UNIVERSITY OF CAMBRIDGE

Assessment of individual differences in online social networks using machine learning

Arman Idani

Department of Psychology

This dissertation is submitted for the degree of
Doctor of Philosophy

Wolfson College

September 2017

I would like to dedicate this thesis to my loving parents and brother …

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text. This dissertation contains about 51,000 words including appendices, bibliography, footnotes and equations and has about 65 tables and figures.

Arman Idani
September 2017

Acknowledgements

I cannot express enough gratitude to my supervisor Professor John Rust for his continuous support, patience and encouragement. This work would not have been possible without his vision and intellectual support. Thank you to Dr Pushmeet Kohli, my advisor at Microsoft Research for his mentorship, advice and support, without which this project would not have been a possibility. This work was supported by a generous scholarship from Microsoft Research and accompanied by guidance and access to Microsoft Research facilities and research resources. I would like to thank Microsoft Research for their generosity and support.

I would like to thank Dr David Stillwell for his friendship, patience, advice and guidance at various stages of this project. David helped me bridge the gap between computer science and psychometrics and has been my most influential advisor for the past four years. I also thank Dr Michal Kosinski for his generous advice and mentorship. Data collected by David and Michal made this project a possibility. My thanks goes to Dr Eiko Yoneki for her support, advice and offer to keep me involved with the Computer Laboratory during my PhD.

My gratitude goes to my tutors at Wolfson College, Dr Anna Bagnoli and Dr Jane McLarty for their guidance, moral support, motivation and continuous encouragement. I thank my friends and colleagues at University of Cambridge, especially Aiden, Arielle, Bartosz, Fiona, Gabriela, Igor, Iva, Joseph, Kalifa, Koen, Luning, Richard, Sandra, Srivigneshwar, Vaishali, Vesselin, Youyou and Zhonghua for the great time that we spent together. The value of the friendships extend far beyond what can be written on papers.

I would like to thank my examiners, Professor Cecilia Mascolo and Dr Almuth McDowall for their time, critical feedback and suggestions which helped me improve the presentation of this thesis.

Finally and most importantly, I thank my loving family whose unconditional care, love, support and attention has enabled me to focus on my studies, interests and ambitions. I am dedicating this thesis to my father, mother and brother.

Abstract

The services that define our personal and professional lives are increasingly accessed through digital devices, which store extensive records of our behaviour. An individual's psychological profile can be accurately assessed using offline behaviour, and I investigate if an automated machine learning system can measure the same psychological factors, only from observing the footprints of online behaviour, without observing any offline behaviour or any direct input from the individual.

Prior research shows that psychological traits such as personality can be predicted using these digital footprints, although current state-of-the-art accuracy is below psychometric standards of reliability and self-reports consistently outperform machine-ratings in external validity. I introduce a new machine learning system that is capable of doing five-factor personality assessments, as well as other psychological assessments, from online data as accurately as self-report questionnaires in terms of reliability, internal consistency and external and discriminant validity, and demonstrate that passive psychological assessment can be a realistic option in addition to self-report questionnaires for both research and practice.

Achieving this goal is not possible using conventional dimensionality reduction and linear regression models. Here I develop a supervised dimensionality reduction method capable of intelligently selecting only useful parts of data for the relevant prediction at hand which also does not lose variance when eliminating redundancies. In the learning stage, instead of linear regression models, I use an ensemble of decision trees which are able to distinguish scenarios where the same observations on digital data can mean different things for different individuals.

This work highlights the interesting idea that similar to how a human expert who is able to assess personality from offline behaviour, an expert machine learning system is able to assess personality from online behaviour. It also demonstrates that big-5 personality are predictors of how predictable users are in social media, with neuroticism having the greatest correlation with unpredictability, while openness having the greatest correlation with predictability.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

It is not yet possible to directly measure psychological latent traits, such as personality; instead, the vast majority of research and practice use self-report psychometric questionnaires. These questionnaires are very effective in predicting important life outcomes [106, 121], such as happiness, spirituality, physical health, peer, family and romantic relationships, occupational choice and performance, political and ideological values, and criminality. As a result, self-report questionnaires continue to be very widely used for both research and in practice.

However, self-report psychometric questionnaires are disadvantageous for two reasons. First, the reliability of a psychometric questionnaire is directly related to the length of the questionnaire. Measuring a broad trait such as conscientiousness typically takes at least 10 questions [58]. As a result, to measure multiple traits researchers often ask participants to complete hundreds of questions. The inventories used in business are also long, auch as NEO-PI-R (240 questions) [29] or MMPI [59]. Brief questionnaires for various psychometric tests do exist however they often have a narrower focus than longer questionnaires. Brief questionnaires with longer focus tend to not be as psychometrically strong as longer questionnaires. Several brief personality questionnaires are developed, such as FIPI with 5 items and TIPI with 10 items [57], and BFI-10 [120]. These tests are not as accurate as longer scales, such as NEO-PI-R or BFI-44 [78] in terms of test-retest reliability, internal consistency (in multi-item tests), and external and discriminant validity. Therefore, brief measures are only used when there are severe time constraints.

The second major disadvantage of self-report psychometric questionnaires is that they are prone to biases and inaccuracies. For instance in recruitment, job-seekers are incentivised to self-enhance to present themselves in a positive light when the outcome depends upon the questionnaire's results [21, 150]. This is known as the self-enhancement bias [89]. Memory is also imperfect, people more easily recall more recent and primary events compared to events that occur in the middle [102]. Also, a test-taker's mood influences their result. There is also a reference-group effect where test takers compare themselves with their own social circle rather than the general population [100].

Personality are patterns of thought and behaviour that are stable over time and distinguish an individual from another [116] and is very informative of an individual. Five-factor personality model, also known as the *Big 5 personality*, is the most widely-used and accepted form of personality representation [26, 53, 138] and has a very strong predictive power of important behaviour and life outcomes [106]. Online social networks store records of user behaviour in the form of pages that users like, people they make friendships with, public figures they follow and interactions they make with each other. So I hypothesise that a machine learning system can make accurate judgements of personality traits by looking at these footprints. In this thesis I perform studies to investigate this hypothesis.

This thesis focuses on development and critical analysis of new proof-of-concept psychometric tests that do not require any active participation from the test-taker. The judgement is done using machine learning to learn patterns from data that the users leave behind on Facebook that can be indicative of how users would respond to a psychometric questionnaire. The tests include five-factor personality, satisfaction with life, depression and self-monitoring. These traits correlate well with life outcomes and online behaviour. Critical analysis of these tests include comparison of machine-rated psychological traits with self-reports in terms of reliability, internal consistency, discriminant validity and external validity.

1.1 Introducing passive psychometrics

Ideally, when accurate psychological assessment is needed, a multi-item well established test from highly valid and reliable inventories should be used. However in reality, longer tests are less convenient and users, especially online, do not stay on a website long enough to respond to long questionnaires [122]. When assessment is part of a service, such as personalising a website or providing recommendations to users, asking users to complete psychometric questionnaires is not optimal in terms of user experience. Users lose interest on services when the initial setup time is longer.

Passive psychometrics is defined as psychological assessments which do not involve active participation from the test-taker and judgements are done only by looking at the data left behind by the user. In the context of this thesis, passive psychometrics is done by making predictions of psychological traits from the digital data that users leave on online social networks, specifically Facebook [1]. Accurate passive psychometrics solves the problem of length with self-report questionnaires, as they do not require active involvement of the participant at all. This also means they do not allow self-enhancement bias, memory and mood effects to influence the prediction of latent traits, as data is left on social networks for a long period of time. In terms of research, making psychometrics faster makes data collection faster, easier to scale and cheaper to run.

Prediction of psychological traits from digital data is not a new concept. Large amounts of data are stored on major social networks which give valuable insight of the way people use such services and various studies [86, 50, 117, 152, 109, 146, 15, 131, 30] have demonstrated that individual differences such as personality, well-being, political views, religious beliefs, values, depression, history of drug use, addiction, and demographics are predictable from digital footprints that users leave on social networks such as Facebook, Twitter [2] and Qzone [3]. While these studies demonstrate that such predictions are possible and are often useful, they do not measure such traits as accurately as long self-report questionnaires.

Youyou et al. [152] demonstrated the most accurate passive assessment of five-factor personality model from the pages that users like on Facebook, with mean correlation of $r = .52$ with self-reports for the five-factor personality model, split-half correlations of $r = .62$ and a lower predictive power of external variables compared to self-reports. These are accurate predictions, however they do not reach the psychometric standards of self-reports in terms of reliability, internal consistency and external validity. Goldberg et al. [50] predicted personality from Facebook pages that the user has liked, status updates and all other personal and friends data, with mean correlations of $r = .56$ with self-reports, however did not report any measurements on internal consistency, discriminant or external validity and used a very small sample.

In this thesis I will introduce methodology which builds upon the existing literature of prediction of desired information from social networks and use them to predict the five-factor personality and will assess the psychometric properties of the assessment in terms of cor-

[1] https://www.facebook.com

[2] https://www.twitter.com

[3] http://qzone.qq.com/

relation with self-reports, internal consistency and external and discriminant validity. This also presents an opportunity to extend the research in the five-factor personality traits, and I will discuss how the traits compare in terms of predictability and their power to explain an individual's predictability.

1.2 Thesis outline

This thesis uses data collected by the MyPersonality project [86]. MyPersonality was a Facebook application which operated from 2007 until 2012 and provided the users with various psychometric tests with feedback. Millions of users on Facebook used the application and consented for their Facebook data to be used for research. In chapter 2, I introduce MyPersonality and its psychometric tests and will introduce the history of relevant concepts to this thesis such as personality traits. There have been prior attempts at prediction of personality traits from Facebook data, some used the same dataset that I do in this thesis. In chapter 3, I examine the findings of these studies, their methodology and I provide a critical analysis.

