Bachelor's Thesis

Annotating, analyzing and modeling of a video based personality trait corpus for mood primitives and likeability

Thomas Mol

Supervised by:

Dr. Heysem Kaya

Prof. Dr. Albert Salah

Utrecht University

Computer Science Department
Utrecht University
The Netherlands

July 2020

Annotating, analyzing and modeling of a video based personality trait corpus for mood primitives and likeability

Thomas Mol

Abstract

Video based job interviews and resumes are increasing in popularity among organizations and corporations. They are a vital part in today's job candidate screening process and prove to be more cost efficient than previous types of interviews and resumes. These advancements make way for the automatic screening of job candidates by analyzing and modeling video based data-sets. To be more precise, the assessment of apparent personality traits and mood primitives give insight into the impressions a job candidate leaves on a job recruiter. In this thesis we describe a video based system that analyzes a video clip of a person and then produces personality trait and mood scores. On top of that, it predicts whether the person should be invited to a job interview or not. This thesis explores several types of classification models to make accurate predictions but also to provide an explainable and interpretable model. The results show that the agreeableness, openness and conscientiousness dimensions have the most influence on the job interview prediction.

Acknowledgments

I would like to thank Dr. Heysem Kaya for this useful insight, feedback and knowledge throughout this project. I would also like to thank Prof. Dr. Albert Salah for his feedback on this thesis. Lastly, I also want to thank Bob Breemhaar for providing his annotations of the data-set.

Contents

1 Introduction

Affective computing and artificial intelligence are progressively becoming more popular and widely adopted by organizations (Davenport and Ronanki 2018, Tao and Tan 2005). Affective computing systems intent to be responsive, interpret and recognize human affects, e.g., emotions. Many applications of such affective systems can be developed for educational tools, mental healthcare and much more. Furthermore, artificial intelligence and machine learning can be utilized in the development of affective systems. The performance capability of today's computer systems are becoming increasingly better to the point at which affective systems can be used in real-time.

Another area where affective computing and machine learning methods can be applied is in the job candidate selection process. Job interviews are an essential part of the job candidate selection process for many corporations and organizations. Affective computing could help both the job seeker and the job recruiter. In this thesis we will explore a system that can predict whether an applicant should be invited to a job interview based on their perceived personality and mood. However, this system should not be used to make an actual job hiring or interview invitation decision, but instead it should be used to evaluate and indicate wrong judgments a recruiter might make during the job candidate screening process. Research shows that people form opinions of others within 100 milliseconds exposure of a new face (Willis and Todorov 2006, Todorov et al. 2009).These judgments are often based on stereotypes, which could introduce biases from the recruiter during the job candidate screening process. With the proposed system the applicant can also rectify their perceived impression in the case that a recruiter shows biases. This system will be modeled based on video modality. Additionally, the job seeker could also benefit from such a system as the system could show them what impression they give the recruiter and even suggest improvements to increase their chances to be invited to the job interview.

Nevertheless, caution should be taken when developing and using such a system. Supervised learning methods will be needed to create such a system. This means the system will rely on human based annotations for its supervision. As a result, the system will also learn the biases and stereotypes that annotators might have. For example, a younger age group is given higher overall job interview ratings compared to an older age group (Morgeson et al. 2008). However, in actual job performance, the older workers are perceived to have better performance (Truxillo et al. 2012). Therefore, first impressions may not be a robust basis to determine actual job performance of a candidate.

In this thesis we will first discuss related literature on apparent personality traits, mood primitives and video resumes and job interview. Then, we will discuss our proposed methodology for our experiments which includes data annotation, feature extraction

from video clips, feature normalization and lastly, classification and explainability models. We will also analyze the inter-rater agreement as we will have two annotators annotating our data-set. Next, we will report and discuss the results of the proposed experiments. After, we will look at the explainability analysis of the models. Lastly we will discuss our findings, shortcomings and describe future research than can be done regarding this research area.

2 Related Literature

2.1 Job Interviews & Job Application Videos

Many organizations and corporations use job interviews to select the most qualified person for an open position. Job interviews typically consist of a short meeting where the applicant is assessed based on their skills and experience but also on their personality, mood, energy and motivation during the interview. A study by Kraus and Kurtis (1990) shows that managers also base their decision heavily on their overall impression of the applicant. These dimensions are assessed to determine and predict if the applicant would be successful at the job they have applied for. Hiring an incompetent worker would not only mean the loss of salary but also the time it takes to search for a new candidate. A study by Dettmar (2004) shows that these implicit and explicit costs of hiring the wrong candidate combined could be as much as that of a years salary of that new hired worker. Despite this risk, job interviews prove to be a reliable method in the job candidate selection process (Weekley and Gier 1987).

However, with the increase in use of online communication tools and online application processes job the approach for job interviews has changed as well. Many organizations choose to use online platforms to recruit new employees for open positions. A study in 2011 showed that 76% of unemployed people search online for new jobs (Faberman, Kudlyak, et al. 2016), which is an indication of how reliant organizations have become on online job recruitment. Research has been done to improve the online recruitment process by increasing market transparency and lower transaction cost with the use of semantic web technologies (Bizer et al. 2005). Further, video resumes have become increasingly popular as a result of the increase of use of online communication platforms. Studies show that video resumes are mostly made by a young demographic who are looking for a junior position (Nguyen and Gatica-Perez 2016). In addition to online video resumes, a new type of online job interviewing has also emerged which is called asynchronous video interviews (Torres and Mejia 2017). This type of interviewing involves an applicant who sends a video recording of them answering interview questions (Toldi 2011). On the other hand, synchronous video interviewing are held on platforms like Skype, Microsoft Teams and other video conference software. These new methods of interviewing can decrease costs as they prove to be more time-efficient (Weber and Silverman 2012).

2.2 Personality Traits

The assessment of personality traits is equally important when analyzing and determining which candidate is the most suitable for the position (Kinsman 2005). However, assessing a person's personality can be a rather diffi-

cult task. Therefore, we must first understand how personality is defined and how to analyze it.

In psychology, the characterization of someone's personality is based on long-term behavior. Personalities have sophisticated structures and are constructed of many different aspects such as habits and the way people think or feel. Therefore, it can be difficult to analyze, measure and categories personalities to divide humans into different personality types. Instead, psychology now mainly focuses on personality traits, where the OCEAN (or CANOE) Big Five personality trait model developed by Paul Costa and Robert R. McCrae (Costa and McCrae 1992) is most widely accepted. These traits are; Openness to experience, Conscientiousness, Extraversion, Agreeableness and lastly, Neuroticism.

The first factor of the 5 factor personality trait model is Openness to experience. People who score high in this factor generally are imaginative and care more about aesthetics, ideas and values (McCrae 1993). The Conscientiousness factor is related to how well-organized and persistent a person is. People with high conscientiousness are considered caring, dependable and organized (Widiger 2017). The traits underlying the Extraversion dimension are related to sociability and activity. Extraverted people are more talkative, adventurous and sociable while introverted people tend to be more silent, cautious and focused on their inner state of mind(Widiger 2017). Agreeableness is a dimension of interpersonal behavior. People with high scores on agreeableness are cooperative, good natured and are seen as friendly (Graziano and Eisenberg 1997). Lastly, the Neuroticism dimension represents a person's tendency to experience emotional anxiety (Widiger 2017). The dimension can also be described as someone's emotional stability. People with high neuroticism are more anxious and nervous and tend to worry more while people with low scores are more calm and composed.

