Layered Regression Analysis on Multimodal Approach for Personality and Job Candidacy Prediction and Explanation

Achmadnoer Sukma Wicaksana

Layered Regression Analysis on Multimodal Approach for Personality and Job Candidacy Prediction and Explanation

by

Achmadnoer Sukma Wicaksana

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday August 24, 2017 at 13:00.

Student number: 4517253
Project duration: January 20, 2017 – August 24, 2017
Thesis committee:

Dr. C. Liem MMus,	TU Delft, supervisor
Prof. dr. A. Hanjalic,	TU Delft
Dr. D. M. J. Tax,	TU Delft
dr. A. M. F. Hiemstra,	Erasmus University Rotterdam

TU Delft

Abstract

Video blogs (vlogs) are a popular media form for people to present themselves. In case a vlogger would be a job candidate, vlog content can be useful for automatically assessing the candidate's traits, as well as potential interviewability. Using a dataset from the CVPR ChaLearn competition, we build a model predicting Big Five personality trait scores and invite to Interview score of vloggers, explicitly targeting explainability of the system output to humans without the technical background. We use human-explainable features as input, and linear models for the systems building blocks of our layered architecture to ensure a transparent system. This multimodal layered architecture model is an enhancement to our initial submission model to the ChaLearn competition. Six multimodal feature representations are constructed to capture facial expression, movement, speaking pattern, and linguistic usage. Each of these representations is treated individually before the late fusion technique to combine each prediction. For each, correlation analysis is done to get the relation between input features and the predicted traits by considering the significance level of Pearson's correlation coefficient. This way, we split each representation into two feature set; full feature set and subset of a high significance level of features. Three regression techniques are fitted to these two datasets per representation to get the best possible model for each. Then, the six predictions are combined on the second layer of regression to ensure the fair weighting. Our layered regression architecture ensures us to have the best possible model for each representation to make a better overall accuracy. As a result, our enhanced model outperform our initial ChaLearn competition submission model and other systems in the competition. Despite our simple linear model that has lower accuracy than the more complex model on the same competition, we have a strength of the more interpretable model and report description.

Preface

My journey of two years has come to this report. This master thesis with the title *Layered Regression Analysis on Multimodal Approach for Personality and Job Candidacy Prediction and Explanation* is a study on predicting human personalities and invite-to-interview likelihood based on self-video representation. This project was conducted as part of graduation requirements of Computer Science - Data Science & Technology program at the Delft University of Technology.

The project itself was originally started as my request to my supervisor, **Cynthia Liem**, to study multimodal approach in machine learning situation. The goal was to deepen my knowledge of data science from the multimedia data mining perspective and not only focused on the analysis part. I am grateful to have her as my supervisor who was very supportive throughout the process, knowing my situation. I also had the pleasure to learn and work alongside her to compete on the CVPR2017 Looking at People Challenge and produce our paper, my *very first* paper submission, to the Explainable Computer Vision Workshop and Job Candidate Screening Coopetition 2017 with hopefully another joint publication by the end of the year.

I also would like to thank my scholarship committee, **Lembaga Pengelolaan Dana Pendidikan (LPDP) Indonesia**, which has been generous to fund me on this academic journey for two years. I would not have been here in the first place without you wonderful people. As a token of gratitude, I aimed high and had been working hard to prove that you invest in a right person and show what I am capable of by obtaining perfect GPA and by producing this master thesis.

Last but not least, I would like to thank my friends and family that fill my world these past years. Especially, my partner in life, **Naviera Yuliasari**, who has been here through our struggles from the day we met until we finish our journey together in the Netherlands. The center of my universe, **Aksara Sabeel Wicaksana**, who color my world with every laughter and cry you made. Also, to my mother, **Zubaida Gani**, who spend most of the first half-year in the Netherlands to help me focus on finishing this thesis.

I hope you enjoy your reading.

Achmadnoer Sukma Wicaksana
Delft, August 2017

Contents

List of Figures

List of Tables

1 Introduction

The digital age and ever growing technology keep on improving every aspect of our lives. From gaming industry to mobile industry, the advances in technology change them in a better way. Long gone are the days we have to communicate to distant relatives using mail. All of it can be done easily from the palm of our hands right now. This is also the case for expressing our mind to the broader audience. The newspaper was one of the earlier popular media to broadcast our thoughts. We can write our opinion and submit them through a proper channel to be published and distributed to the readers. Despite the advantage of wide distribution channel in early days, there is a time restriction compared to recent years' solutions, such as *blogs* [38].

With the rise of the Internet, people began to host their blog to update activities, share and exchange opinions, and also express emotion [40]. A broader audience, time advantage, and also the chance to speak on author's point of view freely made blogging a huge hit [41]. Rather than mere text and small images, in the case of the newspaper, a blog offers ample media choices to work with. Blogging has broadened the marketplace of ideas by allowing more people's voices to enter the communication. Moreover, the development in technology and wider Internet adoption helped the birth of multimedia sharing platforms. Few of these are, for example, SoundCloud for people to listen and upload their covers of songs, Instagram to share images and short videos, and also YouTube for video sharing. The latter enables people to be more expressive and creative in delivering their content to attract other one billion users on that website to view their videos. This user-to-user social experience is a distinguishing factor from the traditional broadcasting which leads to the success of YouTube [58].

Each day, billions of hours of video are watched on YouTube and for each minute that passes by, three hundred hours of new videos are uploaded.[1] These vary in content, from educational videos to a review of the latest technology products. Included in this huge collection of content are video blogs (*vlogs*) that people use to present themselves and share anything to the world. The vlog is one of the most popular video formats on YouTube, even considered as the epitome of the YouTube social phenomenon [53]. Mainly, the characteristics of these consider the *vlogger*'s interest and daily life in an unscripted way. Since most of the vloggers show themselves in the videos, people can judge the quality of the content as well as their personalities by watching. This personality judgment is an interesting thing that happens since the viewers mostly do not know the vlogger directly in real life, but have their own opinion of them by only watching a short duration video.

1.1. Personalities Assessment

Aristotle once said that human beings are naturally a "social animal", which means humans need others in order to live their lives. A human is wired to connect to others because of basic needs to survive. We need to initiate and maintain relationships to live in the social environment. Moreover, the degree of comfort in a relationship is also a distinguishing factor on the longevity of it, such as in mentoring settings [55]. Interestingly, personality can affect the nature of forming such a bond and its quality, but not vice versa [4]. Thus, personality is one of the aspects that people observe in engaging contact with others, and we are trained to process information to form our understanding of a person's character.

[1] https://www.youtube.com/yt/press/statistics.html

People judge at first sight, and that is why you never get a second chance to make a great first impression. Whether the judgment turns out false or not, it shows the importance of first impressions on shaping the images to be perceived by others. Interestingly, a recent study shows that it only needs a tenth of a second for people to correctly judge other personalities based on facial appearance [62]. Within this short period, people can translate visual cues from the subjects to assess their personalities. Moreover, in that research, the addition of time to judge the personality does not increase the accuracy of the judgment but only improve the confidence level of the judgment itself.

People develop personalities over time forming their way of thoughts, behaviors, and emotions. Also, personality traits are usually relatively stable throughout a person's lifetime, making it interesting for traits researcher to study how they affect a person's life, instead of using transient personality states [16, 20]. One of the notable models to define human personality trait is the Big Five Personality.

The Big Five Personality model is the prominent paradigm to define human personality that is a result of a long development for many years. It also is widely used in many different cultures, thus making it represent a global model of personality [19]. The name itself does not imply the greatness that the model offer but rather the broad level of abstractions that each personality dimension represent [32]. The five dimensions of the model are usually agreed to be described with the *OCEAN* acronym, namely Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism [21, 32]. The description of each of the five dimensions is as follows:

- Extraversion: People who are high in Extraversion enjoy being center of attention and meeting new people. While people who are low in Extraversion tend to be more reserved and carefully think things through before they speak.
- Agreeableness: People who are high in Agreeableness tend to feel empathy and concern for other people. While those low in this trait tend to not care about how other people feel
- Conscientiousness: People who are high in Conscientiousness tend to pay attention to details. While people who are low in this trait usually procrastinate important tasks and do not like schedules.
- Neuroticism: People who are high in this trait tend to feel anxious and worry about many different things. Those low in this trait are usually emotionally stable and very relaxed.
- Openness: People who are high in this trait are often very creative and open to new things. People low in this trait are usually dislike changes and not very imaginative.

Since the Big Five Personality can be used to define an individual's overall personality, human resources (HR) professionals often use it to help with the organizational process as it is one of the requirements of a job [30]. According to Cook at [19], personality is one among seven main aspects that are assessed in personnel selection besides mental ability, physical characteristics, interests and value, knowledge, work skills, and social skills. Moreover, usually, these are used as the first filter to select a group of applicants for further steps such as an interview. This process is called a *sifting* procedure. The idea is to prioritize the efforts of HR practitioners to focus on candidates that match with the needs of the organization and the job responsibilities. However, application sifting often takes up a lot of time for HR practitioners to do, especially with a lot of open vacancies. Furthermore, research shows that sometimes sifting is not done effectively and also the self-report information from applicants is inaccurate. This behavior of faking information from applicants can undermine the validity of personality assessment [29]. Therefore, a system for speeding up the process while providing good accuracy will be cherished dearly.

1.2. Video Resume

In parallel, the popularity of video-based content, combined with fast technology development, has also given rise to the video resume as a new type of job screening mechanism. A video resume is a short-duration video that an applicant sends as a replacement or complement to the usual text-based resume. Usually, it is a monologue in which the applicant talks about their traits and past experiences [60]. In fact, some companies have embraced the video resume, asking the candidates to submit video resume as one of the requirements in the hiring process to secure the jobs, for example, the National Institute for the Deaf at the Rochester Institute of Technology [34].

The video resume offers advantages over a mere paper-based resume in getting to know an applicant and showcasing an applicant's quality to get a competitive edge. As a recruiter, you can meet the real person behind the paper-based resume and highlight the applicant's potential that sometimes misses in the paper-based resume. You can obtain information about their personalities, qualifications, and creativity by looking at how they present themselves. On the side of candidates, they have access to greater media richness to show their capabilities that are rather hard to show on paper alone. A video resume can also act as a stand out tool to the employer by showing the uniqueness aspect of the applicants to catch the recruiter's attention. In comparison to the paper-based case, the applicants are aware that their resume is being thoroughly inspected by HR practitioners and tend to provide inaccurate information in their writing [19]. In the form of video, this misleading information can be suppressed by observing directly rather than trusting the self-reported information, for example in the case of communication and presentation skills. Therefore, the video resume might be a good method to assess main aspects that we mention earlier, such as personalities.

On the other hand, looking through all video resume might be a time-consuming activity. Especially, if there are many open jobs with many candidates at the same time to be reviewed. Not to mention legal actions that might occur from unsatisfied candidates who think they were rejected based on discrimination from the video resume. They can claim that they were discounted from the job list from their appearances in the video. Race, gender, and disability are among many that can be the source of bias and subjectivity of the recruiter, whether it is direct or indirect discrimination (*adverse impact*) [19, 28].

1.3. Multimodal Analysis

One of the solutions to both time and bias problems on sifting is by using application (resume) scanning software. This is much faster than conventional sifting and also eliminate bias by ignoring these discriminatory factors from the system [19]. While a human might get tired or careless on doing sifting, the machine counterpart can do it continuously and meticulously. Resumix is one example that was mentioned in the said book [19], but no details on how to generate the results were described because it is a copyrighted software. Also, it is not clear whether the software can generate an output based on video resume data. One can follow up this automation initiative by building their predictive system from video resume data.

Rather than text presented in a traditional resume, the video format offers richer data to analyze. It has a visual aspect which is the video itself and audio aspect which is the sound of the person. Also, if one wants to obtain a text representation of the video, they can get the content of the speech on the video. It is often useful to consider these cues from each modality to build a system for analyzing personality and the likelihood to be invited to further steps of hiring decision [60]. The enthusiasm of applicants on the video and facial expression can be used to infer Extraversion, while Openness may be apparent on verbal eloquence and intellectualism. Agreeableness may be reflected on how friendly the person sounds and smile on the face. Dysfluency of speech and emotions apparent on the video may be related to the Neuroticism. Conscientiousness may be inferred from professional manner on speech.

In efforts such as [9, 11], a large amount of audiovisual vlog content has been shown to be useful for modeling and prediction personality traits of the person. Work at [7] used a large collection of vlog content for their layered regression model, using Support Vector Regression and feeding the output as input to Gaussian Process Regression, to infer interview score from personality scores. Although not necessarily identical, vlog and video resumes have a similar form of one-way communication; the person speaks to the camera, and self-presentation will be an important motivation behind video production and sharing. As far as we know, there is no available dataset yet for personality and hirability computation using the video resumes setting. Thus, the similarity between these two video settings can be utilized as the first step to gather some insight on predicting personalities using video analysis.

1.4. Explainability

Besides the advancement of the algorithm itself, the predictive power of machine learning techniques also depends on the input to the system. However, the input for such model also depends on how developers decide to shape it and sometimes it is not independent from subjective aspects of the human. Even worse, the behavior of embedding human prejudice to the system can make the biases become objectively justified by the model [3]. The concern of accidentally including bias in machine learning

techniques is getting more and more recognition these past years. An initiative such as fatml[2] (fairness, accountability, and transparency in machine learning) raise awareness on ensuring non-discrimination and understandability in machine learning techniques. As the complexity of machine learning grows, the ability (or inability) to describe the automated decision becomes a matter of complexity itself. In sensitive areas, such as hiring process, the importance of this becomes more eminent, especially when the fairness itself is regulated by laws.

Unlike most of the machine learning problems that aim for only optimizing system accuracy, the problem of automatically assessing personality traits—and especially hirability of a potential job candidate—from audiovisual content needs to consider another aspect: explainability [51]. This type of work considers assessments of humans, and typical human decision-makers for this task do not have a technical computer science background. Moreover, as mentioned earlier, there is a possibility of legal actions from unsatisfied candidates in the job hiring process alleging unfair practice from employers. In the United States of America, the employer must abide by anti-discrimination federal laws at each stage of the hiring process, and the applicants have legal rights for it, even before becoming employees. Applicants may file formal complaints of the unfair hiring practices and file a charge to the U.S. Equal Employment Opportunity Commission (EEOC) for investigation. If the employer failed to do so and is proven by the court, they will object to some of these laws: Title VII of the Civil Rights Act of 1964, the Age Discrimination in Employment Act of 1967, and the Americans with Disabilities Act of 1990 [43]. The equivalent for EEOC on the other countries would play the same role; such as HALDE (Haute Autorité de Lutte contre les Discriminations) for France and EHRC (Equality and Human Rights Commission) for the United Kingdom.

Considering these factors, it is critical for the system to not only focus on numbers but to also understand both measurements and the decisions made in a model for further reasoning. Having a transparent system that has power on explaining decision will be a huge factor in this scenario. Based on the result, it could be seen what metrics should be considered to produce the output and how they correlate to the decision making. Also, by referring to the decision made by the automated system, it will be a more solid foundation rather than solely relying on the decision of a single or couple of employers that might have tendencies toward a particular group of people.

1.5. Problem Statement

In the job candidacy setting, personalities become one of the main aspects to be observed by HR practitioners for sifting. However, the practicality aspect of sifting manually usually takes a lot of time, especially when many job positions are available. The last paragraph of Section 1.1 mentions this problem and the needs to speed up process while maintaining accuracy. Also, the popularity of the video resume might decrease the odds of misleading information on the resume, because recruiters can observe the applicants instead of looking through text. On the other hand, it also comes with the increased time spent on sifting and the introduction of judgmental bias from human prejudice. The two sides of the coin when introducing video resume to the job candidacy process is described in Section 1.2. In order to overcome this situation, we can build an automated system that predicts personalities based on objective input from a collection of video data.

Time and subjectivity issues might not be the only requirements for the automated system. As described in Section 1.4, there are commissions and laws that regulate hiring processes. Should there be a suspicion or dispute, people can report to the commissions, and there might be legal issues afterwards. Thus, there is a need of the power of explanation from the model to both recruiters and applicants to understand the reasoning behind the decision. The use of sophisticated and complex model might produce a great accuracy for the system. However, in the job candidacy settings, it might not be the best choice if the model can not describe its decision pipeline.

1.6. Research Objective

The previous section outlines the points that are needed to build an automated system for sifting applicants in job candidacy setting. In order to do that, we formulate our importance of this research by defining our research objectives. Our research aims to build a predictive model for personalities and invitation to interview that has the ability to explain the decision making and is easily understandable

[2] http://www.fatml.org/

by a human.. Based on the research objective presented above, we can formulate research questions for our experiments as follows:

1. How can we emulate personalities and invite-to-interview judgment of people from video data?
2. How can we build a system that predict personalities and interview scores decently and have the power of explainability?

1.7. Scientific Contribution

As an early work of my thesis, I and my supervisor, Cynthia Liem, participated in the qualitative phase of the ChaLearn Looking at People Competition 2017[3]. It aims to help both recruiters and job candidates by using automatic recommendations based on multimedia resumes using a large collection of video and traits data. As the result, we were awarded as the winner for the qualitative phase and published our work [61] at the corresponding conference workshop. In this thesis, we will continue to develop the submission model to increase the performance of the said model. Also, the increased interest in the video resumes led to interest of Erasmus University's Organisational Psychology researchers to collaborate with TU Delft Multimedia Computing Group, and this thesis presents the first contributions of these joint efforts.

