Deciphering non-verbal behaviors based on speech and text

A work by:

- Daniel STULBERG HUF
- Lawson OLIVEIRA LIMA
- Luan ROCHA DO AMARAL
- Lucas VITORIANO DE QUEIROZ LIRA
- Rajeh EL SADDI
- Tomas GONZALEZ VILLARROEL

Outline

- Context and Goals
- Pipeline
- The IEMOCAP Database
- Audio Model
- Text Model
- Late Fusion Model
- Training
- Results
- Visual Interface
- Conclusions

The human factor

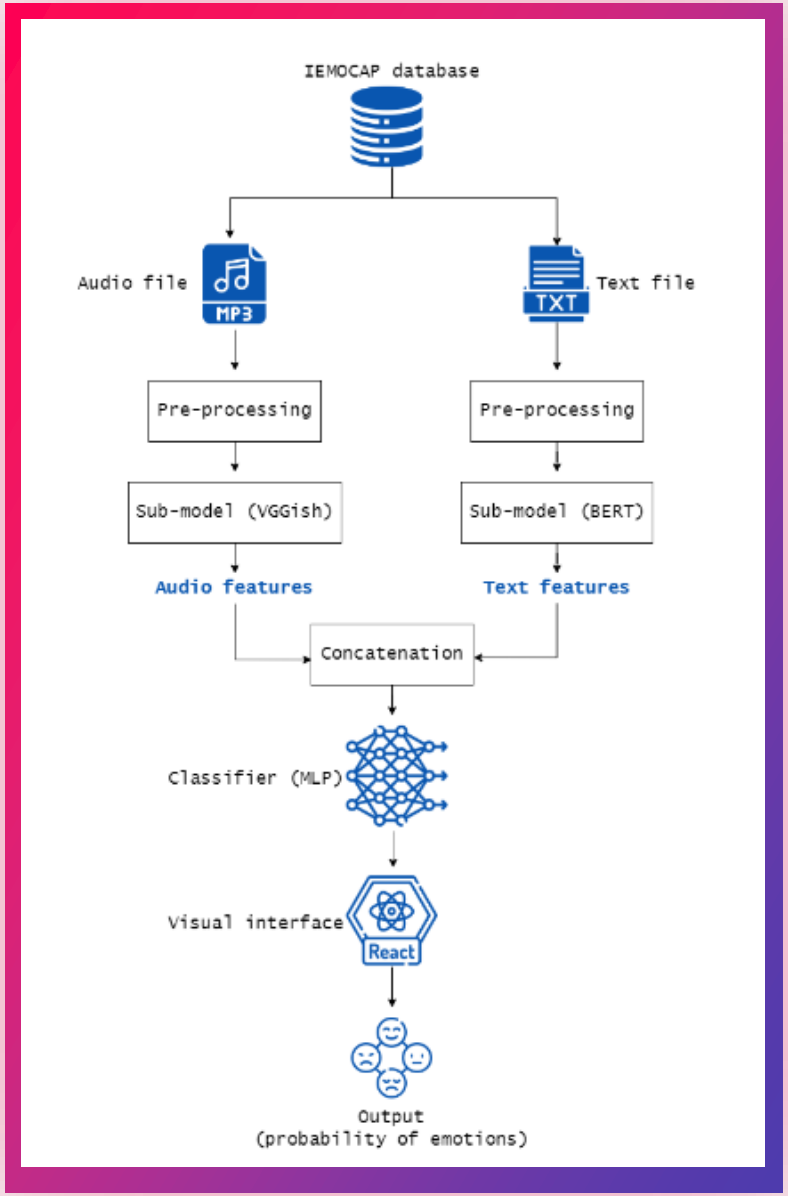
MSA approach

Context and Goals

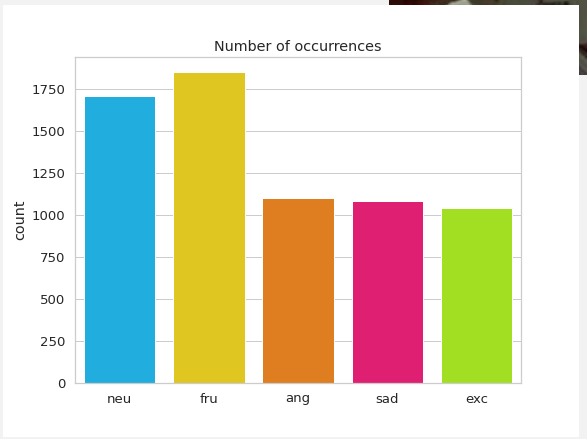
IEMOCAP

Audio+Text model

Pipeline


IEMOCAP database
Audio file
MP3
Text file
TXT
Pre-processing
Pre-processing
Sub-model (VGGish)
Sub-model (BERT)
Audio features
Text features
Concatenation
Classifier (MLP)
Visual interface
React
Output
(probability of emotions)