Several works have used the five factor personality model to analyze human behavior. Quercia et al. (2011) for example, who analyzed relationships with different types of Twitter users. They were able to predict the personality types of Twitters users based on their following, followers and listed counts. Further, Allbeck and Badler (2008) used the OCEAN model to map personalities to crowd simulations in order to improve the realism of their simulations. They used the OCEAN model as a basis for their simulation agents of psychology. The resulting simulations has shown the impact of different personality traits in crowds.

2.3 Apparent Personality Analysis

Another approach to apply the five factor model is to analyze and annotate the apparent personality traits (J. Junior et al. 2018, Chen et al. 2016), rather than the actual personality traits. In this process, an annotator annotates the impression of a subject's personality rather than the actual personality traits of the subject. This impression of a person's personality can also be described as a person's apparent personality. This is easier as the annotator relies on external evaluations without any direct involvement of the subject.

Several studies have shown that the modeling of apparent personality traits is possible from different modalities like text (Gievska and Koroveshovski 2014, Alam et al. 2013), speech (Valente et al. 2012, Madzlan et al. 2014) and also video based (J. Junior et al. 2018, Qin et al. 2016, Escalante, Kaya, Salah, Escalera, Güçlütürk, et al. 2020). On top of that, research has been done on modeling job interview decisions based on apparent personality traits from different modalities, includ-

ing the video based modality (Kaya, Gurpinar, et al. 2017, Kaya and Salah 2018, Yu 2019). These papers prove that the modeling of apparent personality traits is feasible and show high scores in classification accuracy.

Further, deep learning based classifiers are used in this thesis to model apparent personality traits and other dimensions. Compared to other methods of feature extraction, deep learning is becoming more widely used as it proves to be more robust. For example, it has been applied in the recognition of faces (Parkhi et al. 2015), emotion recognition (Kaya, Gürpınar, et al. 2017) and modeling job interview decisions (Kaya and Salah 2018).

2.4 Mood Primitives & Likeability

Another dimension that can be used to classify a person's behavior is to assess their mood or emotions over a short period of time. While the terms emotion and mood seem to be used interchangeably, they are in fact different. In psychology, an emotion is seen as a immediate reaction to a stimulus, while a mood lasts for a longer period (Bower G. H. 2000). For example, someone can be in positive mood throughout the day but still have a negative emotional reaction to a stimulus like a bad smell, or being insulted (Matlin 2012). Additionally, this makes it more difficult to identify and specify the cause of someone's mood as there is no direct cause like in a emotional reaction (Desmet et al. 2016).

In order to simplify the classification of a subject's mood, emotion classification can be used. Instead of trying to determine someone's mood over a short period, the emotions can be used as an indicator or their mood. To do this, a base set of emotions is needed to start classifying a subject's mood. The six basic emotions developed by Paul Ekman in 1992 are widely regarded as the standard to identify emotions (Ekman 1992). Figure 1 illustrates the emotions as described by Ekman. Ekman found that there are nine distinct characteristics that distinguish the basic emotions, for example, the physiological response, the duration and their quick onset. He also found that cultural background had no impact in the identification of emotion. Meaning that the six basic emotions can be used across different cultures to classify emotions. The six emotions that Ekman identified as basic are; anger, happiness, surprise, disgust, sadness and fear (Ekman 1992).

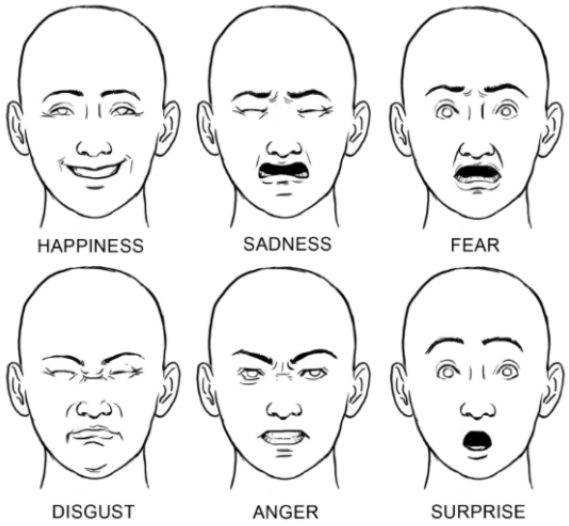

Figure 1: The 6 basic emotions as proposed by Ekman.

Research has also indicated that mood can be defined as a dimensional model with two dimensions: valence and arousal. This model was first proposed by Russel in 1980 (Russell 1980), and is known as the circumplex model of affect. In this model arousal can be seen as the amount of energy a person shows. This dimension ranges from a deactivated state, described as sleepiness, to an activated state, which can be described as aroused. The valence dimension refers to the pleasantness or unpleasantness. This dimension typically ranges from a state of misery to a state of pleasure. Combining these two dimension creates four

quadrants with four basic moods; angry, happy, sad and relaxed. Figure 2 shows the circumplex model and indicates where the basic moods would be positioned on this model.

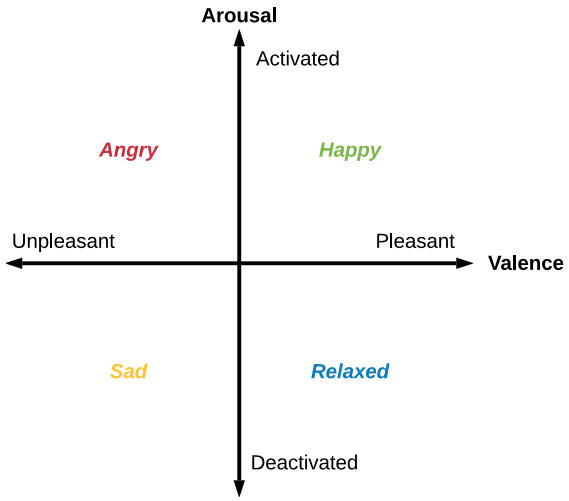

Figure 2: The circumplex model of affect as proposed by Russel.

The dimensional model has been used by Desmet et al. (2016) to develop a method that can be used to identify moods. In their paper, Desmet et al. propose a pictorial scale to report or express a mood. The pictorial scale consists of eight different moods, two for each quadrant on the circumplex model. However, in this thesis, the valence and arousal dimensions will be regarded as two separate dimensions which will be annotated with an ordinal scale as opposed to having a single categorical mood dimension.

Another dimension that can be modeled to automate the job candidate screening process is likeability (Celiktutan and Gunes 2015). Likeability usually refers to whether someone is seen as friendly and social. Although likeability is a highly subjective dimension, research has shown that it is associated with the Big Five personality traits of extraversion and neuroticism, while being positively related to agreeableness (Van der Linden et al. 2010). Studies have also shown that likeability is a significant factor in the hiring processes (Raza and Carpenter 1987). Hayes and Macan (1997) also found that an applicant's attractiveness has an effect on the likeability of the applicant. However, little research has been done on using video features to model the likeability of an applicant and predict whether an applicant should be invited for a job interview. This will be explored further in this thesis.

3 Methodology

Our proposed method analyzes a video clip containing one person. The output is an estimate of the interview invitation variable, the big five OCEAN personality traits, the two mood primitives valence and arousal and lastly, the likeability variable. Only video features are used in our proposed method. This section will describe the main stages of our method, starting with describing our data-set and the process of annotating this data. Then the feature extraction and normalization processes are defined. Lastly, the classification methods are described. Figure 3 shows the pipeline of our proposed method.

