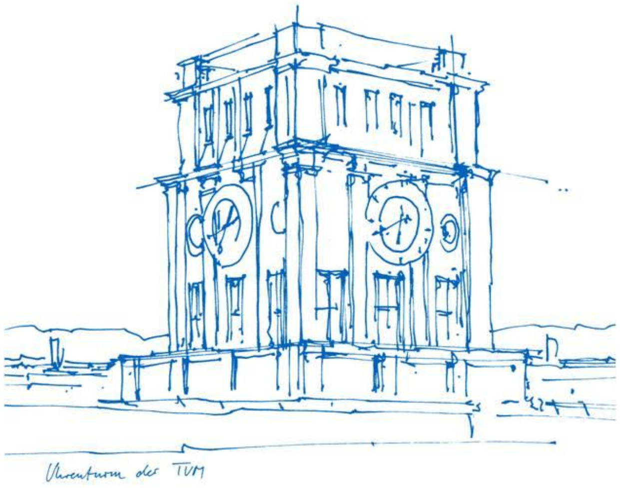
Critical Speech-Analysis with Explanations

Technische Universität München

Fakultät für Informatik

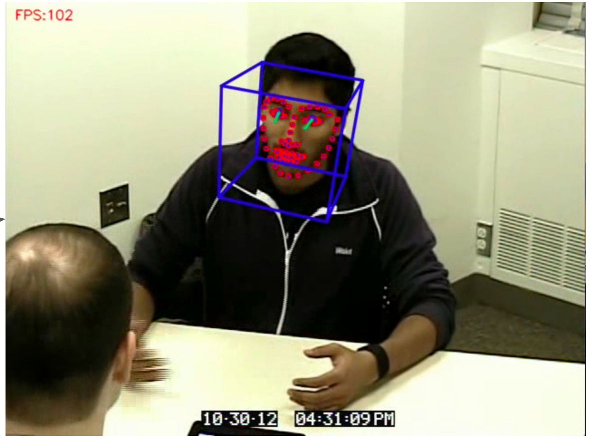
XAI Lab Course, WS21/22

08.02.2022

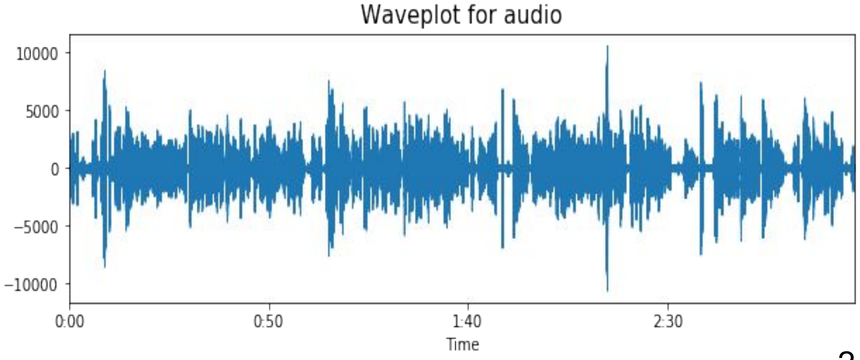


Ahmed, Ruiyun, Stefan

Interview videos



“So ah my interest kinda laid both in a little bit of the health care I imagined I was going be a Doctor growing up and ...”



Textual features

Example text: “So ah my interest kinda laid both in a little bit of the health care I imagined I was going be a Doctor growing up and ...”

- Word count features with **NLTK**
  Unique words in each interview

- **L**inguistic **I**nquiry **W**ord **C**ount (**LIWC**)
  based on psychological research

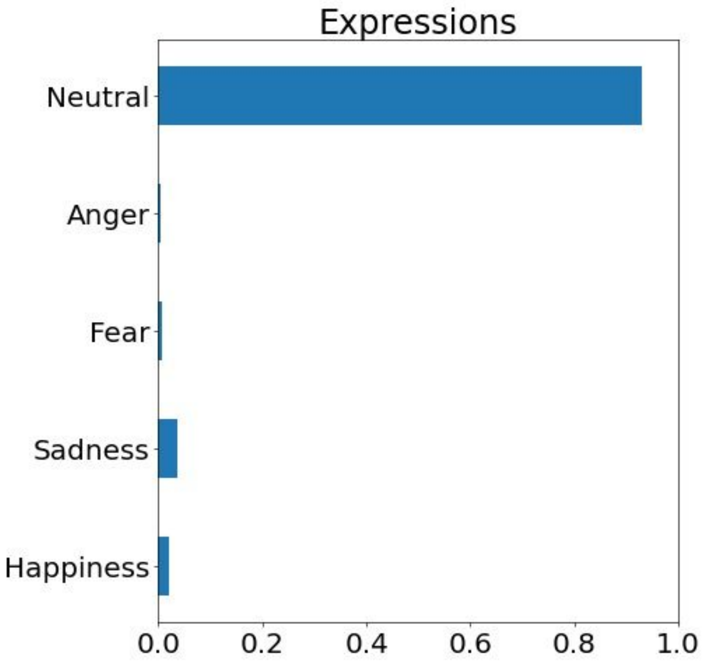
- Sentiment analysis
  of sentences with **BERT**

LIWC Category	Examples
Non-fluencies	uh, umm, well
PosEmotion	hope, improve, kind, love
NegEmotion	bad, fool, hate, lose
Work	project, study, thesis, university

Sentiment analysis with BERT

Example sentence:

“And um as far as extracurriculars go I do a few things.”