Happy

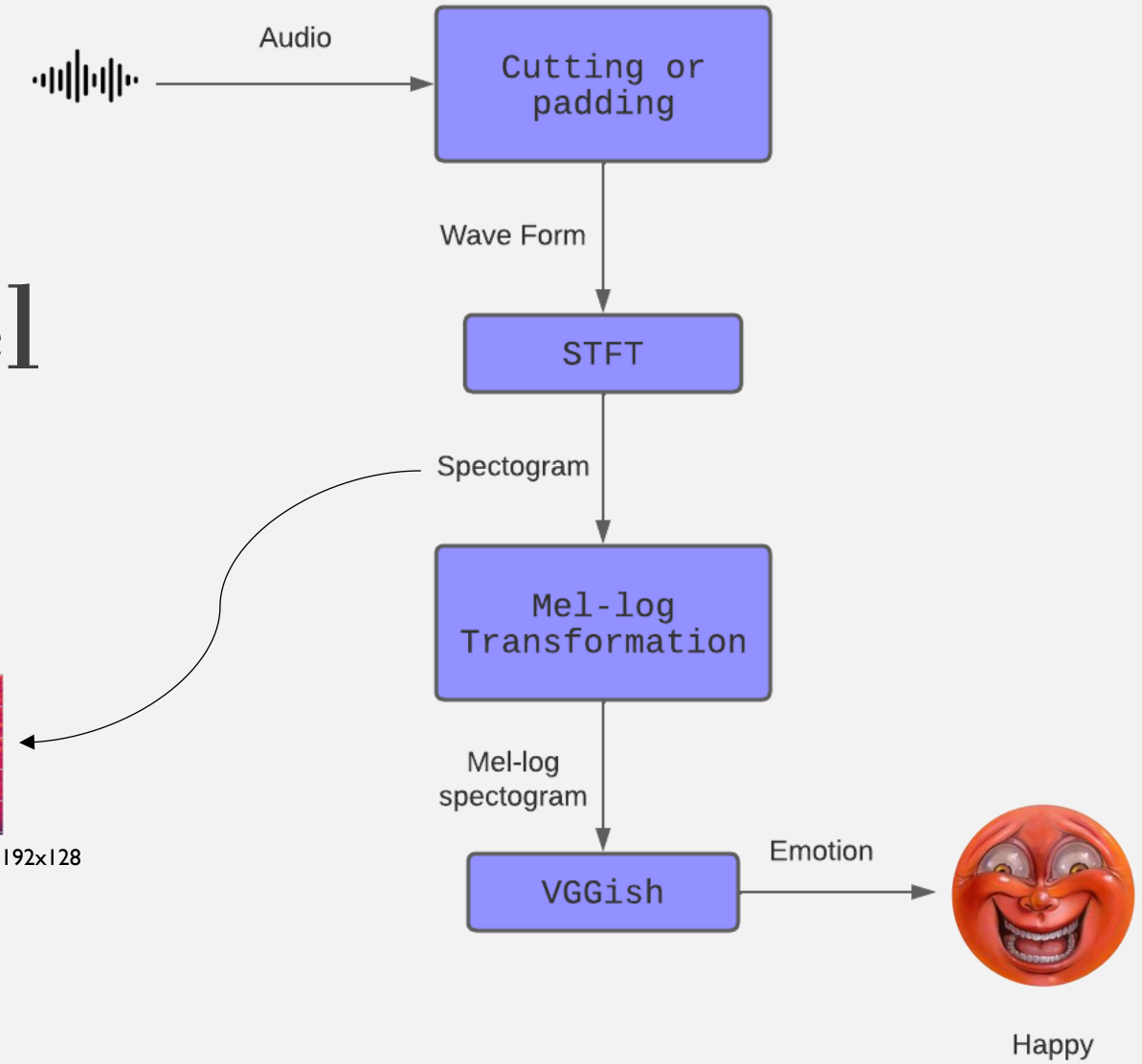


The IEMOCAP database

