Multimodal Fusion and Decision Trees for Automatic Explainable Job Candidate Screening from Video CVs

Heysem Kaya[1], Furkan Gürpınar[2], Albert Ali Salah[2]

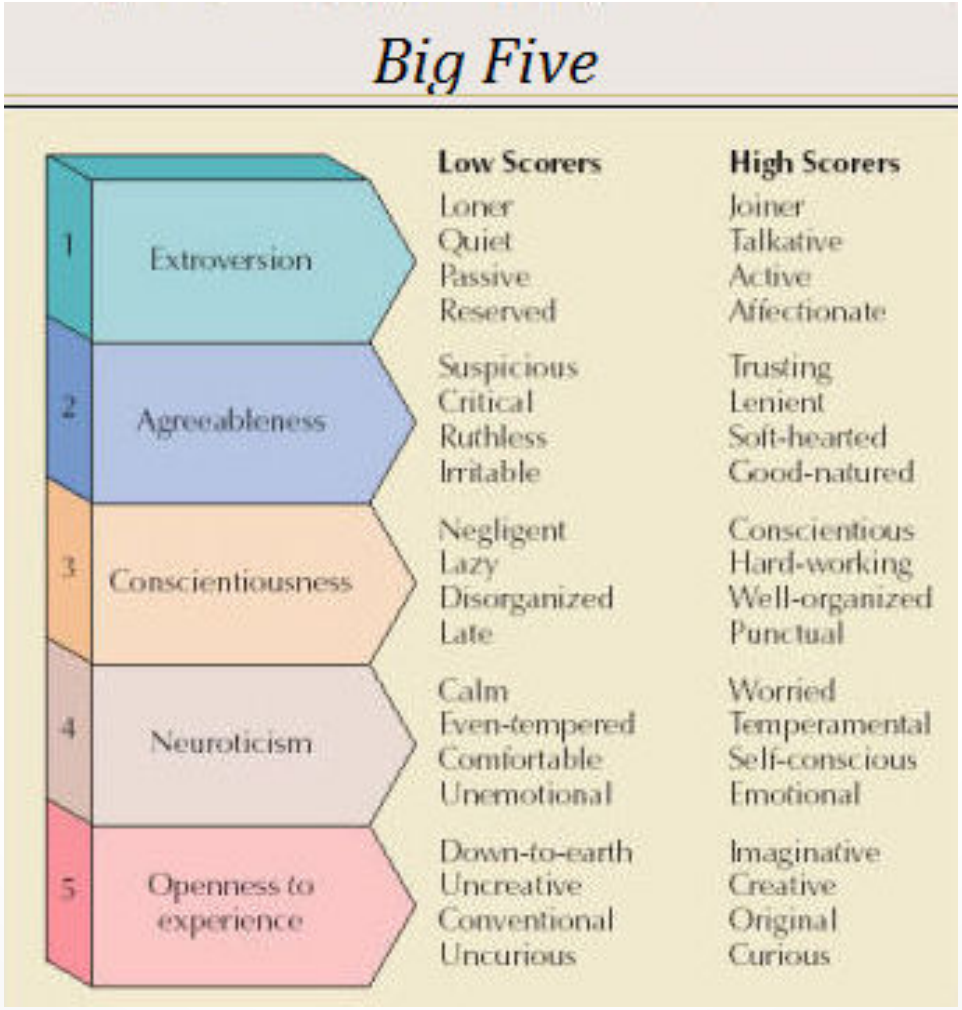
hkaya@nku.edu.tr, gurpinar@boun.edu.tr, salah@boun.edu.tr

[1] Namık Kemal University, Dept. of Computer Eng., Çorlu / TURKEY
[2]Boğaziçi University, Dept. of Computer Eng. İstanbul / TURKEY