3.1 The Data Set

For modeling and experimenting we relied on a publicly available data-set which contains 10,000 video clips with audio with an average duration of 15 seconds[1]. The clips are collected from over 3000 videos available on YouTube. Figure 4 shows screenshots from four samples taken from the video clip data-set. The videos are annotated with the use of Amazon Mechanical Turk for apparent personality traits. These are the 5 OCEAN personality traits variables. Annotators saw a pair of video clips and were asked to rank the clips in order of which clip showed a higher score across the the 5 OCEAN personality trait variables.

Furthermore, the videos are annotated for an additional variable, which is the interview variable. This variable measures whether a candidate should be invited for a job interview or not. Additionally these videos are annotated for ethnicity groups, age groups and gender (Escalante, Kaya, Salah, Escalera, Gucluturk, et al. 2018). From this data-set of 10,000 videos the first 960 were selected for experimenting and further annotations of three more variables, namely, the valence, arousal and likeability variables.

3.2 Annotating Data

The selected 960 video clips were annotated additionally for mood primitives, likeability and a variable describing the presence of background music. The mood primitives, valence and arousal, and likeability were annotated with an ordinal scale with 3 classes. An annotation of class 1 means the variable scored low or negatively, class 2 is the neutral case, and class 3 means the variable scored high or positively. Initially the scale was set with 5 classes, however annotations with class 1 or class 5 were scarce so a smaller scale was chosen to improve the accuracy of the models.

Two annotators annotated the same 960 video clips for these 4 dimensions. After which, the annotations were merged to create two additional data-sets. One data-set used the minimum of the two annotation, while the second one used the maximum of the two annotations. As a result, there were now 4 sets of

[1]The data-set can be found on http://chalearnlap.cvc.uab.es/dataset/24/description/

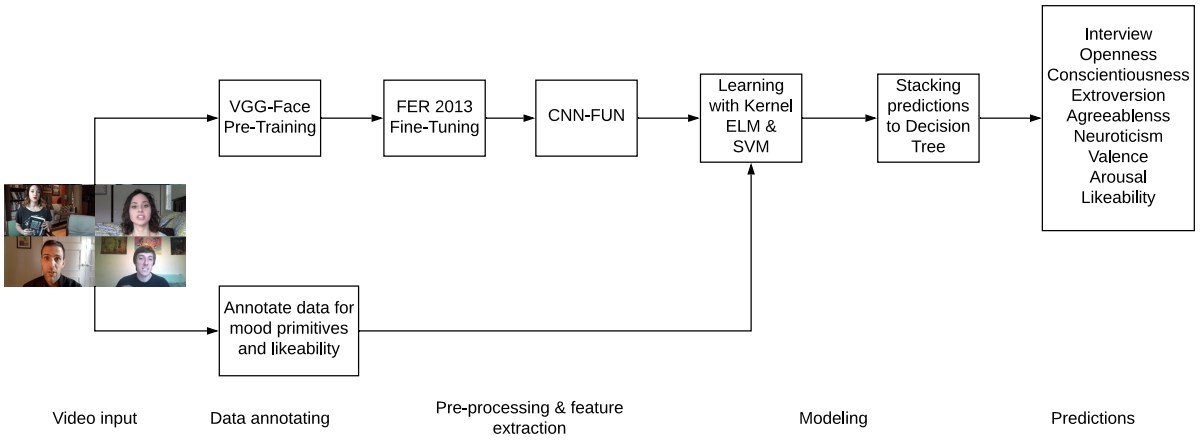

Figure 3: Flowchart of the proposed methodology.

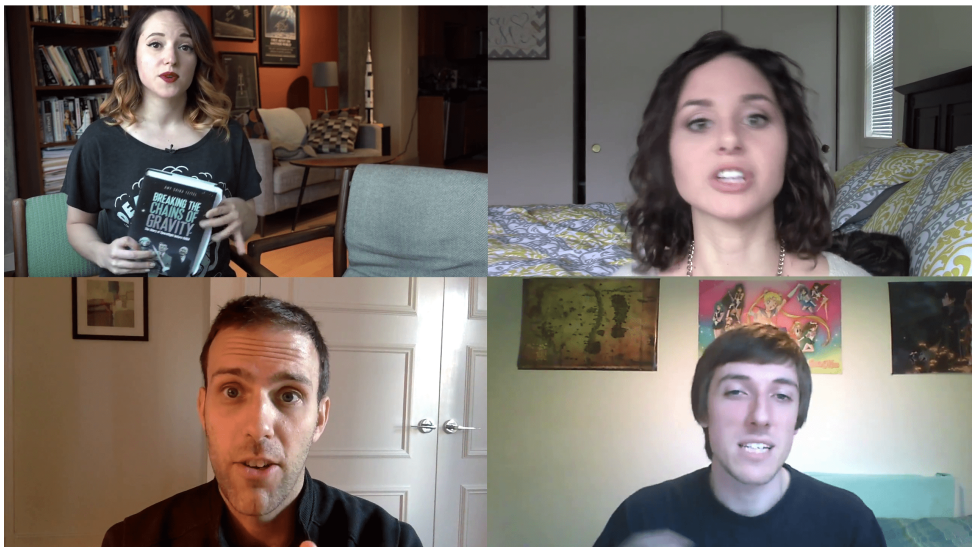

Figure 4: Screenshots from four samples of the video clip data-set.

annotations available to use and model classifiers with.

To measure the inter-rater reliability between the two annotators the Cohen's kappa coefficient was calculated. This coefficient is considered more robust as it takes into account the occurrence of an agreement by chance. The resulting coefficient will determine whether the agreement is accidental or not. Table 1 shows the results of the Cohen's kappa calculations. The observed agreement (po) is the number of agreements divided by the total of number items. The random agreement (pe) is the probability that the annotators agreed on either of the possible annotations. Using the observed and random agreement the Cohen's kappa can be calculated. The coefficient ranges from 0 to 1 and can be interpreted as shown in Table 2.

Since Cohen's kappa takes into account the disagreement but not the degree of disagreement a modified version of Cohen's kappa can be used which is the weighted Cohen's kappa (Cohen 1968). This can be applied well if an ordinal scale is used for the ratings. Since the annotations of the data-set use an ordinal scale, the weighted Cohen's kappa can be applied. The weighted kappa is calculated with the use of a table containing weights, which measures the degree of disagreement. This

Dimension	Weight	po	pe	Kappa	Kappa Error	Agreement	Null hypothesis
Arousal	Unweighted	0.6229	0.4077	0.3634	0.0264	Fair	Rejected
	Linear	0.8073	0.6713	0.4138	0.0387	Moderate	Rejected
	Quadratic	0.8995	0.8031	0.4896	0.0493	Moderate	Rejected
Valence	Unweighted	0.7177	0.5046	0.4302	0.0293	Moderate	Rejected
	Linear	0.8573	0.7371	0.4571	0.0429	Moderate	Rejected
	Quadratic	0.9271	0.8534	0.5027	0.0572	Moderate	Rejected
Likeability	Unweighted	0.5812	0.4230	0.2743	0.0276	Fair	Rejected
	Linear	0.7818	0.6704	0.3380	0.0404	Fair	Rejected
	Quadratic	0.8820	0.7941	0.4272	0.0506	Moderate	Rejected

Table 1: Inter rater agreement results. po: observed agreement, pe: random agreement, Kappa: Cohen's Kappa.

Cohen's Kappa	Agreement
0	Agreement equivalent to chance
0.10 - 0.20	Slight
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Substantial
0.81 - 0.99	Near perfect
1	Perfect

Table 2: Cohen's kappa agreement interpretation.

can be linear or quadratic. Table 1 shows the results of the weighted Cohen's kappa alongside the unweighted versions.