1.8. Outline

The rest of the thesis will be structured as follows. Chapter 2 will discuss about related works that are studied for this project. The following Chapter 3 and Chapter 4 will elaborate on the dataset that we used for the experiment and the methodology on how we formulate the features and system overview. In Chapter 5, we will outline the results of our submission to the ChaLearn competition and the discussion of its results. Chapter 6 will discuss the further enhancement of our submission model to improve system accuracy. Last, the conclusion and remarks for future research are stated on Chapter 7.

[3] http://chalearnlap.cvc.uab.es/challenge/23/description/

2

Related Works

This chapter surveys previous works on personality assessment, both from the psychology and computer science domains. Also, we talk about work on measuring job candidacy assessment from collection of video data, and in particular other submission to the ChaLearn competition.

2.1. Personality Assessment

Personality traits prediction has been a long-time research endeavor in the domain of psychology. Mostly, they based their researches on the original Big Five Personality model from [21] to study personalities. One of the prominent self-reporting techniques to infer personality is the International Personality Item Pool (IPIP) [25]. The questionnaire items are behavioral statements with five possible degrees of agreement to the statement. Rather than using text, work at [57] used gamification by offering an interactive and engaging way to fill in the questionnaire in the form of image-based personality assessment. Instead of using questionnaire, a method developed at [44] by using facial appearances of two different photograph settings to infer personalities by other observers.

The apparent results from psychology research and widely available user generated data caught the attention of machine learning and computer vision researchers to develop techniques for the same purpose. Work at [50] built a model using linguistic features from a huge collection of Facebook messages from 75,000 volunteers that had taken a personality test beforehand. The same model was used by Liu *et al*. to analyze personalities based on tweet data and profile picture of Twitter [37]. Work at [6] used an online questionnaire to obtain personality traits and also short self-presentation videos from the same volunteers to obtain audiovisual features to analyze that personality. Besides self-report information such as questionnaire to measure personality, there were also other efforts on obtaining personality based on first impressions. Biel *et al*. gather a collection of self-talk video from YouTube then annotate the personality using crowdsourcing at Amazon Mechanical Turk. This data then was used to analyze the relation between non-verbal behavior [8, 9], verbal content [11], and facial expression [10] to personality impressions.

2.2. Job Candidacy Assessment

As mentioned in the previous chapter, job candidacy assessment in the psychology domain has been a long time practice. The seven main aspects (personality, mental ability, physical characteristics, interests and value, knowledge, work skills, and social skills) described in the last paragraph of Section 1.1 are among many efforts to find the best practices. While analyzing job suitability has been a long time research topic in organizational psychology, it is not the case for the computer scientist counterpart. The work at [45] tried to predict personalities and hirability from a video resumes dataset, and interest in this task also led to several ChaLearn 'Looking at People' benchmark challenges [48]. To the best of our knowledge, this work was the first of few that infer hirability from video collection data.

Kaya *et al*. [33], from the same ChaLearn competition we enter, use feature level fusion and Extreme Learning Machine with linear kernel on facial, scene, and audio features as the input to predict interviewability from a collection of video data. Then, they use decision tree to explain their result of the fusion using random forest. Work at [7] used image statistical features from facial videos, to infer

five pesonality traits scores using non-linear Support Vector Regressor. These five prediction act as input for Gaussian Process Regression with non-linear kernel to predict Interview score, which would make it even harder to interpret the model. Multilayer Perceptron Neural Network was used at [26] to predict five personality and interview score from video, audio, and text features. They try to find five key frames for each video by clustering the facial pose, and it would be hard to explain the decision based on these key frames.

3

Dataset

This chapter elaborates about the dataset that was used for the experiment. The dataset is the benchmark dataset from ChaLearn competition that we used to build and test our system. The data mining step up to the ground-truth gathering process are described further in this chapter.

The dataset used for our experiments is a collection of ten thousand selected videos, publicly available as the dataset for ChaLearn ECVW (Explainable Computer Vision Workshop and Job Candidate Screening Competition) 2017.[1] This dataset was an enhancement of the previous year ECCVW (European Conference on Computer Vision Workshop and Challenge) 2016 [48], with the addition of speech transcription data and new metric of interview annotation to complement personality annotations. Part of our initial works was submitted to enter the qualitative phase of ECVW 2017 and we also used this dataset for further improvements of the system after the challenge ended. The details about the dataset will be described further in this section.

3.1. Collection and Processing

In order to gather the data, the organizer collected HD 720p YouTube videos of people facing and speaking in English to a camera. Initially, 13,951 videos were collected with several keywords from various channels, with the limitation of 3 videos per channel to maintain the uniqueness aspect of the data. The dataset also consists of other properties such as various gender, age, nationality, and ethnicity that appear in the collection to be a good sample of the population. Videos that did not meet the requirements (too short or non-English speaker) were discarded, resulting in total of 8,581 videos.

From these videos, 32,139 clips of 15 seconds long were automatically generated by searching video segments that have one and only one face with at least one visible eye in 75% of the frames. Face and eye detection were done by using Viola-Jones [56] from OpenCV. In order to maintain robustness, only total six clips were allowed to be generated from each video. Furthermore, a second step of filtering was done to these clips with these criteria:

- One unique person as foreground at a safe distance from the camera.
- Good quality of audio and images.
- Only English speaking.
- People above 13-15 years old. Non-identified babies appearing with the parents might be allowed.
- Not too much camera movement (changing background allowed, but avoid foreground constantly blurred).
- No adult or violent contents (except people casually talking about sex or answering Q&A in an acceptable manner). Discard any libelous, doubtful or problematic contents.
- No nude (except if only parts above shoulders and neck are visible).

[1] http://chalearnlap.cvc.uab.es/challenge/23/description/

- Might have people in the background (crowd, audience, without talking, with low resolution of faces to avoid any confusion with the speaker).
- No advertisement (visual or audio information about products or company names).
- Avoid visual or audio cuts (abrupt changes).

As a result, 10,000 final clips were generated from more than three thousand videos with 3.27 clips per video in average. Then, these clips were split into three sets; 6,000 videos for the training set, 2,000 for the development set, and 2,000 for test set. The total length of these clips was 41.6 hours with approximately 4.5 million frames. In addition, as mentioned earlier, this year version of the dataset has an addition of a speech transcript to complement the existing data. Each clip was transcribed by a professional transcription service which generated 435,984 words, 14,535 unique words, with 43 words per clip on average. The simplified version of the data collection and processing can be seen in Figure 3.1, while the complete data statistics are shown in Table 3.1 [48].

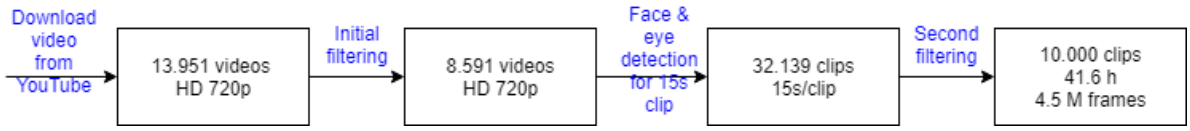


Figure 3.1: Pipeline for data processing

Table 3.1: Benchmark dataset statistics

preparation	Downloaded videos	13,951* (HD 720p @ 30 FPS)
	Remaining videos (supervised from *)	8,581**
	Sampled videos per channel	3 (at most)
	Sampled clips per video	6 (at most)
	Clip length	15 seconds
	Candidate clips (sampled from **)	32,139†
final data set	**Final set of clips** (supervised from †)	**10,000‡**
	Total duration of clips	41.6 hours (4.5M frames)
	Unique channels (originating ‡)	2,764; {1 : 2,584, 2 : 161, 3 : 19}
	Unique videos (originating ‡)	3,060; {1 : 721, 2 : 533, 3 : 464, 4 : 398, 5 : 435, 6 : 509}
	Mean no. clips per video	3.27
	Duration of originating videos	608.7 hours
	Total no. views of originating videos	More than 115M; {0-100 : 27.64%, 100-1K : 34.15%, 1K-10K : 22.68%, 10K-100K : 11.44%,>100K : 4.08%}
	Originating videos' avg. rating	4.6/5.0; {1 : 8, 2 : 11, 3 : 43, 4 : 1340, 5 : 1,395}
	Originating videos' keywords (top 20)	'Q&A', 'q&a', 'vlog', 'questions', 'makeup', 'beauty', 'answers', 'funny', 'Video Blog (Website Category)', 'question and answer', 'answer', 'question', 'fashion', 'Vlog', 'Questions', 'vlogger', 'how to', 'tutorial', 'q and a', 'Answers'
	Total words	435,984
	Total unique words	14,535
	Mean no. words per clip	43

3.2. Ground-truth Estimation

A huge amount of clips gathered from the previous process step created challenges on how to label properly and rapidly. Amazon Mechanical Turk (AMT) was chosen as the crowdsourcing platform as it gained popularity in the computer vision field [36] and has the ability for generating a massive amount of annotations in little time. In order to reduce variance, multiple votes can be cast upon each video. On the other hand, bias can be reduced by introducing pairwise comparison as shown at the costume

interface in Figure 3.2. The small-world algorithm [59] was used to ensure good overall coverage of video pairs to be evaluated by AMT workers, as it provides high connectivity, avoids disconnected regions in the graph, has well-distributed edges, and a minimum distance between nodes [31]. In order to convert pairwise scores to cardinal scores, the Bradley-Terry-Luce (BTL) model [14] is fitted by the use of Maximum Likelihood estimation. The detailed explanations on the procedure on this dataset are described in [17]. As a result, 321,684 pairs were obtained to label 10,000 videos. The type of questions that AMT workers should answer in order to collectively produce these five personalities and interview scores are shown in Figure 3.2.

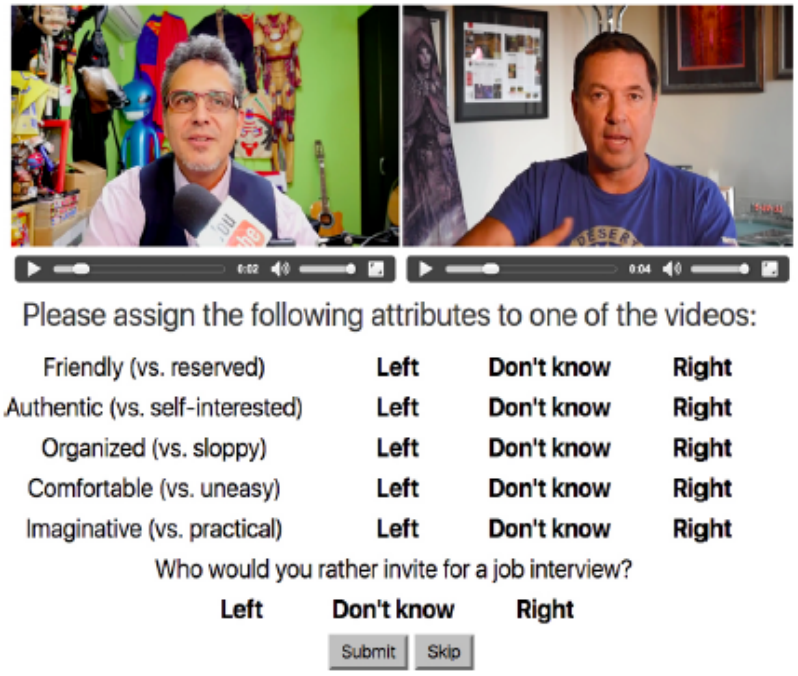
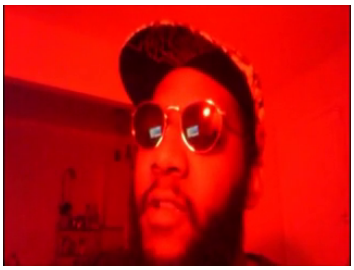
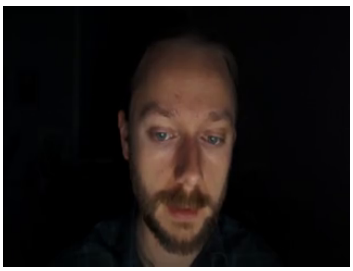
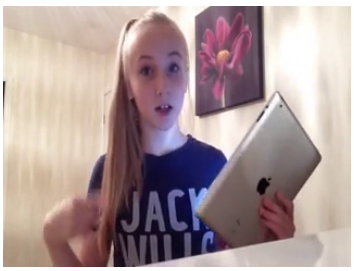
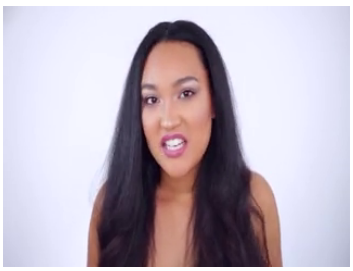


Figure 3.2: Interface of pairwise comparison to collect labels

Scores that were gathered by using crowdsourcing originally consider the Big Five personality model: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. In addition, an invite-to-interview score, which describes a person's likelihood to get invited to a job interview, were introduced in this year's dataset to complement the personality traits scores. All of these values were represented in continuous values within the range [0, 1] as a result of the crowdsourcing. Example snapshots of videos that have low or high values for six of the measurement dimension are shown in Table 3.2.

Table 3.2: Snapshots of videos with high and low values for each dimension

Traits	Extraversion	Agreeableness	Conscientiousness
			
score	0.046729	0.000000	0.048544
			
score	0.925234	0.912088	0.951456

Traits	Neuroticism	Openness	Interview
			
score	0.031250	0.111111	0.149533
			
score	0.937500	0.977778	0.915888

4

Methodology

This chapter elaborates the structural way of developing the system for prediction. First and foremost, we define how we extract features from the available dataset to satisfy our multimodal approach. Lastly, an overall system overview for our experiment is described.

4.1. Features

As explained in the previous chapter, one important aspect that we have to keep in mind when predicting human personality scores is that the ground truth of assessment was done by a human. Also, the final decisions on whether a person should be invited for a job interview will usually also be made by a human, who likely does not have a technical background. This means that the model has to be as transparent and explainable as possible to mimic their decision of judging.

Considering this, we carefully select features that can easily be interpreted by a human. We do this by utilizing three modalities – visual, audio, and textual – to extract features in our model. In the visual modality, we consider features to capture a person's facial movement and expression, as they are one of the best indicators for personality [13, 44, 60]. On the audio modality, we capture on how the person speaks by measuring their emphasis patterns as they show correlation with person personalities [9]. Also, we want to analyze the content of the speech by using textual features described in the additional transcription data to capture the comprehensiveness of speech as it might correlate to intelligence [19].

4.1.1. Visual Features

For the visual representation, the system was not built to focus on the video in general, but particularly on facial expression and movement. In order to do this, we used OpenFace tools to segment only the face from each video, standardizing the segmented facial video to 112x112 pixels as shown in Figure 4.1. OpenFace is an open source toolkit which does not only segment faces but offers a feature extraction library that can extract and characterize facial movements and gaze [5]. We use the output from this tools to further shape our three features representations: Action Unit representation, Emotion representation, and Motion Energy Image representation.

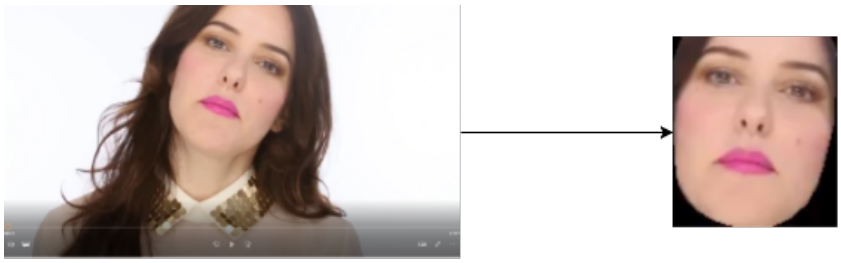


Figure 4.1: Full video to face segmented video

Action Unit representation

On this first representation, we use Action Units measurements produced by OpenFace. OpenFace is able to recognize a subset of individual Action Units (AU) that construct facial expressions encoded in the Facial Action Code System (FACS) as shown in Table 4.1 [22, 23]. These AUs then can be described in two ways: in terms of presence (indicating whether a certain AU is detected in a given time frame) and intensity (indicating how intense an AU is at a given time frame). This means for each time frame; each AUs will have two values that we want to use; the presence and the intensity values.

Action Unit	Description
AU1	Inner Brow Raiser
AU2	Outer Brow Raiser
AU4	Brow Lowerer
AU5	Upper Lid Raiser
AU6	Cheek Raiser
AU7	Lid Tightener
AU9	Nose Wrinkler
AU10	Upper Lip Raiser
AU12	Lip Corner Puller
AU14	Dimpler
AU15	Lip Corner Depressor
AU17	Chin Raiser
AU20	Lip stretcher
AU23	Lip Tightener
AU25	Lips part
AU26	Jaw Drop
AU28	Lip Suck
AU45	Blink

Table 4.1: Action Units that are recognized by OpenFace and its description

For each of these AUs, we construct three features for our input to the system. First, we consider the percentage of time frames during which the AU was visible in a video to show how much movement of each AUs is detected in the video. Second, we store the maximum intensity of the AU in the video to show how strong each AUs occurs in the video. Lastly, we also store the mean intensity of the AU over the video to capture the average strength of AUs in the video. These three features per AU add up to 52 features in total for the Action Unit representation.