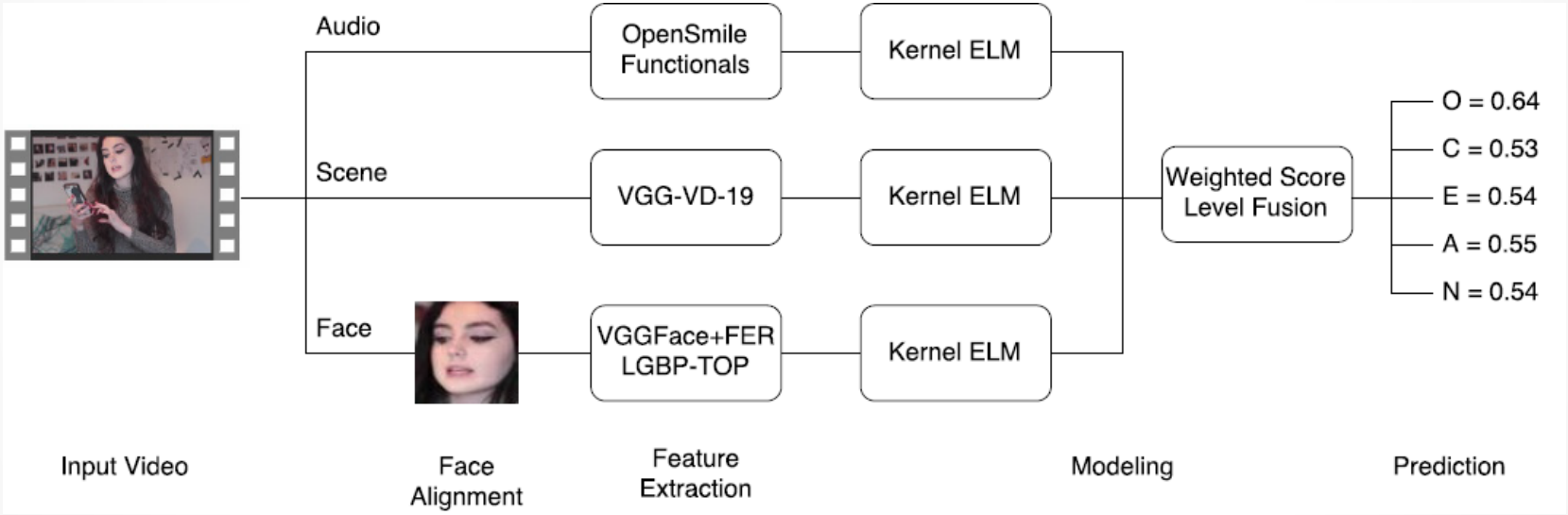
Motivation

- YouTube: Millions of personal videos uploaded
- People are eager to introduce themselves via video CVs
- Other videos of people can also be used to evaluate personality traits
- Automatic recognition of personality traits (OCEAN)
- Using personality traits to decide on job interview recommendation
- Verbally and visually explaining the decision of job interview selections
- Illustrating biases in the people selecting candidates!

The Big Five Traits – OCEAN



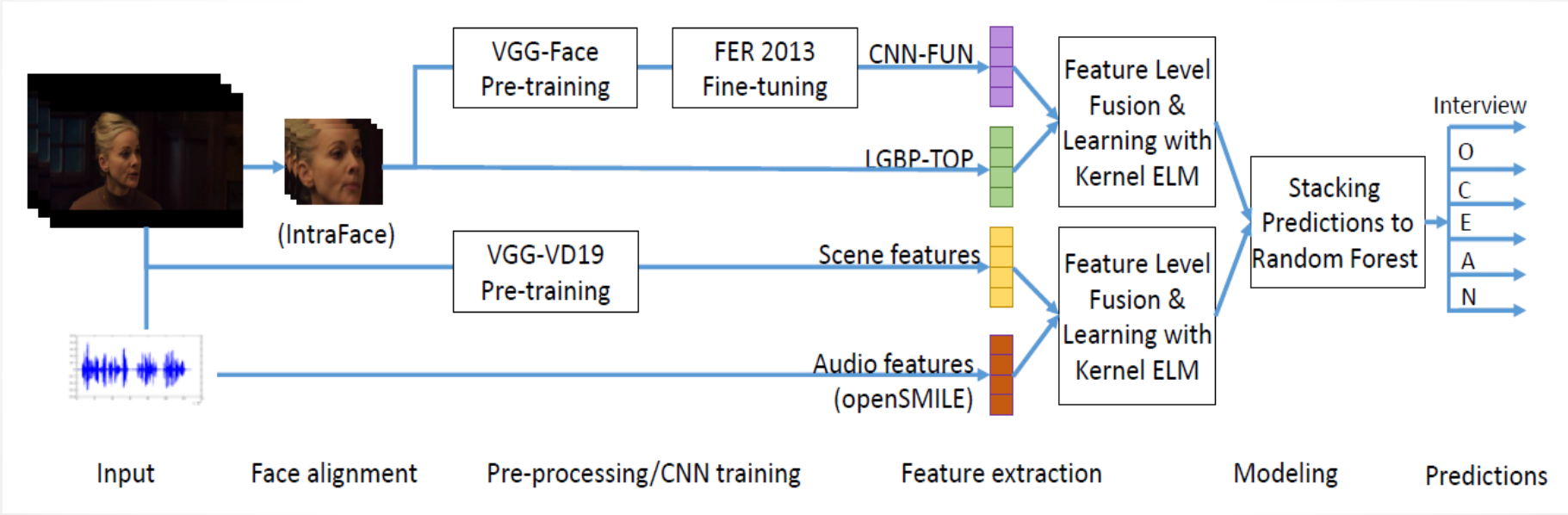
Apparent personality trait estimation



*F. Gürpınar, H. Kaya, and A. A. Salah. Multimodal Fusion of Audio, Scene, and Face Features for First Impression Estimation. *ICPR 2016*, Cancun, Mexico, December 2016.

[#] Kaya, H., Gürpınar, F. and Salah, A.A., Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing*, 2017.

Proposed Approach



Results: Quantitative Challenge

Table 1: Validation set performance of the proposed framework (System 8) and its sub-systems. FF: Feature-level fusion, WF: Weighted score-level fusion, RF: Random Forest based score-level fusion

SysID	System	INTER	AGRE	CONS	EXTR	NEUR	OPEN	MEAN TRAITS
0	ICPR 2016 Winner	N/A	0.9143	0.9141	0.9186	0.9123	0.9141	0.9147
1	Face: VGGFER33	0.9095	0.9119	0.9046	0.9135	0.9056	0.9090	0.9089
2	Face: LGBPTOP	0.9112	0.9119	0.9085	0.9130	0.9085	0.9103	0.9104
3	Scene: VD_19	0.8895	0.8954	0.8924	0.8863	0.8843	0.8942	0.8905
4	Audio: OS_IS13	0.8999	0.9065	0.8919	0.8980	0.8991	0.9022	0.8995
5	FF(Sys1, Sys2)	0.9156	0.9144	0.9125	0.9185	0.9124	0.9134	0.9143
6	FF(Sys3, Sys4)	0.9061	0.9091	0.9027	0.9013	0.9033	0.9068	0.9047
7	WF(Sys5, Sys6)	0.9172	**0.9161**	0.9138	0.9192	0.9141	0.9155	0.9157
8	RF(Sys5, Sys6)	**0.9198**	**0.9161**	**0.9166**	**0.9206**	**0.9149**	**0.9169**	**0.9170**