With success and popularity of online social networks such as Facebook and Twitter, these websites contain a rich collection of data for most of their active users. They include user profiles, status updates, interactions of users with each other, expressions towards data generated by other users (sharing, liking or commenting), expressing interests towards people or pages (following, or liking), and many others. This data is extremely large, usually messy and unstructured. Even after typical cleaning and structuring, the data is still very large, highly sparse and usually a large portion of it provide little value. Chapter 4 addresses challenges in using large datasets from social network for prediction of valuable information, it reviews the current methodologies of reducing dimensions of large data, explains their shortcomings and introduces a new method of dimensionality reduction and feature selection that I name *Entropic Component Allocation*, or *ECA* that removes redundancy, noise and preserves a high degree of relevant variance to the task at hand.

Reduced dimensioned data needs to be matched with personality scores from the MyPersonality project to train the predictive models. The trained models will then be able to provide accurate predictions based on the patterns that they have learned from the data. Chapter 5 explains how the predictive models work, the design decisions and how each one can contribute to a better prediction. This model is compared with models in literature and other alternative choices such as deep learning. This model, combined with reduced dimensioned data using ECA is then used to predict five-factor personality scores for users of Facebook, based on the pages that they have liked.

Chapter 6 assesses the validity of the personality predictions as a psychometric test using the same metrics that are used to assess new self-report psychometrics tests: reliability, internal consistency, external validity and discriminant validity. Not all personality traits can be predicted with the same level of accuracy, and not all personality traits offer the same predictive power to explain why a user is more or less predictable. In chapter 7, I explore these differences in detail and run further studies in order to investigate why these observations are the way they are. I also investigate the per-item predictability of the five-factor models to learn answers to which items are more predictable and answers to which items are the hardest to predict.

To investigate if other commonly used outcome variable are also predictable from data from Facebook, in chapter 8 I expand the scope of the thesis and focus on developing three new passive psychometric tests for self-monitoring, depression and satisfaction with life, using the same methodology that I developed for the passive five-factor personality test. Personality itself is often used to predict these variables and I investigate if they can be accurately assessed directly from pages that users like on Facebook. I examine the validity of these passive tests by measuring their agreement with self-reports, by examining their internal consistency and external validity.

Chapter 9 provides general discussion about the findings in prior chapters, it also outlines limitations of the passive method compared to self-reports, as well as concerns for privacy, and discusses an outlook for how passive psychometrics can be incorporated into both research and practice, alongside traditional self-report questionnaires. I also discuss the implications of this work for personality theory, how these findings can add to our understanding of the five-factor model.

Finally, chapter 10 summarises the contributions of this thesis.

Chapter 2

Data Source: MyPersonality

MyPersonality was a Facebook application developed by Dr David Stillwell and Dr Michal Kosinski [86, 85], initially released in 2007 and collected data until 2012. The application provided well-known psychometric tests to users of Facebook and offered immediate feedback, as well as offering the users with the option to opt into sharing their Facebook data for research. Over 7.5 million users completed at least one psychometric test on MyPersonality, and over 3.5 million users shared information on their Facebook profiles with the application.

Data collected from the profile of Facebook users include the user's demographics (age, gender, relationship status, location, hometown, religious views, political views and education), their Facebook network (friendships), the pages that the user has liked on Facebook (referred to as Facebook Likes) and their status updates.

Psychometric tests provided by the MyPesonality application to Facebook users include a five-factor personality test, IQ test, satisfaction with life, Schwartz's Values, CES-D depression, self-monitoring, delay discounting, sensational interests, and morality foundations.

Creation of MyPersonality application, collection and validation of the dataset have been done by David Stillwell and Michal Kosinski, and are not part of this dissertation. All analysis, predictive models in the following sections and chapters and all text in this chapter are produced by Arman Idani.

2.1 Personality

Personality are patterns of thought and behaviour that are stable over time and distinguish an individual from another [116] and is very informative of an individual. The way to assess personality has not always been clear. Even in modern times we are still unable to directly measure psychological traits such as personality, in a way that we measure a lot of physical characteristics such as weight, height or blood pressure. We have not yet invented an MRI machine for personality [1]. Therefore, there has always been a debate about how personality should be measured.

Early attempts at measuring personality focused mostly on physical measurements. In fact, there was a widely subscribed idea during the 1800s called *phrenology* which claimed that measurements of a human skull are the key to unlocking the mysteries how the individual's mind works and how individuals differ from one another. Of course, phrenology is now widely considered to be a pseudoscience. Another historical attempt at measuring personality was centred on physical appearance. Sheldon et al. [129] argued that body types determine behaviours and interests such as being assertive or intellectual. This is also now widely considered to be a pseudoscience.

More recently however, personality is measured and discussed in terms of traits [3, 24]. Traits can be seen as measurements of personality that are stable throughout an individual's life. To make personality traits as useful and informative as possible, they should ideally be independent from each other and correlate well with behaviour and life outcomes. Together, traits are able to predict how an individual behaves and thinks in various circumstances and explain differences in behaviour among different individuals.

Allport [4], one of the pioneers of trait theory, looked through words commonly used in language to describe people and grouped them into three families, cardinal, central and secondary traits, each describing the level of importance of each trait. However, they were still a very large number of overlapping traits. In the following years, by removing redundancy and looking for latent factors, modern models of personality were developed. For a personality trait to be useful, it needs to be reliably measurable across different people varying in age, gender, culture or life experience, and be able to predict their behaviour, thoughts and outcomes.

[1] This is an area of research, however [143, 22].

In recent decades, the five-factor personality model has been widely accepted as the preferred form of representing personality [51].

2.1.1 Five-factor personality model

Five-factor personality model, also known as the *Big 5 personality*, is the most widely-used and accepted form of personality representation [26, 53, 138] and has a very strong predictive power of important life outcomes [106]. The five personality traits are shown to be relatively stable overtime [28, 27] and are found to be universal among multiple cultures [99]. The five traits are called *Openness to experience*, *Conscientiousness*, *Extraversion*, *Agreeableness* and *Neuroticism*.

Openness to experience

Openness to experience shows an appreciation for new experiences, adventures, curiosity and ideas. It represents a need for variety and novelty. Individuals with high openness to experience have more diverse interests and are more creative, and individuals with low openness to experience show interest towards more familiar and known experiences. Some word markers used to describe openness are *creative vs. uncreative*, *curious vs. uninquisitive*, *knowledgeable vs. ignorant* and *deep vs. shallow* [52].

In terms of social network usage, individuals with higher levels of openness tend to share more information about themselves [6], and are move likely to have online activities such as blogging, posts to page timelines and general social media use [25, 56]. As a hypothesis, I expect openness to be correlated with predictability, as having more activities on social networks will leave more digital footprints, which makes the data that will be used for predictions larger and richer.

Conscientiousness

Conscientiousness shows a need for being efficient, disciplined, neat and thorough. Individuals with high conscientiousness show desire to perform a task well, are more organised in their daily lives and are more successful in academic and workplace performance, and have higher job satisfaction [69, 79]. Some word markers used to describe conscientiousness are *organised vs. disorganised*, *economical vs. wasteful*, *thorough vs. careless*, *cautious vs rash* and *reliable vs unreliable* [52].

In terms of social network usage, individuals with higher levels of conscientiousness use social networks less often [149, 125]. This gives us the expectation that conscientiousness will be negatively correlated with predictability.

Extraversion

Extraversion shows energetic behaviour, being talkative and outgoing. Extraverts have a preference for social stimulation and are generally gregarious. They usually have good social skills, have a lot of friends, and enjoy social experiences such as parties, demonstrations and games. Individuals with low extraversion (i.e. introverts) have a more solitary behaviour and prefer activities such as reading, writing and fishing. Some word markers used to describe extraversion are *talkative vs. silent*, *assertive vs. compliant*, *demonstrative vs. reserved*, *enthusiastic vs apathetic* and *energetic vs lethargic* [52].

In terms of social network usage, extraverts use Facebook more often than introverts [149, 56]. This also can indicate that extraverts, similar to individuals with high degrees of openness, are easier to predict than introverts.

Agreeableness

Agreeableness is the tendency to be cooperative and loving with others rather than being suspicious and competitive. People with high agreeableness have a forgiving attitude and compliant behaviour, meaning that they get along well with others. They consider intention of others to be more noble and trust people easier. Those scoring low on agreeableness tend to be more suspicions of others and consider them more dishonest. Some word markers used to describe agreeableness are *warm vs. cold*, *kind vs. unkind*, *flexible vs stubborn*, *cooperative vs. uncooperative* and *agreeable vs. quarrelsome* [52].

In terms of social network usage, agreeableness does not correlate with more or less usage of social networks [123]. Agreeable individuals are however, more likely to present a truer picture of themselves [91].

Neuroticism

Neuroticism is the tendency to experience negative emotions such as depression, anger, anxiety and moodiness. Individuals with high neuroticism are more susceptible to stress and easily feel hopeless, irritated, worried and threatened. They generally have low self-esteem, pessimistic attitudes and irrational beliefs. Some word markers used to describe neuroticism are *relaxed vs. nervous*, *peaceful vs. volatile*, *uninhibited vs fearful*, *imperturbable vs high-strung* and *unemotional vs emotional* [52].

In terms of social network usage, neuroticism is positively correlated with acceptance seeking on Facebook [127] and neurotic people have a tendency to present a more ideal image of themselves on Facebook [91]. I hypothesise that neurotic individuals are more difficult to predict than individuals with low neuroticism, because manipulating your Facebook data to present a different image adds bias, which makes the data less useful for predictions.