Sentiment analysis with BERT

1. Finetune BERT for sentence sentiment classification on a balanced dataset (classes: neutral, joy, anger, fear, sadness)
2. 83% accuracy for the test set
3. Predict sentiment of each interview sentence
4. Average sentiments over the interview

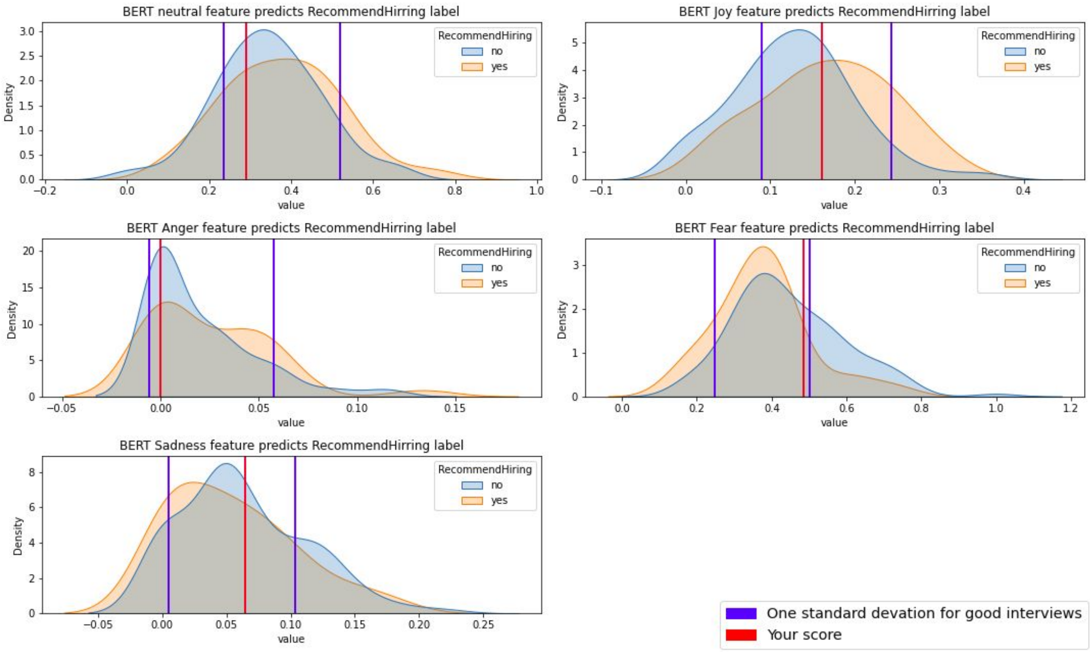
Example:

Interview p89 with a total of 31 sentences
neutral: 9, joy: 5, anger: 0, fear: 15, sadness: 2
avg: neutral: 0.29, joy: 0.16, anger: 0.0, fear: 0.48, sadness: 0.06

Sentiment analysis with BERT

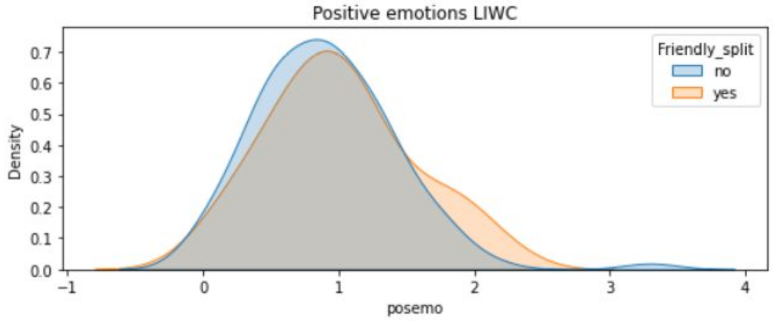
Interview p89 with a total of 31 sentences (bad example interview)