Test Set : Quantitative

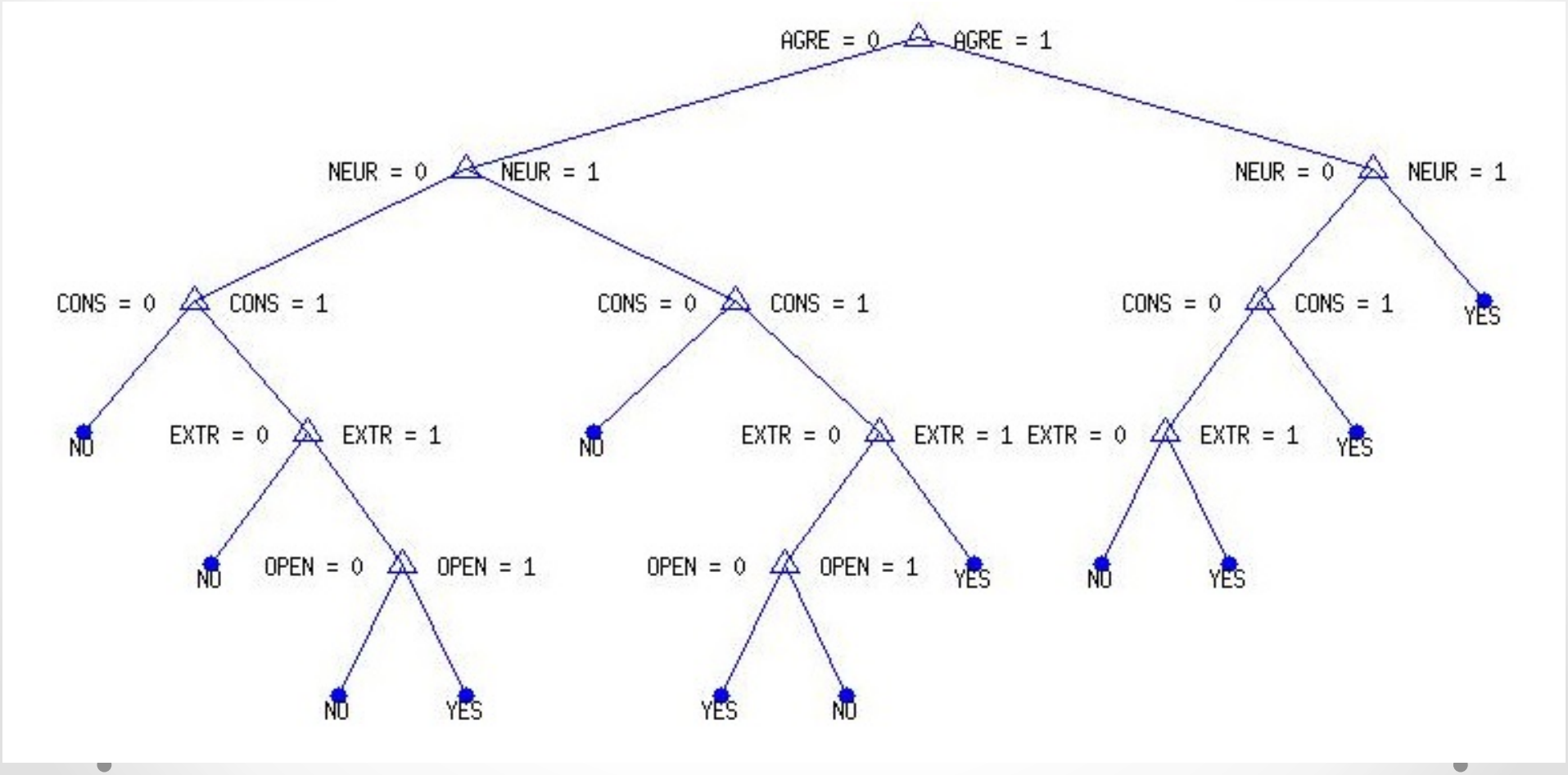
Table 2: Test set performance of top systems in the CVPR 2017 Coopetition - Quantitative Stage

Participant	INTER	AGRE	CONS	EXTR	NEUR	OPEN	MEAN TRAITS
Ours	**0.9209**	**0.9137**	**0.9198**	**0.9213**	**0.9146**	**0.9170**	**0.9173**
Baseline	0.9162	0.9112	0.9152	0.9112	0.9104	0.9111	0.9118
First Runner Up	0.9157	0.9103	0.9138	0.9155	0.9083	0.9101	0.9116
Second Runner Up	0.9019	0.9032	0.8949	0.9027	0.9011	0.9047	0.9013