All the hypotheses and expectations in this section will be tested against observed data in chapter 7.

2.1.2 Personality assessment

Self-reports are the most widely-used method of five-factor personality assessment. They are highly reliable longitudinally [27], are internally consistent [51] and have excellent power in prediction of external factors and life outcomes [121].

Traditionally, most psychological assessments were done using pen-and-paper questionnaires in face-to-face settings. These tests required the test-giver to fill the questionnaire, and an administrator to score and interpret the results for the test-taker. With the invention and widespread adoption of computers in academic institutions in the 20th century, they became accessible to psychologists and psychometricians as tools for data analysis and computations. The first step of incorporating computers into psychological assessment was the introduction of Computer-based Test Interpretations, also known as CBTIs [46]. This is when the computer took over the interpretation part of the job from the administrator. At the time computers did not even have monitors, and the interpretations were done by scanning input data from cards and printing the scores and descriptive statements on paper. The first of such systems was developed at the Mayo Clinic [2] which computerised the MMPI personality test [112].

When computers became faster and more user-friendly, they began to be able to take over the administrative part of the test as well. Elmwood et al. [37] developed a test for assessing intelligence without an examiner in 1972. This was the first online computerised test, the first time that computers took over the role of the test administrator completely. Shortly later in 1974, Lushene et al. [96] developed an online MMPI personality test. This test demonstrated excellent correlation between online and pen-and-paper tests and excellent test-retest reliability.

Computerising psychological assessment provided a lot of benefits compared to traditional pen-and-paper tests. The psychometricians no longer needed to spend any time scoring and interpreting the results of a test manually for each individual, and this made psychological assessment cheaper, faster and easier to scale. Computerised tests provided new opportunities to make the psychological assessment process even more efficient. One of the major progress in this area is the introduction of Computer Adaptive Testing, or CAT [144], where the questions adapt to the individual's answers. CATs use paradigms such as Item Response Theory in order to present the individual with an item that captures the maximum amount of information useful to measure the required trait while avoiding items where the amount of information captured is minimum [38].

While in the 20th century, automation of personality assessment actually meant that the role of the administrator is automated, passive psychometrics aims to automate the role of both administrator and test-taker. Computerising psychological assessment provided a lot of benefits compared to traditional pen-and-paper tests. The psychometricians no longer needed to spend any time scoring and interpreting the results of a test manually, and this made psychological assessment cheaper, faster and easier to scale. Passive psychometrics is a progress in the same direction, by not asking the test-taker of any commitment of their time or energy, the assessment can be even cheaper, even faster and even easier to scale.

Passive psychometrics is the complete automation of personality assessment.

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

2.2 MyPersonality's assessment of five-factor personality

MyPersonality used the International Personality Item Pool (IPIP) proxy [55] for the NEO-PI-R questionnaire [29] to provide a test for five-factor personality model. Questionnaires of 20 items (4 items per trait) to 100 items (20 items per trait) were presented to users.

Users responded to questionnaires by choosing one of multiple-choice answers. Table 2.1 shows potential responses to an item on MyPersonality's personality test.

I tend to vote for liberal political candidates...				
Very inaccurate	Somewhat inaccurate	Neither inaccurate nor accurate	Somewhat accurate	Very accurate

Table 2.1 Potential responses to an item on the personality test

2.2.1 Openness to experience

The 20 items used to assess the openness to experience trait are outlined in table 2.2.

Scoring is done by assigning a scale of 1 to 5 to potential answers. When necessary, items are keyed negatively. The personality scores are calculated as the mean of the scores for each individual item.

There are several items present from the IPIP artistic sensibilities scale that capture the aesthetics NEO-PI-R facet of openness (items 5-8). The fantasy facet is also captured in items 1 and 2, which are present on the IPIP imagination scale. There are three items that capture the values facet (items 18-20). An individuals sense of adventurousness is also captured in items 9 and 10, which relate to the actions facet of openness. Items 11-17 are from the intellect scale and capture the ideas facet. However, these items do not capture the feelings facet, as they include no items from the emotionality scale. This shows that the openness trait captured by these items are more weighed towards values, intellect and aesthetics facets compared to the feelings and actions facets.

The 20-item IPIP openness scale correlates highly with the NEO-PI-R scale, at $r = .92$ [54].

Number	*Item*
1	Have a vivid imagination
2	Enjoy wild flights of fantasy
3	Rarely look for a deeper meaning in things
4	Believe in the importance of art
5	Do not like art
6	Do not like poetry
7	Do not enjoy going to art museums
8	Can say things beautifully
9	Get excited by new ideas
10	Enjoy hearing new ideas
11	Have difficulty understanding abstract ideas
12	Enjoy thinking about things
13	Avoid philosophical discussions
14	Am not interested in theoretical discussions
15	Am not interested in abstract ideas
16	Have a rich vocabulary
17	Carry the conversation to a higher level
18	Tend to vote for conservative political candidates
19	Tend to vote for liberal political candidates
20	Believe that too much tax money goes to support artists

Table 2.2 MyPersonality's items to assess *Openness to experience*

2.2.2 Conscientiousness

The 20 items used to assess the conscientiousness trait are outlined in table 2.3.

Scoring is done by assigning a scale of 1 to 5 to potential answers. When necessary, items are keyed negatively. The personality scores are calculated as the mean of the scores for each individual item.

IPIP Scale	*Item*
1	Complete tasks successfully
2	Follow through with my plans
3	Finish what I start
4	Don't see things through
5	Mess things up
6	Make a mess of things
7	Do things according to a plan
8	Make plans and stick to them
9	Shirk my duties
10	Leave things unfinished
11	exacting in my work
12	Do just enough work to get by
13	Waste my time
14	Find it difficult to get down to work
15	Get chores done right away
16	Need a push to get started
17	Am always prepared
18	Carry out my plans
19	Pay attention to details
20	Don't put my mind on the task at hand

Table 2.3 MyPersonality's items to assess *Conscientiousness*

All six NEO-PI-R facets of conscientiousness are captured by the 20 items. The self-efficiency scale (items 1-4) measures the competence facet of conscientiousness. The order facet is measured using the orderliness scale, by items 5-8. Items 9 and 11 measure dutifulness while item 10 measures achievement starving. Self-discipline is measured by items 13-18. items 19 and 20 measure the cautiousness facet. While there are items measuring all facets, the measured conscientiousness is weight more towards self-discipline and self-efficiency compared to achievement starving and cautiousness.

The 20-item IPIP conscientiousness scale correlates highly with the NEO-PI-R scale, at $r =$ .88 [54].

2.2.3 Extraversion

The 20 items used to assess the extraversion trait are outlined in table 2.4.

Scoring is done by assigning a scale of 1 to 5 to potential answers. When necessary, items are keyed negatively. The personality scores are calculated as the mean of the scores for each individual item.

Number	*Item*
1	Feel comfortable around people
2	Make friends easily
3	Am hard to get to know
4	Keep others at a distance
5	Cheer people up
6	Warm up quickly to others
7	Avoid contact with others
8	Talk to a lot of different people at parties
9	Am the life of the party
10	Don't talk a lot
11	Start conversations
12	Keep in the background
13	Have little to say
14	Don't like to draw attention to myself
15	Know how to captivate people
16	Find it difficult to approach others
17	Am skilled in handling social situations
18	Would describe my experiences as somewhat dull
19	Do not mind being the centre of attention
20	Retreat from others

Table 2.4 MyPersonality's items to assess *Extraversion*

There are several items from the friendliness scale that capture the warmth facet of extraversion such as items 1-6. Items 7-11 capture gregariousness. An individual's assertiveness is captured in items 12-14 and 20. Positive emotions facet is captured in items 15, 16 and 18.

The activity level and excitement seeking facets are less present in the items, although some items (such as item 9) can be seen as multifaceted as well.

The 20-item IPIP extraversion scale correlates highly with the NEO-PI-R scale, at $r = .88$ [54].

2.2.4 Agreeableness

The 20 items used to assess the agreeableness trait are outlined in table 2.5.

Number	*Item*
1	Believe that I am better than others
2	Insult people
3	Get back at others
4	Am out for my own personal gain
5	Have a sharp tongue
6	Trust what people say
7	Make people feel at ease
8	Have a good word for everyone
9	Contradict others
10	Respect others
11	Sympathise with others feelings
12	Am concerned about others
13	Suspect hidden motives in others
14	Cut others to pieces
15	Treat all people equally
16	Believe that others have good intentions
17	Hold a grudge
18	Make demands on others
19	Accept people as they are
20	Am easy to satisfy

Table 2.5 MyPersonality's items to assess *Agreeableness*

Scoring is done by assigning a scale of 1 to 5 to potential answers. When necessary, items are keyed negatively. The personality scores are calculated as the mean of the scores for each individual item.

The trust facet of agreeableness is captured by items on the trust scale of the IPIP item pool (items 6, 13 and 16). The altruism facet is captured by items 7, 8 and 12. The compliance facet is captured by the items from the cooperation scale (items 3, 5, 9, 14, 17, 18 and 20). Items 1 captures the modesty facet. Items 11, 12 and 15 capture the tender-mindedness facet of agreeableness. The morality facet is not explicitly captured, although item 15 can be seen as relevant. This demonstrates that the test is weighted more towards the cooperation and sympathy facets of agreeableness than the morality or modesty facets.

The 20-item IPIP agreeableness scale correlates highly with the NEO-PI-R scale, at $r = .90$ [54].

2.2.5 Neuroticism

The 20 items used to assess the neuroticism trait are outlined in table 2.6.