avg: neutral: 0.29, joy: 0.16, anger: 0.0, fear: 0.48, sadness: 0.06



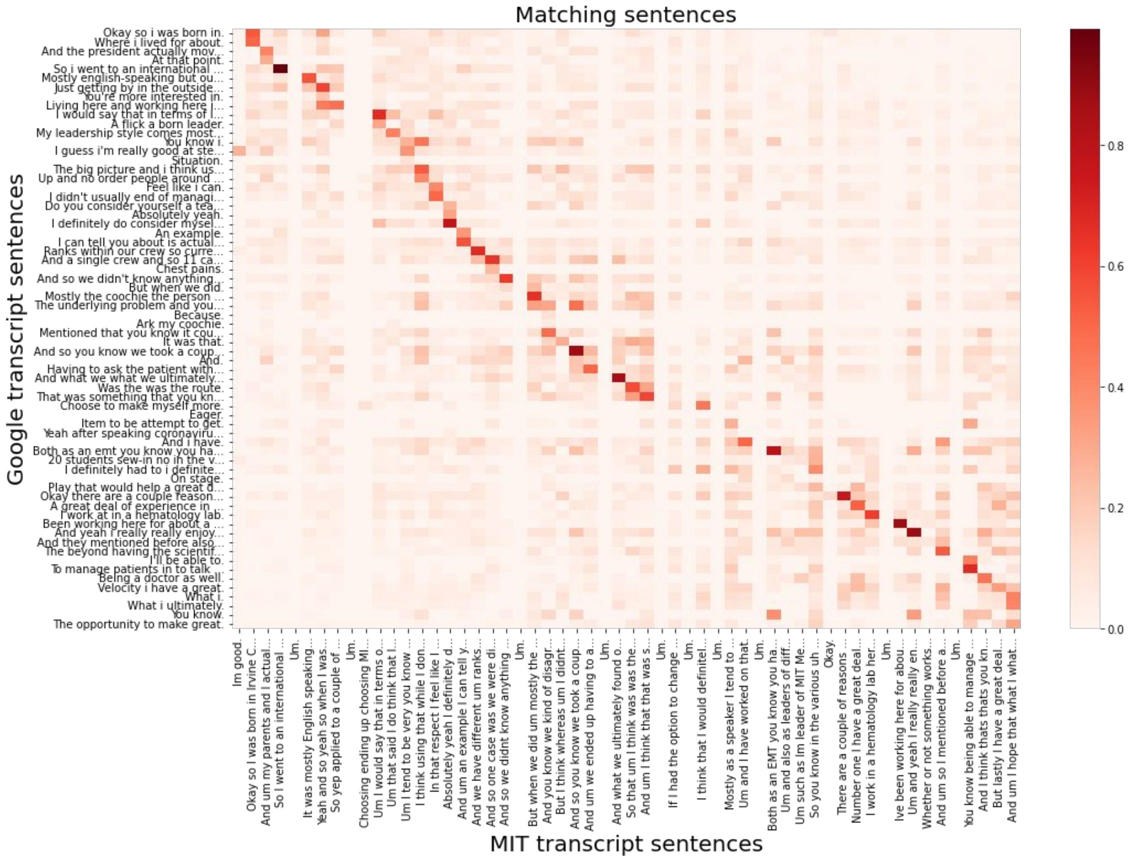
Unsuccessful attempt

- google speech to text output as basis for the textual analysis
- finetune BERT on the sentiment labels given in the dataset
- most categories of the LIWC (4/90 categories have been valuable)



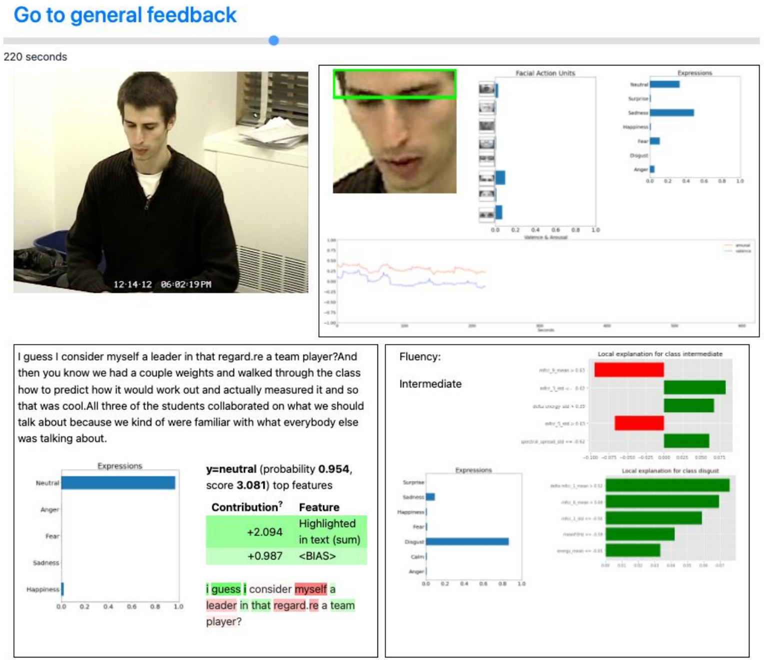
Timestamp creation

1. Separation of speakers by voice clustering
2. Speech to text with google API
3. Matching between google sentences and transcript sentences



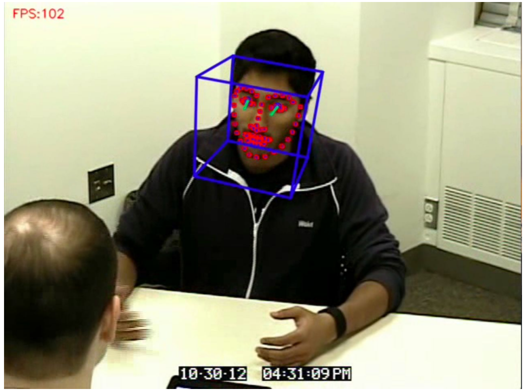
Dashboard

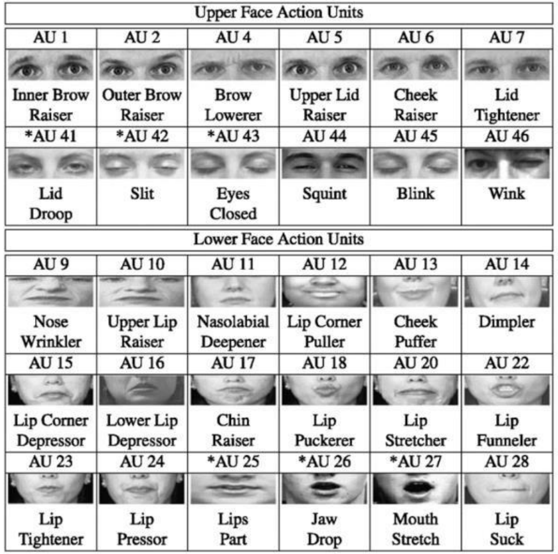
- Flask Application
- Overview with on-the-fly updates
- Detailed feedback for each domain
- General feedback with score