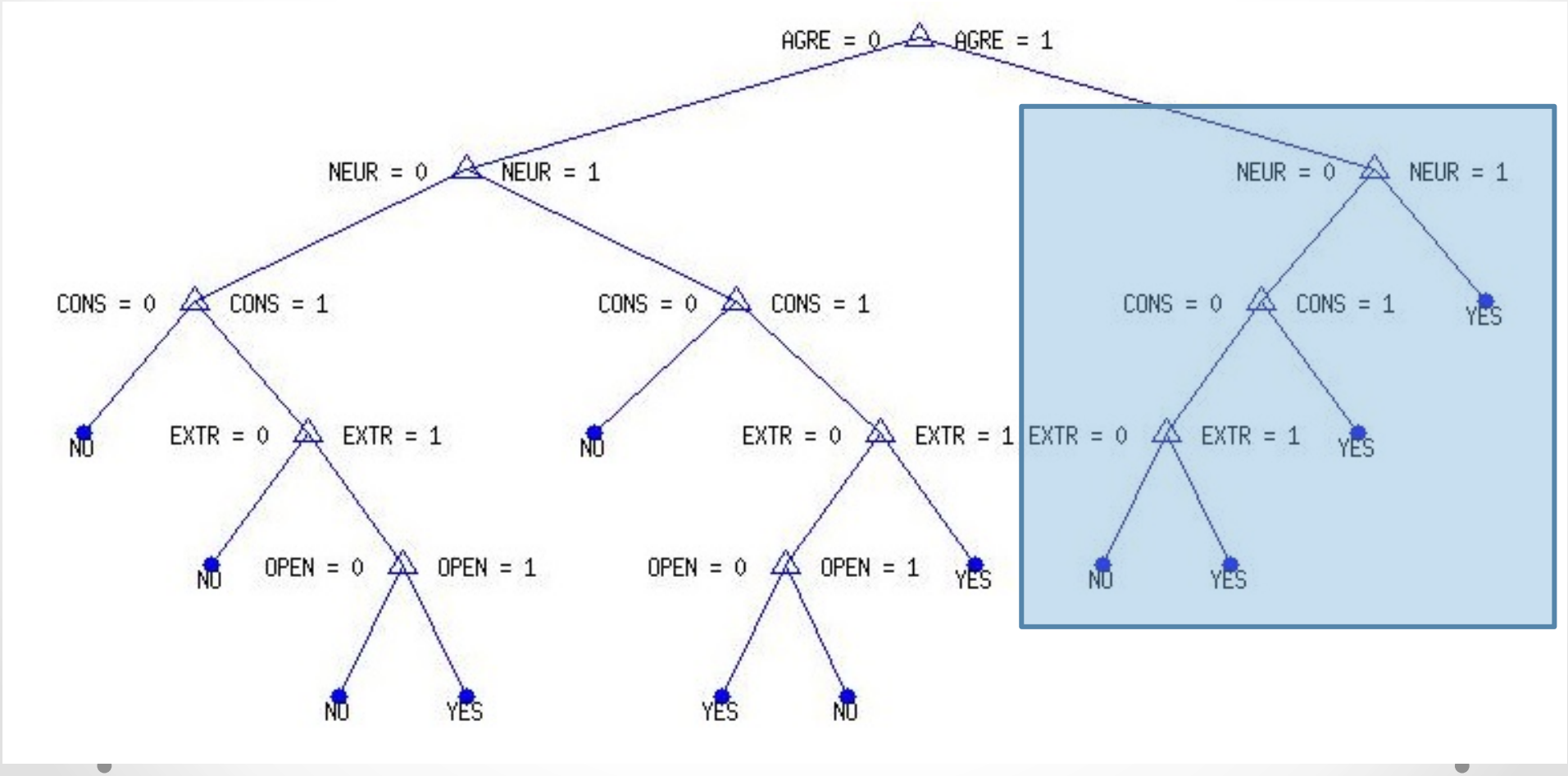
The Qualitative Challenge

- Use predictions from the quantitative stage as input
- Binarize predictions w.r.t. training set mean
- Use a single decision tree to map the OCEAN predictions to the interview variable
- Convert the tree into explanations via if-then rules
- Provide a descriptive image including:
  - The first detected face from the video
  - The mean normalized predicted scores
  - The automatically generated explanation