Scoring is done by assigning a scale of 1 to 5 to potential answers. When necessary, items are keyed negatively. The personality scores are calculated as the mean of the scores for each individual item.

Items 8, 13, 14 and 20 capture the anxiety facet of neuroticism. Anger hostility facet is represented by items 10, 15 and 18. Items 2-4, 7, 9, 17 and 19 measure the depression facet. The vulnerability facet is measured by items 1, 12 an 16. This shows that the test is more weighed towards measuring depression, anxiety and anger scales of neuroticism compared to impulsivity and self-consciousness.

The 20-item IPIP neuroticism scale correlates highly with the NEO-PI-R scale, at $r = .93$ [54].

Number	*Item*
1	Panic easily
2	Dislike myself
3	Have frequent mood swings
4	Am often down in the dumps
5	Am filled with doubts about things
6	Remain calm under pressure
7	Feel comfortable with myself
8	Worry about things
9	Am very pleased with myself
10	Rarely get irritated
11	Feel threatened easily
12	Am not easily frustrated
13	Fear for the worst
14	Am relaxed most of the time
15	Seldom get mad
16	Get stressed out easily
17	Often feel blue
18	Rarely lose my composure
19	Seldom feel blue
20	Am not easily bothered by things

Table 2.6 MyPersonality's items to assess *Neuroticism*

2.2.6 MyPersonality's internal consistency

Table 2.7 shows the comparison of the internal consistency of the five-factor personality scores of the MyPersonality project, reported by Kosinski [84] compared to the 100-item IPIP, as reported by author [54]. The sample size of MyPersonality was 182,922 users who responded to the 100-item questionnaire.

	Cronbach's Alpha Reliability	
Trait	*MyPersonality*	*IPIP*
Openness	.85	.89
Conscientiousness	.92	.90
Extraversion	.93	.91
Agreeableness	.88	.85
Neuroticism	.93	.91

Table 2.7 Cronbach's Alpha Reliability of the MyPersonality dataset [84] compared to IPIP as reported by author [54].

The measurements of personality by MyPersonality are as internally consistent as measurements by Goldberg et al. [54].

2.3 Inclusion criteria

Overall, 182,922 users responded to a 100-item questionnaire (20 items per trait) and their personality scores have been computed based on their responses to the items on the questionnaire. The criteria for inclusion of users in the upcoming studies in this thesis was based on the number of pages that they liked, referred to as Facebook Likes. Only users with more than 100 Likes were included. This is to ensure that users without much experience on the platform are not picked. The average number of Liker per included user is 196, which is less than the average number of Likes of Facebook users, as reported by Youyou et al. [152], so in a similar practical test most users can be included.

Users who did not have any information about their age and gender were also excluded. The demographics collected for the users are: Age, gender, relationship status, education, location, hometown, and number of friends. Overall, 98,515 worldwide volunteers were passed the inclusion criteria and were used in the studies in upcoming chapters.

2.4 Ethics of using online data

Collection of data on a large scale from social networks presents challenging questions in terms of ethics and respecting the privacy of users, and issues relating to consent.

The goal of all social networks is to provide a medium which facilitates the sharing of information between individuals. More people are using social networks now than ever before. With over 2 billion active users on Facebook [42] and record highs on other social networks [139, 76], more and more footprints of behaviour are being left online at a lightning fast rate and at a great volume. Data on social networks is now becoming the largest and perhaps even most detailed records of human behaviour in existence, as users spend an average of two hours per day on social media and messaging applications and websites in 2017 [49].

Social networks are often considered to be the natural custodians of all of these records, keeping them safe in their data centres. This presents an equilibrium that allows the users to continue having access to services and features that they rely on, in exchange for the social network to use their data in in order to present them targeted advertisement to generate the revenue they need to run the social network and be profitable. As users continue to use social networks at an ever-increasing rate, it is a reasonable conclusion that such a trade-off seems reasonable to most users.

Making online social data sources publicly available is becoming more and more common. We can use data from social networks to predict very sensitive information from users, which is the central aim of this thesis, which also has been shown in literature [86]. Therefore, making a dataset public can practically be the same as sharing a list of individuals with their private and sensitive information on the internet. The individual might have consented for their data to be used for research however they often are not aware of how rich and informative of their private information that data is.

Recently, security breaches in social networks have become justifiably publicised and users are increasingly worried about their data finding its way into the public domain and into the wrong hands, with great concern over the release of sensitive information (such as name, demographics, phone number, address and passwords) and a rather lack of focus on footprints, usually available publicly, that can be used to predict many of the same sensitive information very accurately. Recent papers [86, 152] have brought public and media attention to this overlooked area of online privacy.

However, there are research advantages in making data from social networks publicly available. As the data is usually large, analysis and interpreting the data is often a major task and it is benefited when more researchers have access to the data. Consent remains a complex and difficult to explore subject of discussion about data from social networks. I argue that a traditional understanding of consent in research involving human subjects cannot apply to research done on footprints of users on social networks. Currently, consent is usually done to the time of data collection. The users allow their data to be used for research and researchers will feel free to use it for any research in the future, or to share it with other researchers. This is probably the only practical way when it comes to data from social networks. It is unfeasible for researchers to ask tens or hundreds of thousands of users for consent before doing every single study on the same data.

One way to address this challenge is to process the data in a way that reconstruction of the identifiable information becomes impossible or very computationally intensive. These processes include anonymising and randomising of the datasets. I argue that this should be done at the time of collection and it should be standard in all social network data collection tasks and ethics guidelines should emphasise its importance for research on social networks and online data collection in general. This way, the users can consent their data to be used for a wide range of research while having the peace of mind that what is learned from the data will not be released with identifiable information that can track back to them. This process is discussed in further detail in section 9.4.3.

The American Psychological Association [3] [87] and the American Association for the Advancement of Science [4] [5] have guidelines about internet-based research involving human subjects, however issues relating to the challenging aspects of consent and general privacy of users are not discussed. These guidelines were published in 2004 and 1999 respectively, belonging to a different era of the internet before widespread popularity of social networks. The British Psychological Association [5] has updated their ethics guidelines about internet-mediated research in 2017 [135] where issues related to consent are discussed. They suggest, in accordance to their own guidelines about non-internet based research involving human subjects [134], that consent needs to be sought for all observations of non-public behaviour.

The classification of online data as public or private is not straight-forward. Offline behaviour is often considered public when it is expected to be observed by strangers [134]. However,

[3] http://www.apa.org

[4] https://www.aaas.org

[5] http://www.bps.org.uk

individuals usually expect a degree of privacy on social networks. When a user shares something on social media to their friends, the post is still visible to the social network itself, who are technically strangers. There are also legal considerations, as often footprints stored on social networks or online discussion forums are legally owned by the social network itself. This is not the case for Facebook [41], however it is the case for Twitter who claims legal ownership of the tweets that people send [140]. Can individuals consent to sharing data for research that they don not legally own? Can the social network company itself release data that they do not own for research? Should they release data that they own, without consulting the users? These questions further demonstrate the complexities of the issue of consent in internet-based research involving human subjects.

2.5 Summary

In this chapter I introduced MyPersonality, a Facebook application developed by Dr David Stillwell and Dr Michal Kosinski. It offered psychometric tests to Facebook users and provided feedback. It operated for about 6 years and millions of users took the tests and opted into sharing data on their Facebook profiles for research. MyPersonality is a dataset that represents Facebook users well in terms of demographics, and its psychometric tests are shown to be as reliable and valid as traditional pen-and-paper tests.

Later in this thesis, I use the MyPersonality dataset to evaluate the methods introduced in chapters 4 and 5. The passive personality test, introduced and examined in chapter 6, as well as the self-monitoring, depression and satisfaction with life tests introduced in chapter 8 use data from the MyPersonality project.

Chapter 3

Literature Review

In this chapter, I review the literature of passive personality assessment and explain the published work that focus on prediction of personality traits from social media.

The first major work of predicting five factor personalty traits from Facebook was done by Kosinski et al. [86]. This was done as part of MyPersonality project, which is the same data used in this thesis. This study used a narrower inclusion criteria and limited their users to American ones. Therefore, their sample size was limited to 58,000 users. There is however a large amount of overlap between the users in this thesis and the users in Kosinski's study. Table 3.1 reports the correlation between the predicted personality traits in Kosinski's study and the self-report personality scores.

Trait	*Correlations*
Openness	.43
Conscientiousness	.29
Extraversion	.40
Agreeableness	.30
Neuroticism	.30

Table 3.1 Correlations between predicted personality scores and self-report personality scores, reported by Kosinski et al. [86].

In terms of methodology, this study reduced the dimensioned of the Facebook Likes using Singular Value Decomposition [32] and kept the 100 most informative components. The reduced dimensioned data was used to train a linear regression model which made the predictions. This study was the first of its kind to investigate if Facebook data can be used to predict a wide range of private traits and other attributes, such as age, gender, sexual orientation, intelligence, satisfaction with lie and substance use from a wide range of users (58,000). They demonstrated that Facebook as a social communications platform is also a useful source of information for psychological assessment. They did not study further psychometric properties of the predicted personality scores, such as internal consistency, external validity or discriminant validity.

Youyou et al. [152] further studied the MyPersonality dataset and examined how accurate predicted personality is compared to judgements made by friends, co-workers and spouse. They found their predictions to be as accurate as a spouse judgement of personality, and less accurate than self-reports. This study is also limited to Facebook users in the United States and their sample size is 86,220 throughout the study. There is a large overlap between the users in this study and the users studied in this thesis. Table 3.2 reports the correlation between the predicted personality traits in Youyou's study and the self-report personality scores, for users with more than 200 Facebook Likes.

