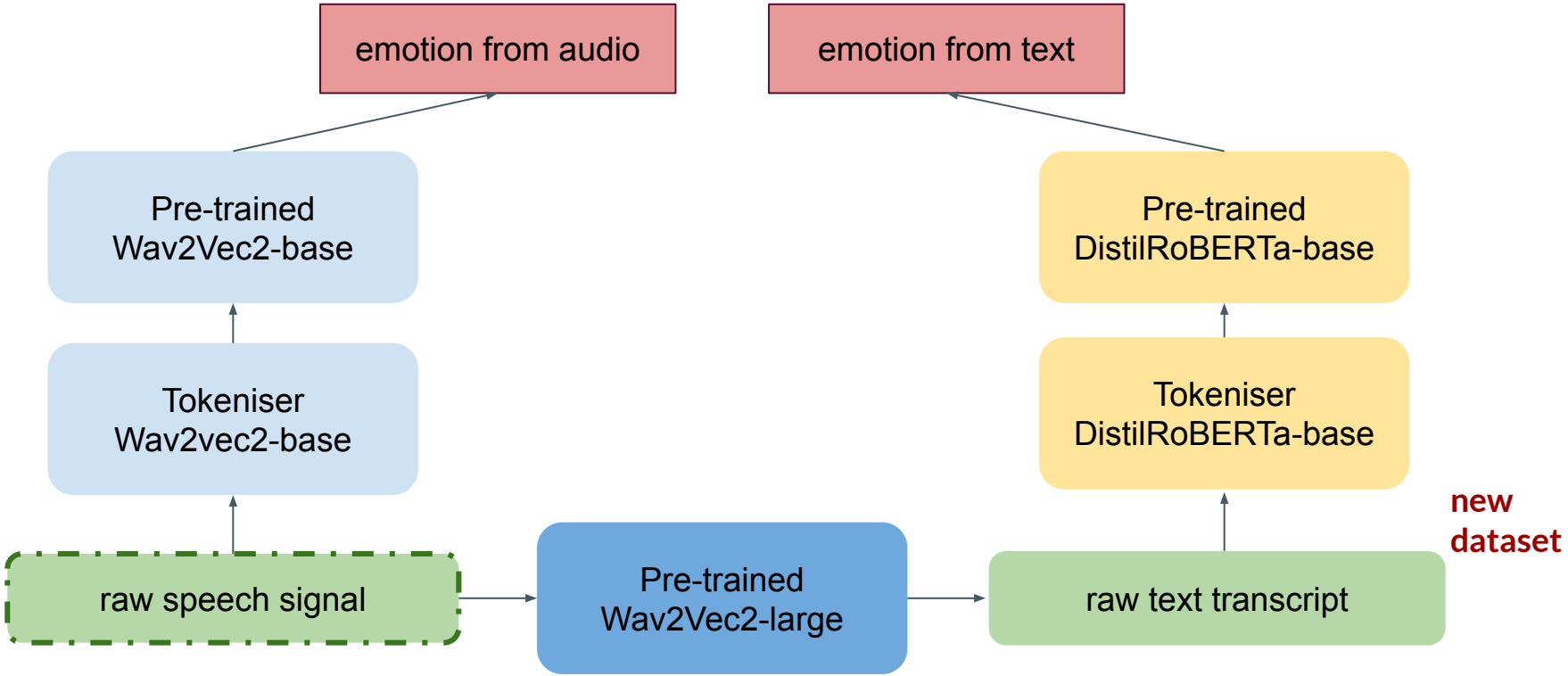
A distilled version of the RoBERTa-base model.

It was pretrained on the raw texts only, with no humans labelling them.