The Decision Tree for Qualitative Challenge



The Decision Tree for Qualitative Challenge



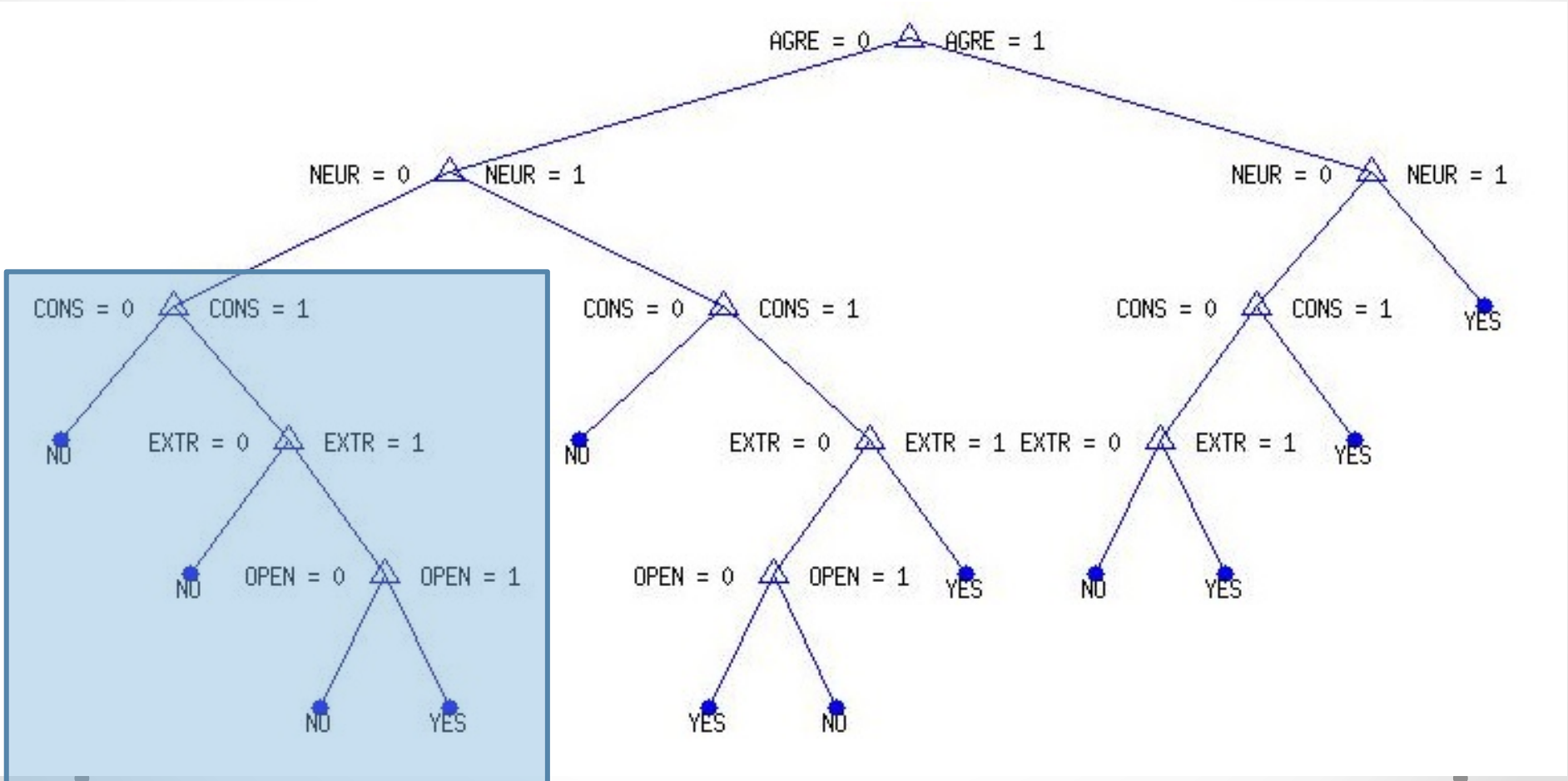
Probabilities

- p(YES | not Agreeable) = 0.198 (complement: 0.802)
- p(NO | Agreeable) = 0.186 (complement: 0.814)

Example 1

- High agreeableness, everything else is low.
- Interview Decision: NO

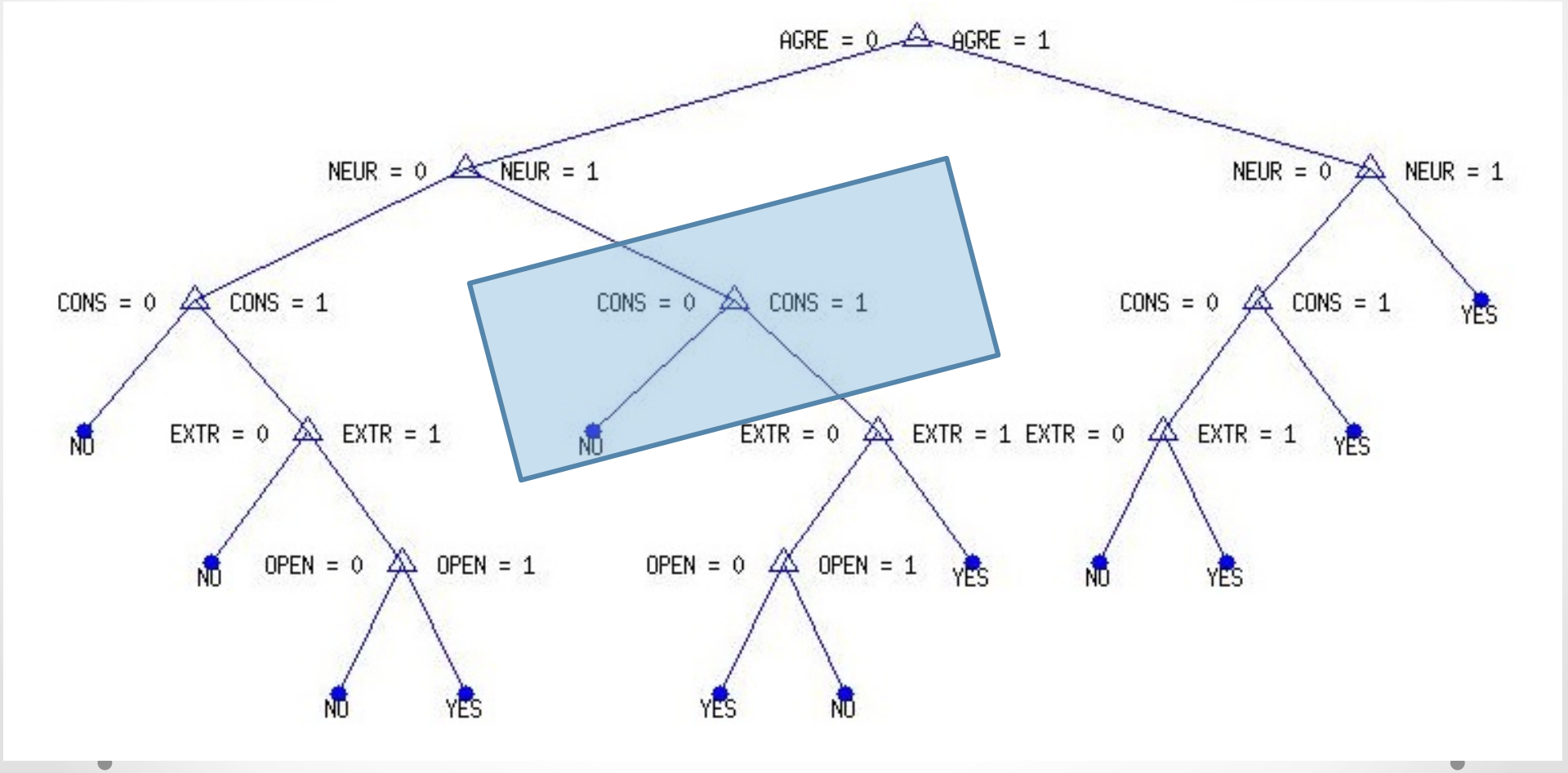
The Decision Tree for Qualitative Challenge