Trait	*Correlations*
Openness	.65
Conscientiousness	.52
Extraversion	.55
Agreeableness	.56
Neuroticism	.49

Table 3.2 Correlations between predicted personality scores and self-report personality scores, reported by Youyou et al. [152].

In terms of methodology, this study used a LASSO feature selection method embedded with a linear regression model. LASSO method is useful at extracting non-zero parameters from highly sparse datasets and demonstrated an improvement over the unsupervised dimensionality reduction method used by Kosinski et al. [86]. With mean correlation of $r = .56$ between predicted personality traits and self-report personality traits, this study demonstrated that improvements in methodology can lead to more accurate predictions of personality from online social networks. This study also reported internal consistency of the predicted five-factor personality, at mean split-half correlations of $r = .62$ for the five personality traits. In terms of external validity, the predicted personality scores were less able to predict outcome variables such as values, sensational interests, life satisfaction and self-monitoring compared to self-report personality.

Park et al. [109] also studied MyPersonality dataset and investigated the possibility of automation of personality assessment by using social media language as the basis of predictions. They captured the status updates of over 66,000 Facebook users from MyPrsonality who had completed the five-factor personality test, performed linguistic feature extraction on the texts that users posted on social media, reduced dimensions and used the features with linear regression models to produce personality predictions, which were compared against the self-reports. This study did not perform cross-validation, however they used separate training and testing subsamples of users. Table 3.3 reports the correlation between the predicted personality traits in Park's study from social media language and the self-report personality scores.

Trait	*Correlations*
Openness	.43
Conscientiousness	.37
Extraversion	.42
Agreeableness	.35
Neuroticism	.35

Table 3.3 Correlations between predicted personality scores and self-report personality scores, reported by Park et al. [109].

This study did not report an assessment of internal consistency of the predicted personality scores, but they reported external and discriminant validity of the predicted personality compared to self-report personality. Table 3.4 shows the discriminant validity of the predicted personality scores from language on social media compared to self-reports.

With a mean discriminant validity of $r = .28$, the personality traits predicted in Park's study are more related to each other compared to self-reports, at mean $r = .18$. Park also reported the external validity of the predicted personality traits compared to the self-report personality, and found that their predictions of personality from language on social media had great agreement with the predictions of self-report personality in terms of predicting external variables. Table 3.5 outlines the level of agreement between the predictive power of predicted personality from language on Facebook and self-report personality in Park's study.

		Self-Reports					*Predicted*				
		O	C	E	A	N	O	C	E	A	N
Self-Reports	O										
	C	.00									
	E	.13	.19								
	A	.07	.17	.19							
	N	-.08	-.31	-.34	-.36						
Predicted	O	**.43**	-.12	-.08	-.05	.00					
	C	-.25	**.37**	.16	.17	-.17	-.25				
	E	-.07	.12	**.42**	.10	-.15	-.17	.33			
	A	-.07	.17	.13	**.35**	-.14	-.12	.44	.27		
	N	.05	-.17	-.18	-.13	**.35**	.06	-.41	-.43	-.34	

Table 3.4 Comparison of the discriminant validity of personality assessed through language on social media with self-report personality, from Park et al. [109]. Values represent correlation (r). O, Openness; C, Conscientiousness; E, Extraversion; A, Agreeableness; N, Neuroticism;

Trait	*Correlations*
Openness	.83
Conscientiousness	.86
Extraversion	.83
Agreeableness	.90
Neuroticism	.96

Table 3.5 Agreement between the predictive power of predicted personality from language on Facebook and self-report personality, reported by Park et al. [109].

MyPersonality is the only publicly available Facebook dataset that contains a wide range of Facebook data of the users as well as the results of their psychometric questionnaires. However, there are smaller studies that have looked into the prediction of personality from social network data. One of the earlier works of predicting big-5 personality from Facebook was done by Golbeck et al. [50]. In this study, they predicted big-5 personality traits from a wide collection of Facebook data. They made a Facebook application to administer a 45-item BFI

[78] questionnaire. The application collected all profile information of the participants including personal information, activities, Facebook network structure (friendships and pages they liked) and status updates. The size of their sample was only 279 users. Table 3.6 reports the correlation between the predicted personality traits in Golbeck's study and the self-report personality scores.

Trait	*Correlations*
Openness	.65
Conscientiousness	.59
Extraversion	.55
Agreeableness	.48
Neuroticism	.53

Table 3.6 Correlations between predicted personality scores and self-report personality scores, reported by Golbeck et al. [50].

At a mean correlation of $r = .56$ between the predicted and self-report personality tests, this is a very large level of agreement. However, there are several important points to consider. First, the number of users were only limited to 279. This is much lower than MyPersonality studies and more prone to include biases. The amount of data used per user was also significantly more. Second, in their methodology description, it appears that they performed dimensionality reduction on all users before splitting the samples into training and testing subsets for their cross-validation. This increases the risk of overfitting, as the dimensionality reduction method can cause data leakage between the training and testing subsets. This suspicion is increased with the knowledge that the study was only limited to 279 users. Moving on from the risk of overfitting in these results, studies by Kosinski et al. [86] and Youyou et al. [152] primarily focused on using Facebook Likes, and the study by Park et al. [109] focused on social media language, Golbeck's study used a combination of features from Facebook. Therefore, the results are not comparable. Golbeck did not report internal consistency, external validity or discriminant validity.

Wald et al. [145] administered a 45-item BFI [78] test to 537 Facebook users and collected their demographics and status updates from Facebook. They constructed a word count method of dimensionality reduction and feature extraction from natural language [115] to prepare data for analysis. Instead of predicting personality scores, they only tried to predict whether a user belongs to the bottom 5% or the top 10% of the users when ranked based on their self-report personality scores, and reported an overall successful allocation rate of 37% of the users.

Predicting personality from social media is a new area of research. The ideas of a social network have existed for a long time, but modern social networks have only become widely used with the increased access to computing devices and high-speed internet. At the time of writing this thesis, MyPersonality is the only dataset that has high quality Facebook data as well as highly valid self-report data for Facebook users. None of the published studies were able to demonstrate reliability, internal consistency, external validity or discriminant validity for the predicted personality scores in a similar standard to the self-report personality scores.

In the following chapters, I introduce new methodology to enhance the predictive power of the machine learning models, and investigate if the enhanced predictions made in this thesis demonstrate psychometric properties in terms of reliability, internal consistency, external and discriminant validity.

Chapter 4

Dimensionality Reduction

Data captured on social networking websites is very large and is often referred to as Big Data. There are hundreds of millions of active users on widely-used online social networks and with the incredibly large amount of content on these websites and services, potential for variance among users is effectively limitless. As a result, such data needs to be made smaller when preparing them for machine learning. This is done by simply removing obviously unwanted pieces of information often referred to as cleaning, but after cleaning, more sophisticated dimensionality reduction methods are often necessary.

Section 4.1 explains the existing methods of dimensionality reduction that are widely used in research and practice. They are very effective at performing certain tasks, but they also have their shortcomings which are explained. Section 4.2 introduces a dimensionality reduction method named *Entropic Component Allocation*, or *ECA*, which overcomes the shortcomings of existing methods for the purpose of accurate prediction of individual differences from large and highly sparse datasets such as data from social networks. In sections 4.3 and 4.4, I perform several studies to evaluate ECA compared to two widely used existing methods of dimensionality reduction on two separate datasets from social networks. All methods have their advantages and weaknesses, in section 4.5 I review ECA critically, where it excels and where it does not. Finally, I summarise this chapter in section 4.6.

4.1 Conventional methods

Signal-to-noise ratio is a concept often used in communication engineering where signals are being transmitted through wires, optical cables or antennas. All mediums of information transfer add a certain amount of noise to the information. When the signal is weak, the noise becomes stronger than finer points of the signal and an information loss occurs. The aim is always to build mediums where the amount of noise induced onto the signal is minimised, or use algorithms to correct for errors induced during transmission. While originally aimed for communication channels, the concept has been used in many fields such as photography, biology, and data science. In datasets with very large amounts of data per user, most of the data is often classified as noise, or irrelevant for a specific task. The data that is useful for each task is often classified as signal. While actual calculation of a signal-to-noise ratio is not a common practice in data science, the terms provide a useful method of communication about the relevant and non-relevant parts of any dataset.

In terms of the methodology used for reducing dimensions, dimensionality reduction methods can be categorised as supervised or unsupervised methods. Unsupervised methods reduce the dimension of the data without requiring a class variable to be tested against. Their aim is often to preserve variance of the original data while reducing the dimensions by eliminating redundancies. Supervised methods require a class label, a variable by which they assess if any specific part of the data is relevant or not. This usually involves measuring the statistical relationship between the class variable and other variables which the aim is to reduce. They do not aim to preserve variance however, the aim is to improve the predicting power, reduce noise, and sometimes reduce redundancy.

In terms of the outputs of the reducing dimension methods, they can be categorised as either feature extraction methods, or feature selection methods.

Feature extraction methods aim to convert the initial data into a new domain where the data is compressed into fewer dimensions, often referred to as components. They are used in a wide variety of applications such as image processing, speech recognition, natural language processing, EEG and fMRI analysis of brain activity and social network analysis. Widely-used algorithms are Principal Component Analysis [73], Latent Semantic Indexing with Singular Value Decomposition [32], Latent Drichlete Allocation [16, 61] and Non-Negative Matrix Factorization [93]. Typically, feature extraction methods are unsupervised, although a supervised variant of Latent Drichlete Allocation has been introduced for classification of topics for documents [98]. Several supervised variants of semantic indexing and principal components have also been developed [153, 13, 8, 7, 132], however they are simply picking linearly correlated components after unsupervised reduction. Linear Discriminant Analysis [45] can provide supervised feature extraction by maximising the separation between predictors of the class variable and minimising variance within predictors of the class variable, however its utility is limited to predictors with linear correlations to the class variable.