Emotion representation

For the second representation, we want to capture a more interpretable facial expression rather than the underlying AUs that describe particular facial movement. In addition to a single AU as a feature, we consider its combinations that can produce a basic set of emotions as shown in Table 4.2. As we can see on the table, we consider seven basic emotions that are a result of a combination of two or more AUs. In order to do this, we track the occurrence of each of these AUs on each time frame along the video.

The emotion exists if and only if all of the corresponding Action Units exist within the same time frame. For example, we only count **happiness** visibility if AU6 and AU12 occur at the same time. Also, we do not count **sadness** as visible if AU1 and AU4 are present while AU15 is missing in a given time frame. This way, the presence of emotion can be obtained and furthermore we can compute the percentage of emotion occurrence just like we did with the Action Unit representation. Likewise, we can also obtain the intensity of the emotion by averaging the intensity of the constructing action units. By doing so, the maximum and average intensity over the video can be obtained the same way as the previous representation. Thus, we get 21 features in total for Emotion representation as our next visual representation.

Emotion	Action Units
Happiness	6 + 12
Sadness	1 + 4 + 15
Surprise	1 + 2 + 5 + 26
Fear	1 + 2 + 4 + 5 + 7 + 20 + 26
Anger	4 + 5 + 7 + 23
Disgust	9 + 15
Contempt	12 + 14

Table 4.2: Emotions and its corresponding Action Units that construct them

MEI representation

The resulting face segmented video from OpenFace also is used for another video representation. In order to capture overall movement of the vlogger's face, a Weighted Motion Energy Image (wMEI) is constructed from the resulting face segmented video. MEI is a grayscale image that shows how much movement happens on each pixel throughout video, with white indicating a lot of movement and black indicating less movement [12]. wMEI was proposed in the work by Biel *et al*. [9] as a normalized version of MEI, by dividing each pixel values with the maximum pixel value. The calculation of wMEI is as follow:

$$MEI = \sum_{t=0}^{T} (D_t), \quad wMEI = \frac{MEI}{max(MEI)},$$

with D_t is a binary image that shows moving pixels in frame t with T is the duration of time frame. wMEI is obtained by dividing all pixel values of MEI with the maximum pixel values.

Our method is inspired by the aforementioned work with improvement on background noise reduction. In the said work, the whole video frame is used as an input to compute wMEI, which makes background movement contribute to the overall wMEI measurements. Thus, there are cases in which the resulting wMEI is all white due to background movement, rather than the movement of a human subject. For example, this happens when the vlogger recorded the video in a public space or while on the road as we can see on the example in Figure 4.2. In this example, the vlogger was self-recording while walking on an outdoor landscape and there was a person on the background following her. As a result, the moving landscape and other person movement are captured to MEI calculation and made it all white which indicates a lot of movement on every pixel, while the video itself only shows the face of the vlogger. There are other cases where the vlogger and were not recording while moving but the wMEI produced is still all white. This is because the video was taken at a crowded public space where people were moving a lot on the background.

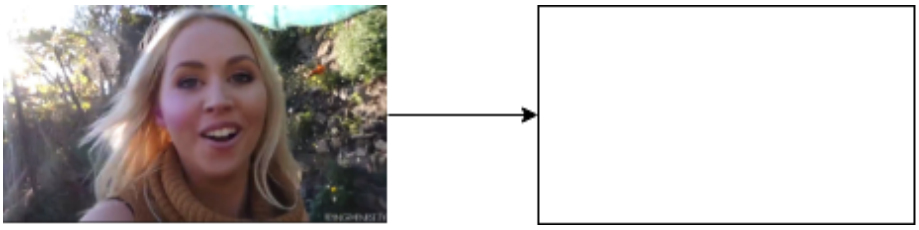


Figure 4.2: wMEI for full video with a lot of background movement

Knowing this situation, we want to limit the influence of the background by performing the wMEI technique on the face-segmented video. By departing from our face segmented video instead of a whole video frame, we minimize the involvement of background in our calculation and thus get a better representation of the subject's true movement, as we can see in Figure 4.3. In this example, the vlogger movements were concentrated around mouth, nose, eyes, and cheeks area and better representing the object's true movement.

In order to create wMEI, we obtain the base face image of each video and iterate over the video time frames to compute the overall movement for each pixel. The resulting grayscale image has a possible

Figure 4.3: wMEI for face segmented video

0 - 255 values for each pixel, with 0 indicates black and 255 indicates white. Thus, these values can be used to compute three statistical features; mean, median, and entropy, to constitute a Motion Energy Image representation.

The current dataset [48] that we are working on for this problem has been carefully selected so that only one unique foreground person faces the camera in the video. However, the current OpenFace implementation has limitations when the video still contains other visual sources with faces, such as posters or music covers in the background. While the situation is rare, we occasionally noticed that a poster was detected and segmented as 'main face' rather than the subject's actual face. For such misdetections, no movement will be detected at all, so this corner case is easily captured by our system and reported on in our feature description.

4.1.2. Audio Features

Audio representation

In the audio modality, we obtain the speaking emphasis patterns to measure tone of voice in the video as it is one of the most powerful social signals [42]. In order to capture the prosodic emphasis patterns, we used speech features extraction code in MATLAB developed by MIT Media Lab [15, 47]. It uses two Hidden Markov Model (HMM) to separate voice/unvoiced and speech/non-speech section of the video and then compute several features on audio tone. Each of the features will be presented as two value, the mean and mean scaled standard deviation. Table 4.3 shows six main features of the audio mode and its corresponding description. In total, there are 12 features for audio representation composed from these six.

Table 4.3: Audio features and its description

Audio Features	Description
F0	Main frequency of audio
F0 conf.	Confidence of F0
Loc. R0 pks	Location of autocorrelation peaks
# R0 pks	Number of autocorrelation peaks
Energy	Energy of the voice
D Energy	Derivative of the energy

4.1.3. Textual Features

Based on findings in organizational psychology, personality traits are not the only (and neither the strongest) predictors for job suitability. In fact, GMA (General Mental Ability) tests, such as intelligence tests, have the highest validity at the lowest application cost [19, 49]. While we do not have formal GMA assessments for subjects in our dataset, we consider that language use of the vlogger may indirectly reveal GMA characteristics, such as the use of difficult words. This is why we also consider textual features, both considering speaking density, as well as linguistic sophistication.

Textual features are generated by using transcripts that were provided as the extension of the [48] dataset. For a handful of videos, transcript data was missing; we manually annotated those videos, such that all videos have transcript data for our purposes, with exception of one video that has no transcript because the person speaks in sign language in the video.

As reported in literature [19, 49] and confirmed in private discussions we had with organizational psychologists, assessment of GMA (intelligence, cognitive ability) is important for many hiring decisions. While this information is not reflected in personality traits, we felt that the linguistic usage of the subjects may reveal some related information.

Readability representation
To assess the linguistic usage of the vlogger, we employed several Readability indexes on the transcripts. This was done by using open source implementations of various readability measures in the NLTK-contrib package of the Natural Language Toolkit (NLTK). More specifically, we used 8 measures as features for the Readability representation: ARI [52], Flesch Reading Ease [24], Flesch-Kincaid Grade Level [35], Gunning Fog Index [27], SMOG Index [39], Coleman Liau Index [18], LIX, and RIX [2]. While these measures are originally developed for written text (and officially may need longer textual input than a few sentences in a transcript), our expectation still would be that they would consistently reflect complexity in linguistic usage.

Text Count representation
In addition, we also used two simple statistical features for an overall Text representation: total word count in the transcript, and the amount of unique words within the transcript.

4.2. System Overview

In this section, we would like to elaborate the overall building blocks of our system. The first one being the system architecture to build our model. Each of the blocks then will be discussed subsequently up until the method to explain the system outcome.

4.2.1. System Architecture

The overall diagram of our system to predict each trait can be seen in Figure 4.4. We select features from each modality that can be interpreted easily by a human, making up six features representation in total. Then, we process each representation individually up until before the fusion stage. We separate these representations initially because we want to see the accuracy of individual predictions and keep our model as interpretable as possible by explaining each feature's significance. We want our system to be able to trace back the prediction scores to each underlying features, to explain the result. Therefore, linear models are best suited for our purposes. It also should be noted that linear regression is a commonly seen model in social sciences literature. From this perspective, we apply linear models on all of prediction blocks, from the prediction for each representation and also the fusion method. By incorporating all linear models on our system, we can obtain representation significance to the final prediction by looking at the regression coefficients from fusion step. Likewise, each feature significance from each representation can also be obtained by looking at the coefficients for each model on representation.

The correlation of each feature to the prediction traits will be checked to see whether a subset of selected features will produce better accuracy rather than using all features. Not only that, but we will also check correlation among features to see the evidence of multicollinearity. Several dimension reduction and regression techniques will be applied to get the best model for trait prediction. At this stage, we will get the best subset of features and the best regression technique to use further on. The output of those selected techniques then will be used as input to a second layer of regression techniques for late fusion. As the final output, a textual based report will be generated detailing our decision making reasoning.

4.2.2. Correlation Analysis

In the pipeline of data analytics, sometimes it is inevitable to have relations among variables. This relationship sometimes benefits the analysis process but also can bring harm if it is not handled properly. We can explore this relationship by using pairwise Pearson's correlation coefficient with this following calculation:

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} ,$$

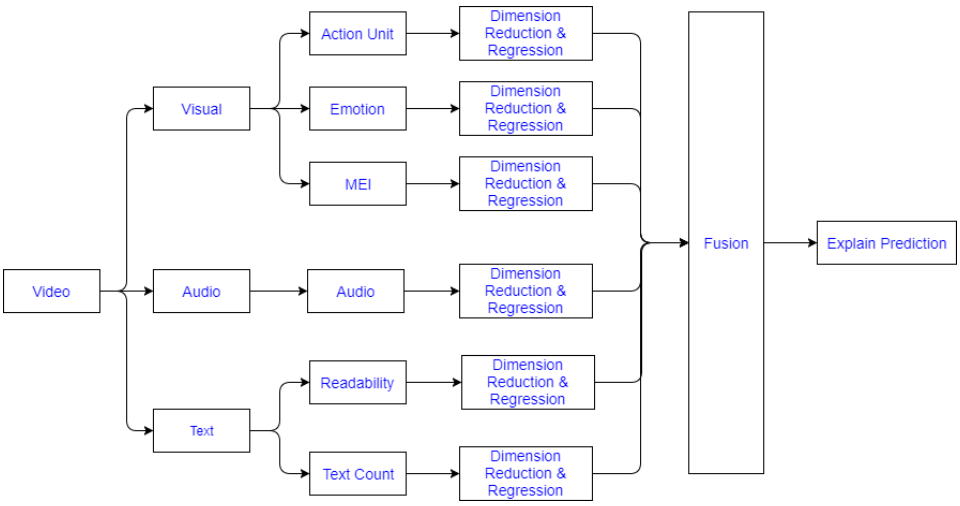


Figure 4.4: Overall system diagram to predict each trait

with $\rho_{X,Y}$ indicates the correlation between variable X and Y, σ_{XY} is covariance of these two variables, and σ shows standard deviation of each variable.

Pearson's correlation coefficient (ρ) measure the strength of linear relationship between a pair of continuous quantitative variables, ranging between [-1, 1]. If there is no relationship between the two variables, the correlation coefficient would be zero. A value greater than 0 means they have positive associations; as one variable value increases the other will increase as well. Likewise, the value less than 0 indicates negative correlation, which means the value of the two variables will be the opposite in magnitude.

In personality studies, correlation analysis is a common technique to determine the significance of features to be utilized to build the model [9, 10, 50]. In our case, we use correlation analysis to check feature significance to the trait prediction and further compare the accuracy of this subset and the model with a full number of features. Not only that, we use it to check whether there is an evidence of multicollinearity among input features.

4.2.3. Dimensionality Reduction & Regression

In the regression analysis, multicollinearity between input variables is not necessarily wanted. It might come from the early stages of the data collection process and also from feature the generation process. It can affect the standard deviation of of the regression coefficients which can make significant variable appear to be insignificant, and vice versa. This also means that we can not determine the precise effect of each feature if we fit regression model with highly correlated features as the input.

In order to mitigate the effect of multicollinearity in our model, several techniques have been considered. The first one is by using the prominent Principal Component Analysis (PCA) technique before feeding the results to Ordinary Least Square (OLS) Regression. The next two are Ridge and Lasso Regression, which incorporate $l2$ and $l1$ regularization technique on the linear regression model, respectively.

Principal Component Regression (PCR)

PCA is a linear transformation that converts a set of correlated variables into uncorrelated variables called principal components. This technique also make sure that the highest principal component accounts for the highest variation of data. Thus, by selecting several principal components, we can maintain variability of data while reducing the amount of variables we have to work with significantly. The transformation from original feature vectors to new principal components can be expressed as follows:

$$Y_{NxM} = X_{NxK} * W_{KxM} \ ,$$

with X is the original feature matrix with N number of observations and K number of original feature dimensions transformed into matrix Y expressing the same N observations with M principal components. The transformation matrix W is composed by eigenvectors with M dimensions.

These principal components then will be fed as input to OLS Regression. This regression technique is a simple linear regression technique that estimates the coefficients by minimizing a loss function with a least square method:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^P}{\operatorname{argmin}} \|y - X\beta\|_2^2 ,$$

Ridge Regression
Rather than using PCA to reduce the dimension to prevent multicollinearity, Ridge regression incorporate a penalizing function to the least square regression model. By doing so, it tries to shrink coefficient towards zero, so that the significance of a subset of input features will be eminent by the value of the coefficients. The estimation of the coefficients then will be as follow:

$$\hat{\beta}^{ridge} = \underset{\beta \in \mathbb{R}^P}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_2^2}_{\text{Penalty}} ,$$

with λ is the tuning parameter. When λ equal to zero, this becomes a least square regression and when λ is infinity the $\hat{\beta}^{ridge}$ is 0. For other value of λ, we balance between fitting linear model and shrinking the coefficients.

Lasso Regression
The difference between Ridge Regression and Lasso Regression is the ridge uses l_2 penalty while lasso uses l_1 penalty. The difference on this penalty function will make the lasso regression to select a subset of important features and make the other to have zero coefficient. In the case of Ridge, the non-important features will still have value a smaller portion larger than zero, and not necessarily perform feature selection.

$$\hat{\beta}^{lasso} = \underset{\beta \in \mathbb{R}^P}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}} ,$$

4.2.4. Fusion Scheme

As explained before, we trained separate model for all of the feature representations and combined them at a later stage. We consider two different fusion mechanisms, the first one being averaging and the last one being weighting using regression.

Averaging
Averaging technique means we put a fixed-weight value for each of the representations. This method was mainly used at the early stage of thesis development and for the ChaLearn competition. For each of the trait, the notation for averaging predictions is as follow:

$$p_{iF} = \frac{1}{M} \sum_{R=1}^{M} p_{iR} ,$$

with p_{iF} is the final prediction score for video i, M is the total number of feature representation, and p_{iR} is the predicted value of video i using representation R.

Regression
The difference between this approach and the aforementioned one is that we do not assign the same value for the weight used for each representation. Instead, we feed the prediction output of each representation as the input to another regression model. Thus, the weight should be adjusted based on the coefficients of the regression and follow this rule:

$$p_{iF} = \sum_{R=1}^{M} \beta_R * p_{iR} ,$$

with β_R being the regression coefficient of the representation R.

4.2.5. Output Explanation

In the Qualitative phase of the ChaLearn CVPR 2017 challenge, the goal was to explain predictions by a human-understandable text. We implemented a simple text description generator, departing from the following thoughts:

Natural Language
As explained before, each of our visual, audio, textual features were picked to be explainable in natural language to a non-technical human. However, we do not have any formal proof which of our features are fully valid predictors of personality traits or interviewability. While our model gives indicators on the strongest linear coefficients, the assessments it was trained on are made by external observers (crowdsourcing workers), which poses a very different situation from the assessment settings in the formal psychology studies as reported in [49]. Therefore, we will not make a hard choice for 'good' features yet, but rather provide a comprehensive report on each observed feature, also indicating acknowledgment of potential feature weaknesses (e.g. indicating that several readability scores were developed for larger texts).

Subject Assessment
As our feature measurements did not formally get tested yet in terms of psychometric validity, it is debatable to consider feature measurements and predicted scores as absolute indicators of interviewability. However, for each person, we can indicate whether the person scores 'unusually' with respect to a larger population of 'representative subjects' (formed by the vloggers represented in the 6000-video training set). Therefore, for each feature measurement, we report what the **typical range** of the feature is, and at what **percentile** the feature score of the subject is, compared to scores of the subjects in the training set.

Feature Influence
Finally, to still reflect major indicators from our linear model in our description, for each representation, we pick the two linear regression coefficients that are largest in the absolute sense. In the case of PCR, we obtain the PCA dimensions corresponding to these coefficients and trace back which two features contributed most strongly to this PCA dimension, and whether the features contribute positively or negatively to the linear model. For these features, a short notice is added to the description, expressing how the feature commonly affects final scoring (e.g. 'In our model, a higher score on this feature typically leads to a higher overall assessment score' for a positive linear contribution.)