Overall, every dimension shows a moderate agreement when quadratic weighting is applied. Further, for every dimension and weight the null hypothesis is rejected, where the null hypothesis is that the observed agreement is accidental. Therefore it is safe to say the inter-rater reliability is higher than chance level. The data was then split up in a train set of 660 video clips and a test set of 300 video clips to be used for model training, optimization and testing.

3.3 Video Feature Extraction

To train our models we use video features that are extracted from the 15 second YouTube clips provided by the ChaLearn Competition. The features are facial features which are extracted over the entire video clip. These features are summarized by functionals. First, the faces were detected and then aligned using the Supervised Descent Method (Xiong and De la Torre 2013). Every detected face is rotated using the roll angle which is estimated from the eye corners. Also, there are 49 landmarks located on each face to which a 20% margin is applied to corp the image to a size of 64 x 64 pixels. Image-level features are extracted from a convolutional neural network which was trained for emotion recognition (Kaya, Gürpınar, et al. 2017). This network is a pre-trained VGG-Face network (Parkhi et al. 2015), which is optimized for large data-sets. The final layer of the network was changed to a 7-dimensional emotion recognition layer. The resulting network was then fine-tuned using the FER-2013 dataset (Goodfellow et al. 2015), which contains over 35,000 images. The final trained network has 37 layers from which the 33rd response was used. This descriptor has 4096

dimensions which were then summarized using four statistical functions. Those functions were the mean, standard deviation, slope and offset from a linear polynomial.

3.4 Feature Normalization

Z normalization transforms the input vector into a vector where the mean is 0 or very close to 0 while the standard deviation is close to 1. The number for a single data point of the computed output vector can also be called the z-score. This score essentially is the number of standard deviations a vector is away from the mean. It can indicate how familiar a data point is relative to the others.

While Z normalization is applied at feature level L2 normalization is applied at feature vector level. L2 normalization is applied at each row of the feature set so that if the values are squared and then summed, they add up to exactly 1.

Figure 5 shows the flowchart of the normalization steps that were applied at feature and vector level.

3.5 Classification Algorithms

Several types of models were used for classification from the extracted video features. These models include:

1. Extreme Learning Machines (ELM)
2. Support Vector Machines (SVM)
3. Decision Trees (DT)
4. Ridge Classification

In the first stage we use the ELM and SVM classifiers, while in the second stage we use the predictions of the ELM classifier to train a decision tree and linear model. The second stage classifications will give an indication of the feature importance and contribute to the explainability of our classification models.

Originally proposed by Huang in 2004 (Huang et al. 2004), the learning algorithm for Extreme Learning Machine uses single-hidden layer feedforward neural networks (SLFN). It chooses the nodes randomly and analytically determines the the output weights for the neural network. As a result, the training speed of this algorithm is noticeably fast. The technical details of the classifiers can be found in Huang et al. 2004 and Huang et al. 2006. However, in this thesis we use Kernel ELM, where the first level transformation is done implicitly in the kernel space and the second layer weights (mapping the kernel matrix to outputs) are learned analytically using regularized least squares. We use a linear kernel which has a regularization coefficient. We optimize this coefficient during training and with threefold subject independent cross-validation of the training set.

Just like ELM, the SVM algorithm is used for regression and classification. SVMs is a supervised learning method which is effective when the number of features is high or even higher than the number of samples. The algorithm finds the optimal separating hyperplane in a dimensional space of size N, where N is the number of features. It then uses the hyperplane to classify the data-points. To obtain the most accurate model the algorithm finds the hyperplane where the margins of the hyperplane are the largest. Just like the ELM algorithm a linear kernel is used and a regularization coefficient is optimized.

The models that are trained using the ELM and SVM algorithms will create a set of predictions of the personality trait, mood and likeability dimensions. The predictions of the ELM classifier will be used in the second stage of our classification. During this stage, the predictions are stacked to a Decision Tree classifier as high level features. The DT classifier is trained using the predictions and the in-

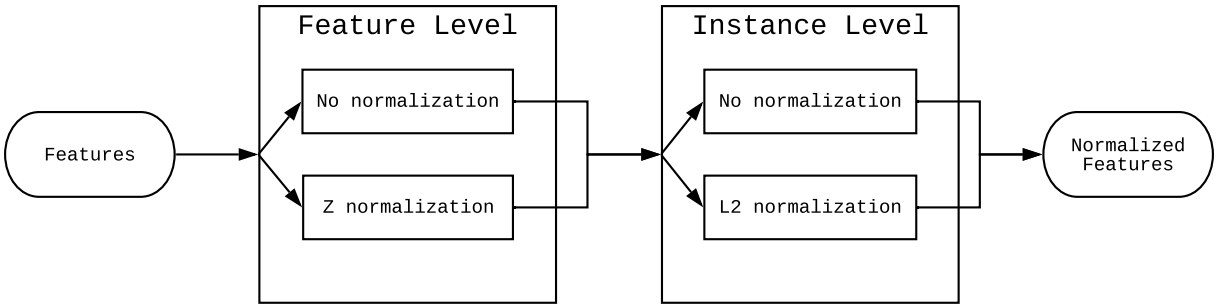

Figure 5: Feature normalization pipeline

terview invitation variable as labels. Then, the trained DT can be visualized which contributes to the explainability and interpretability of our classification model.

Lastly, the same prediction data set to train the Decision Tree classifier is used to train a linear model. We will be using the Ridge Classifier which converts the target values to -1 and 1. Then, the model treats the problem as a regression problem. The trained model will provide us with the coefficient of the features of the decision function. In other words, the feature importance can be derived from these models which will show us what feature is the most or least important for determining the class.

4 Experiment Results

This section will provide, analyze and discuss the results of the annotations and of the experiments done with the data-set. First, general statistics will be provided regarding the mood and likeability annotations. Then the new annotations in conjunction with the interview variable will be analyzed to identify any statistical relationship. Next, we will provide the results of the experiments. The performance of the models created during the experiments will be expressed in their Unweighted Average Recall or UAR. The UAR is the mean of the recall scores of each class. The recall is the ratio between the number of true positives and false negatives. If all predictions would be correctly classified the UAR would be 1. The UAR metric is a better indication of prediction capacity than precision accuracy as our data-set is imbalanced. The UAR does not favor the majority class and is therefore a better indicator of the prediction capabilities of our models.

4.1 Annotation Analysis

As described in the methodology, 960 video clips of the data-set were annotated for the Big Five personality traits, two mood primitives (valence and arousal), likeability, background music and the interview variable. The personality and interview dimensions were annotated using Amazon Mechanical Turk. All dimensions, with the exception of background music, were annotated for apparent presence or absence. The interview variable indicates whether the annotator would invite the subject to an interview or not. Table 3 and Table 4 show the number of annotations for each class and dimension.

The apparent personality trait and interview dimensions were post-processed to create cardinal scores for each video clip (Escalante, Kaya, Salah, Escalera, Gucluturk, et al. 2018). These scores were also binarized by taking the mean of each dimension and assigning a 2 if the score was equal or higher than the mean and assigning a 1 if the score was lower than the mean. Table 5 shows the number of annotations based on the binarized data-set.