Example 2

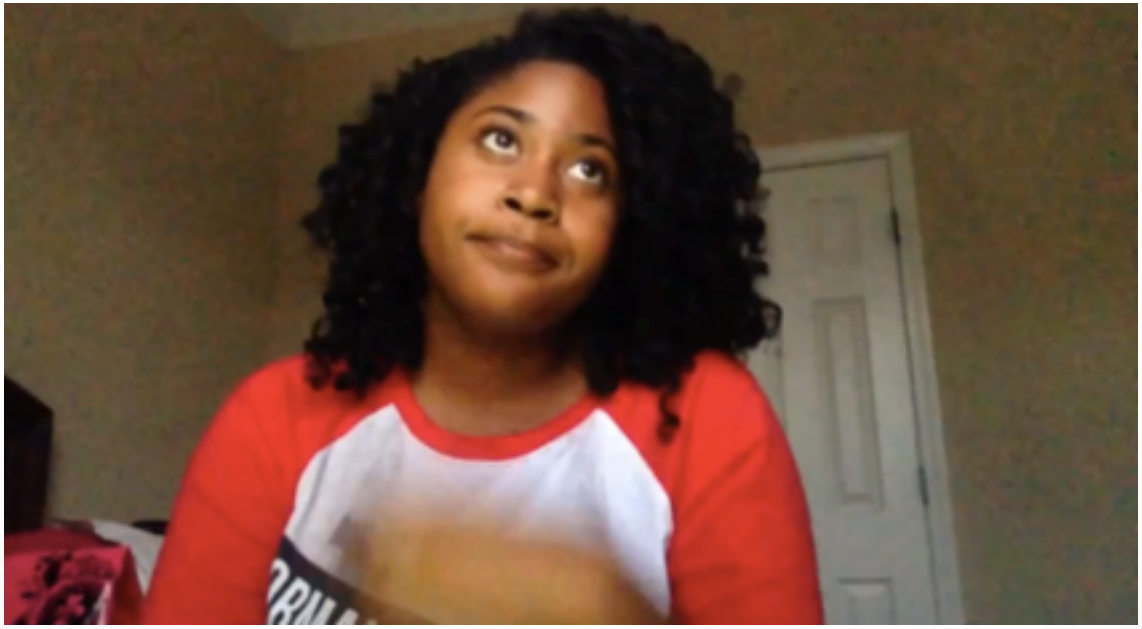
- Low Agree & Neuro, all others positive
- Interview Decision: YES