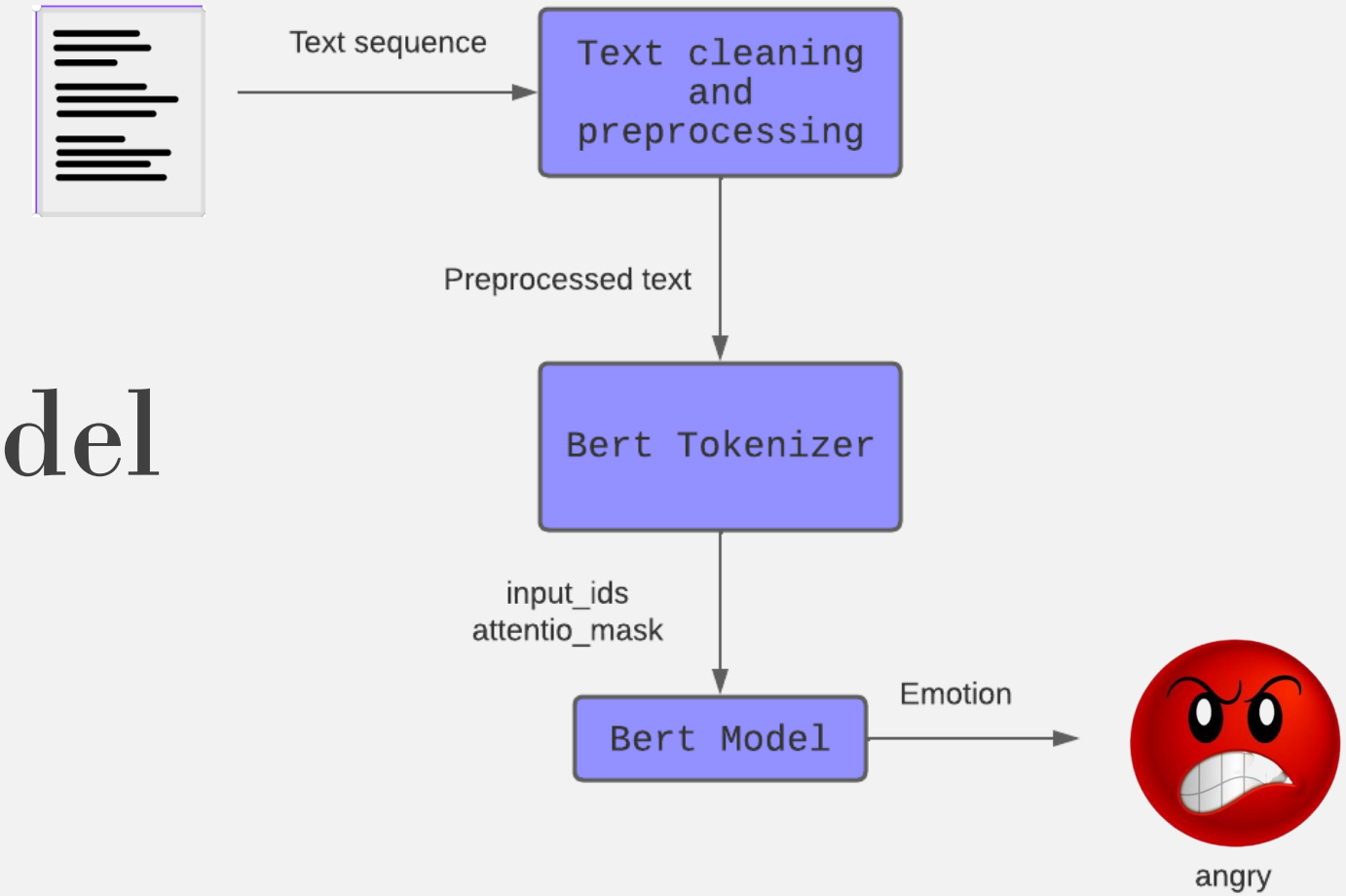
Interactive Emotional Dyadic Motion Capture