4.3. Evaluation

In order to measure the performance of the developed model, the competition organizers use Mean Absolute Error (MAE) to measure the error for each of the personality traits and interview value. MAE is a common evaluation metric to measure accuracy for continuous variable and is a negatively-oriented score, meaning the lower the score the better. MAE also can be interpreted easily as it measures the average of the absolute difference between predicted and true value. The accuracy is then computed by subtracting 1 with the MAE. Thus, the accuracy of the model is defined as follow:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |p_{iT} - g_{iT}|,$$

$$A_T = 1 - MAE,$$

with A_T being the system accuracy for trait T, N is the total number of test set, p_{iT} is the predicted value for each traits, and g_{iT} is the ground truth value. This measurement also will be used to measure the system accuracy for further enhancement.

5

ChaLearn Submisson

Upon deciding on our methods, we worked to build the system for the ChaLearn submission. The first section discuss our initial experiment for predicting personalities and invite-to-inerview scores. After that, we discuss on how we interpret the results of the resulting a text-based explanation.

5.1. Quantitative Measurement

The building blocks of our predictive model for the competition are slightly simpler than the proposed diagram for this thesis. We were late to get the information of the competition, and due to the submission time constraints, we only encompassed four feature representations: Action Unit, MEI, Readability, and Text, as seen in Figure 5.1.

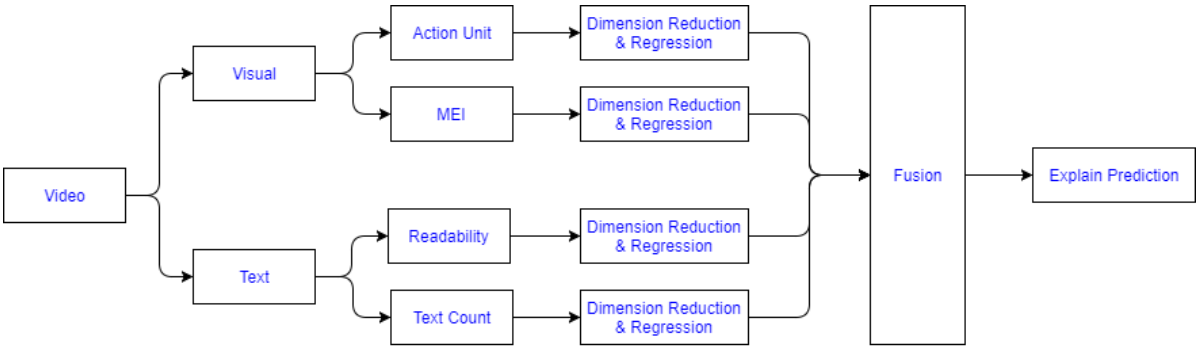


Figure 5.1: Diagram for ChaLearn competition system

Employing the 6000 training set videos, for each representation, we train a separate model to predict personality traits and interview scores. At that time, we were too late to formally enter the quantitative phase of the competition and could compete in the qualitative phase. Thus, we built our model with the explainability in mind rather than ensuring as high accuracy as possible.

We apply Principal Component Analysis (PCA), which performs an orthogonal linear transformation of features, as dimensionality reduction technique for each representation, retaining 90% variance. The resulting transformed features then act as input for a simple linear regression model to predict the scores. For a final prediction score, we apply late fusion by averaging the predictions made by the four different models.

While we did not formally participate in the quantitative phase of the ChaLearn CVPR2017 competition[1], Table 5.1 shows the overall quantitative accuracy of our system on the 2000 videos in the benchmark test set, for each of the Big Five personality traits and the interview invitation assessment. For each predicted class, we compare our scores to the lowest and highest scores (from all of the participants) in the ChaLearn CVPR 2017 Quantitative challenge.

[1] http://chalearnlap.cvc.uab.es/challenge/23/description/

Categories	Our System	Lowest	Highest
Interview	0.887744	0.872129	0.920916
Agreeableness	0.896825	0.891004	0.913731
Conscientiousness	0.880077	0.865975	0.919769
Extraversion	0.887040	0.878842	0.921289
Neuroticism	0.884847	0.863237	0.914613
Openness	0.890314	0.874761	0.917014

Table 5.1: Accuracy (1 - Mean Absolute Error) comparison between our proposed system and the lowest and highest accuracy for each prediction category in the ChaLearn CVPR 2017 Quantitative challenge.

As expected, our system does not yield optimal accuracy, but comparing our scores to the officially reported scores in the Quantitative challenge, our proposed system would consistently outperform the lowest-scoring system for each category. This comes with the benefits of low computational power demands for model fitting, and the earlier discussed advantages for explainability due to our linear models.

We note that our submission model did not include the audio modality, which might improve the model further. Such further improvements will be discussed in the next chapter.

5.2. Qualitative Phase

In the qualitative phase, the organizer set some rules on how the output will be judged. Each of the qualitative phase ChaLearn competition participants must submit a textual description that explains the reasoning behind the Interview prediction scores. Then, the jury will evaluate the output on a 0 to 5 scale based on following criteria:

- **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?

We only used linear models to ensure that we can trace the importance of each original feature for our final prediction. By considering the linear regression coefficients, we know for each PCA dimension whether it contributes positively or negatively to the prediction. Furthermore, considering the PCA transformation matrix, we can trace back how strongly each original feature contributed to each PCA dimension. As a result, for each video on the validation and test set, a fairly long but consistent textual description was generated. An example fragment of the description is given in Figure 5.2 while the full report of subject assessment can be seen in Appendix A.

We start by explaining each feature representation separately. In this example, we report on the subject's use of language. For each feature, we report the description of the feature on the natural language, and for this example is the amount of spoken words we get from the transcript. We then also report the range of this feature from the distribution and the actual feature value of the subject. We also report the percentile score of the feature so that the subject know how they rank. In this example, the subject is at 62th percentile which means the subject has higher feature value than 62% of the distribution. Lastly, another short explanation is added if the feature has one of the highest significance to the model. In this example, this feature has strong positive significance to the model, thus the addition of the last sentence.

The qualitative scores for our description submission are reported in Table 5.2. While our average scores were slightly lower than that of the other submitted system in the challenges, our system led to higher standard deviations (possibly indicating stronger jury responses). Ultimately the differences

```
******************
* USE OF LANGUAGE *
******************

Here is the report on the person's language use:


** FEATURES OBTAINED FROM SIMPLE TEXT ANALYSIS **
Cognitive capability may be important for the job. I looked at a few very
simple text statistics first.

*** Amount of spoken words ***
This feature typically ranges between 0.000000 and 90.000000. The score
for this video is 47.000000 (percentile: 62).
In our model, a higher score on this feature typically leads to a higher
overall assessment score.
```

Figure 5.2: Example description fragment.

between the systems were not deemed statistically significant, and our system was proclaimed winner of the challenge. The competitor [33] explained the result of Random Forest prediction by using binarization with a threshold. If the score value is higher than the distribution mean value, they will assign 1 or 'high' as the description, and 0 for a score lower than mean. In my opinion, this binarization method can be dangerous to explain the interview score prediction. There will be no distinction between a high score subject and a subject that has a score just above average; they lie within the same 'high' group. The subject can not know their actual placement in the distribution, they only know either they are in the 'low' or 'high' group. Moreover, the value around the mean will be sensitive, because a slight score difference will result in getting invited to the interview or not, although they are basically 'just average.' Also, they use the same binarization method on personality traits to infer the Interview decision. This adds another problem because we do not know what makes the model to label the subject 'high' or 'low' on the personality traits. In our model, we report the placement of the subject in the distribution by using percentile. We also report if the feature has high significance value to the model. This way, I believe the output of our system is more interpretable than the aforementioned approach.

Evaluation	Scores
Clarity	3.33 ± 1.43
Explainability	3.23 ± 0.87
Soundness	3.43 ± 0.92
Interpretability	2.4 ± 1.02
Creativity	3.4 ± 0.8

Table 5.2: Explainability scores

Our proposed model employs features that can easily be described in natural language, employs a linear (PCA) transformation to reduce dimensionality, and uses simple linear regression models for predicting scores, such that scores can be traced back to and justified with the underlying features. While using hand-crafted features and models of this simplicity are not what typically is seen in state-of-the-art automatic content processing solutions, we consider this explainability a clear strength and it was verified from the appreciation by human judges. Thus, for the future experiments, we will maintain linear models but will identify further representations and other linear regression techniques for further improvement of the model.

6

Enhanced Multimodal System

By reflecting on the ChaLearn competition result, we worked to find the even better results in term of accuracy by considering several further setups. We add the audio modality and other representations for our model and do exploratory data analysis to it. Not only that, but we also try other linear regression techniques for prediction and also for final fusion. In this chapter, we discuss how these setups affect prediction results. The first section discusses our initial exploratory analysis on correlation analysis. Then, we discuss several regression techniques and their effect to the model's predictive power. Lastly, we choose the best model for each representation to fuse it for final prediction.

6.1. Correlation Analysis

6.1.1. Feature Utilization

We compute Pearson's correlation for each feature on each representation to the personality trait and interview score to analyze feature utilization to prediction. This can be used to infer what judges (crowdsource workers) saw when they made their judgments. The results of the computation are shown in Table 6.1-6.6. As we can see, most of the correlation scores have low magnitude, but there are some that have higher magnitude with also high significance levels. For example, on the Action Unit representation, features that are based on AU12_c and AU12_r (Code: AU25, AU26, AU27) have high correlations with all six traits at high significance levels. As a result, Emotion features that are based on these AUs also have high correlations at high significance levels, namely 'happy' and 'contempt'. This means, human judges (or crowdsource workers) highly regard happiness to infer all six traits. Likewise, we find that sad, fear, and anger emotions (which are rather the opposite of the happy emotion) have negative correlations to all six traits.

If we want to compare our correlation results for audio and MEI features with work at [9] that presented the same approach, we will see different result in term of amount of features that have high significance level ($p < 0.05$). These differences might come from the differences of the dataset itself. In their case, they used one-minute vlogs data with a total of 442 videos, with five annotations from crowdsourcing per video. In the ChaLearn dataset, they have 10,000 15-second excerpts of vlogs that had been assessed by more people from crowdsourcing (321,684 pairs to label 10,000 videos). In addition, the ChaLearn dataset was obtained from a series of filtering criteria to ensure the quality of video and audio. Furthermore, they obtained MEI calculation using full frames of video while we used only facial-segmented video. Thus, these underlying differences of the data might have caused the differences.

We also can see that all readability indexes have a high significance level to all traits, except for Extraversion. One interesting thing to note here is that all readability indexes have positive correlations to all six traits with exception of Flesch Reading Ease. This is expected because, in the Flesch Reading Ease, lower values indicate more complicated speech (or text), while for the other indexes, higher values indicate more complicated speech. For Readability and Text feature representations, the features all have high significance level for all traits, except three readability indexes for Extraversion.

The summary of feature utilization is shown in Table 6.7. We take all features that have a high significance level ($p < 0.05$) regardless of the correlation scores. The code in each cell encodes the

utilized feature for each representation, and also with the information on the number of utilized features for this representation. As we can see, we have various feature subset for each trait, and later we will test whether this subset performs better or not than the full set of features.

Table 6.1: Pearson's correlation coefficient between Action Unit features and Personality and Interview score († p<0.05, * p<0.01, ** p<0.001, *** p<0.0001)

Code		A	C	E	I	N	O
AU1	% presence (AU01_c)	0.047**	0.002	0.127**	0.063***	0.098***	0.128***
AU2	max intensity (AU01_r)	0.040*	-0.046**	0.081***	0.033*	0.067***	0.065***
AU3	mean intensity (AU01_r)	0.037*	-0.019	0.079***	0.040*	0.072***	0.076***
AU4	% presence (AU02_c)	0.033*	-0.019	0.079***	0.039*	0.069***	0.090***
AU5	max intensity (AU02_r)	0.014	-0.073***	0.015	0.005	0.026†	0.007
AU6	mean intensity (AU02_r)	0.027†	-0.047**	0.037*	0.016	0.046**	0.041*
AU7	% presence (AU04_c)	-0.137***	-0.171***	-0.235***	-0.194***	-0.187***	-0.179***
AU8	max intensity (AU04_r)	-0.096***	-0.171***	-0.147***	-0.143***	-0.110***	-0.115***
AU9	mean intensity (AU04_r)	-0.134***	-0.159***	-0.218***	-0.184***	-0.180***	-0.189***
AU10	% presence (AU05_c)	0.063***	0.072***	0.043**	0.063***	0.059***	0.089***
AU11	max intensity (AU05_r)	0.016	-0.021	0.093***	0.028†	0.046**	0.080***
AU12	mean intensity (AU05_r)	0.015	-0.018	0.100***	0.031†	0.051***	0.88***
AU13	% presence (AU06_c)	0.076***	0.025	0.196***	0.104***	0.121***	0.111***
AU14	max intensity (AU06_r)	0.073***	-0.043**	0.161***	0.069***	0.094***	0.080***
AU15	mean intensity (AU06_r)	0.045**	-0.058***	0.073***	0.018	0.026†	-0.010
AU16	% presence (AU07_c)	0.01	-0.03†	0.09***	0.02	0.04*	0.06***
AU17	max intensity (AU07_r)	0.04*	-0.04*	0.09***	0.03†	0.06***	0.06***
AU18	mean intensity (AU07_r)	0.02	-0.03†	0.03†	0.01	0.01	0.00
AU19	% presence (AU09_c)	0.01	-0.03†	0.09***	0.03†	0.07***	0.09***
AU20	max intensity (AU09_r)	0.01	-0.05**	0.10***	0.02	0.08***	0.11***
AU21	mean intensity (AU09_r)	0.01	-0.04*	0.09***	0.02	0.08***	0.10***
AU22	% presence (AU10_c)	0.07**	0.05**	0.06***	0.06***	0.06***	0.02
AU23	max intensity (AU10_r)	0.03†	-0.05**	0.02	0.01	0.03†	-0.01
AU24	mean intensity (AU10_r)	0.02	-0.03†	-0.05**	-0.02	-0.02	-0.09***
AU25	% presence (AU12_c)	0.13***	0.08***	0.35***	0.19***	0.22***	0.23***
AU26	max intensity (AU12_r)	0.16***	0.09***	0.36***	0.22***	0.25***	0.25***
AU27	mean intensity (AU12_r)	0.14***	0.08***	0.33***	0.19***	0.21***	0.20***
AU28	% presence (AU14_c)	0.03†	-0.04*	0.14***	0.05**	0.05***	0.06***
AU29	max intensity (AU14_r)	0.01	-0.12***	0.03†	-0.03†	0.00	-0.04*
AU30	mean intensity (AU14_r)	0.00	-0.10***	0.03†	-0.02	-0.02	-0.06***
AU31	% presence (AU15_c)	-0.05**	-0.12***	0.00	-0.07***	-0.01	-0.01
AU32	max intensity (AU15_r)	-0.01	-0.10***	-0.04*	-0.05**	-0.01	-0.05**
AU33	mean intensity (AU15_r)	-0.02	-0.09***	-0.05***	-0.05***	-0.03†	-0.06***

AU34	% presence (AU17_c)	-0.03†	-0.06***	0.03†	-0.02	0.02	0.02
AU35	max intensity (AU17_r)	-0.03†	-0.11***	-0.02	-0.05***	0.00	-0.02
AU36	mean intensity (AU17_r)	-0.04*	-0.13***	-0.07***	-0.08***	-0.04*	-0.08***
AU37	% presence (AU20_c)	-0.01	-0.04**	-0.01	-0.02	0.02	0.00
AU38	max intensity (AU20_r)	0.00	-0.08***	0.01	-0.02	0.03†	0.03†
AU39	mean intensity (AU20_r)	0.00	-0.08***	0.00	-0.02	0.01	0.01
AU40	% presence (AU23_c)	-0.01	-0.10***	-0.16***	-0.10***	-0.11***	-0.17***
AU41	max intensity (AU23_r)	0.01	-0.10***	0.03†	-0.02	0.04*	0.03*
AU42	mean intensity (AU23_r)	-0.01	-0.12***	0.02	-0.04*	0.02	0.01
AU43	% presence (AU25_c)	-0.02	-0.06***	0.11***	0.01	0.06***	0.10***
AU44	max intensity (AU25_r)	0.04*	-0.03*	0.05**	0.02	0.07***	0.06***
AU45	mean intensity (AU25_r)	0.07***	-0.01	0.11***	0.06***	0.10***	0.10***
AU46	% presence (AU26_c)	0.00	-0.03	0.11***	0.02	0.07***	0.09***
AU47	max intensity (AU26_r)	0.00	-0.06***	0.02	0.00	0.04**	0.03*
AU48	mean intensity (AU26_r)	0.02	-0.03*	0.04*	0.02	0.06***	0.04*
AU49	% presence (AU45_c)	0.02	-0.01	0.09***	0.03†	0.04**	0.08***
AU50	max intensity (AU45_r)	0.13***	0.13***	0.19***	0.16***	0.16***	0.18***
AU51	mean intensity (AU45_r)	0.07***	0.10***	0.17***	0.12***	0.12***	0.16***
AU52	% presence (AU28_c)	-0.04**	-0.12***	-0.15***	-0.11***	-0.12***	-0.17***

Table 6.2: Pearson's correlation coefficient between Audio features and Personality and Interview score († p<0.05, * p<0.01, ** p<0.001, *** p<0.0001)