Class	Arousal	Valence	Likeability
1	77	46	99
2	364	707	505
3	519	207	356
Total:	960	960	960

Table 3: Number of annotations (self) per class of the mood primitive and likeability dimensions

Class	Arousal	Valence	Likeability
1	108	82	197
2	593	705	602
3	259	173	161
Total:	960	960	960

Table 4: Number of annotations (gold min) per class of the mood primitive and likeability dimensions

Class	OPEN	CONS	EXTRA	AGRE	NEUR	INTER
1 (Low)	528	458	443	437	441	426
2 (High)	432	502	517	523	519	534
Total:	960	960	960	960	960	960

Table 5: Number of annotations per class of apparent personality trait and interview invitation dimensions. AGRE: agreeableness, CONS: conscientiousness, EXTRA: extroversion, NEUR: neuroticism, OPEN: Openness to Experience, INTER: interview invitation

4.2 Statistical Relationships

Some interesting statistical experiments can be done regarding the statistical relationship of the mood primitives and likeability dimensions and the interview invitation. For these experiments we will use the cardinal values of the interview variable and the self annotations of the arousal, valence and likeability dimensions. Three box-plots are created and are shown in Figure 6, Figure 7 and Figure 8. For these box-plots the y-axis indicates the cardinal interview invitation variable, where a higher score means they would be more likely to be invited for an interview. The x-axis indicate the three ordinal classes of the arousal, valence or likeability dimensions. The orange horizontal line in the box-plots indicate the mean. As the plots show, there is a upwards trend in the mean of all three box-plots. Meaning that if the subject is classified in the higher class, they would be more likely to have a higher score in the interview invitation variable.

However, to see if there is a correlation between the interview variable and mood and likeability variables we must look at the Spearman's Rank-Order Correlation coefficient. This coefficient is used to calculate correlation for ordinal categorical variables. To calculate this coefficient we will be using the binarized dataset of the interview variable. The coefficient is calculated as follows: $r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$, where d_i is the difference between the two ranks of each observation and n is the number of observations. Table 6 shows the correlation and p value between the three dimensions and the binarized interview variable. The results show a low p value for all dimensions, with every p value being close to 0. This means that the null hypothesis should be rejected, where the null hypothesis is that two sets of data are uncorrelated. By rejected the null hypothesis we can conclude that the sets of data are not uncorrelated, however it does not prove that they are strongly correlated. If we look at Table 6 we can see that the correlation values are between 0.2 and 0.4. For the arousal dimension this means that the strength of the correlation is small, while for the valence and likeability dimensions there is medium strong correlation. A perfect correlation would be indicated with a value of 1 or -1 (for a negative correlation). Typically, a correlation is considered strong if the correlation coefficient value is higher than 0.7 or lower than -0.7.

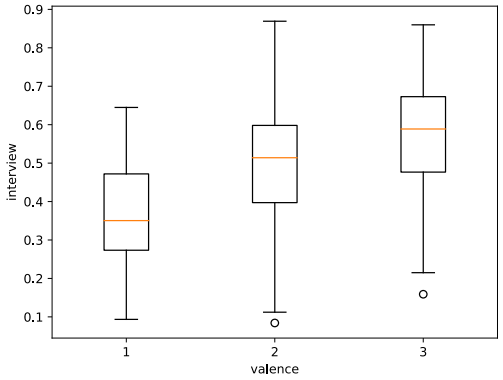

Figure 6: Box-plot of the valence classes and interview invitation variable

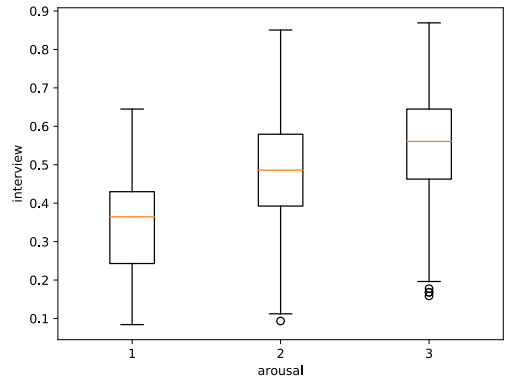

Figure 7: Box-plot of the arousal classes and interview invitation variable

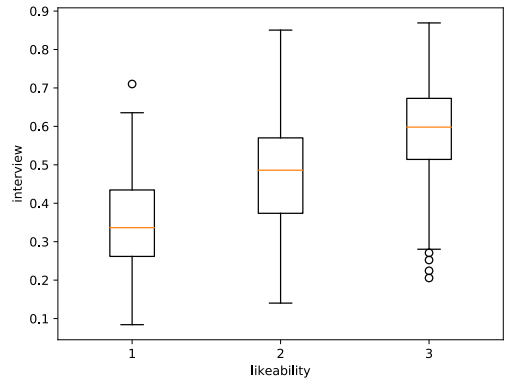

Figure 8: Box-plot of the likeability classes and interview invitation variable

4.3 Mood and Likeability Model Experiments

In this section, we will discuss the performance of the models that are trained using the mood and likeability dimensions and the extracted video features. As stated in the methodology, the data-set was split into a training set of 660 instances and a test set of 300 instances. Both the self annotation and the gold standard annotations were used for training the classifiers. Both the SVM and ELM algorithms as described in section 3.5 were used to the train models. Table 7 shows the results in UAR of the trained ELM models. A linear kernel was used to train these models. Also, hyper-parameters, like the C parameter, were optimized to achieve the best scores. The highest score of each dimension is highlighted bold. The normalization column refers to what type of normalization was applied to the feature set. "NN" stands for no normalization, "ZN" for z normalization and "L2" stands for l2 normalization. These are the normalization techniques as described in section 3.4. As the results show, the best results of each dimension is higher than 33%, which indicates that the models are more accurate than chance level. Chance level is 33% in the case of our data-set as the dimensions are labeled with three classes. Further, the results for the weighted ELM are higher than non-weighted ELM for all dimensions. Also, the gold standard annotation dimensions score higher than the self annotated labels. This implies that more annotators are needed for the same data to increase the performance of the models.

	Arousal	Valence	Likeability
Cor	0.296	0.204	0.395
p	6.51E-21	1.76E-10	3.09E-37

Table 6: Results of the Spearman Rank Correlation coefficient calculations. Cor: Spearman's Rank-Order Correlation coefficient, p: statistical difference

The results of the SVM models trained using the extracted video features are shown in Table 8. Again, the results are expressed in their UAR, and the highest score for each dimension is highlighted in bold. The results show moderately good performance. The highest scores for each dimension is higher than chance level, but they are slightly worse than the ELM models.

Other types of kernels, like the RBF and sigmoid kernel, were explored for the ELM and SVM models. However, these models performed worse on all dimensions compared to the linear models.

Weighted	Norm	self_aro	gold_aro	self_val	gold_val	self_like	gold_like
No	**NN**	39.79%	42.67%	47.91%	47.78%	38.47%	42.61%
	ZN	43.76%	49.88%	49.11%	46.21%	51.56%	49.11%
	L2	39.80%	42.49%	47.53%	47.19%	39.84%	43.25%
	ZN+L2	39.68%	47.63%	49.95%	46.64%	48.68%	49.98%
Yes	**NN**	41.60%	47.15%	49.00%	52.25%	51.33%	53.04%
	ZN	**45.13%**	**52.46%**	**51.92%**	55.92%	**53.19%**	49.32%
	L2	38.11%	51.08%	49.36%	**56.77%**	52.18%	**53.49%**
	ZN+L2	43.91%	52.37%	50.05%	52.38%	53.08%	50.35%

Table 7: Linear Kernel Extreme Learning Machine model performance results expressed in UAR. Highest score highlighted in bold. Norm: normalization, aro: arousal, val: valence, like: likeability.