Video Features

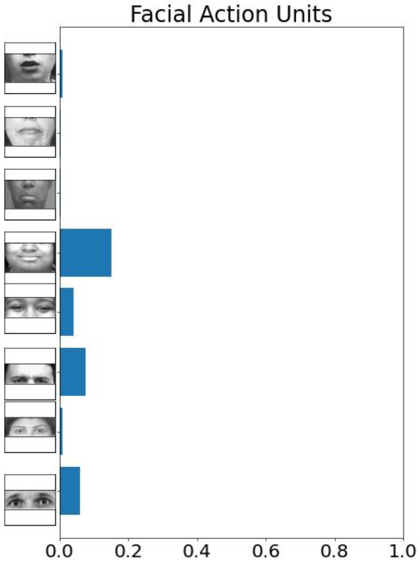
- Facial Action Units detection
- Emotion Recognition
- Valence and Arousal level





Facial Action Units

- **OpenFace -** 18 Facial Action Units



AU	Full name	Prediction
AU1	Inner brow raiser	I
AU2	Outer brow raiser	I
AU4	Brow lowerer	I
AU5	Upper lid raiser	I
AU6	Cheek raiser	I
AU7	Lid tightener	P
AU9	Nose wrinkler	I
AU10	Upper lip raiser	I
AU12	Lip corner puller	I
AU14	Dimpler	I
AU15	Lip corner depressor	I
AU17	Chin raiser	I
AU20	Lip stretched	I
AU23	Lip tightener	P
AU25	Lips part	I
AU26	Jaw drop	I
AU28	Lip suck	P
AU45	Blink	P

Emotion recognition

- 1st Approach:

  Rule-based approach based on EMFACS (Emotional Facial Action Coding System) and FACSAID (Facial Action Coding System Affect Interpretation Dictionary)

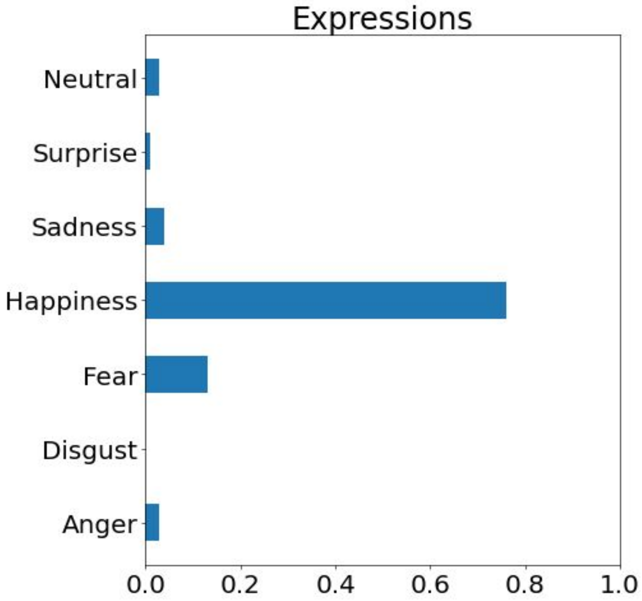
  Problem:

  Biased, since some emotions needs much more AUs to be detected

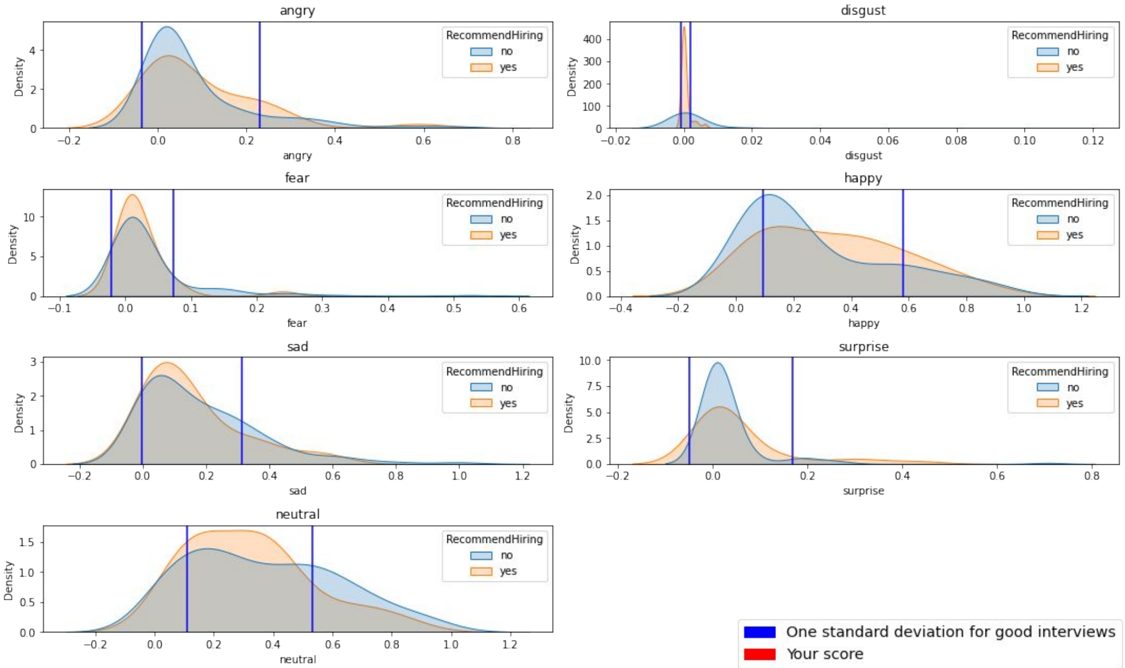
Emotion ⬍	Action units ⬍
Happiness	6+12
Sadness	1+4+15
Surprise	1+2+5B+26
Fear	1+2+4+5+7+20+26
Anger	4+5+7+23
Disgust	9+15+17
Contempt	R12A+R14A