Feature selection methods on the other hand simply aim to select relevant variables from the main data. The variables that are related to the class variable are kept, and variables that are not related or those that are redundant are omitted. Feature selection methods are almost always supervised.

Feature selection algorithms can be categorised in three groups: filters [95], wrappers [82] and embedded. Filter-based feature selection methods select features based solely on the relationship between the variable and the class. The relationship can be a simple correlation [65, 83], information gain with respect to class [151] or any other relationship or goodness of fit measurement. Wrapper systems typically pick a subset of variables, develop a learning model and evaluate the predictability of the class using that model, and use a searching algorithm to look for new variables to add or subtract to the subset in order to optimise the results of the evaluations. Embedded systems work in a similar way to wrappers, however instead of treating the predictive model as a black learning box, they allow feedback between the model and variable selection during the training phase. This presents the advantage of optimisations, and disadvantage of being more restrictive in the choice of predictive model as the one used in the embedded algorithm must also be used for training. Wrapper systems offer better compatibility with various predictive models. The main advantage of filter-based feature selection methods is speed and simplicity. Their main disadvantage is their inability to remove redundancies from the data. Wrapper and embedded algorithms do remove redundancies as variables that add little to the evaluation of the model are not picked, but they are significantly more time consuming, and not picking variables chosen to be redundant can also result in loss of variance among signal parts of the data especially in highly sparse datasets, and they often only respond well to the model they have been evaluated with.

Comparatively, feature extraction algorithms offer a higher degree of versatility due to their unsupervised nature. Their output is useful for any analysis and prediction, in fact some datasets release the extracted features to other researchers instead of unreduced data. On the other hand, the unsupervised nature preserves variance found in the data non-discriminantly which for most applications increases noise. The noises confuses machine learning models and they will usually either need a lot more internal complexity to account for them, or they are completely unable to. Feature selection methods on the other hand, are better at removing noise from the data, as they only keep variables which have a direct relationship with the class variable and eliminate redundancies. However, by eliminating redundancies, a trade-off is made where non-negligible portions of signal are also lost as those variables chosen to be redundant often can include useful information.

4.1.1 Usage in literature

There are several studies that looked into prediction of individual differences from online social networks. All of the studies use some form of dimensionality reduction method, as the data is too large to analyse otherwise.

In studies that involved the MyPersonality dataset, Kosinski et al. [86] used a latent semantic indexing with singular value decomposition feature extraction algorithm to reduce Facebook Likes from MyPersonality dataset into 100 components, and used the reduced dimensioned data combined with a linear regression model to predict various private traits such as age, gender, martial status, substance use history and relationship status. Youyou et al. [152] used an embedded LASSO feature selection method, combined with linear regression as its embedded model to directly predict personality scores from MyPersonality data. Park et al. [109] predicted personality scores from status updates, and reduced the dimensions of the status updates by using bag-of-word reduction methods, widely-used in natural language processing.

In studies that involved other datasets, Golbeck et al. [50] used language features, demographics and preferences to predict personality scores. Language features were reduced using the Linguistic Inquiry and Word Count (LIWC) method [115]. It's a method that analyses the text and provides various features in different categories, such as word counts, psychological processes, personal concerns and relativity of the tenses with respect to time.

4.2 Entropic Component Allocation

The aim is to develop a new dimensionality reduction method that is supervised and preserves only the variance relevant to the class variable (removes noise), is able to not only keep variables which are linearly related to the class variable, but also variables that can be used to discriminate between the users to generate decision trees, and does not lose useful information even from redundant variables while also eliminates redundancies. In this section, *class* refers to the variable which we are trying to predict. *Attribute* refers to the features that exist in the data, and *user* refers to each sample of the data.

Entropic Component Allocation is a supervised hybrid feature selection and extraction system, designed for the specific purpose of maximising predictability from data captured in online social networks such as Facebook. ECA aims to excels at three key tasks:

1. Minimising redundancy, while preserving relevant variance, even if hidden within seemingly redundant variables
2. Maximising removal of noise from the data
3. Maximising capture of both features that correlate with the class, and features that are useful to discriminate between the users in order to build decision trees.

ECA can be computed in an exhaustive way (easier to understand, but more computationally intensive), or in a randomised way which is faster to compute. Algorithm 1 presents the pseudocode for the exhaustive ECA. A step-by-step explanation of each step is provided in section 4.2.1. Section 4.2.3 explains how optimisations can make ECA faster to compute.

Algorithm 1 Exhaustive ECA Pseudocode

```
procedure EXHAUSTIVE-ECA
    $n \leftarrow$ number of attributes
    $c \leftarrow$ class variable
    $C \leftarrow n \times n$ correlation matrix for each two attributes
    for all attributes $(m)$ do:
        $i \leftarrow InfoGain\ (m, c, all)$
        if $i > threshhold_a$ then
            Preserve attribute $m$
        else
            Split users based on $m$
            if $InfoGain\ (m, c, split) - InfoGain\ (m, c, all) > threshold_b$ then
                Preserve attribute $m$
    for all Preserved attributes $(p)$ do
        for Preserved attribute $(q)$ do
            if $C(p, q) > threshold_c$ then
                Add $q$ to vector $P$
        $R \leftarrow$ Compression of $P$ using SVD to preserve $threshhold_d$ of variance.
        Output $R$ as a component
```

4.2.1 Entropic Component Allocation: Step-by-Step

In this section I explain the high-level step by step process of exhaustive ECA.

We have n number of attributes per user. First we compute the $n \times n$ correlation matrix, this is used to find redundancies among attributes (line 4).

For each attribute, we need to see if the existence of that attribute adds something to the useful data with respect to class c, in other words, is the attribute in question part of signal, or can it be dismissed as noise. There are several ways make this assessment, such as simply looking for a correlation [65]. However the concept of mutual information [107] continuously outperforms other methods in detecting relationships between variables [66, 80].

Mutual information of two variables (or sets of variables) refers to the mutual dependence among the two sets of variables. It can be used in the context of finding whether attribute m adds something to the whole dataset within the context of predicting the class c. First, mutual information between c and all parameters, including m, is calculated. Then mutual information between c and all parameters, excluding m, is calculated. They can then be compared and the amount of *Information Gained* by having m within the data can be computed (simply the ratio of two previous data). This is done in line 6. If the information gain is more than $threshhold_a$, attribute m can be preserved. In information theory terminology, variable m adds meaningful *entropy* to the dataset. Using information gain ratios for feature selection are not novel to this algorithm however, other implements do exist [114], however this is where they usually end the task.

While the introduced method up until this stage, as well as the one reported in literature [114], finds attributes that are directly relevant to the class, it is unable to find attributes that have value in discriminating the users into different groups, which can be useful for predictive models to predict the class variable. Those data are especially useful in training decision trees, where based on different attributes, the data can be split into smaller parts and different models can be trained for different parts of the data. Part of the novelty of ECA is in its power to preserve attributes that are useful to train decision trees. To achieve this, the same concept of information gain ratio is used, in a slightly different way.

For each attribute m, the users can be split into two parts. If the attribute is nominal (*true* or $false$), the users can easily be split based on the value of m, and information gain ratios of the splits can be calculated and compared to the information gain ratio of the whole dataset. This is the same principle that is used to train decision trees [118], however I use it to detect variables that are useful for decision trees. Any attribute which can split the data while increasing information gain more than *threshhold*$_b$ is preserved. If the attribute is numeric, the median is used to discriminate the users into two equal subgroups, and information gain ratios on the split are computed. Using the Newton–Raphson method, the optimal position of the discrimination is iteratively found to maximise information gain ratio, and the maxima is compared to the information gain ratio of the whole dataset. If the gain is more than *threshhold*$_b$, the variable is preserved. This section is embedded into line 11.

Finally, redundancy is a major problem with typical feature selection systems. The algorithm, until this point, has not accounted for any redundancies in the preserved attributes. Multiple attributes might be well-correlated with each other, and they can be removed to avoid redundancy, however removing the variables can cause useful information to be lost. For example, on Facebook there can be multiple pages dedicated to various politicians in the UK Labour Party. These pages will be correlated with each other, as the same users who like some of these pages are likely to like the rest, therefore a lot of the pages might seem redundant. However data is imperfect, users do not like all the pages and whichever page is chosen as the representative, it will miss users who have liked only other pages. In this context, dimensionality reduction algorithms which preserve most of variance do have a theoretical advantage.

To use this advantage, among the preserved attributes, I look into attributes that are correlated with each other, more than *threshhold*$_c$, and they are compressed into fewer variables using conventional unsupervised dimensionality reduction algorithms, here latent semantic indexing with singular value decomposition (stochastic implementation in Apache Mahout [105]) while *threshhold*$_d$ of variance is preserved, represented in lines 15-17.

ECA is an intuitive dimensionality reduction method designed specifically for big social network datasets which are highly sparse, to be used with state-of-the-art decision tree classification or regression models.

4.2.2 Tuning parameters

Like most machine learning methods, ECA has parameters which need to be tuned. In algorithm 1, four thresholds were introduced that are the parameters of ECA.

threshold$_a$, introduced in line 7 is the amount of information gain from a variable that warrants it being preserved. This is assigned as a percentage of the overall information gain. The right percentage is usually a function of how large the dataset is in terms of the number of attributes, and what level of compression is being aimed for, which should be judged by the level of sparsity in the data. If a target compression ratio is desired, all attributes can be ranked based on their information gain and the number of desired attributes can be chosen.