Normalization	self_aro	gold_aro	self_val	gold_val	self_like	gold_like
NN	42.86%	**49.05%**	49.66%	47.92%	41.90%	47.25%
ZN	**43.90%**	48.18%	43.57%	43.47%	**43.29%**	**48.44%**
L2	43.16%	48.68%	**50.83%**	**48.22%**	40.42%	45.20%
ZN+L2	42.01%	46.06%	45.23%	44.66%	40.86%	47.71%

Table 8: Support Vector Machine model performance results expressed in UAR. Highest score highlighted in bold. aro: arousal, val: valence, like: likeability.

4.4 Personality Trait Experiments

In this section, the ELM models are trained to predict the big five personality traits and the interview invitation variable. For this experiment, the same 660 training and 300 test instances were used as the mood primitives and likeability model experiments. Also, the binarized version of the big five personality trait impression and interview variables were used to train the model. The experiment UAR scores are shown in Table 9, and the highest score for each dimension is highlighted in bold. As the results show, the scores are substantially better than the scores of the mood and likeability dimensions. The best UAR scores are higher than 60% and some are over 70%. Overall, the Extraversion and Openness to Experience dimensions achieve the highest UAR scores in our trained models.

4.5 3-Fold Cross Validation

The results of the experiments in sections 4.3 and 4.4 show that the models can reach an accuracy level of around 55% for the mood and likeability dimensions and upward of 72% for the binary personality trait and interview dimensions. These models were optimized by utilizing weighted models, feature normalization and hyperparamater optimization. During the process of hyperparameter optimization parameters like the C value are optimized to increase performance. However, this means that the model is optimized for the test set, as this set is used to compute the performance score. This means that the real-life performance of the model may not be assessed reliably. To solve this problem the k-fold cross validation method can be applied. In our case, we did a 3-fold cross validation, which means that our original 660 instances of our training set was split up in a new train set of 440

Weighted	Normalization	AGRE	CONS	EXTRA	NEUR	OPEN	INTER
No	NN	60.12%	66.45%	70.90%	62.16%	64.23%	67.94%
	ZN	58.79%	64.55%	69.36%	64.42%	69.43%	68.04%
	L2	60.79%	66.34%	71.88%	**65.12%**	68.51%	68.19%
	ZN+L2	58.33%	63.46%	69.02%	63.78%	70.32%	67.49%
Yes	NN	**62.05%**	64.29%	69.64%	63.97%	70.77%	66.62%
	ZN	58.95%	62.90%	70.00%	64.42%	70.05%	**69.56%**
	L2	60.07%	**66.61%**	**72.28%**	63.92%	**72.58%**	67.16%
	ZN+L2	59.27%	62.06%	69.67%	64.43%	68.87%	68.95%

Table 9: Kernel Extreme Learning Machine results for the personality trait and interview variables (binary). AGRE: agreeableness, CONS: conscientiousness, EXTRA: extroversion, NEUR: neuroticism, OPEN: Openness to Experience, INTER: interview invitation.

samples and test set of 220 samples. The split was done three times to train and test three models per dimension. Then the optimal hyperparameters were chosen based on the highest performance scores of each model on the respective validation set. The prediction output of the test sets were combined to create a data-set of 660 high level features. These instances were also compared with the original labels to compute the UAR, which can be found in Table 10 in the column of "3-Fold CV". "Weighted" refers to if the weighted version of the ELM classifier was used or not. Next, the last 300 video features were deployed on the optimal models that were trained during the 3 Fold Cross Validation. This resulted in 3 times 300 predictions per dimension. Using majority voting, or choosing the medium class in case of a draw, the final prediction set was constructed for each dimension. These predictions were also compared to the original labels to compute the UAR. The scores can be found in Table 10 in the "Test set" column. Both L2 and ZN normalization was applied to the video feature set to train the models in the cross validations.

	3-Fold CV		Test set	
Weighted	Yes	No	Yes	No
INTER	65.75	67.05	66.87	66.75
AGRE	62.67	61.68	59.35	58.95
OPEN	67.64	68.23	70.34	67.65
NEUR	65.71	66.29	60.38	59.36
EXTRA	70.62	70.07	71.02	69.35
CONST	65.20	65.36	59.39	58.56
ARO	47.67	41.38	44.25	37.26
VAL	56.17	52.74	43.38	37.21
LIKE	45.97	43.16	43.95	42.54

Table 10: The UAR (%) scores of the 3 Fold Cross Validation experiment and the predicted test set using ELM.

4.6 Explainability Analysis and Decision Trees

The performance results of the trained models are overall good, as they show UAR scores of around 65%. Now to explain the predictions made by our model we will be training and visualizing a decision tree. However, a decision tree trained with over 16000 video features may theoretically result in a tree with over 2^{16000} leaves in the worst case, given sufficient amount of instances. Therefore, to ensure interpretability, we will choose 5 to 8 high level features to model the interview in-

	Weighted	Unweighted
Personality (Big 5 train impression predictions)	63.86%	63.87%
Personality + valence	62.83%	63.20%
Personality + arousal	62.47%	65.10%
Personality + likeability	61.68%	64.80%
Personality + valence + arousal	62.04%	**65.78%**
Personality + valence + likeability	63.26%	63.87%
Personality + likeability+ arousal	**64.56%**	63.39%
Personality + valence + arousal + likeability	64.50%	63.15%

Table 11: UAR scores of the modeled Decision Trees per dimension combination.

vitation variable. We will use the predictions of the 3 fold cross validation models, as described in section 4.5, as the high level features. Meaning, the openness, conscientiousness, extroversion, agreeableness, neuroticism, valence, arousal and likeability predictions were used as the high level features. Meanwhile, the interview dimension was used as the label.

Table 11 show the UAR scores of the trained decision tree models for each dimension combination. Personality refers to the big five personality traits, while weighted and unweighted refers to the type of prediction set that was used. Note that since the predicted dimension is binary, the chance level is 50%. To increase the explainability and interpretability of the decision tree, the mood and likability dimensions were, like the personality traits, also binarized. The mood and likeability dimensions had few samples for the low class so the decision was to combine them with the medium class.

As the results indicate for the unweighted predictors, adding arousal to the dimensions increases the performance score of the models, except when adding the likeability dimension. For the weighted prediction set, the likeability and arousal combination was the best scorer. The overall top scorer, with a UAR score close to 66%, used the unweighted prediction set of the personality traits, valence and arousal dimensions.

Figure 9 is the visualization of the decision tree that has the highest UAR score. The constructed decision tree can be interpreted as follows: a node will contain a high level feature, agreeableness for example. Then traversing to the left of this node means the feature scored low while traversing to the right branch means the feature scored high. The leaf nodes indicate the predicted class, in our case whether or not a person is invited to a job interview. Neuroticism correlates negatively with the interview variable, therefore it is represented with its opposite; non-neuroticism. In the decision tree, this means that the right branch indicates a high score in non-neuroticism. As the tree shows, openness to experience is the most important feature as it is the top level node.