Emotion recognition

- Use pre-trained FER model based on a CNN architecture



Emotions distribution



Good interviewees have a more happy and less neutral or sad facial emotion

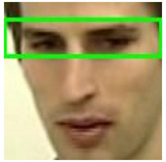
Relationship FAU, Emotions and MIT labels

Smiled, Friendly, Authentic

Happy

Good interview

Not Authentic

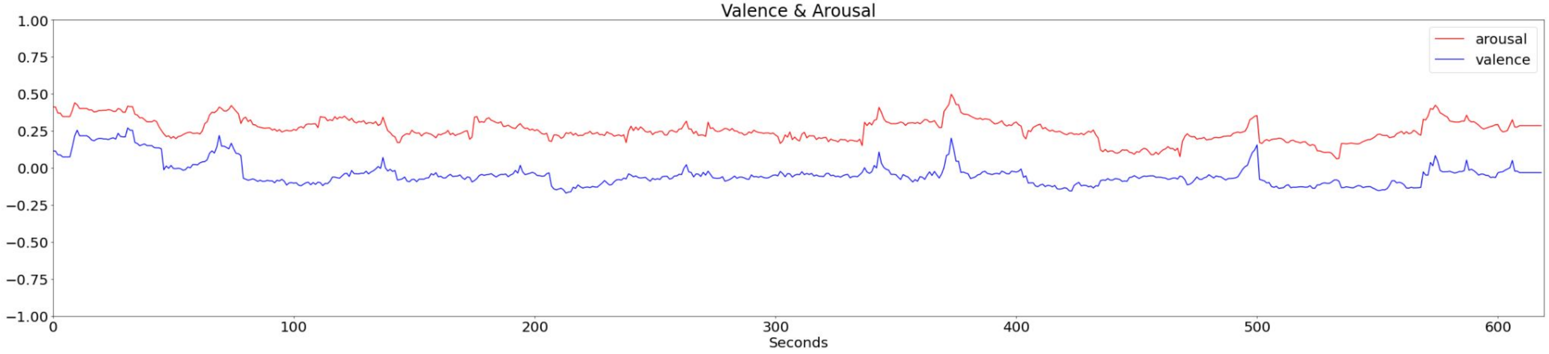


Neutral

Bad interview

Valence and Arousal level

- Use pre-trained model with a CNN-RNN architecture



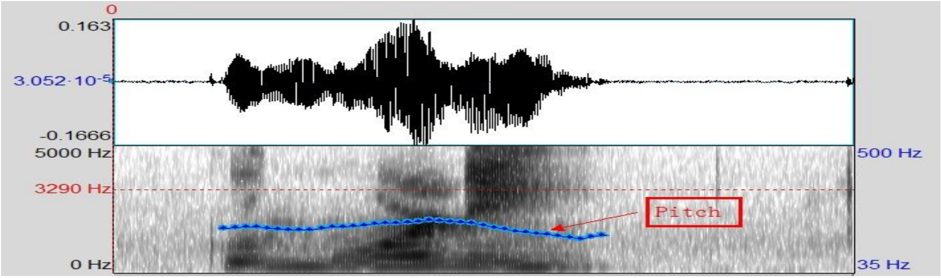
Unsuccessful attempts

- Rule-based approach for emotion detection
- Train a classifier:
  - from emotion to recommended hiring label
  - from facial action units to facial emotion
- Use a smile detection model

Audio Features

General preprocessing steps:

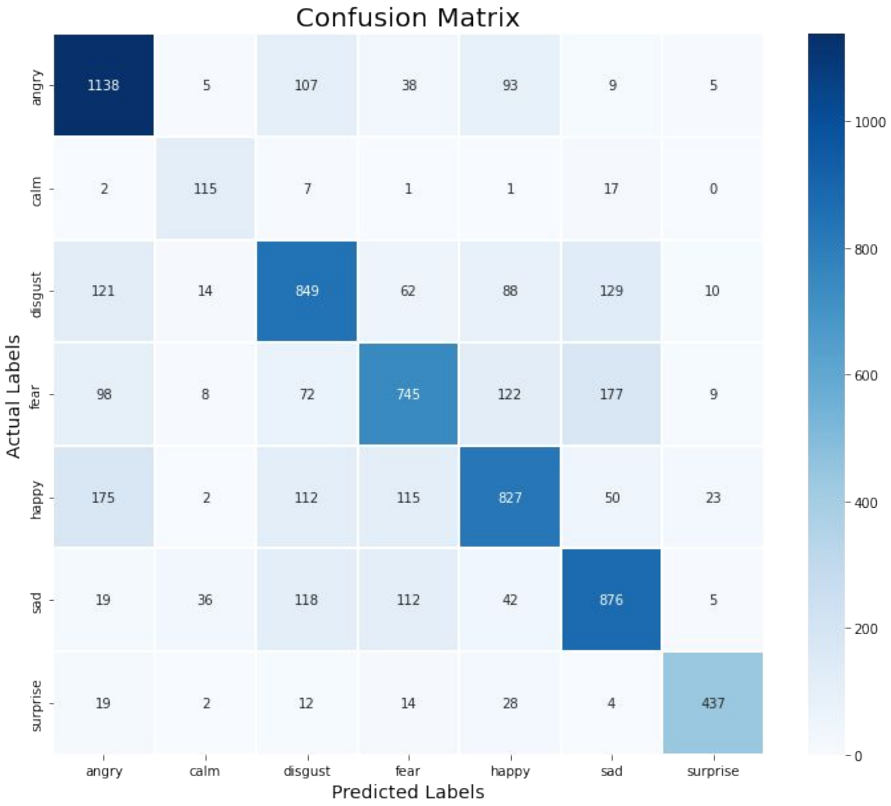
- Separating Speakers (interviewer/interviewee) using unsupervised clustering with PyAudioAnalysis
- Separating each audio into chunks of 3s
- Extract 150 low level features with PRAAT and PyAudioAnalysis
- Use these features to train models for further analysis



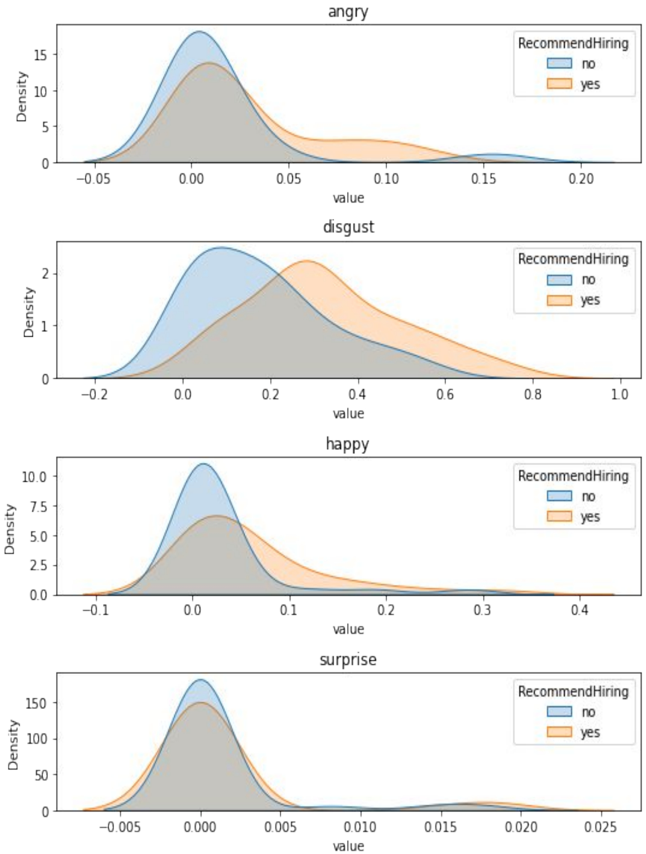
Prosodic Feature	Description
Energy	Mean spectral energy.
F0 MEAN	Mean F0 frequency.
F0 MIN	Minimum F0 frequency.
F0 MAX	Maximum F0 frequency.
F0 Range	Difference between F0 MAX and F0 MIN.
F0 SD	Standard deviation of F0.
Intensity MEAN	Mean vocal intensity.
Intensity MIN	Minimum vocal intensity.
Intensity MAX	Maximum vocal intensity.
Intensity Range	Difference between max and min intensity.
Intensity SD	Standard deviation.
F1, F2, F3 MEAN	Mean frequencies of the first 3 formants: F1, F2, and F3.
F1, F2, F3 SD	Standard deviation of F1, F2, F3.
F1, F2, F3 BW	Average bandwidth of F1, F2, F3.
F2/F1 MEAN	Mean ratio of F2 and F1.
F3/F1 MEAN	Mean ratio of F3 and F1.
F2/F1 SD	Standard deviation of F2/F1.
F3/F1 SD	Standard deviation of F3/F1.
Jitter	Irregularities in F0 frequency.
Shimmer	Irregularities in intensity.