The Decision Tree for Qualitative Challenge

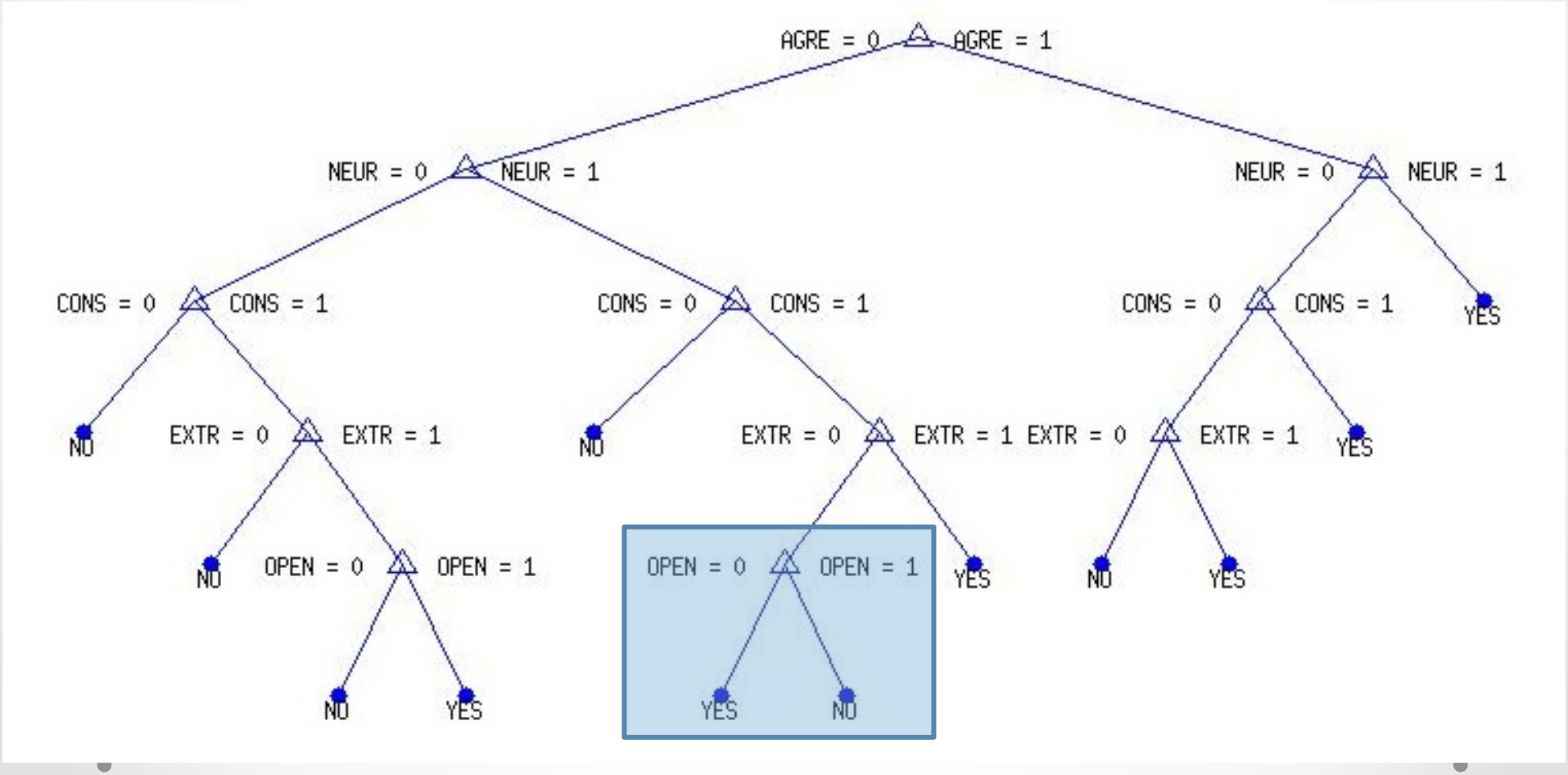


Example 3

- Low Agree & Consc., all others positive
- Interview Decision: NO



The Decision Tree for Qualitative Challenge



Example 4

- Low Agr. & Ext.; High Con. & Neu. + LOW openness
- Interview Decision: YES

Example 5

- Low Agr. & Ext.; High Con. & Neu. + HIGH openness
- Interview Decision: NO

Explanations

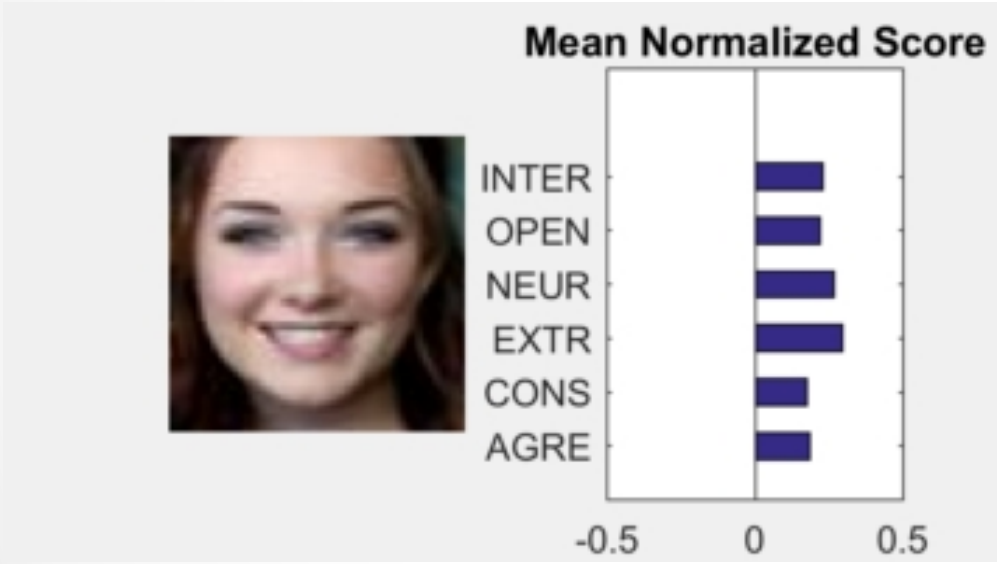
- If invite decision is 'YES'
  - 'This [gentleman/lady] is invited due to [his/her] high apparent {*list of high scores on the trace*}' [optional depending on path:', although low {*list of low scores on the trace*} is observed.']
- If invite decision is 'NO'
  - This [gentleman/lady] is not invited due to [his/her] low apparent {*list of low scores on the trace*}' [optional depending on path: ' although high {*list of high scores on the trace*} is observed.']
- If the direct and indirect predictions get in conflict
  - The directly predicted interview score and the classification based on traits are not consistent, the [gentleman/lady] may be re-evaluated. Following explanation is based on predicted traits.

More on Explanations

- The metadata of the corpus does not contain gender annotations
- We manually annotated 6000 videos from images
- We also check which modality is dominant
  - If the face system has the same sign with the final results: it is visual
  - Else (speech has higher effect than scene): it is audio system

Example 6

Visual Explanation

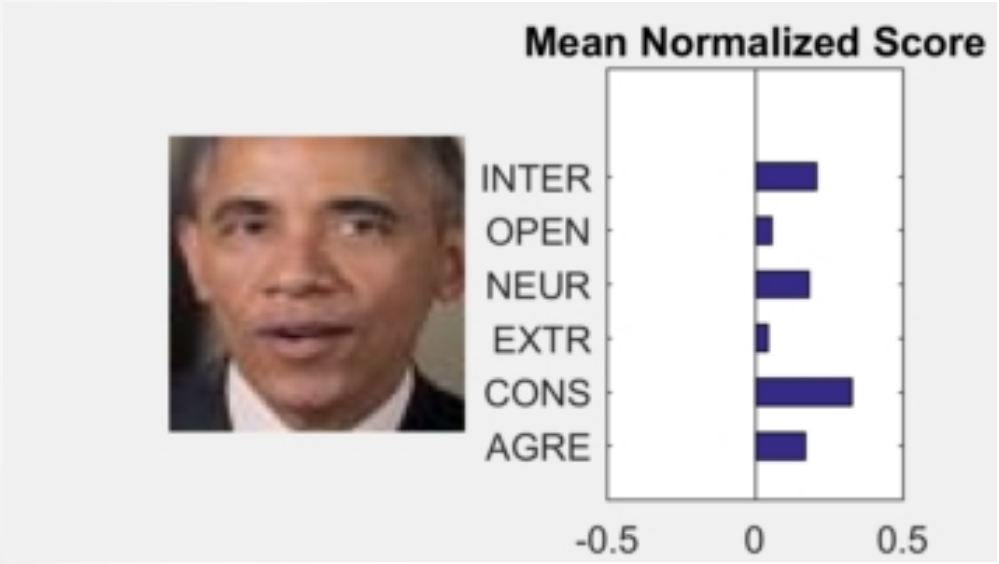


Automatic Verbal Explanation

This lady is invited for an interview due to her **high apparent agreeableness and neuroticism** impression. The impressions of agreeableness, conscientiousness, extroversion, neuroticism and openness are primarily gained from facial features.