To further examine feature importance a ridge classifier model was trained using the same high level feature prediction set that was used for the decision tree. Figures 10 and 11 show the distribution of the feature importance as modeled by the ridge classifier. As both distributions show, the agreeableness, openness and conscientiousness dimensions stand out as the most important features. This is comparable to the decision tree in Figure 9, as the three top level nodes

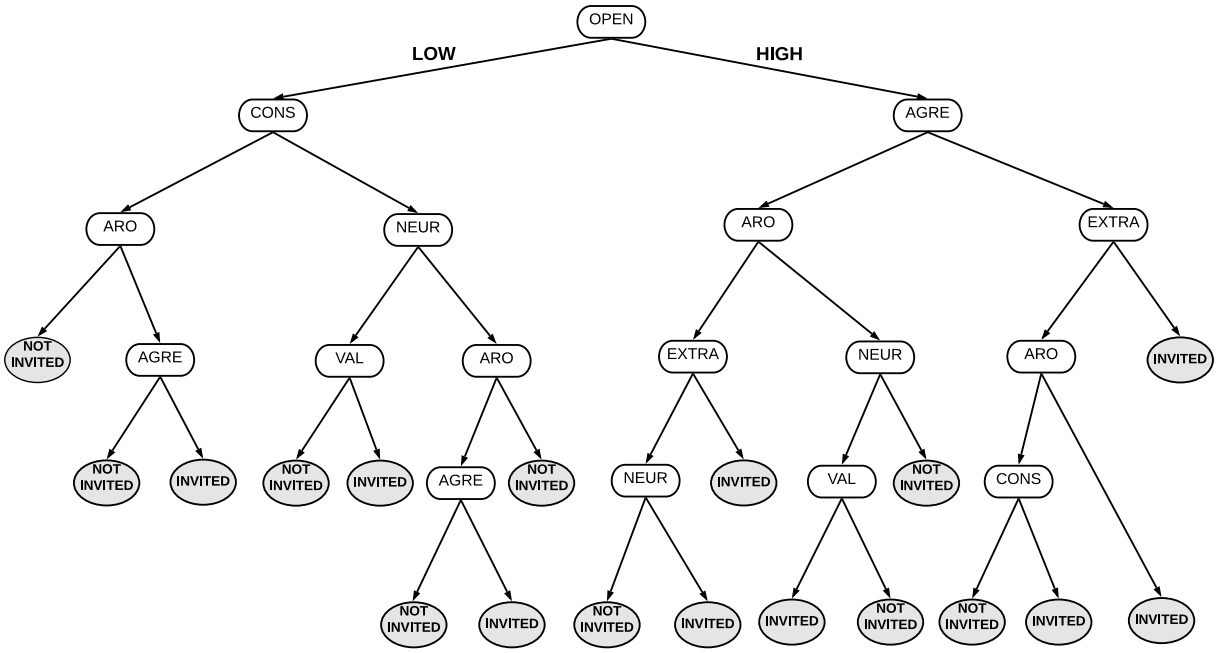

Figure 9: Decision Tree of the personality traits, arousal and valence dimensions, predicting the interview invitation variable. Right branches are always labeled as HIGH, while left branches are always labeled as LOW.

contain the same dimensions, implying their importance.

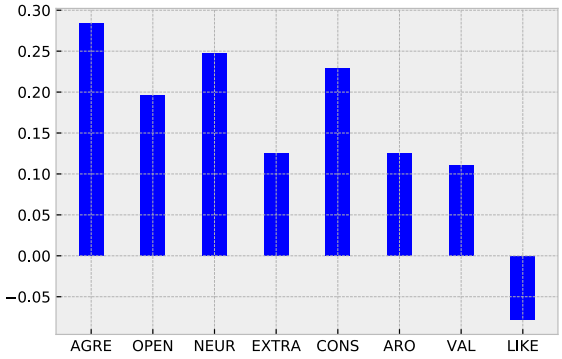

Figure 10: Feature importance distribution using a Ridge Classifier with the weighted prediction set.

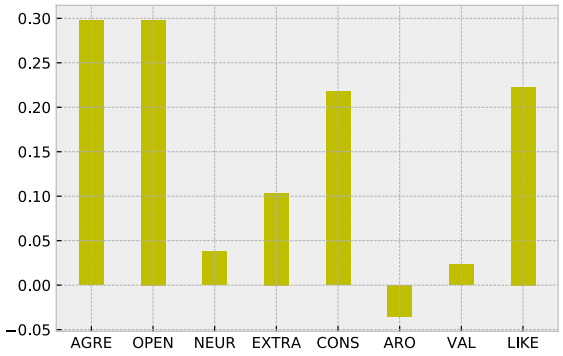

Figure 11: Feature importance distribution using a Ridge Classifier with the unweighted prediction set.

5 Discussion & Future Work

Video based job interviews and resumes are becoming increasingly popular methods for organizations to adopt in their hiring process. Automatic personality trait and mood could vastly accelerate and support this process. As the literature illustrates, personality traits and mood are important factors that influence the decision to invite an applicant to an interview. Creating such automated systems can only be done however, if the trained models are highly accurate and do not show bias. Therefore, the experiments done in this thesis are to compare and improve classification models using extracted video features which can lead to better automated job screening systems.

Extensive experiments have been carried out using our data-set, ranging from classification of dimensions to creating explainable models like decision trees. The Weighted Kernel ELM algorithm has proven to be accurate to model mood and likeability dimensions from a set of video features. With UAR scores of up to 57%, the performance was considerably above the chance level of 33%. Further, the results show ELM models have also proven to be more accurate than the widely used Support Vector Machine algorithm. Also, a moderate correlation has been found between the interview invitation variable and the mood and likeability dimensions. Despite this correlation, the decision tree model and feature importance experiment show that valence and arousal play a minor role in the prediction of the interview invitation variable. The personality traits, especially openness, agreeableness and conscientiousness show to be more important factors in predicting the interview invitation variable. The resulting decision tree could be used to help job recruiter in the job screening process. However, it is strongly recommended that more modalities and dimensions should be explored before such a classification system should be used in the wild. The experiments done in this thesis only utilized video features from the video clips in the data-set. However, video resumes obviously also include acoustics and linguistics, which should be used in combination of video features to create more accurate models.

A great deal of future research can be done with the data-set that was used for this thesis. For example, the data-set is also annotated for ethnicity, age groups, gender and background music. These variables are annotated using a categorical scale. Therefore, the data-set could be split into different categories regarding these dimensions, which could result in interesting new decision trees and model results. It could also expose biases towards the different categories of these dimensions from the annotators, or expose a bias a job recruiter might have. Also, in this thesis, only the video features are used to train the models. Furthermore, for the modeling stage only 960 samples were used of the available 10,000 video clips. If a larger sample size, if not all video clips, would also be annotated for the

mood and likeability dimensions, the performance of the classifications model could be improved. This could also lead to better explainability models. The bias of the annotators regarding ethnicity, age groups and gender was also not explored. Further analysis of this data could expose biases and therefore also illustrate biases in the classification models.

References

Alam, Firoj, Evgeny A Stepanov, and Giuseppe Riccardi (2013). "Personality traits recognition on social network-facebook". In: *Seventh International AAAI Conference on Weblogs and Social Media*.

Allbeck, Jan M and Norman I Badler (2008). "Creating crowd variation with the ocean personality model". In:

Bizer, Christian et al. (2005). "The impact of semantic web technologies on job recruitment processes". In: *Wirtschaftsinformatik 2005*. Springer, pp. 1367–1381.

Bower G. H., Forgas J. P (2000). "Affect, memory, and social cognition". In: *Cognition and emotion*. New York: Oxford University Press. Chap. 3, pp. 87–168.

Celiktutan, Oya and Hatice Gunes (2015). "Automatic prediction of impressions in time and across varying context: Personality, attractiveness and likeability". In: *IEEE Transactions on Affective Computing* 8.1, pp. 29–42.

Chen, Baiyu et al. (2016). "Overcoming calibration problems in pattern labeling with pairwise ratings: application to personality traits". In: *European Conference on Computer Vision*. Springer, pp. 419–432.

Cohen, Jacob (1968). "Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit." In: *Psychological bulletin* 70.4, p. 213.