$threshold_b$, introduced in line 11 works in a similar way. Its intention is to find attributes that can discriminate between other attributes with respect to a class, which is useful in training decision trees. Tuning $threshold_b$ is about how deep and complex we wish our decision trees to be in the next stages. Big learner trees typically require a lower level of threshold here because they need more find-tuned discriminatory variables, whereas smaller or weaker learner trees do not need a low threshold and will function well with a higher threshold, therefore fewer features will be selected.

$threshold_c$, introduced in line 15 chooses the level of acceptable correlation between two attributes to consider them potentially redundant. It needs to be viewed within the context of the sparsity of the data. The more sparse the data is, the higher the correlation limit needs to be.

$threshold_d$, introduced in line 17 chooses the percentage of the variance that is desired to be kept among potentially redundant attributes. It needs to be selected as high as possible since the items are already correlating. Usually 95% of variance is a reasonable rule of thumb however, based on the sparsity of the dataset, it might need adjusting as well.

As is the case with tuning parameters of any dimensionality reduction method, an understanding of the properties of the data at hand is necessary, and usually it will also involve some degree of trial-and-error.

4.2.3 Computational complexity and optimisations

Dimensionality reduction using singular value decomposition (SVD) is currently widely-used in analysing big datasets from social networks. Computational complexity of typical a SVD compression of an $m \times n$ matrix (m users by n attributes) is $O(min(m.n^2, m^2.n))$. This is considered a time consuming computation, especially in datasets where there are hundreds of thousands, or millions of users and attributes with many non-zero values. ECA has a computational complexity of $O(m.n^2)$, and in comparison with SVD, it is less computationally intensive when the the number of users (m) are more than attributes (n), and it is more computationally intensive when there are more attributes than users.

There are various implementations of SVD where the computation is much more optimised [72, 155, 71]. They work by using approximation methods to compute the matrix instead of directly calculating it and they improve speeds of calculations by orders of magnitude. The most computationally intensive part of ECA is the computation of the correlation matrix and the correlation matrix can be approximated by randomisation instead of being directly computed [154], or we can compute it from an estimated covariance matrix [34].

Another way to speed up computation of ECA is to only compute correlations between attributes when necessary (more formally, in a *lazy* way). This ensures that correlations between non-preserved attributes are not computed. However, due to the supervised nature of ECA, multiple runs of ECA are required if the aim is to predict multiple classes, and constructing the correlation matrix beforehand which can be used for all future runs of ECA might be more optimal. Algorithm 2 provides the pseudocode for lazy ECA.

Algorithm 2 Lazy ECA Pseudocode

```
procedure LAZY-ECA
    $n \leftarrow$ number of attributes
    $c \leftarrow$ class variable
    for all attributes ($m$) do:
        $i \leftarrow InfoGain\ (m, c, all)$
        if $i > threshhold_a$ then
            Preserve attribute $m$
        else
            Split users based on $m$
            if $InfoGain\ (m, c, split) - InfoGain\ (m, c, all) > threshhold_b$ then
                Preserve attribute $m$
    for all Preserved attributes ($p$) do
        for Preserved attribute ($q$) do
            if $Correlation(p, q) > threshhold_c$ then
                Add $q$ to vector $P$
        $R \leftarrow$ Compression of $P$ using SVD to preserve $threshhold_d$ of variance.
        Output $R$ as a component
```

4.3 Study 1: ECA evaluation on Facebook

In this section, I evaluate ECA using a big social dataset, and compare it with an unsupervised conventional method: randomised SVD, and a wrapper subset selection system [82] using a random forest decision tree as its learning model [20], and using a greedy forward stepwise selection and backwards stepwise elimination [23]). This search algorithm is an improvement over traditional sequential feature selection search algorithms [2], due to its added randomness which helps avoid the search algorithm getting trapped in a local minima. The randomised SVD is implemented using Apache Mahout [105] while the wrapper feature selection method is implemented in Weka [64]. Appendix A includes further details about these frameworks.

The purpose of this study is to investigate if, and to what extent, ECA provides an advantage over existing widely-used methods.

4.3.1 Sample

The data used for the evaluation is part of MyPersonality project. It includes Facebook Likes of 40,000 American users and is used to predict age, gender, relationship status, and voting preferences. The users satisfy the inclusion criteria outlined in section 2.3. The reduction in the number of users from 98,515 users has been made only to speed up machine learning experiments. The decision to keep American users was made to keep the users contextually relevant to each other, as the same number of users from worldwide would have made the data less contextually rich.

4.3.2 Method

As for the predictive model, classification and regression by random forest decision trees are used [20]. The model is implemented using the Scikit-learn framework [113]. Appendix A includes further details about this framework. For continuous classes such as age, prediction accuracy is measured using Pearson's correlation between the predicted values and the actual values from the dataset. For nominal classes such as gender and relationship status, area-under-curve is used to measure prediction accuracy. To avoid overfitting bias, all analysis have been 10-fold cross validated. The users have been split into 10 bins, 9 are used for training and 1 for testing, and the process is repeated until all bins have been tested. The outcomes of all folds are averaged to produce the final result.

The criteria for comparison is the number of features required to make accurate predictions. The methods were tested for 10, 50, 100, 250 and 500 features. Feature refers to the size of the output of the dimensionality reduction method. 100 features is a parameter choice for SVD dimensionality reduction of Facebook Likes used in literature [86]. 500 features is usually the maximum number of features that can reasonably be incorporated in linear models from online social networking data. It also corresponds to capturing about 90% of variance from this dataset using an SVD dimensionality reduction method. 250 features is used to investigate how models perform as the number of features are increased, on their way to 500. 10 and 50 features are picked to examine how well different methods perform where the aim is to reduce features and simplify models.

4.3.3 Results

Figure 4.1 shows the correlation between the predicted age and the user's self-report age, using the three dimensionality reduction methods. Age is a very predictable variable from online social networks, because people of different ages often have different interests and this is visible from the pages that they like on Facebook.

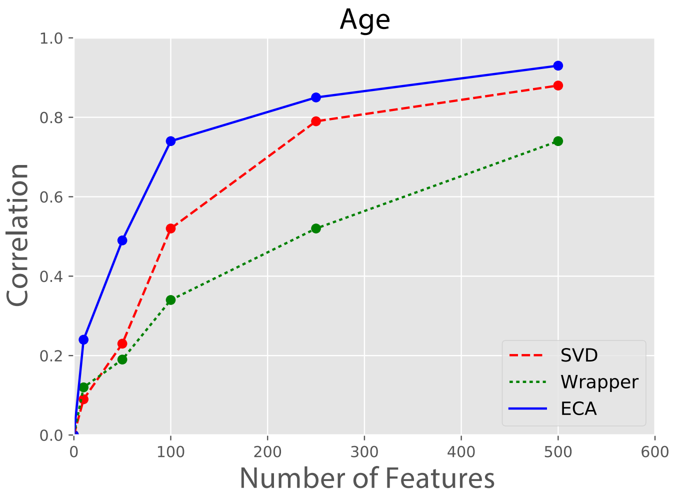


Fig. 4.1 Comparison of SVD, Wrapper and ECA in predicting age, Correlation.

All methods are able to reduce dimensions of the data in a way that the learner models can make accurate predictions on age. However, the number of features needed to make accurate predictions varies significantly. For example, ECA at 250 features is able to provide the same level of prediction accuracy for age as SVD is at 500 features.

Figure 4.2 shows the area-under-curve between the predicted gender and the user's self-report gender, using the three dimensionality reduction methods. Like age, gender is also highly predictable from Facebook Likes. For example, liking pages aboutvideo games is an indicator of someone being male while liking pages about cosmetics is an indicator of someone being female. This is observable from the data directly, and is reported in literature as well [86].

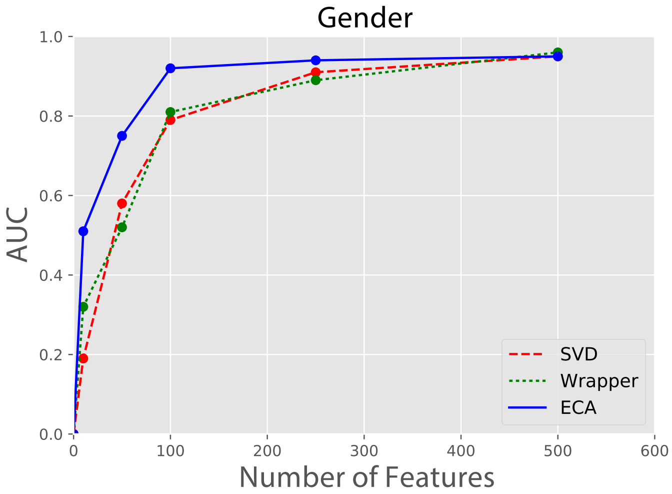


Fig. 4.2 Comparison of SVD, Wrapper and ECA in predicting gender, Area Under Curve.

At 500 features, all methods are able to achieve a similar and very high level of prediction accuracy. However, the same pattern is observed as ECA is able to achieve its highest level of accuracy with fewer features.

Figure 4.3 shows the area-under-curve between the predicted relationship status and the user's self-report relationship status, using the three dimensionality reduction methods.

Predicting relationship status is often a complicated process, because it is a variable that can change. Digital footprints on online social networks are accumulated over a long period of time, often years or even more than a decade in the case of Facebook. During this time frame, the relationship status of people can change. However, the results show that it is still highly predictable. Critically though, the predictions should be seen as the prediction of the relationship status during the majority of the time that the user has been part of Facebook, rather than the relationship status at the immediate moment of data collection.