An automatic process generates inputs and labels

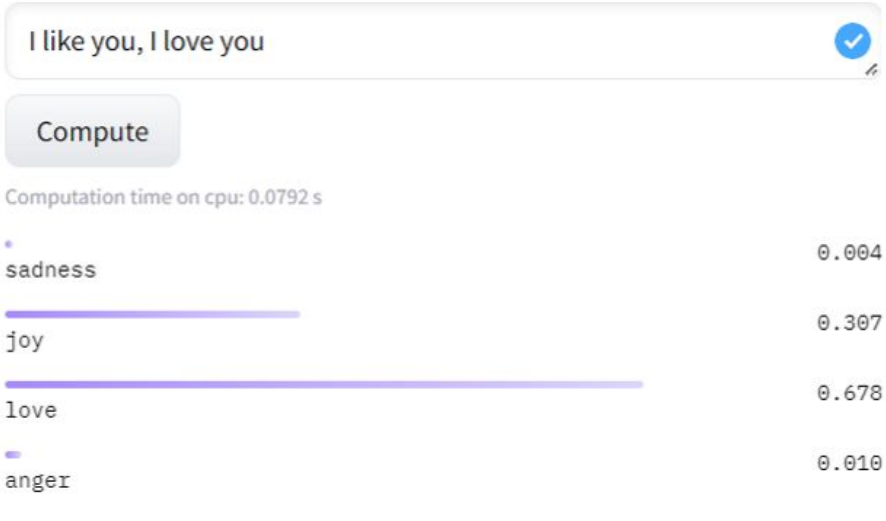


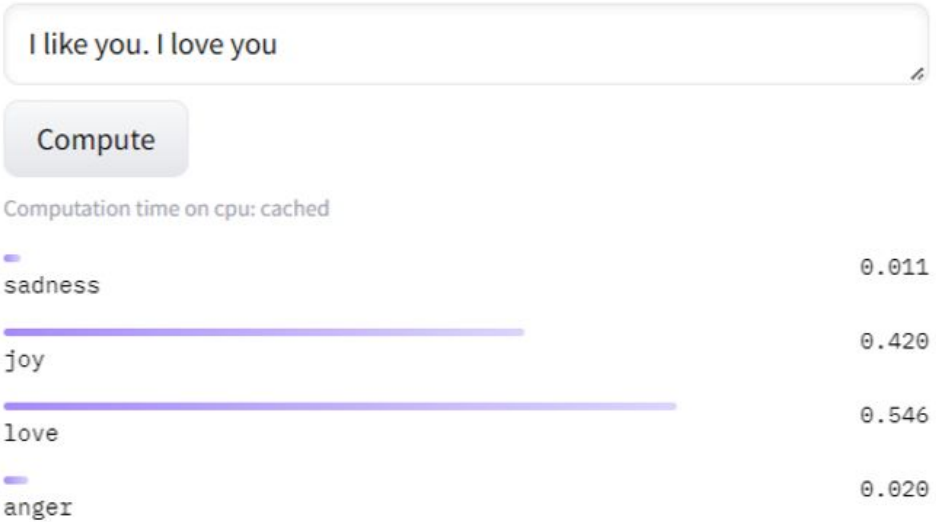
ARCHITECTURE



DISTILROBERTA-BASE FINE-TUNED ON EMOTIONS

distilRoBERTa-base model cannot properly differentiate semantic differences based on punctuations. The following example shows a case where **using comma instead of dot** has a ~13% difference over same labels





RESULTS

Models	Accuracy
Wav2Vec2-base 960h	0.06
Fine-tuned Wav2Vec2-base 960h	**0.29**
distilRoBERTa base	0.44
Fine-tuned distilRoBERTa base	0.35

Wav2Vec only found predictions for ⅛ labels.

We believe that with proper resources ($) the model would outperform existing models in each label.

Results for wav2vec2 are derived from 3 epochs. Model required large memory and batch size=1 in able to be fine-tuned in Colab.