Costa, Paul T and Robert R McCrae (1992). *Neo personality inventory-revised (NEO PI-R)*. Psychological Assessment Resources Odessa, FL.

Davenport, Thomas H and Rajeev Ronanki (2018). "Artificial intelligence for the real world". In: *Harvard business review* 96.1, pp. 108–116.

Desmet, Pieter MA, Martijn H Vastenburg, and Natalia Romero (2016). "Mood measurement with Pick-A-Mood: review of current methods and design of a pictorial self-report scale". In: *Journal of Design Research* 14.3, pp. 241–279.

Dettmar, Kevin JH (2004). "What we waste when faculty hiring goes wrong". In: *Chronicle of Higher Education* 51.17.

Ekman, Paul (1992). "An argument for basic emotions". In: *Cognition & emotion* 6.3-4, pp. 169–200.

Escalante, Hugo Jair, Heysem Kaya, Albert Ali Salah, Sergio Escalera, Yagmur Gucluturk, et al. (2018). "Explaining first impressions: modeling, recognizing, and explaining apparent personality from videos". In: *arXiv preprint arXiv:1802.00745*.

Escalante, Hugo Jair, Heysem Kaya, Albert Ali Salah, Sergio Escalera, Yağmur Güçlütürk, et al. (2020). "Modeling, Recognizing, and Explaining Apparent Personality from Videos". In: *IEEE Transactions on Affective Computing*.

Faberman, Jason, Marianna Kudlyak, et al. (2016). "What does online job search tell us about the labor market?" In: *FRB Chicago Economic Perspectives* 40.1.

Gievska, Sonja and Kiril Koroveshovski (2014). "The impact of affective verbal content on predicting personality impressions in youtube videos". In: *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition*, pp. 19–22.

Goodfellow, Ian J et al. (2015). "Challenges in representation learning: A report on three machine learning contests". In: *Neural Networks* 64, pp. 59–63.

Graziano, William G and Nancy Eisenberg (1997). "Agreeableness: A dimension of personality". In: *Handbook of personality psychology*. Elsevier, pp. 795–824.

Hayes, Theodore L and Therese Hoff Macan (1997). "Comparison of the factors influencing interviewer hiring decisions for applicants with and those without disabilities". In: *Journal of Business and Psychology* 11.3, pp. 357–371.

Huang, Guang-Bin, Qin-Yu Zhu, and Chee-Kheong Siew (2004). "Extreme learning machine: a new learning scheme of feedforward neural networks". In: *2004 IEEE international joint conference on neural networks (IEEE Cat. No. 04CH37541)*. Vol. 2. IEEE, pp. 985–990.

— (2006). "Extreme learning machine: theory and applications". In: *Neurocomputing* 70.1-3, pp. 489–501.

Junior, JCSJ et al. (2018). "First impressions: A survey on computer vision-based apparent personality trait analysis". In: *arXiv preprint arXiv:1804.08046*.

Kaya, Heysem, Furkan Gurpinar, and Albert Ali Salah (2017). "Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video cvs". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–9.

Kaya, Heysem, Furkan Gürpınar, and Albert Ali Salah (2017). "Video-based emotion recognition in the wild using deep transfer learning and score fusion". In: *Image and Vision Computing* 65, pp. 66–75.

Kaya, Heysem and Albert Ali Salah (2018). "Multimodal personality trait analysis for explainable modeling of job interview decisions". In: *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer, pp. 255–275.

Kinsman, Michael (2005). "Hiring the Right Person; Interviews and references aren't enough". In: *Network Journal* 13.3, p. 33.

Kraus, Richard G and Joseph E Curtis (1990). *Creative management in recreation, parks, and leisure services*. St. Louis; Toronto: Times Mirror/Mosby College Pub.

Madzlan, N et al. (2014). "Towards automatic recognition of attitudes: Prosodic analysis of video blogs". In: *Speech Prosody, Dublin, Ireland*, pp. 91–94.

Matlin, M.W. (2012). *Cognition, 8th Edition*. Wiley Global Education. ISBN: 9781118476925. URL: `https://books.google.nl/books?id=XOfGngEACAAJ`.

McCrae, Robert R (1993). "Openness to experience as a basic dimension of personality". In: *Imagination, cognition and personality* 13.1, pp. 39–55.

Morgeson, Frederick P et al. (2008). "Review of research on age discrimination in the employment interview". In: *Journal of Business and Psychology* 22.3, pp. 223–232.

Nguyen, Laurent Son and Daniel Gatica-Perez (2016). "Hirability in the wild: Analysis of online conversational video resumes". In: *IEEE Transactions on Multimedia* 18.7, pp. 1422–1437.

Parkhi, Omkar M, Andrea Vedaldi, and Andrew Zisserman (2015). "Deep face recognition". In:

Qin, Rizhen et al. (2016). "Modern physiognomy: an investigation on predicting personality traits and intelligence from the human face". In: *arXiv preprint arXiv:1604.07499*.

Quercia, Daniele et al. (2011). "Our twitter profiles, our selves: Predicting personality with twitter". In: *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*. IEEE, pp. 180–185.

Raza, Susan M and Bruce N Carpenter (1987). "A model of hiring decisions in real employment interviews." In: *Journal of applied psychology* 72.4, p. 596.

Russell, James A (1980). "A circumplex model of affect." In: *Journal of personality and social psychology* 39.6, p. 1161.

Tao, Jianhua and Tieniu Tan (2005). "Affective computing: A review". In: *International Conference on Affective computing and intelligent interaction*. Springer, pp. 981–995.

Todorov, Alexander, Manish Pakrashi, and Nikolaas N Oosterhof (2009). "Evaluating faces on trustworthiness after minimal time exposure". In: *Social Cognition* 27.6, pp. 813–833.

Toldi, Nicole L. (2011). "Job applicants favor video interviewing in the candidate-selection process". In: *Employment Relations Today* 38.3, pp. 19–27. DOI: 10.1002/ert.20351. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ert.20351. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ert.20351.

Torres, Edwin N and Cynthia Mejia (2017). "Asynchronous video interviews in the hospitality industry: Considerations for virtual employee selection". In: *International Journal of Hospitality Management* 61, pp. 4–13.

Truxillo, Donald M et al. (2012). "Perceptions of older versus younger workers in terms of big five facets, proactive personality, cognitive ability, and job performance". In: *Journal of Applied Social Psychology* 42.11, pp. 2607–2639.

Valente, Fabio, Samuel Kim, and Petr Motlicek (2012). "Annotation and recognition of personality traits in spoken conversations from the ami meetings corpus". In: *Thirteenth annual conference of the international speech communication association*.

Van der Linden, Dimitri et al. (2010). "Classroom ratings of likeability and popularity are related to the Big Five and the general factor of personality". In: *Journal of Research in Personality* 44.5, pp. 669–672.

Weber, Lauren and RE Silverman (2012). "Your resume vs. oblivion". In: *The Wall Street Journal* 24.

Weekley, Jeff A and Joseph A Gier (1987). "Reliability and validity of the situational interview for a sales position." In: *Journal of Applied Psychology* 72.3, p. 484.

Widiger, Thomas A (2017). *The Oxford handbook of the five factor model*. Oxford University Press.

Willis, Janine and Alexander Todorov (2006). "First impressions: Making up your mind after a 100-ms exposure to a face". In: *Psychological science* 17.7, pp. 592–598.

Xiong, Xuehan and Fernando De la Torre (2013). "Supervised descent method and its applications to face alignment". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 532–539.

Yu, Jianguo (2019). "Multimodal Information Fusion based on Deep Learning". PhD thesis. The University OF Aizu.