Code		A	C	E	I	N	O
A1	F0 (m)	-0.033†	-0.011	-0.137***	-0.054***	-0.083***	-0.1232***
A2	F0 (s)	0.130***	0.176***	0.167	0.177***	0.173***	0.168***
A3	F0 conf. (m)	-0.056***	-0.065***	-0.161	-0.088***	-0.110***	-0.157***
A4	F0 conf. (s)	0.052***	0.080***	-0.023	0.058***	0.030†	-0.018
A5	Loc R0 pks (m)	0.092***	0.116***	0.144***	0.130***	0.145***	0.164***
A6	Loc R0 pks (s)	-0.016	-0.010	-0.119***	-0.042**	-0.065***	-0.115***
A7	# R0 pks (m)	0.010	0.031†	0.034*	0.022	0.015	0.030†
A8	# R0 pks (s)	0.051***	0.071***	0.051***	0.062***	0.051***	0.050**
A9	Energy (m)	0.100***	0.093***	0.120***	0.118***	0.145***	0.141***
A10	Energy (s)	0.100***	0.086***	0.120***	0.113***	0.144***	0.139***
A11	D Energy (m)	0.005	0.015	0.012	0.009	0.006	0.003
A12	D Energy (s)	0.112***	0.085***	0.074***	0.109***	0.131***	0.107***

Table 6.3: Pearson's correlation coefficient between Emotion features and Personality and Interview score († p<0.05, * p<0.01, ** p<0.001, *** p<0.0001)

Code		A	C	E	I	N	O
E1	% presence (happy)	0.111***	0.055***	0.305***	0.161***	0.183***	0.188***
E2	max intensity (happy)	0.135***	0.057***	0.315***	0.176***	0.207***	0.207***
E3	mean intensity (happy)	0.124***	0.049***	0.274***	0.155***	0.180***	0.174***
E4	% presence (sad)	-0.101***	-0.144***	-0.091***	-0.126***	-0.087***	-0.064***
E5	max intensity (sad)	-0.078***	-0.133***	-0.076***	-0.105***	-0.069***	-0.057***
E6	mean intensity (sad)	-0.069***	-0.119***	-0.069***	-0.096***	-0.066***	-0.0557***
E7	% presence (surprise)	0.032†	-0.011	0.091***	0.037*	0.069***	0.102***
E8	max intensity (surprise)	0.078***	-0.004	0.129***	0.074***	0.110***	0.125***
E9	mean intensity (surprise)	0.071***	0.001	0.107***	0.067***	0.089***	0.103***
E10	% presence (fear)	-0.028†	-0.051***	-0.037*	-0.042*	-0.025†	-0.024
E11	max intensity (fear)	-0.040*	-0.071***	-0.044**	-0.057***	-0.040*	-0.031†
E12	mean intensity (fear)	-0.040*	-0.071***	-0.045**	-0.057	-0.041	-0.031
E13	% presence (anger)	-0.025	-0.064***	-0.088***	-0.058***	-0.050***	-0.073***
E14	max intensity (anger)	-0.055***	-0.124***	-0.102***	-0.090***	-0.052***	-0.066***
E15	mean intensity (anger)	-0.054***	-0.116***	-0.097***	-0.086***	-0.050***	-0.062***
E16	% presence (disgust)	-0.014	-0.054***	0.036*	-0.016	0.017	0.043**
E17	max intensity (disgust)	0.011	-0.038*	0.078***	0.010	0.060***	0.078***
E18	mean intensity (disgust)	0.022	-0.012	0.086***	0.026†	0.069***	0.084***
E19	% presence (contempt)	0.102***	0.037*	0.310***	0.154***	0.174***	0.195***
E20	max intensity (contempt)	0.120***	0.036*	0.295***	0.158***	0.187***	0.191***
E21	mean intensity (contempt)	0.102***	0.026†	0.240***	0.130***	0.154***	0.147***

Table 6.4: Pearson's correlation coefficient between Readability features and Personality and Interview score († p<0.05, * p<0.01, ** p<0.001, *** p<0.0001)

Code		A	C	E	I	N	O
R1	ARI	0.101***	0.170***	0.038*	0.130***	0.091***	0.071***
R2	Flesch Reading Ease	-0.090***	-0.168***	-0.035*	-0.120***	-0.078***	-0.065***
R3	Flesch-Kincaid	0.111***	0.179***	0.059***	0.142***	0.101***	0.091***
R4	Gunning Fog Index	0.115***	0.174***	0.050***	0.137***	0.096***	0.086***
R5	SMOG Index	0.111***	0.171***	0.039*	0.128***	0.092***	0.071***
R6	Coleman Liau Index	0.065***	0.146***	-0.003	0.091***	0.066***	0.025***
R7	LIX	0.081***	0.151***	0.023	0.104***	0.063***	0.052†
R8	RIX	0.082***	0.151***	0.0226	0.105***	0.065***	0.050**

Table 6.5: Pearson's correlation coefficient between MEI features and Personality and Interview score († p<0.05, * p<0.01, ** p<0.001, *** p<0.0001)

Code		A	C	E	I	N	O
M1	Mean	0.024	-0.072***	0.111***	0.031†	0.102***	0.130***
M2	Median	0.057***	-0.007	0.150***	0.079***	0.133***	0.164***
M3	Entropy	0.07512***	0.176***	0.014	0.093***	0.017	-0.01019

Table 6.6: Pearson's correlation coefficient between Text features and Personality and Interview score († p<0.05, * p<0.01, ** p<0.001, *** p<0.0001)

Code		A	C	E	I	N	O
T1	Word count	0.234***	0.229***	0.222***	0.275***	0.303***	0.238***
T2	Unique word	0.221***	0.231***	0.202***	0.266***	0.288***	0.215***

Table 6.7: Summary of cue utilization for Personality Trait and Interview from six feature representations with the code of features.

	Action Unit	Emotion	MEI	Audio	Readability	Txt
A	AU1-4, AU6-10, AU13-15, AU17, AU22, AU23, AU25-28, AU31, AU34-36, AU44, AU45, AU50-52 # = 28	E1-12, E14, E15, E19-E21 # = 17	M2, M3 # = 2	A1-5, A8-10, A12 # = 9	R1-8 # = 8	T1, T2 # = 2
C	AU2, AU5-10, AU14-44, AU47, AU48, AU50-52 # = 43	E1-6, E10-17, E19-E21 # = 17	M1, M3 # = 2	A2-5, A7-10, A12 # = 9	R1-8 # = 8	T1, T2 # = 2
E	AU1-4, AU6-22, AU24-30, AU32-34, AU36, AU40, AU41, AU43-46, AU48-52 # = 43	E1-21 # = 21	M1, M2 # = 2	A1, A5-10, A12 # = 8	R1-5 # = 5	T1, T2 # = 2
I	AU1-4, AU7-14, AU17, AU19, AU22, AU25-29, AU31-33, AU35-36, AU40, AU42, AU45, AU49-52 # = 32	E1-11, E13-15, E18-21 # = 18	M1-M3 # = 3	A1-6, A8-10, A12 # = 10	R1-8 # = 8	T1, T2 # = 2
N	AU1-17, AU19-23, AU25-28, AU33, AU36, AU38, AU40, AU41, AU43-52 # = 41	E1-9, E11, E13-21 # = 19	M1, M2 # = 2	A1-6, A8-10, A12 # = 10	R1-8 # = 8	T1, T2 # = 2
O	AU1-4, AU6-14, AU16, AU17, AU19-21, AU24-30, AU32, AU33, AU36, AU38, AU40, AU41, AU43-52 # = 41	E1-9, E11, E13-21 # = 19	M1, M2 # = 2	A1-3, A5-10, A12 # = 10	R1-8 # = 8	T1, T2 # = 2

6.1.2. Multicollinearity

Other than to infer feature utilization, we also use correlation analysis to get an idea of the correlation between features within a representation. The resulting correlation heatmaps for Audio and Readability features is in Figure 6.1 and 6.2, respectively. The other correlation heatmaps, which also show high correlations between features, are shown in the Appendix B.

From Figure 6.1, we can see that many features correlate with each other: for example, features that represent pitch such as F0, F0 conf., and Loc R0 pks. We can see that F0 (m) has a high positive correlation with F0 conf. (m). On the contrary, Loc R0 pks (m) has high negative correlations to both F0 (m) and F0 conf. (m). This is natural because we can estimate pitch by dividing signal rate by the location of the first autocorrelation peaks. The clearer evidence of multicollinearity is shown in Figure 6.2. All of the readability indexes have high positive correlations to each other with the exception of Flesch Reading Ease. The reason is the same as we explained before: lower Flesch Reading Ease score indicates higher text complexity, while lower values on other readability indexes indicates lower text complexity.

These results confirm our suspicion that is inevitable to have features that are correlated with each other on data analysis. Thus, the dimension reduction is indeed justified.

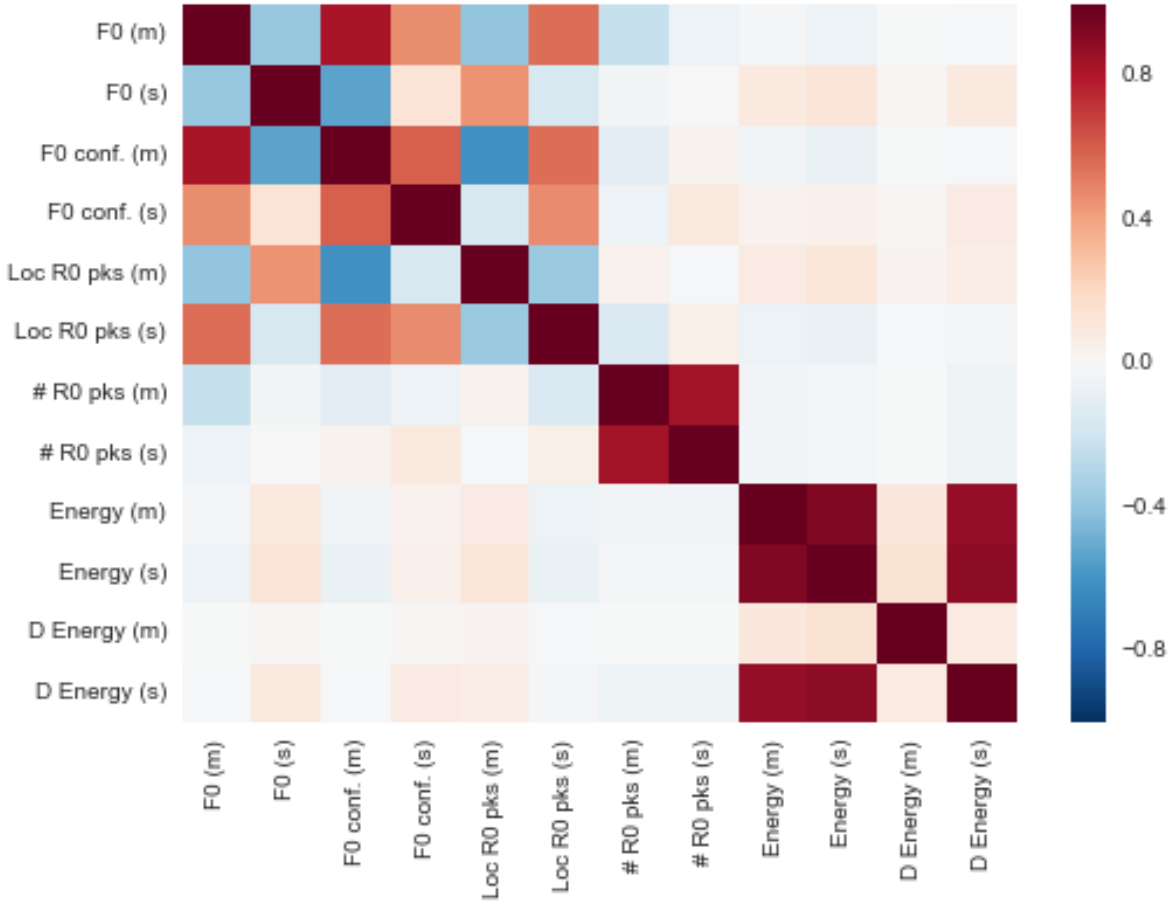


Figure 6.1: Heatmap for Audio correlation

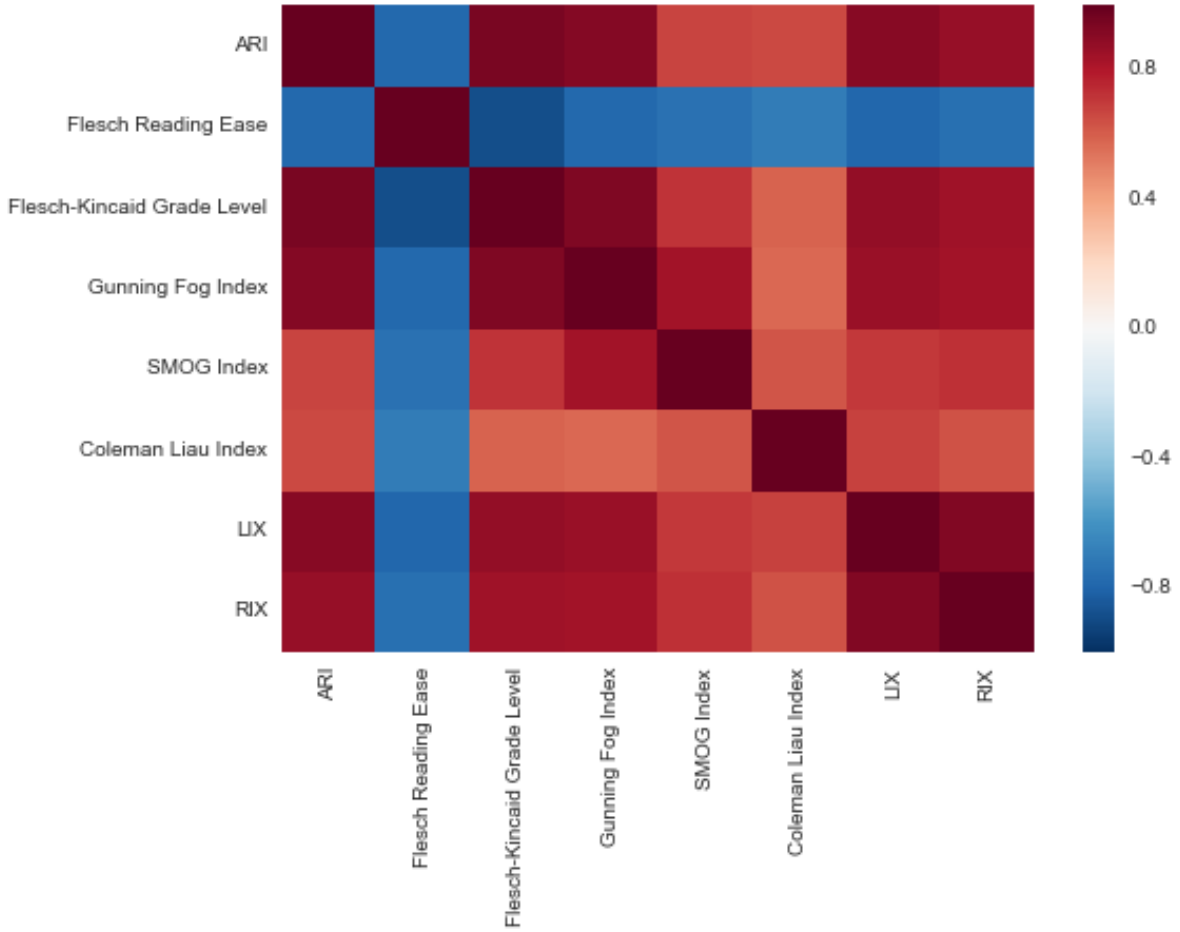


Figure 6.2: Heatmap for Readability correlation

6.2. Prediction

In this experiment, we want to test out individual feature representation accuracies in several setups. We consider three regression techniques. For each, we test out two different subsets of features to infer each trait. We use 10-fold cross validation for the training set to get the accuracy of each model with low bias. Then, we compare the result between full features and a subset of features (from correlation analysis) for each representation and trait.

For the PCR experiment, we apply PCA to retain 90% variance, both for full feature and subset of feature, before feeding it to OLS regression. This is the same approach as our competition submission. The resulting accuracy for all the model using this regression technique is shown in Table 6.8. We can see that reducing the features by using correlation analysis does not necessarily mean the model always produces the highest accuracy, similarly to the experiment that was done at [10]. In fact, for Action Unit and MEI, most of the higher accuracy comes from the full feature set. On the other hand, the effect of selecting a group of significant features provides a better accuracy for some traits using Emotion and Audio. This might be because while the significance level is high, the correlation score itself are mostly low, and also by not selecting features that high correlation but low significance the accuracy might reduce. For Readability and Text, all models have the same accuracy because the correlation analysis indicates that all features are significant, with one exception on Extraversion.

Table 6.8: Model accuracy using PCR for five personalities and Interview scores for two different feature subset per representation from the training set by using 10-fold cross validation. Bold number indicates the higher score between two subset of feature for each representation.