Sentiment Analysis

- Training a multiple ML models on a classification dataset with 7 emotions (Anger, Happiness, Fear, Sadness, Disgust, Calmness, Surprise)
- Accuracy of the SVM model 70%
- Compute a class for each chunk of the interview and aggregate the results
- Test it on the MIT dataset



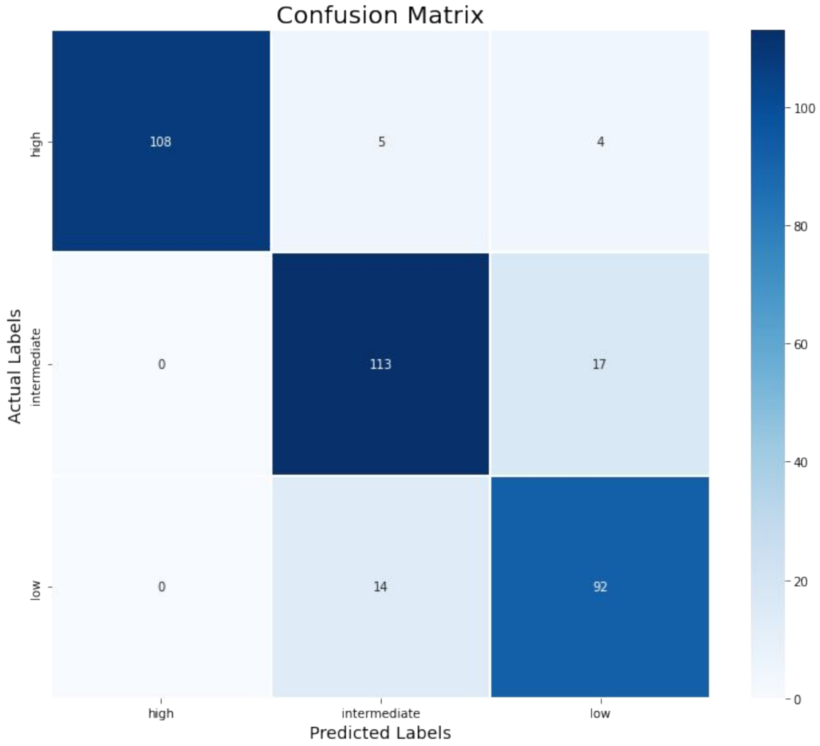
Sentiment Analysis



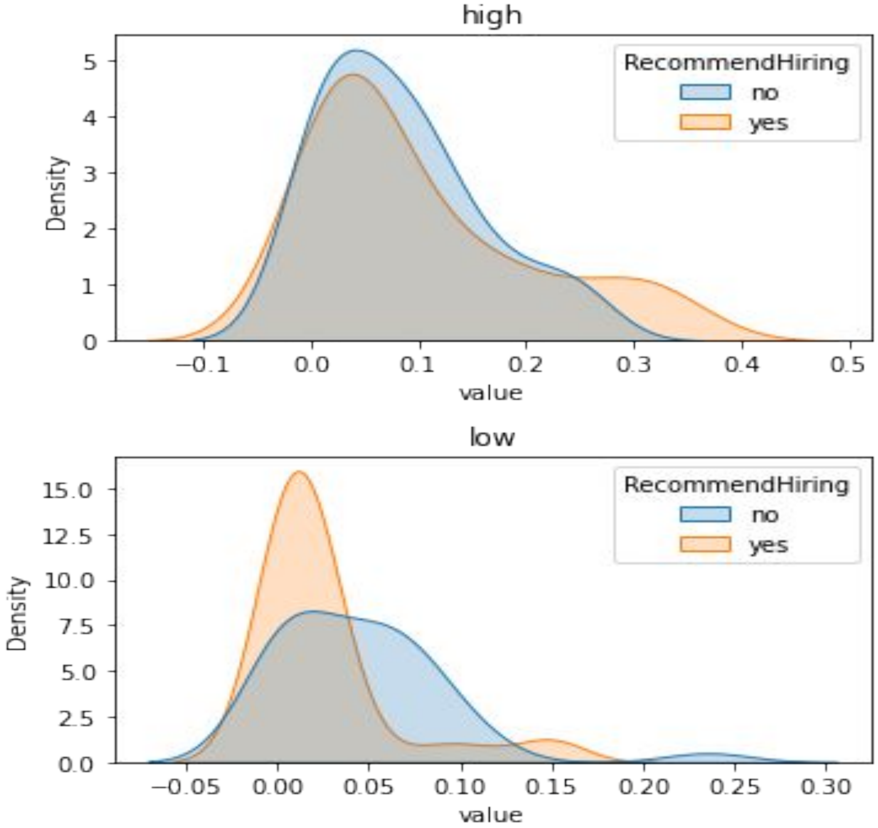
Sentiment	Correlations with scores	P-values
Angry	0.1307	0.1264
Calm	-0.0005	0.9948
Disgust	0.2584	0.0022
Fear	-0.3242	0.0001
Happy	0.2464	0.0035
Sad	-0.1226	0.1517
Surprise	-0.0031	0.9710

Fluency classification

- Using a dataset containing 1409 audio files classified into 3 classes (low_fluency, intermediate_fluency, high_fluency)
- Training a SVM model for the classification. Obtained accuracy: 88%
- After testing this model on the MIT dataset, most audio are classified into intermediate and high fluency.