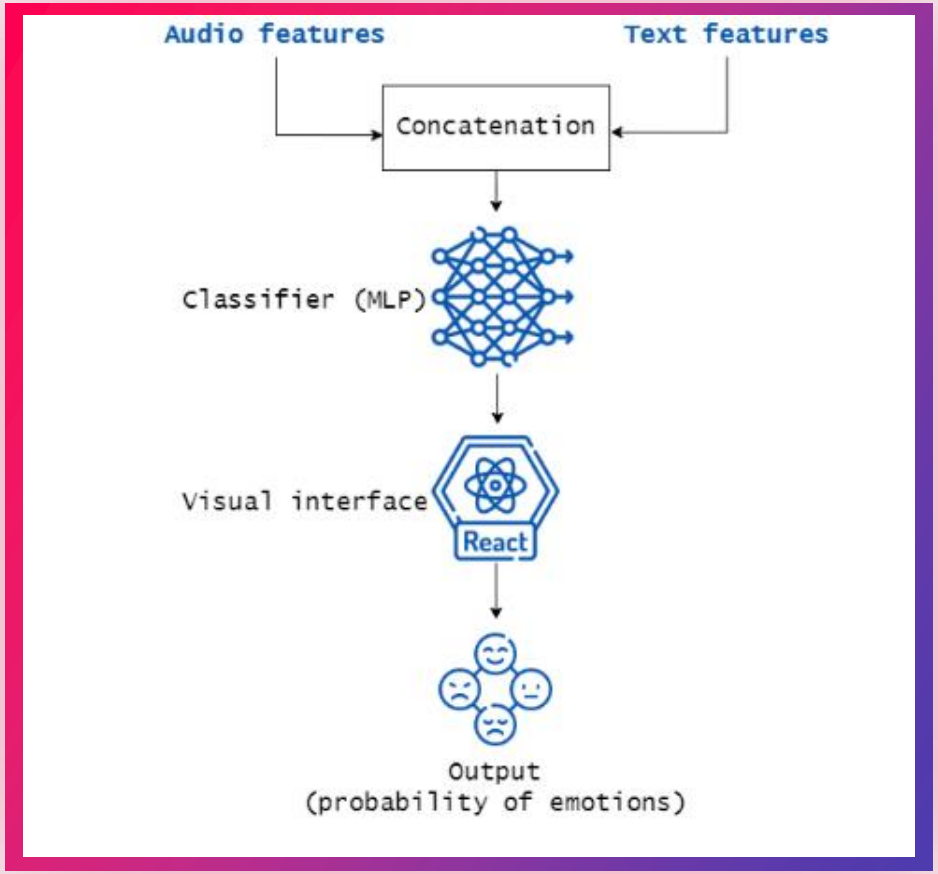
- 11 GB University of Southern California
- 10 Emotions
- Audio, Text, and Motion Capture

The audio model



The text model





The late fusion model

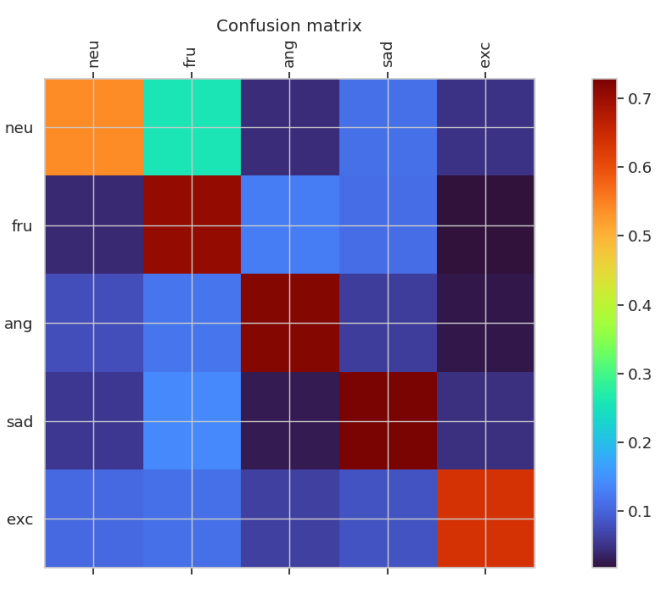
Block	Components	Output shape
Input	Concatenated features	4096(VGGish) + 768(Bert)
Block-1	Dense (4864, 512) → Relu	512
Block-2	Dense (512, 512) → Relu	512
Block-3	Dense (512, 5) → Softmax	5