	A	C	E	I	N	O
action unit	**0.895761**	0.881743	**0.894954**	**0.889237**	**0.886030**	**0.894302**
action unit selection	0.895683	**0.882018**	0.894268	0.889143	0.884878	0.893413
mei	**0.894217**	**0.878105**	**0.880526**	0.883548	**0.880047**	**0.886034**
mei selection	0.894201	0.877694	0.879964	0.883548	0.879130	0.885433
emotion	0.894556	0.876993	0.889838	**0.885686**	**0.882745**	**0.888197**
emotion selection	**0.894572**	**0.877261**	0.889838	0.885667	0.882731	0.888191
audio	0.895243	**0.877885**	**0.881318**	0.884841	0.880414	**0.885235**
audio selection	**0.895312**	0.877773	0.881286	**0.884997**	**0.880605**	0.885230
readability	0.893425	0.872676	0.878186	0.881020	0.877359	0.882928
readability selection	0.893425	0.872676	**0.878335**	0.881020	0.877359	0.882928
text	0.897544	0.878843	0.881097	0.887194	0.884015	0.887275
text selection	0.897544	0.878843	0.881097	0.887194	0.884015	0.887275

For the Ridge and Lasso regression, we want to test alternate regression techniques to reduce multicollinearity, just like what PCA and OLS did. For these two experiments, we optimize the λ for each model by using 10-fold cross validation to ensure we pick the λ that produce the lowest error. The resulting accuracy for Ridge and Lasso model are shown in Table 6.9 and 6.10, respectively. In both cases, the selection of a subset of significant features also does not impact the accuracy similar to the previous experiment, and most of the better accuracy are from full feature set.

If we look at each trait individually on the column of Table 6.8, 6.9, and 6.10, we can see that our new additions of feature representations – Emotion and Audio – have a higher accuracy than Mei and Readability for most of the traits, and sometimes outperforming Text representation. This means that our decision to add these two might produce a better accuracy overall.

From all these experiments, we fit our data to find the best model for all representations and traits. By comparing Table 6.8, 6.9, and 6.10, we can obtain which model suits the best for which representation and trait, as summarized in Table 6.11. As can be seen from the table, introducing two alternative regression methods shows some improvement on all of the traits, as there is no trait that only uses the PCR method. These models are the ones that we are going to use for late fusion to obtain the final

Table 6.9: Model accuracy using Ridge Regression for five personalities and Interview scores for two different feature subset per representation from the training set by using 10-fold cross validation. Bold number indicates the higher score between two subset of feature for each representation.

	A	C	E	I	N	O
action unit	**0.896483**	**0.882807**	**0.895445**	**0.889712**	**0.886446**	**0.895020**
action unit selection	0.896051	0.882710	0.895236	0.889278	0.885074	0.894191
mei	**0.894273**	**0.877782**	**0.880502**	0.883388	**0.879886**	**0.886039**
mei selection	0.894226	0.877676	0.880408	0.883388	0.879699	0.885706
emotion	**0.894763**	**0.877184**	0.889968	**0.885717**	**0.882760**	0.888651
emotion selection	0.894638	0.877086	0.889968	0.885660	0.882741	0.888651
audio	**0.895480**	**0.878227**	0.881647	**0.885026**	**0.880650**	0.885356
audio selection	0.895430	0.878093	**0.881676**	0.884820	0.880393	**0.885450**
readability	0.893342	0.872445	0.877964	0.880769	0.876674	0.882450
readability selection	0.893342	0.872445	**0.877991**	0.880769	0.876674	0.882450
text	0.897434	0.878840	0.881112	0.887195	0.883977	0.887250
text selection	0.897434	0.878840	0.881112	0.887195	0.883977	0.887250

prediction.

Table 6.10: Model accuracy using Lasso Regression for five personalities and Interview scores for two different feature subset per representation from the training set by using 10-fold cross validation. Bold number indicates the higher score between two subset of feature for each representation.

	A	C	E	I	N	O
action unit	**0.896503**	**0.882742**	**0.895544**	**0.889737**	**0.886428**	**0.895073**
action unit selection	0.896151	0.882667	0.895251	0.889267	0.885212	0.894277
mei	**0.894270**	**0.877775**	**0.880500**	0.883381	**0.879868**	**0.886058**
mei selection	0.894220	0.877684	0.880309	0.883381	0.879613	0.885732
emotion	**0.894835**	**0.877207**	0.889769	**0.885660**	**0.882812**	0.888781
emotion selection	0.894728	0.877139	0.889769	0.885595	0.882807	**0.888785**
audio	**0.895495**	**0.878146**	0.881645	**0.885014**	**0.880533**	0.885397
audio selection	0.895422	0.878058	**0.881647**	0.884763	0.880289	**0.885465**
readability	0.893327	0.872453	0.877862	0.880781	0.876556	0.882325
readability selection	0.893327	0.872453	**0.877870**	0.880781	0.876556	0.882325
text	0.897423	0.878838	0.881106	0.887193	0.883969	0.887267
text selection	0.897423	0.878838	0.881106	0.887193	0.883969	0.887267

Table 6.11: Summary of the best regression and feature subset for each representation and trait.

	A	C	E	I	N	O
action unit	Lasso full	Ridge full	Lasso full	Lasso full	Ridge full	Lasso full
mei	Ridge full	PCR full	PCR full	PCR full	PCR full	Lasso full
emotion	Lasso full	PCR select	Ridge full	Ridge full	Lasso full	Lasso select
audio	Lasso full	Ridge full	Ridge select	Ridge full	Ridge full	Lasso select
readability	PCR full	PCR full	PCR select	PCR full	PCR full	PCR full
text	PCR full	PCR full	Ridge full	Ridge full	PCR full	PCR full

6.3. Fusion

The last step in our model for predicting the traits is generating a final prediction from six the prediction values obtained on the different representation. These predictions values are the results of the best model for each feature representation as we described previously. We conduct the fusion using four methods. The first one is simple averaging just like we did on the competition submission, and the three others are the regression methods that we used on the previous step. The procedure of each regression method is also the same with the previous step; we use 90% variance retention for PCR, and find the best λ for Ridge and Lasso regression. The accuracy of each method is shown in Table 6.12.

As we can see from the result, the simple averaging fusion shows the lowest accuracy for all of the traits. PCR shows promising results on Interview prediction, while the big five traits are better predicted by using Ridge or Lasso. Interestingly enough, for these five traits, the difference between accuracy using Ridge and Lasso is really low, with 0.000005 difference at most. This might come from the difference of penalty function ($l1$ and $l2$) thay they use that cause this slight difference.

Table 6.12: Accuracy for 5 traits and Interview scores using three different regression techniques and averaging for fusion on the test data. Bold number indicates the highest accuracy for each trait.

	A	C	E	I	N	O
PCR	0.900535	0.886589	0.899530	**0.895019**	0.894293	0.898255
Ridge	0.900816	**0.887389**	**0.900123**	0.893086	**0.894517**	**0.899134**
Lasso	**0.900819**	0.887388	0.900119	0.893022	0.894515	0.899129
Avg	0.896853	0.880222	0.888320	0.887341	0.885234	0.890587

As right now we already have our final accuracy score, we want to compare how our model fares to others. The summary of our best model for each trait from Table 6.12 and the comparison to other models can be seen in Table 6.13. The work at [26] used a multilayer Perceptron Neural Network on three separate modalities – video, audio, and speech –, and do a final fusion using weighted average. If we are allowed to compare our initial work to theirs, their system outperforms our initial model on all of the traits. After further enhancement on this thesis, the table is turned and our layered regression model outperform their model on all of the traits, except Agreeableness. However, our model is still has a lower accuracy than the work at [33], the quantitative winner of the ChaLearn competition. They use thousands facial features, thousands scene features, and audio feature for the Extreme Learning Machine regression, which makes their model far more complex than ours.

This layered regression architecture allowed us to get insight on individual performance of feature representations, rather than combining all features together into one big regression problem. This way, we can choose the best suitable model for each representation as well as maintaining system transparency for explanation purpose. The fusion using regression also ensures we can produce better results than with averaging. Because, rather than distributing the weight evenly, this way the strong features will have a higher weight, and vice versa.

Table 6.13: Accuracy between our system and others from the same challenge.

Categories	Enhanced	Initial	[26]	[33]
Interview	0.895019	0.887744	0.894	0.9198
Agreeableness	0.900819	0.896825	0.902	0.9161
Conscientiousness	0.887389	0.880077	0.884	0.9166
Extraversion	0.900123	0.887040	0.892	0.9206
Neuroticism	0.894517	0.884847	0.885	0.9149
Opennes	0.899134	0.890314	0.896	0.9169

7

Conclusion & Future Direction

We presented a system for personality trait and interviewability prediction, which was designed such that the system's underlying features and decision-making processes were as transparent as possible. Despite the simplicity of our features and models, reasonable quantitative system accuracy scores were obtained. Qualitative natural language descriptions generated from our model also were judged positively by the jury members in the challenge.

We ensure that the input features are the one that can be easily interpreted by human so that we can mimic peoples judgment while maintaining system transparency. We develop linear models on all of the architecture to ensure that we can trace back the decision-making to the underlying features. Moreover, we develop a layered architecture so that the best regression model for each representation can be obtained and also ensure better overall accuracy by fusion using regression. The accuracy of our layered architecture shows promising result compare to our initial competition submission. It also outperforms work at [26] that used multilayer perceptron for the prediction. However, our layered linear architecture still has lower accuracy than the more complex model at [33].

If we are allowed to compare our method on explaining the prediction to another work [33] from the same qualitative competition, we believe our work has the better explainability. We report the natural description of our features, present the actual comparison of a subject to the distribution, and also indicate which features that have the strong influence on the model. The work at [33] use Random Forest for stacking the prediction and binarization with thresholding for the explanation part. This binarization method possesses problem to explain the actual placement of a subject in the distribution, other than assigning it to 'high' or 'low' class. Moreover, they based their Interview score decision from the personality traits binarization. This leads to another problem because the model can not explain why a subject has a 'high' or 'low' personality traits score in their model.

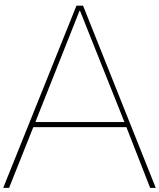
As for our textual descriptions, we currently did not select any strong features but provided a full report on every single feature. This may have made our current report somewhat long and overwhelming to a human reader. We expect that our explanations will allow for a better user experience when presented in a less textual way, e.g. in the form of graphs. Next to this, to avoid information overload, smarter information selection can be performed. However, in order to do this, it is important to validate our features more strongly in the psychometric sense, and it will be useful to obtain further qualitative input from human judges on what parts of our explanation were understood and appreciated, and what parts were deemed less interpretable. It will particularly be important to receive such feedback from organizational psychologists and HR specialists, as those will be the most likely users and final decision-makers for a system like ours.

Lastly, it should be noted that the current dataset considers vlogs, but not official video resumes. Although there are similarities between vlogs and video resumes, video resumes might have distinct differences in term of content delivery and preparation [45]. For example, it is safe to assume that when people want to apply for a job, they will want to maximally impress a potential employer, rather than presenting themselves casually and more spontaneously, which may be the case in vlogs. Furthermore, the vlogs also were not targeted at a specific job vacancy, while job-specific demands may, in reality, be important for candidate assessment. From the initiated collaboration with the Erasmus University of Rotterdam, we now have more realistic video resume data at our disposal on which our model can be

tested. However, up until this report was finalized, there still was no ground truth annotation available. It will be interesting to further pursue this direction and see the result of this experiment.

For reproducibility, the code used in our submission to the ChaLearn challenge is made available on GitHub[1].

[1] https://github.com/sukmawicaksana/CVPR2017



ChaLearn Textual Report

****
ASSESSMENT REPORT FOR VIDEO 7qGYGbIg45c.001.mp4:
****
On a scale from 0.0 to 1.0, I would rate this person's interviewability as 0.471230. Below, I will report on linguistic and visual assessment of the person. Percentiles are obtained by comparing the person against scores of 6000 earlier assessed people.

********** USE OF LANGUAGE **********
Here is the report on the person's language use:

** FEATURES OBTAINED FROM SIMPLE TEXT ANALYSIS **
Cognitive capability may be important for the job. I looked at a few very simple text statistics first.

*** Amount of spoken words ***
This feature typically ranges between 0.000000 and 90.000000. The score for this video is 31.000000 (percentile: 25).
In our model, a higher score on this feature typically leads to a higher overall assessment score.

*** Amount of unique words ***
This feature typically ranges between 0.000000 and 61.000000. The score for this video is 26.000000 (percentile: 25).

** FEATURES OBTAINED FROM READABILITY ANALYSIS **
As slightly more sophisticated measure, I ran several readability metrics. Note that several of these were originally designed for larger, written texts. This is why metrics may disagree.

*** US grade level required for comprehension according to the ARI score ***
This feature typically ranges between -11.010000 and 33.042676. The score for this video is 2.730575 (percentile: 30).
In our model, a higher score on this feature typically leads to a higher overall assessment score. According to the ARI score, the estimated educational level needed to understand this person is Second Grade.

*** US grade level required for comprehension according to the Coleman Liau score ***
This feature typically ranges between -28.130000 and 31.182500. The score for this video is 5.265900 (percentile: 31).

*** US grade level required for comprehension according to the Flesch-Kincaid score ***

This feature typically ranges between -15.200000 and 28.054900. The score for this video is 3.382100 (percentile: 29).
In our model, a higher score on this feature typically leads to a higher overall assessment score.

*** Years of reading required to understand the text according to the SMOG score ***
This feature typically ranges between 0.000000 and 21.165902. The score for this video is 7.472136 (percentile: 30). In our model, a higher score on this feature typically leads to a higher overall assessment score.

*** Readability assessment according to the Lix score ***
This feature typically ranges between 0.000000 and 101.000000. The score for this video is 20.011494 (percentile: 28). According to the LIX score, the estimated educational level needed to understand this person is Fourth Grade.

*** Readability assessment according to the RIX score ***
This feature typically ranges between 0.000000 and 16.000000. The score for this video is 1.000000 (percentile: 29). According to the RIX score, the estimated educational level needed to understand this person is Fourth Grade.

*** US grade level required for comprehension according to the Gunning-Fog score ***
This feature typically ranges between 0.000000 and 32.133300. The score for this video is 6.358600 (percentile: 30). According to the Gunning Fog Index, the estimated educational level needed to understand this person is Seventh Grade.

*** Reading ease according to the Flesch score ***
This feature typically ranges between -8.725000 and 205.820000. The score for this video is 86.844800 (percentile: 61). According to the Flesch-Kincaid reading ease score, this person's text is easy. Conversational English for consumers.

********** VISUAL FEATURES **********
Here is the report on what I could 'see':

** FEATURES OBTAINED FROM MOTION ENERGY ANALYSIS **
I focused on the person's face and verified how much movement was detected over time.

*** Motion energy entropy: how varied is the degree of movement across the person's face? ***
This feature typically ranges between 0.000000 and 3.831083. The score for this video is 2.364628 (percentile: 76). In our model, a higher score on this feature typically leads to a higher overall assessment score. It looks like the person is consistently facing the camera.

*** Median motion energy: what is the typical degree of movement of this person? ***
This feature typically ranges between 0.000000 and 1.000000. The score for this video is 0.000000 (percentile: 0). In our model, a higher score on this feature typically leads to a higher overall assessment score. When taking the median of the degree of movement, this person moves a lot.

*** Mean motion energy: how much does the person move on average? ***
This feature typically ranges between 0.000000 and 0.978636. The score for this video is 0.134822 (percentile: 0). In our model, a higher score on this feature typically leads to a higher overall assessment score. When averaging the degree of movement, this person moves a lot.

** FEATURES OBTAINED FROM FACIAL ACTION UNIT ANALYSIS **
I focused on Action Units in the person's face: activity of dedicated face muscles. These values may say something about how expressive the person is.

FEATURES FROM THE EYES

*** Action Unit 1: how often was the inner brow raised? ***
This feature typically ranges between 0.000000 and 0.652742. The score for this video is 0.017429 (percentile: 0).

*** Action Unit 1: how much was the inner brow raised at most? ***
This feature typically ranges between 0.000000 and 5.000000. The score for this video is 1.109180 (percentile: 0).

*** Action Unit 1: how much was the inner brow raised on average? ***
This feature typically ranges between 0.000000 and 2.008451. The score for this video is 0.108935 (percentile: 0).

*** Action Unit 2: how often was the outer brow raised? ***
This feature typically ranges between 0.000000 and 0.749455. The score for this video is 0.026144 (percentile: 0).

*** Action Unit 2: how much was the outer brow raised at most? ***
This feature typically ranges between 0.000000 and 5.000000. The score for this video is 0.527766 (percentile: 0).

*** Action Unit 2: how much was the outer brow raised on average? ***
This feature typically ranges between 0.000000 and 0.767348. The score for this video is 0.052774 (percentile: 0).

*** Action Unit 4: how often was the brow lowered? ***
This feature typically ranges between 0.000000 and 1.000000. The score for this video is 0.000000 (percentile: 3).

*** Action Unit 4: how much was the brow lowered at most? ***
This feature typically ranges between 0.000000 and 5.000000. The score for this video is 0.000000 (percentile: 0). In our model, a higher score on this feature typically leads to a lower overall assessment score.

*** Action Unit 4: how much was the brow lowered on average? ***
This feature typically ranges between 0.000000 and 3.965490. The score for this video is 0.000000 (percentile: 0).

*** Action Unit 5: how often was the upper lid raised? ***
This feature typically ranges between 0.000000 and 1.000000. The score for this video is 1.000000 (percentile: 100). In our model, a higher score on this feature typically leads to a lower overall assessment score.

*** Action Unit 5: how much was the upper lid raised at most? ***
This feature typically ranges between 0.000000 and 5.000000. The score for this video is 0.441528 (percentile: 0).