Fluency analysis

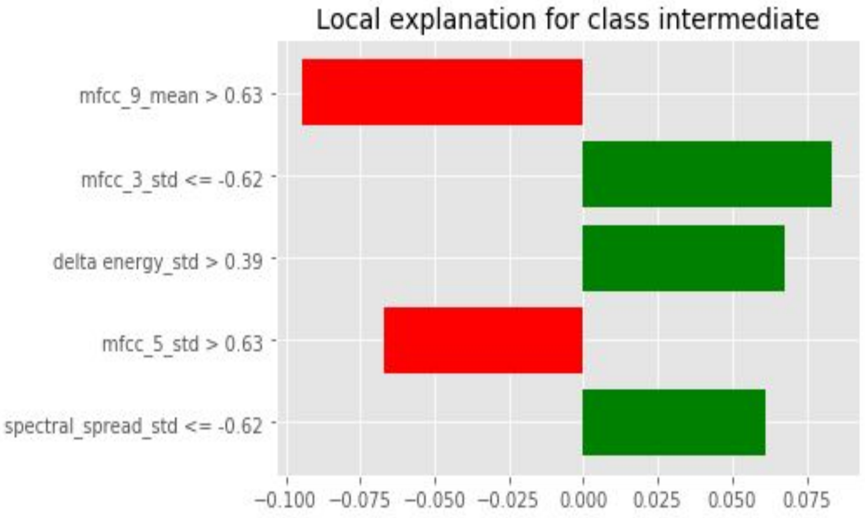


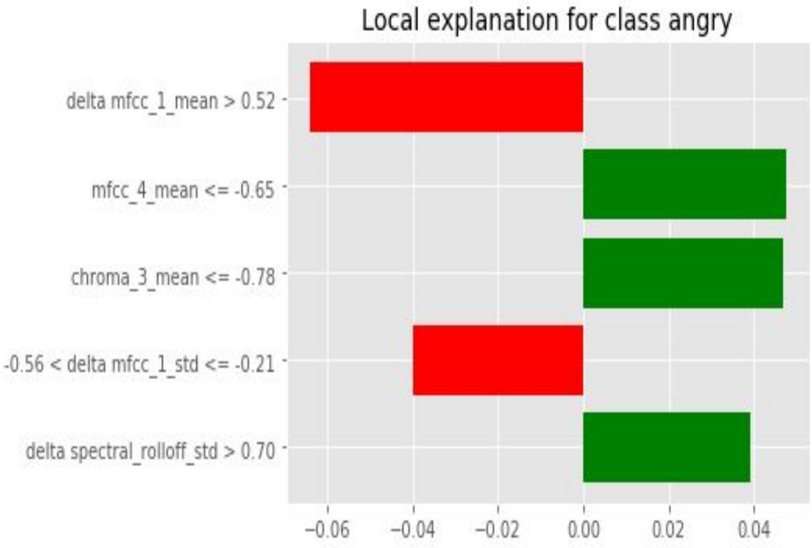
	Correlations with scores	P-values
High	0.0395	0.6452
Intermediate	0.0462	0.5899
Low	-0.2501	0.0030

Additional high level features

- Using the libraries Myprosdoy and Praat
- Features:
    - number_ of_syllables
    - number_of_pauses
    - rate_of_speech
    - speaking_duration
    - articulation_rate
    - balance
- For feedback: compute the mean and standard deviation for interviews with good score and check if the new interview is in the 50% percentile around the mean

LIME explainer





Unsuccessful attempts

Clustering:

- Used different clustering algorithms: kmeans, mean shift, Gaussian mixture, spectral clustering
- Used only extreme data for fitting the algorithms
- Results: no clustering results where good scores are together and bad score are together

Regression:

- Used different models: NN, SVR, random forest, Gradient Boosting
- Results: bad MSE scores, models predicting always the average

Live Demo

Issues with our current approach

- Unreliable annotations
- Lack of data (138 interviews)
- Biased data (no really bad interviews)
- Averaging scores for an entire interview is not optimal
- No rigorous way to assess the generalizability of the models on the MIT dataset

References

- Ribeiro, Marco & Singh, Sameer & Guestrin, Carlos. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. 1135-1144. 10.1145/2939672.2939778.
- Deng, Didan & Chen, Zhaokang & Shi, Bert. (2020). Multitask Emotion Recognition with Incomplete Labels. 592-599. 10.1109/FG47880.2020.00131.
- Baltrusaitis, Tadas & Zadeh, Amir & Lim, Yao & Morency, Louis-Philippe. (2018). OpenFace 2.0: Facial Behavior Analysis Toolkit. 59-66. 10.1109/FG.2018.00019.
- I. Naim, M. I. Tanveer, D. Gildea and M. E. Hoque, "Automated Analysis and Prediction of Job Interview Performance," in *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 191-204, 1 April-June 2018, doi: 10.1109/TAFFC.2016.2614299.
- Devlin, Jacob & Chang, Ming-Wei & Lee, Kenton & Toutanova, Kristina. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
- Agrawal, Anumeha & George, Rosa & Ravi, Sunitha & Kamath S, Sowmya & Kumar, M.. (2020). Leveraging Multimodal Behavioral Analytics for Automated Job Interview Performance Assessment and Feedback.