FUTURE WORK

- Dataset Augmentation: AugLy text and audio augmentation

- Combine Different Approaches

COMBINATION OF DIFFERENT APPROACHES

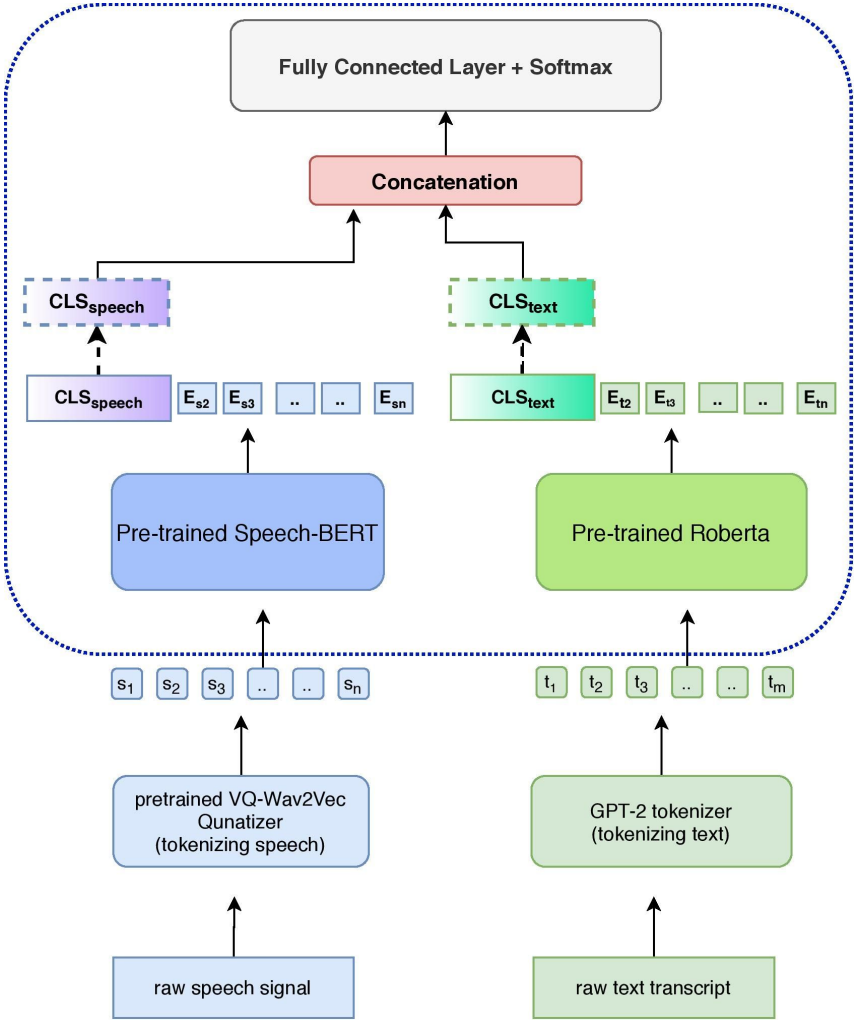
- Combines text and audio emotion prediction to a unified single-label prediction
- Discrete labels should not be combined to output a single discrete label

Happy voice🙂: Everyone is going to war!
(Sad context)😞

Attempt: Map discrete emotions to continuous values of Activation, Dominance and Valence **but** Reverse inference can be sensitive to overlaps in the values across the three features, where wrong emotion prediction may occur

Solution: make model multi-label prediction with Activation, Valence and Dominance

Jointly Fine-Tuning "BERT-like" Self Supervised Models to Improve Multimodal Speech Emotion Recognition



DISCUSSION & THOUGHTS

Valence, Activation and Dominance seem promising metrics in emotion recognition over psychological disorders. Using these well-defined metrics to analyse emotions quantitatively with Data Science, will help dive deeper into finding patterns under which specific behavioural disorders occur.

An indication that this expectation can be met is that already “macroscopically” has been observed that:

text: negative valence in semantic text → suicidal ideation
Audio: negative valence + negative arousal → depression

a depressive person might be spiralling to a suicide attempt

In this work it was very apparent the lack of a good-enough dataset in the context mentioned above, that correlate behavioural disorders with analysed emotions
(Podcast dataset used here is far too neutral for the task described above)