*** Action Unit 5: how much was the upper lid raised on average? ***
This feature typically ranges between 0.000000 and 0.538660. The score for this video is 0.036469 (percentile: 0).

*** Action Unit 7: how often was the eyelid tightened? ***
This feature typically ranges between 0.000000 and 1.000000. The score for this video is 0.000000 (percentile: 0).

*** Action Unit 7: how much was the eyelid tightened at most? ***
This feature typically ranges between 0.000000 and 5.000000. The score for this video is 0.000000 (percentile: 0).

*** Action Unit 7: how much was the eyelid tightened on average? ***
This feature typically ranges between 0.000000 and 3.819835. The score for this video is 0.000000 (percentile: 0).

FEATURES FROM THE MOUTH

*** Action Unit 10: how often was the upper lip raised? ***
This feature typically ranges between 0.000000 and 1.000000. The score for this video is 0.000000 (percentile: 25).

*** Action Unit 10: how much was the upper lip raised at most? *** This feature typically ranges between 0.000000 and 5.000000. The score for this video is 0.000000 (percentile: 0).

*** Action Unit 10: how much was the upper lip raised on average? ***
This feature typically ranges between 0.000000 and 3.136602. The score for this video is 0.000000 (percentile: 0).

*** Action Unit 12: how often was the lip corner pulled? ***
This feature typically ranges between 0.000000 and 1.000000. The score for this video is 0.000000 (percentile: 40).

*** Action Unit 12: how much was the lip corner pulled at most? ***
This feature typically ranges between 0.000000 and 4.584670. The score for this video is 0.239716 (percentile: 0).

*** Action Unit 12: how much was the lip corner pulled on average? ***
This feature typically ranges between 0.000000 and 2.880709. The score for this video is 0.005225 (percentile: 0).

*** Action Unit 15: how often was the lip corner depressed? ***
This feature typically ranges between 0.000000 and 0.762943. The score for this video is 0.021786 (percentile: 0).

*** Action Unit 15: how much was the lip corner depressed at most? ***
This feature typically ranges between 0.000000 and 5.000000. The score for this video is 0.617131 (percentile: 6).

*** Action Unit 15: how much was the lip corner depressed on average? ***
This feature typically ranges between 0.000000 and 1.472110. The score for this video is 0.084487 (percentile: 6).

*** Action Unit 20: how often was the lip stretched? ***
This feature typically ranges between 0.000000 and 0.539510. The score for this video is 0.067538 (percentile: 4).

*** Action Unit 20: how much was the lip stretched at most? ***
This feature typically ranges between 0.000000 and 4.201070. The score for this video is 0.983754 (percentile: 26). In our model, a higher score on this feature typically leads to a higher overall assessment score.

*** Action Unit 20: how much was the lip stretched on average? ***
This feature typically ranges between 0.000000 and 0.579620. The score for this video is 0.109793

(percentile: 26).

*** Action Unit 23: how often was the lip tightened? ***
This feature typically ranges between 0.000000 and 1.000000. The score for this video is 1.000000 (percentile: 100).

*** Action Unit 23: how much was the lip tightened at most? ***
This feature typically ranges between 0.000000 and 5.000000. The score for this video is 1.153310 (percentile: 0).

*** Action Unit 23: how much was the lip tightened on average? ***
This feature typically ranges between 0.000000 and 1.098906. The score for this video is 0.142789 (percentile: 0).

*** Action Unit 25: how often did the lips part? ***
This feature typically ranges between 0.000000 and 0.675381. The score for this video is 0.091503 (percentile: 0).

*** Action Unit 25: how much did the lips part at most? ***
This feature typically ranges between 0.000000 and 5.000000. The score for this video is 2.241340 (percentile: 60).

*** Action Unit 25: how much did the lips part on average? ***
This feature typically ranges between 0.000000 and 1.907661. The score for this video is 0.534212 (percentile: 58).

*** Action Unit 28: how often was the lip sucked? ***
This feature typically ranges between 0.000000 and 1.000000. The score for this video is 0.000000 (percentile: 0).

FEATURES FROM THE CHIN

*** Action Unit 17: how often was the chin raised? ***
This feature typically ranges between 0.000000 and 0.736383. The score for this video is 0.241830 (percentile: 27).

*** Action Unit 17: how much was the chin raised at most? ***
This feature typically ranges between 0.000000 and 5.000000. The score for this video is 1.369790 (percentile: 7).

*** Action Unit 17: how much was the chin raised on average? ***
This feature typically ranges between 0.000000 and 1.992418. The score for this video is 0.365476 (percentile: 0).

*** Action Unit 14: how often was the dimple present? ***
This feature typically ranges between 0.000000 and 1.000000. The score for this video is 0.019608 (percentile: 25).

*** Action Unit 14: how much was the dimple present at most? ***
This feature typically ranges between 0.000000 and 5.000000. The score for this video is 0.293882 (percentile: 0).

*** Action Unit 14: how much was the dimple present on average? ***
This feature typically ranges between 0.000000 and 3.293953. The score for this video is 0.005195 (percentile: 0).

*** Action Unit 26: how often did the jaw drop? ***
This feature typically ranges between 0.000000 and 0.614379. The score for this video is 0.202614 (percentile: 64).

*** Action Unit 26: how much did the jaw drop at most? ***
This feature typically ranges between 0.000000 and 5.000000. The score for this video is 2.458890 (percentile: 73). In our model, a higher score on this feature typically leads to a lower overall assessment score.

*** Action Unit 26: how much did the jaw drop on average? ***
This feature typically ranges between 0.000000 and 1.945176. The score for this video is 0.562125 (percentile: 74).

FEATURES FROM OTHER AREAS

*** Action Unit 6: how often was the cheek raised? ***
This feature typically ranges between 0.000000 and 1.000000. The score for this video is 0.000000 (percentile: 43).

*** Action Unit 6: how much was the cheek raised at most? ***
This feature typically ranges between 0.000000 and 4.056480. The score for this video is 0.000000 (percentile: 0).

*** Action Unit 6: how much was the cheek raised on average? ***
This feature typically ranges between 0.000000 and 2.727962. The score for this video is 0.000000 (percentile: 0).

*** Action Unit 9: how often did the nose wrinkle? ***
This feature typically ranges between 0.000000 and 0.562092. The score for this video is 0.000000 (percentile: 0).

*** Action Unit 9: how much did the nose wrinkle at most? ***
This feature typically ranges between 0.000000 and 5.000000. The score for this video is 0.303294 (percentile: 0).

*** Action Unit 9: how much did the nose wrinkle on average? ***
This feature typically ranges between 0.000000 and 0.906702. The score for this video is 0.046448 (percentile: 0).

*** Action Unit 45: how often did the person blink? ***
This feature typically ranges between 0.000000 and 0.612200. The score for this video is 0.008715 (percentile: 0).

*** Action Unit 45: how much did the person blink at most? ***
This feature typically ranges between 0.000000 and 5.000000. The score for this video is 0.321554 (percentile: 0).

*** Action Unit 45: how much did the person blink on average? ***
This feature typically ranges between 0.000000 and 1.756510. The score for this video is 0.049599 (percentile: 0).

B

Correlation Heatmap

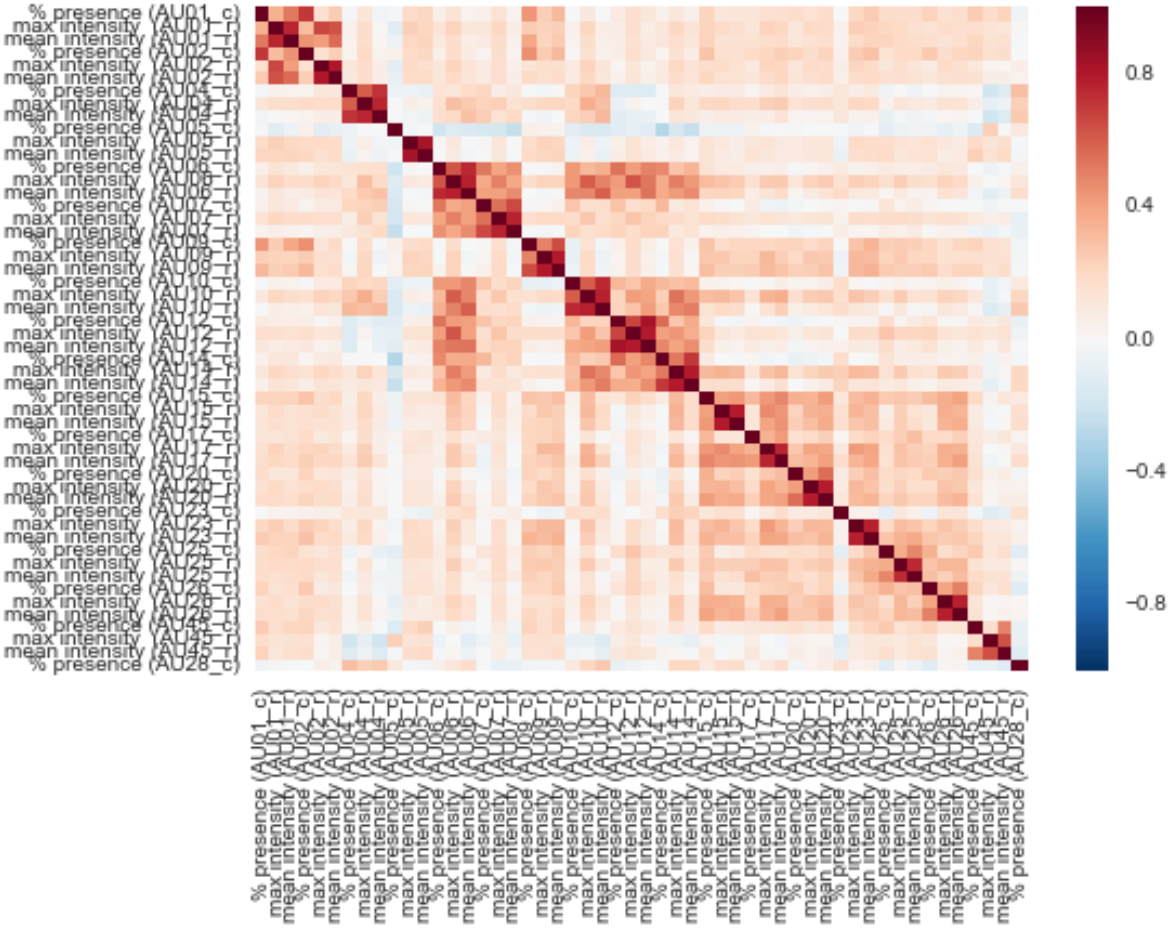


Figure B.1: Heatmap for Action Unit correlation

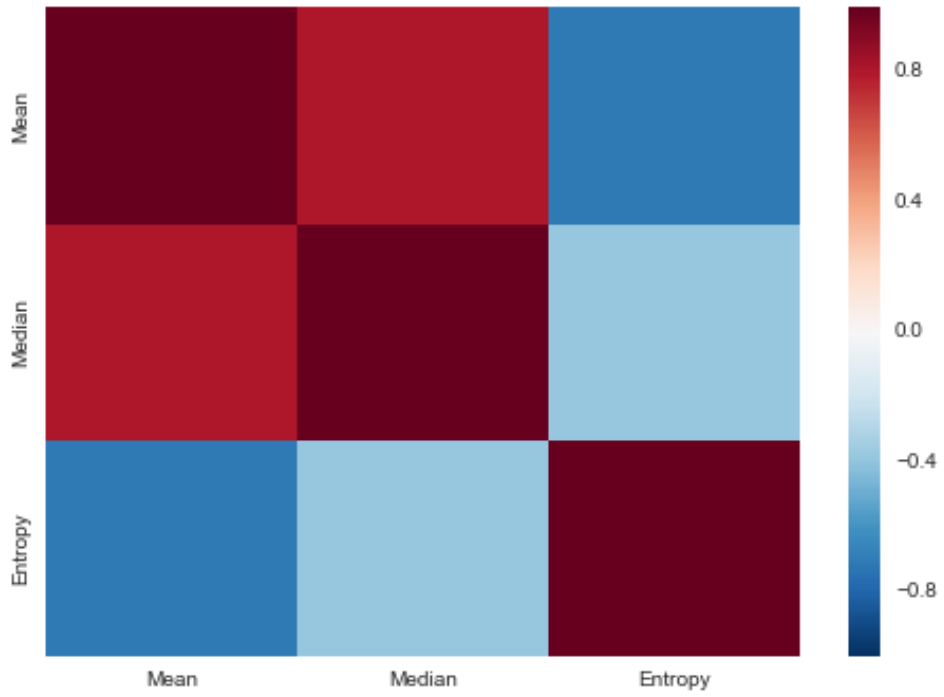


Figure B.2: Heatmap for MEI correlation

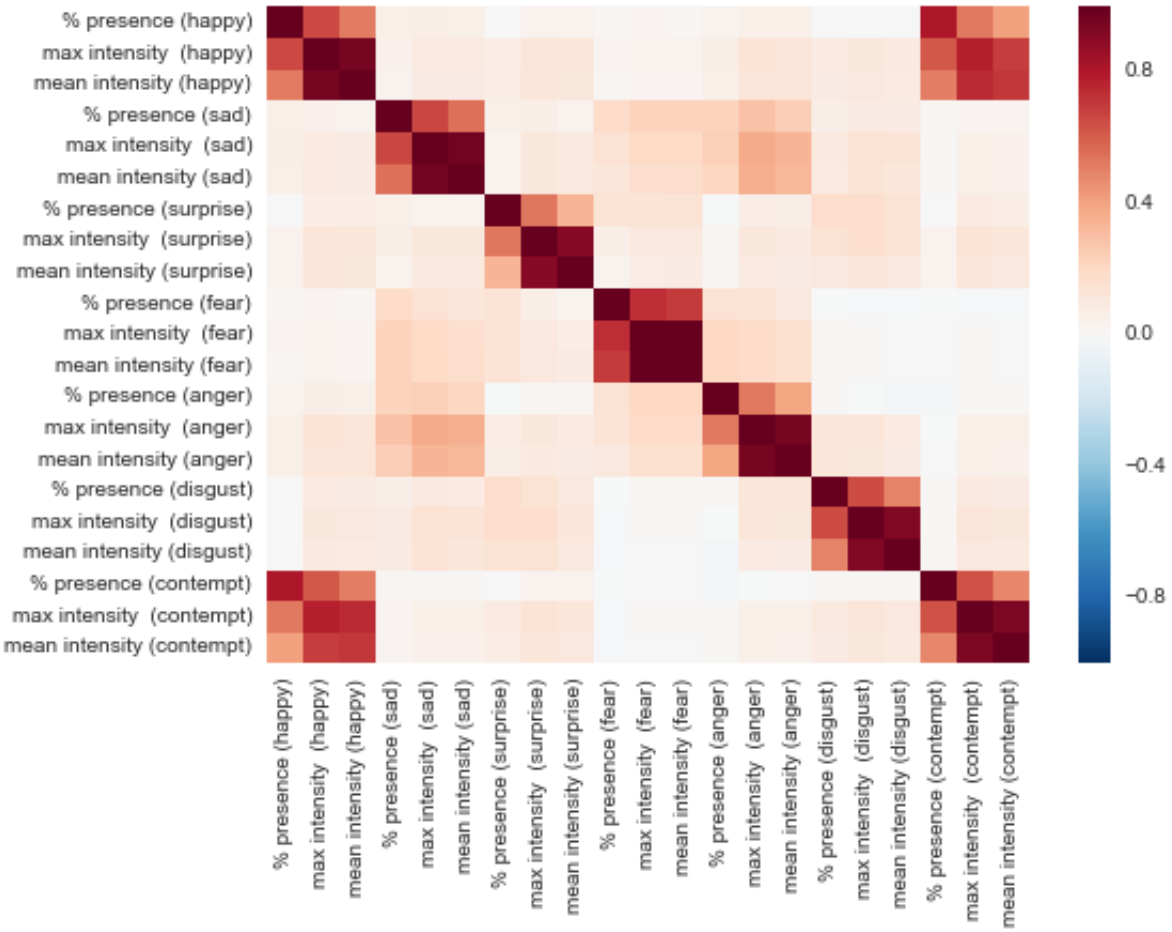


Figure B.3: Heatmap for Emotion correlation

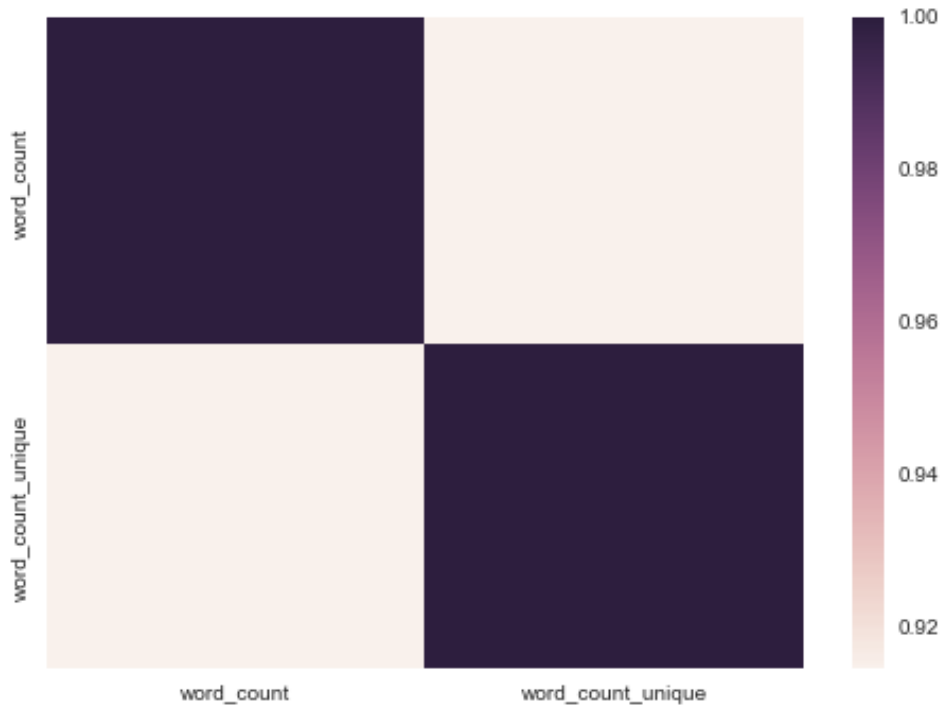


Figure B.4: Heatmap for Text correlation

Bibliography

[1] Hervé Abdi and Lynne J. Williams. Principal component analysis, 2010. ISSN 19395108.