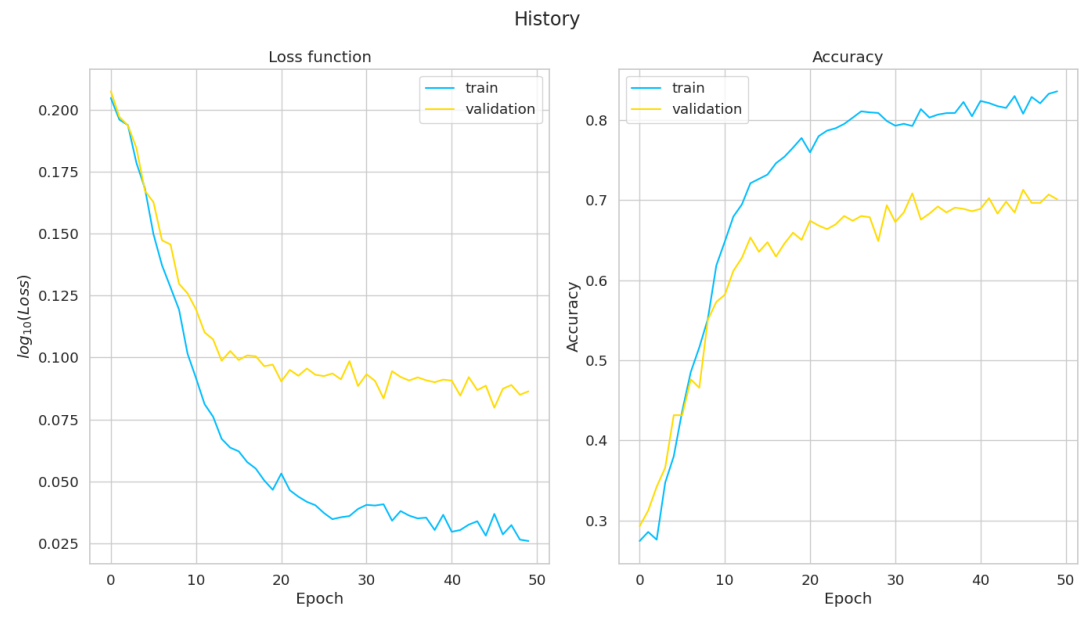
Training

- 50 epochs
- Learning rate: 2*10^-5
- Scheduler (20% Warmup + 80% linear decaying)
- Transfer learning + fine-tuning
- Weighted Cross Entropy Loss
- Around 3h in Colab GPU