Example 7

Visual Explanation

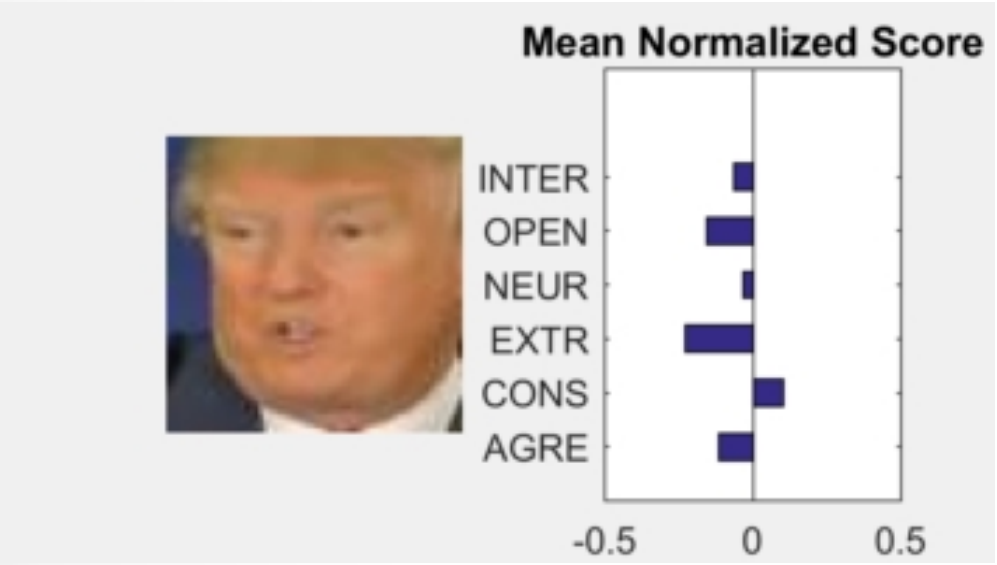


Automatic Verbal Explanation

This gentleman is invited for an interview due to his **high apparent agreeableness and neuroticism** impression. The impressions of agreeableness, conscientiousness, extraversion, neuroticism and openness are primarily gained from facial features

Example 8

Visual Explanation



Automatic Verbal Explanation

This gentleman is not invited due to his **low apparent agreeableness, neuroticism, extroversion and openness** scores. The impressions of agreeableness, conscientiousness, extroversion, neuroticism and openness are primarily gained from facial features.

Evaluation Measures for Qualitative Challenge

- **Clarity**: Is the text understandable / written in proper English?
- **Explainability**: Does the text provide relevant explanations to the hiring decision made?
- **Soundness**: Are the explanations rational and, in particular, do they seem scientific and/or related to behavioral cues commonly used in psychology.
- **Model interpretability**: Are the explanation useful to understand the functioning of the predictive model?
- **Creativity:** How original / creative are the explanations?

Qualitative Test Set

Table 3: Qualitative stage test stage winner teams' scores (mean±std scores of committee members)

Participant	**Our Team**	First Runner Up
Clarity	4.31±0.54	3.33±1.43
Explainability	3.58±0.64	3.23±0.87
Soundness	3.40±0.66	3.43±0.92
Interpretability	3.83±0.69	2.40±1.02
Creativity	2.67±0.75	3.40±0.8
Mean Score	**3.56**	3.16

Algorithmic Accountability

- Does our algorithm incorporate any systematic biases?
  - Do we favor women over men?
  - Do we favor younger subjects over older subjects?
  - Do we favor Caucasians over other ethnicities?

Some Bias for Females

Dimension	Pearson Corr	p Value
'agreeableness'	-0.0228	0.0412937
'conscientiousness'	0.0814	3.12E-013
'extraversion'	0.2072	2.69E-078
'neuroticism'	0.0542	1.24E-006
'openness'	0.1691	2.05E-052
'interview'	0.0692	5.63E-010

Some Bias for Apparent Ethnicity

	Asian		Caucasian		African-American	
Dimension	**Pearson C**	**p Value**	**Pearson Corr**	**p Value**	**Pearson Corr**	**p Value**
'agreeableness'	-0.0024	0.8283	0.060919	4.95E-008	-0.06750	1.51E-009
'conscientiousness'	0.0177	0.1135	0.055778	5.98E-007	-0.07376	3.97E-011
'extraversion'	0.0393	0.0004	0.039351	0.00043	-0.06814	1.06E-009
'neuroticism'	-0.0017	0.8824	0.047332	2.28E-005	-0.05259	2.53E-006
'openness'	0.0099	0.3754	0.083112	9.66E-014	-0.10003	3.02E-019
'interview'	0.0145	0.1946	0.051978	3.30E-006	-0.06754	1.47E-009

Conclusions

- Transfer learning and multi channel fusion are promising.
- Decision trees for decision fusion offer simple ways of explaining outcomes.
- The predictions are consistent over clips from the same video.

- Bias in dataset composition and label distribution is directly learned by the algorithm.
- Shouldn't invitation to job interview depend on the job as well?
- Experimental protocol should make sure no subjects are overlapping between dev & test sets.