[2] Jonathan Anderson. Lix and Rix: Variations on a Little-known Readability Index. *Journal of Reading*, 26(6):490–496, 1983. ISSN 00224103. URL http://www.jstor.org/stable/40031755.

[3] Blaise Agüera y Arcas, Margaret Mitchell, and Alexander Todorov. Physiognomy's New Clothes, 2017. URL https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a.

[4] Jens B. Asendorpf and Susanne Wilpers. Personality effects on social relationships. *Journal of Personality and Social Psychology*, 74(6):1531–1544, 1998. ISSN 1939-1315. doi: 10.1016/j.yhbeh.2008.01.006.

[5] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. Cross-dataset learning and person-specific normalisation for automatic Action Unit detection. In *FG*, volume 06, pages 1–6, 2015. ISBN 978-1-4799-6026-2. doi: 10.1109/FG.2015.7284869. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7284869.

[6] Ligia Maria Batrinca, Nadia Mana, Bruno Lepri, Fabio Pianesi, and Nicu Sebe. Please, tell me about yourself: Automatic personality assessment using short self-presentations. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 255–262, 2011. ISBN 9781450306416. doi: 10.1145/2070481.2070528. URL http://dl.acm.org/citation.cfm?doid=2070481.2070528.

[7] Salah Eddine Bekhouche. Personality Traits and Job Candidate Screening via Analyzing Facial Videos Abdelmalik Taleb-Ahmed. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1660–1663, 2017. doi: 10.1109/CVPRW.2017.211.

[8] Joan Isaac Biel and Daniel Gatica-Perez. The YouTube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Transactions on Multimedia*, 15(1):41–55, 2013. ISSN 15209210. doi: 10.1109/TMM.2012.2225032.

[9] Joan-Isaac Biel, Oya Aran, and Daniel Gatica-Perez. You Are Known by How You Vlog: Personality Impressions and Nonverbal Behavior in YouTube. *Artificial Intelligence*, pages 446–449, 2011. URL http://www.idiap.ch/{~}jibiel/pubs/BielAranGaticaICWSM11.pdf.

[10] Joan-Isaac Biel, Lucía Teijeiro-Mosquera, and Daniel Gatica-Perez. FaceTube: Predicting Personality from Facial Expressions of Emotion in Online Conversational Video. *Icmi'12*, pages 1–4, 2012. doi: 10.1145/2388676.2388689. URL http://publications.idiap.ch/downloads/papers/2012/Biel{_}ICMI-MLMI{_}2012.pdf.

[11] Joan-Isaac Biel, Vagia Tsiminaki, John Dines, and Daniel Gatica-Perez. Hi YouTube! Personality impressions and verbal content in social video. *Proceedings of the 15th ACM on International conference on multimodal interaction - ICMI '13*, pages 119–126, 2013. doi: 10.1145/2522848.2522877. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84892584123{&}partnerID=tZOtx3y1.

[12] Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001. ISSN 01628828. doi: 10.1109/34.910878.

[13] Peter Borkenau, Steffi Brecke, Christine Möttig, and Marko Paelecke. Extraversion is accurately perceived after a 50-ms exposure to a face. *Journal of Research in Personality*, 43(4):703–706, 2009. ISSN 00926566. doi: 10.1016/j.jrp.2009.03.007.

[14] RA Bradley and ME Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 00063444. doi: 10.2307/2334029. URL http://www.jstor.org/stable/10.2307/2334029.

[15] Ron Caneel. *Social Signaling in Decision Making*. PhD thesis, Massachusetts Institute of Technology, 2005. URL http://groupmedia.media.mit.edu/datasets/Social{_}Signaling{_}in{_}Decision{_}Making.pdf.

[16] Avshalom Caspi, Brent W Roberts, and Rebecca L Shiner. Personality Development: Stability and Change. *Annual Review of Psychology*, 56(1):453–484, 2005. ISSN 0066-4308. doi: 10.1146/annurev.psych.55.090902.141913.

[17] Baiyu Chen, Sergio Escalera, Isabelle Guyon, Victor Ponce-Lopez, Nihar Shah, and Marc Oliu Simon. Overcoming calibration problems in pattern labeling with pairwise ratings: Application to personality traits. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9915 LNCS, pages 419–432, 2016. ISBN 9783319494081. doi: 10.1007/978-3-319-49409-8_33.

[18] Meri Coleman and T. L. Liau. A Computer Readability Formula Designed for Machine Scoring. *Journal of Applied Psychology*, 60(2):283–284, 1975. ISSN 0021-9010. doi: 10.1037/h0076540. URL http://content.apa.org/journals/apl/60/2/283.

[19] Mark Cook. *Personnel Selection: Adding Value Through People*. Wiley-Blackwell, fifth edition, 2009. ISBN 9780470986455. doi: 10.1002/9780470742723.

[20] Paul T Costa and Robert R. McCrae. *Longitudinal Stability of Adult Personality*. 1997. ISBN 0-12-134645-5. doi: 10.1016/B978-012134645-4/50012-3. URL http://www.sciencedirect.com/science/article/pii/B9780121346454500123{%}5Cnhttp://linkinghub.elsevier.com/retrieve/pii/B9780121346454500123.

[21] John M. Digman. Personality Structure: Emergence of the Five-Factor Model. *Annual Reviews of Psychology*, 41:417–40, 1990. ISSN 0066-4308. doi: 10.1146/annurev.ps.41.020190.002221.

[22] Paul Ekman and Wallance V. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. 1978. ISBN 0931835011.

[23] Paul Ekman and Erika Rosenberg. *What the face reveals*. 2005. ISBN 0-19-510447-1.

[24] Rudolf Flesch. A New Readability Yardstick. *The Journal of Applied Psychology*, 32(3):221–233, 1948. ISSN 0021-9010. doi: 10.1037/h0057532.

[25] LR R. Goldberg. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models, 1999. ISSN 00926566. URL http://projects.ori.org/lrg/PDFs{_}papers/Abroad-bandwidthinventory.pdf.

[26] Jelena Gorbova and Andre Litvin. Automated Screening Of Job Candidate Based On Multimodal Video Processing. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1679–1685, 2017. doi: 10.1109/CVPRW.2017.214.

[27] R Gunning. *The Technique of Clear Writing*. McGraw-Hill, 1952. URL https://books.google.nl/books?id=ofI0AAAAMAAJ.

[28] Annemarie M F Hiemstra. *Fairness in Paper and Video Resume Screening*. PhD thesis, Erasmus University Rotterdam, the Netherlands, 2013.

[29] Joyce Hogan, Paul Barrett, and Robert Hogan. Personality measurement, faking, and employment selection. *Journal of Applied Psychology*, 92(5):1270–1285, 2007. ISSN 0021-9010. doi: 10.1037/0021-9010.92.5.1270. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/0021-9010.92.5.1270{%}0Ahttp://files/33928/Hoganetal{_}2007{_}Personalitymeasurement,faking,andemploymentselection.pdf.

[30] Robert Hogan, Joyce Hogan, and Brent W. Roberts. Personality measurement and employment decisions. *American Psychologist*, 51(5):469–477, 1996. ISSN 0003-066X. doi: 10.1037/0003-066X.51.5.469. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/0003-066X.51.5.469.

[31] M.D Humphries, K Gurney, and T.J Prescott. The brainstem reticular formation is a small-world, not scale-free, network. *Proceedings of the Royal Society B: Biological Sciences*, 273(1585): 503–511, 2006. ISSN 0962-8452. doi: 10.1098/rspb.2005.3354. URL http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2005.3354.

[32] Op P John and S Srivastava. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(510):102–138, 1999. ISSN 00031224. doi: citeulike-article-id:3488537. URL http://books.google.com/books?hl=en{&}lr={&}id=b0yalwi1HDMC{&}oi=fnd{&}pg=PA102{&}dq=The+big-five+trait+taxonomy:+History,+Measurement,+and+Theoretical+Perspectives.{&}ots=756zS6ZtPk{&}sig=-3pfI7eNKlyZLlJYEmwdDYeJ82Y{%}5Cnhttp://scholar.google.de/scholar?hl=de{&}q=john+sri.

[33] Heysem Kaya, G Furkan, and Albert Ali Salah. Multi-modal Score Fusion and Decision Trees for Explainable Automatic Job Candidate Screening from Video CVs. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1651–1659, 2017. doi: 10.1109/CVPRW.2017.210.

[34] Katie J Kemp, L Michelle Bobbitt, Michelle Bednarz Beauchamp, and Elizabeth Ann Peyton. Using one-minute video résumés as a screening tool for sales applicants. *Journal of Marketing Development and Competitiveness*, 7(1):84–92, 2013.

[35] J P Kincaid, R P Fishburne, R L Rogers, and B S Chissom. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. *Technical Training*, Research B(February):49, 1975. doi: ERIC#:ED108134. URL http://www.dtic.mil/dtic/tr/fulltext/u2/a006655.pdf.

[36] Andrew S I D Lang and Joshua Rio-Ross. Using Amazon Mechanical Turk to Transcribe Historical Handwritten Documents. *Code4Lib Journal*, (15):1–10, 2011. ISSN 19405758. URL http://search.ebscohost.com/login.aspx?direct=true{&}db=lih{&}AN=67494814{&}site=ehost-live.

[37] Leqi Liu, Daniel Preot, and Lyle Ungar. Analyzing Personality through Social Media Profile Picture Choice. *The AAAI DIGITAL LIBRARY*, (Icwsm):211–220, 2016.

[38] L. Lloyd, L. Lloyd, P. Kaulgud, P. Kaulgud, S. Skiena, and S. Skiena. Newspapers vs. blogs: Who gets the scoop. *Proceedings of the AAAI Symp. Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*, page 8, 2005. URL http://scholar.google.com/scholar?hl=en{&}btnG=Search{&}q=intitle:Newspapers+vs.+Blogs:+Who+Gets+the+Scoop?{#}0.

[39] G.H. McLaughlin. SMOG grading: A new readability formula. *Journal of reading*, 12(8):639–646, 1969. ISSN 0022-4103. doi: 10.1039/b105878a. URL http://www.jstor.org/stable/40011226.

[40] Ba Nardi, Dj Schiano, and Michelle Gumbrecht. Blogging as social activity, or, would you let 900 million people read your diary? *… of the 2004 ACM conference on …*, pages 222–231, 2004. ISSN 1-58113-810-5. doi: 10.1145/1031607.1031643. URL http://doi.acm.org/10.1145/1031607.1031643{%}5Cnhttp://dl.acm.org/ft{_}gateway.cfm?id=1031643{&}type=pdf{%}5Cnhttp://dl.acm.org/citation.cfm?id=1031643.

[41] Bonnie a. Nardi, Diane J. Schiano, Michelle Gumbrecht, and Luke Swartz. I'm blogging this: A closer look at why people blog. *Communications of the ACM*, pages 1–16, 2004. URL http://www.dourish.com/classes/ics234cw04/nardi.pdf.

[42] Clifford Nass and Scott Brave. Wired for Speech: How Voice Activates and Advances the Human Computer Relationship. *Computational Linguistics*, 32(3):451–452, 2005. ISSN 0891-2017. doi: 10.1162/coli.2006.32.3.451. URL https://www.lib.uwo.ca/cgi-bin/ezpauthn.cgi?url=http://search.proquest.com/docview/1037392793?accountid=15115{%}5Cnhttp://vr2pk9sx9w.search.serialssolutions.com/?ctx{%}7B{_}{%}7Dver=Z39.88-2004{%}7B{&}{%}7Dctx{%}7B{_}{%}7Denc=info:ofi/enc:UTF-8{%}7B{&}{%}7Drfr{%}7B{_}{%}7Did=info:sid/ProQ{%}7B{%}25{%}7D3Aeducation{%}7B{&}{%}7Drft{%}7B.

[43] National Institute of Health. Conducting a Fair Selection Process. URL https://www.edi.nih.gov/sites/default/files/public/EDI{_}Public{_}files/guidance/toolkits/managers/manager-fair-selection-toolkit01.pdf.

[44] Laura P Naumann, Simine Vazire, Peter J Rentfrow, and Samuel D Gosling. Personality judgments based on physical appearance. *Personality and social psychology bulletin*, 35(12): 1661–1671, 2009. ISSN 0146-1672. doi: 10.1177/0146167209346309.

[45] Laurent Son Nguyen and Daniel Gatica-Perez. Hirability in the Wild: Analysis of Online Conversational Video Resumes. *IEEE Transactions on Multimedia*, 18(7):1422–1437, 2016. ISSN 15209210. doi: 10.1109/TMM.2016.2557058.

[46] Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin E P Seligman. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6):934–952, 2015. ISSN 1939-1315 (Electronic). doi: 10.1037/pspp0000020.

[47] Alex Pentland. Social Dynamics : Signals and Behavior. *Proceedings of the 3rd International Conference on Developmental Learning, Oct 2004*, 5:263–267, 2004. URL http://vismod.media.mit.edu//tech-reports/TR-579.pdf.

[48] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H.J. Escalante, and S. Escalera. Chalearn LAP 2016: First round challenge on first impressions - Dataset and results. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9915 LNCS(November), 2016. ISSN 16113349 03029743. doi: 10.1007/978-3-319-49409-8_32.

[49] F L Schmidt and J E Hunter. The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological bulletin*, 124 (2):262–274, 1998. ISSN 0033-2909. doi: Doi10.1037//0033-2909.124.2.262.

[50] H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E.P. Seligman, and Lyle H. Ungar. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*, 8(9), 2013. ISSN 19326203. doi: 10.1371/journal.pone.0073791.

[51] Galit Shmueli. To explain or to predict? *Statistical Science*, 25:289–310, 2010. ISSN 0883-4237. doi: 10.1214/10-STS330.

[52] E A Smith and R J Senter. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories (6570th)*, pages 1–14, 1967.

[53] M Strangelove. *Watching YouTube: Extraordinary Videos by Ordinary People*. Digital futures. University of Toronto Press, 2010. ISBN 9781442641457. URL https://books.google.nl/books?id=WqMUyFbaMusC.

[54] Ying Li Tian, Takeo Kanade, and Jeffrey F. Conn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, 2001. ISSN 01628828. doi: 10.1109/34.908962.

[55] Daniel B Turban and Felissa K Lee. The role of personality in mentoring relationships. In *The Handbook of Mentoring at Work: Theory, Research, and Practice*, pages 21–50. 2007. ISBN 1452211256.

[56] Paul Viola and Michael J. Jones. Robust Real-Time Face Detection. *International Journal of Computer Vision*, 57(2):137–154, 2004. ISSN 09205691. doi: 10.1023/B:VISI.0000013087.49260.fb.

[57] VisualDNA. Using Visual Questionnaires to Measure Personality Traits. pages 1–10, 2014.

[58] Mirjam Wattenhofer, Roger Wattenhofer, and Zack Zhu. The YouTube Social Network. *Artificial Intelligence*, pages 354–361, 2012. ISSN 07475632. doi: 10.1016/j.chb.2016.09.024.

[59] Dj J Watts and Sh H Strogatz. Collective dynamics of'small-world'networks. *Nature*, 393(6684): 440–442, 1998. ISSN 0028-0836. doi: 10.1038/30918. URL http://202.121.182.16/Course/slides2012/NetSci-2012-7.pdf.

[60] Marie Waung, Robert W. Hymes, and Joy E. Beatty. The Effects of Video and Paper Resumes on Assessments of Personality, Applied Social Skills, Mental Capability, and Resume Outcomes. *Basic and Applied Social Psychology*, 36(3):238–251, 2014. ISSN 0197-3533. doi: 10.1080/01973533.2014.894477. URL http://www.tandfonline.com/doi/abs/10.1080/01973533.2014.894477.

[61] Achmadnoer Sukma Wicaksana and Cynthia C S Liem. Human-Explainable Features for Job Candidate Screening Prediction. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1664–1669, 2017. doi: 10.1109/CVPRW.2017.212.

[62] Janine Willis and Alexander Todorov. First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, 17(7):592–598, 2006. ISSN 09567976. doi: 10.1111/j.1467-9280.2006.01750.x.