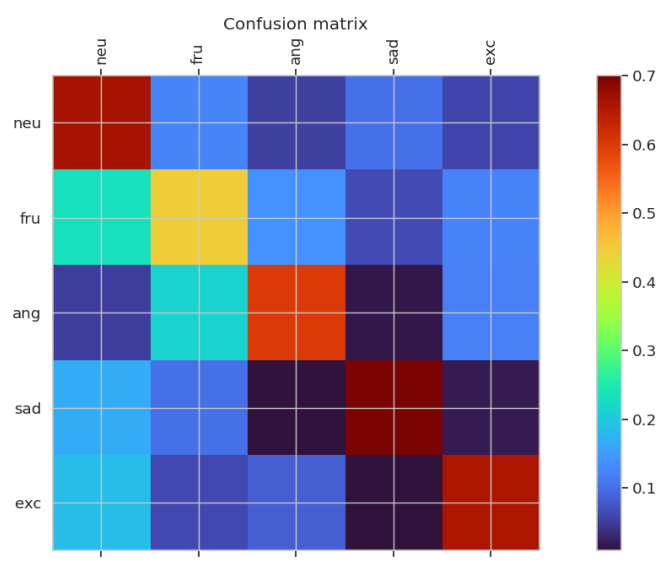
Results

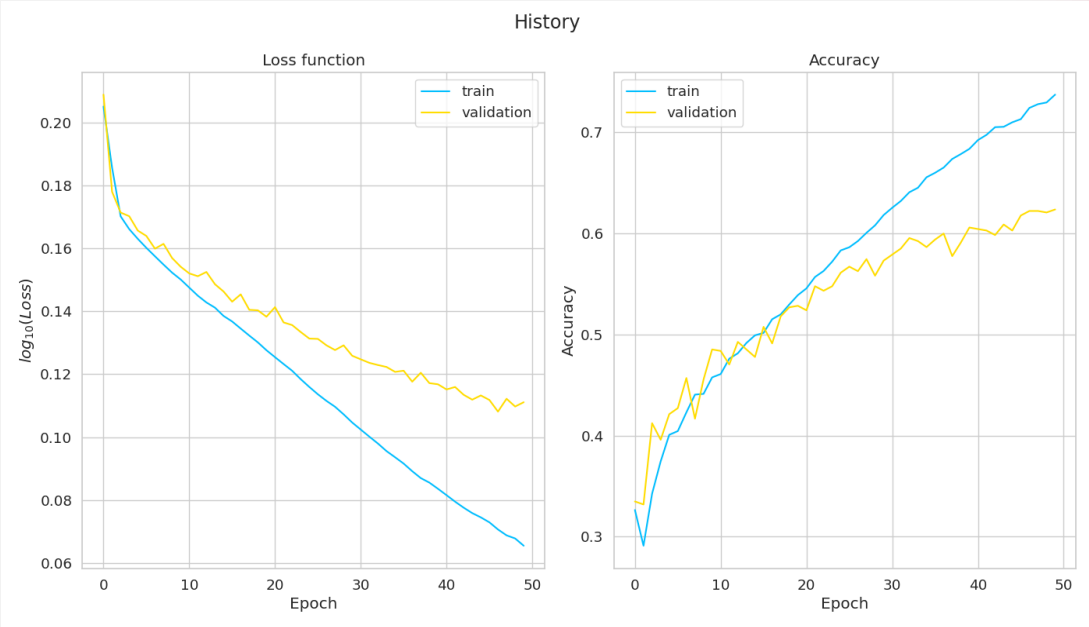
Text model:





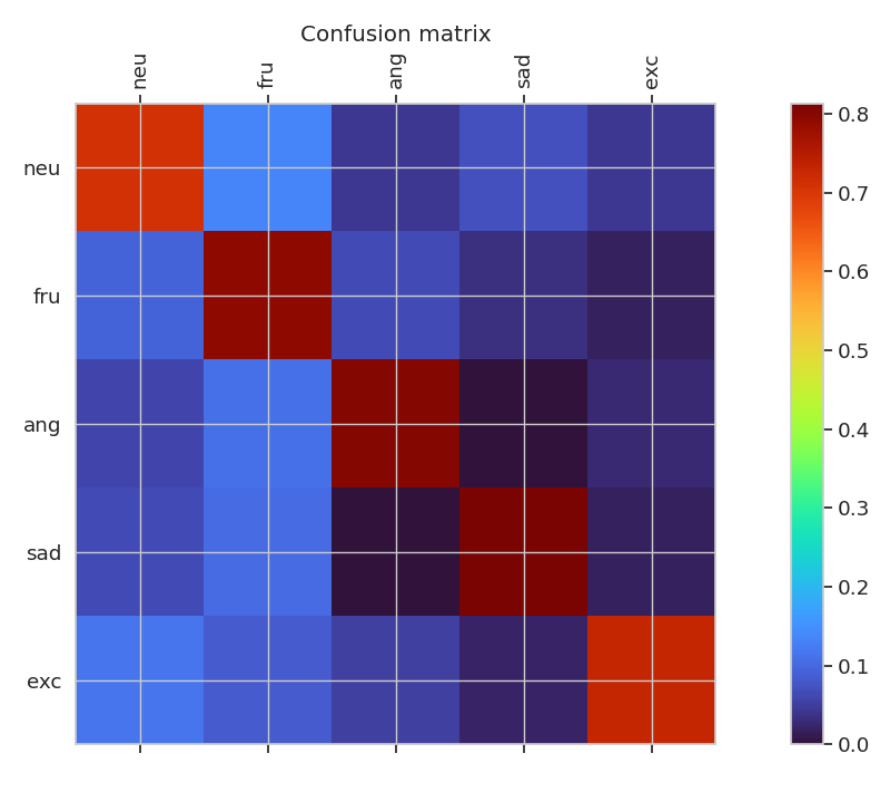
Audio model:

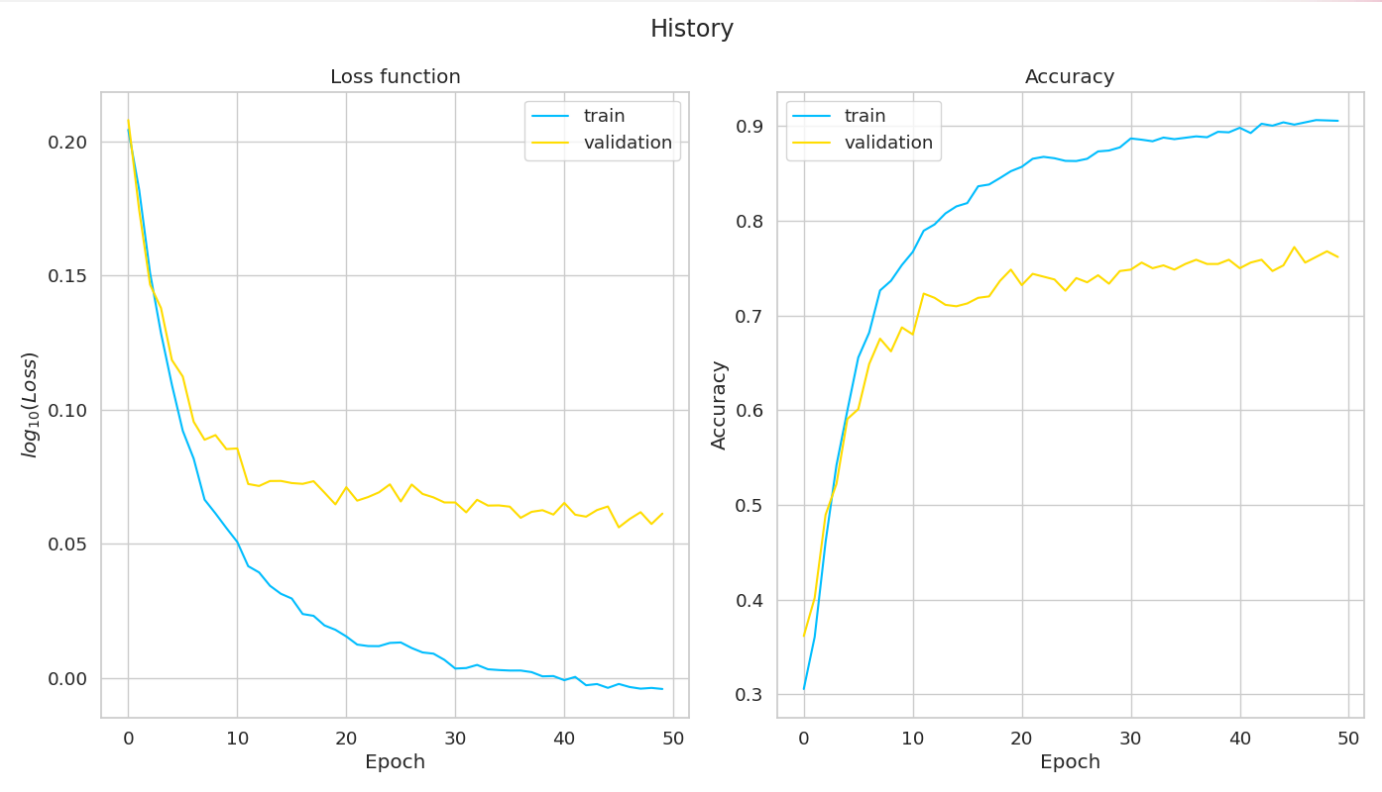




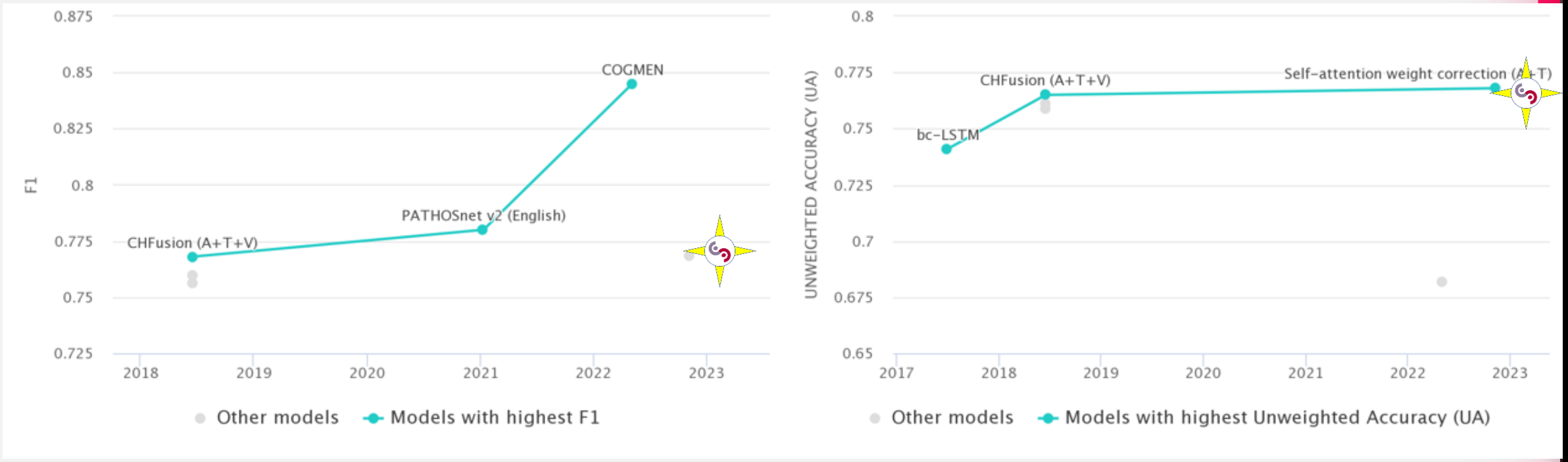
Results

Late fusion model:



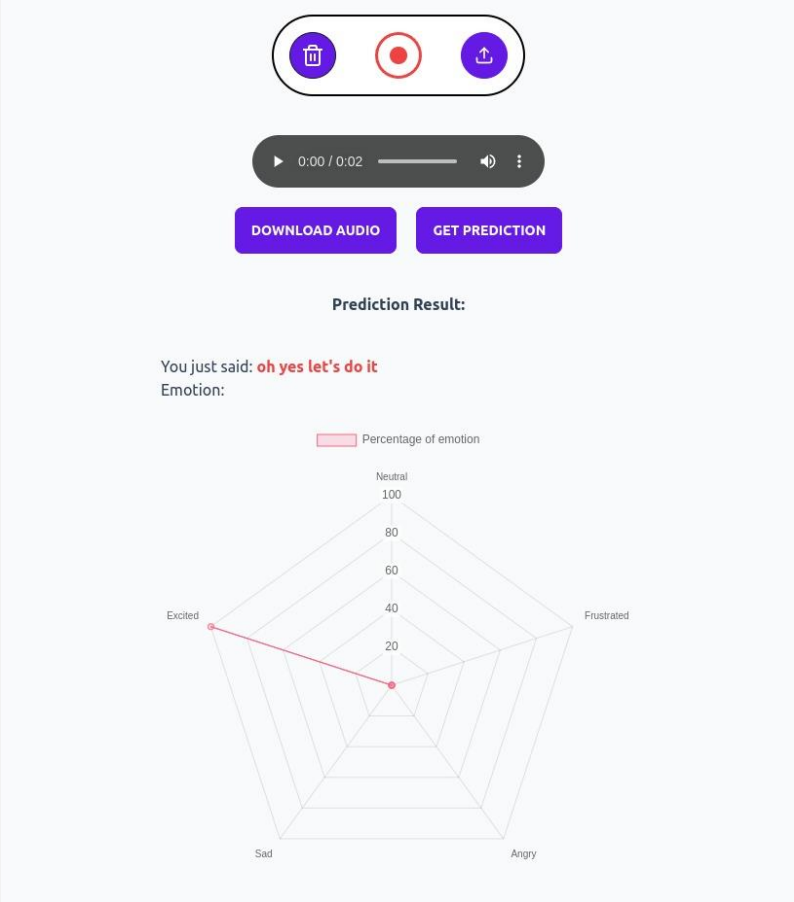


Results



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

Visual web interface



- Possibility to record or upload custom audio to the model
- Usage of HuggingFace AI model to perform Speech to Text conversion
- Radar chart visualization of predicted emotions, along with recognized text

Conclusion

Improvements:

- Incorporate all 3 modalities;
- Enhance algorithms and computational power;
- Deal with person-indepency;
- Scaling deployment of the application to a cloud server.

Thank You