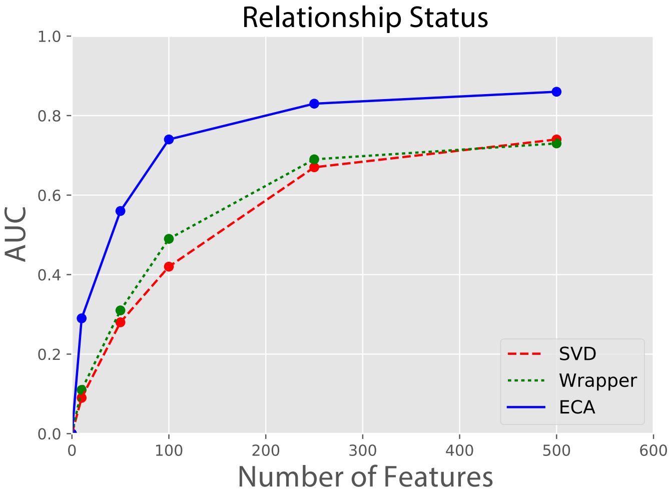


Fig. 4.3 Comparison of SVD, Wrapper and ECA in predicting relationship status (single or not single), Area Under Curve.

Even at 100 features, ECA is able to outperform other methods at 500 features. This is the effect of noise in the data in the case of SVD dimensionality reduction. It preserves vast majority of variance at higher number of features, but the noise in the data makes it unfeasible or impossible for the learner model to extract the relevant useful information. This is where ECA excels, by feeding clearer inputs to the learner models, they can perform better while being more efficient as well, as they are working with fewer features.

Figure 4.4 shows the area-under-curve between the predicted voting preferences and the user's self-report voting preferences, using the three dimensionality reduction methods. Voting preferences are very predictable from online social networks because they are often the most widely-used mediums of communication between the people and their favourite political parties, politicians and news sources.

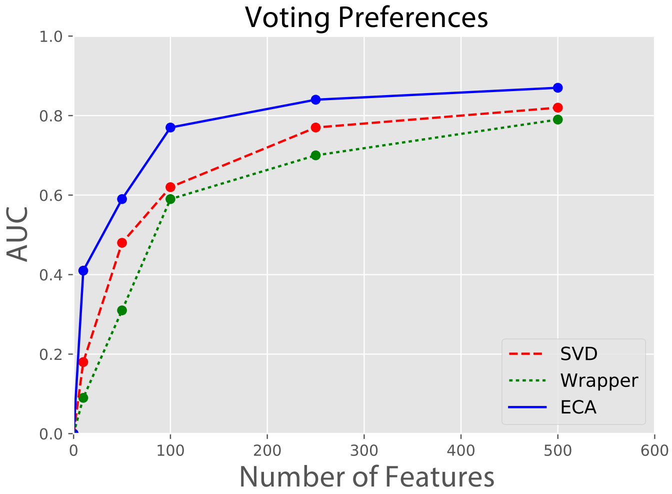


Fig. 4.4 Comparison of SVD, Wrapper and ECA in predicting voting preferences (liberal or conservative), Area Under Curve.

The same pattern follows here as well. ECA, due to its intelligent method of not preserving noise, achieves higher levels of accuracy at fewer features. It performs better than feature selection methods because it does not dismiss redundant attributes completely, instead it compresses them into new features that preserve the variance even among the redundant variables.

As all tests demonstrate, ECA is capable of arriving at accurate predictions with significantly fewer features compared to SVD or wrapper feature selection systems. This is intuitive. SVD is an unsupervised dimensionality reduction system, therefore the variables are optimised to carry most of the original variance, rather than most of the signal (compared to noise) when it comes to predicting a specific class. ECA in comparison only preserves variance relevant to the class variable which it is trying to predict, as explained in section 4.2.1. Therefore it is able to arrive at accurate predictions with fewer features, and outperform SVD in most instances.

Compared to wrapper-based feature selection systems, ECA is more accurate because wrapper feature selection systems remove seemingly redundant attributes from the data which dismisses some variance of signal, on the other hand ECA combines seemingly redundant attributes while preserving this variance.

4.4 Study 2: ECA evaluation on MovieLens

MovieLens [1] is a non-commercial personalised movie recommendation service. Users are able to log into the website, rate different movies, and the website provides recommendations of new movies based on the ratings.

The purpose of this study is to investigate if the advantages of ECA, shown in section 4.3 when it came to a dataset from Facebook, also apply to a non-Facebook, yet sparse, dataset.

4.4.1 Sample

A dataset of 1 million ratings, of 6000 users on 4000 movies has been provided by their developers [68]. The dataset includes not only movie ratings, but also age group, gender and occupation of the users.

MovieLens records age in the ranges as described in table 4.1 and records occupation in one of 20 options, as listed in table 4.2.

Age Group
Under 18
18-24
25-34
35-44
45-49
50-55
56+

Table 4.1 MovieLens datasaet, ranges of age

4.4.2 Method

In this section I use the same ECA, SVD and wrapper-based dimensionality reduction systems and the same predictive model as in section 4.3 to predict age and occupation of users from their movie ratings.

[1] https://www.movielens.org

Writer	Unemployed	Tradesman/craftsman	Technician/engineer
Self-employed	Scientist	Sales/marketing	Retired
Programmer	Lawyer	K-12 student	Homemaker
Executive/managerial	Doctor/health care	Customer service	College/grad student
Clerical/admin	Artist	Academic/educator	Other

Table 4.2 MovieLens dataset, list of occupations

To predict age, I use the mean of each age group for training, and use regression to predict a numeric age value. The value is classified as correct if it is in the correct range and false if it is in the wrong range. I use classification to predict each occupation. I use the percentage of correctly classified instances as the evaluation measurement for age and occupation.

The criteria for comparison is the number of features required to make accurate predictions. The methods were tested for 2, 5, 10, 25 and 50 features. 50 features is where SVD collects 99% of the variance in the dataset, and at 25 features it collects 80% of the variance. 2, 5 and 10 features are picked to compare the efficacy of methods at extracting useful information in as few features as possible.

4.4.3 Results

Figure 4.5 shows the classification accuracy between the predicted gender and the user's self-report age, using the three dimensionality reduction methods. Age is found to be more difficult to predict from movie preferences than it is from online social networks. This is intuitive because online social networks include data not only about movies, but about many other areas of potential interest as well.

ECA is able to achieve the maximum accuracy with as few as only 5 features, while with SVD it only achieves the same level at 50 features. This is the same pattern that was observed in section 4.3.

Table 4.6 shows the classification accuracy between the predicted occupation and the user's self-report occupation, using the three dimensionality reduction methods.

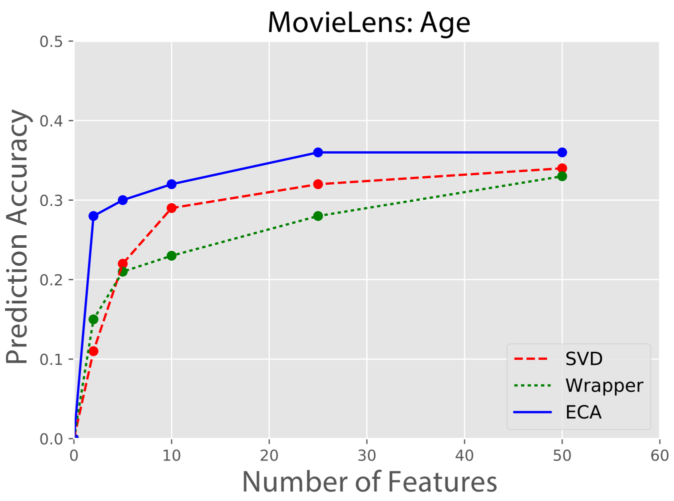


Fig. 4.5 Comparison of SVD, Wrapper and ECA in predicting age, Prediction Accuracy.

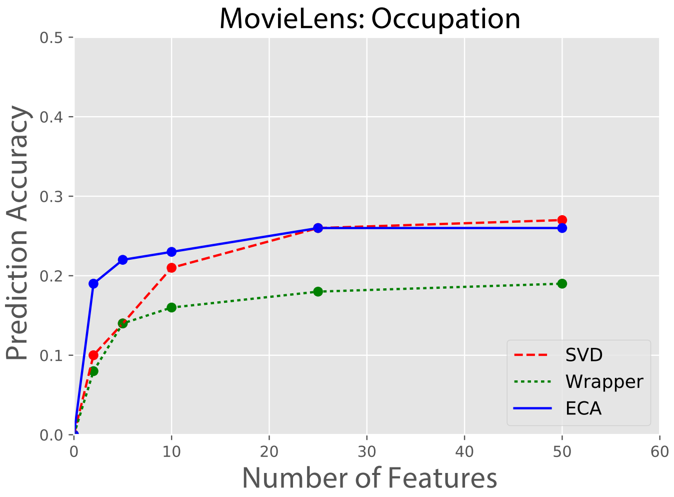


Fig. 4.6 Comparison of SVD, Wrapper and ECA in predicting occupation, Prediction Accuracy.

The same patterns can be observed here as ECA is capable of reaching more accurate predictions with fewer features, while the overall accuracy with maximum number of features is similar between ECA and SVD, ECA reaches them with fewer features. This is because this MovieLens is a much smaller dataset compared to Facebook, therefore noise in the data will not become a major factor in terms of prediction accuracy, and most of the benefits of ECA become limited to decreasing the number of features for predictive models.

4.5 ECA advantages and weaknesses

Being able to achieve the same level of accuracy with a fewer number of features is very beneficial in machine learning, as a lower number of features translate to needing less computational resources for training the models. This is a major systems advantage as training decision trees usually has a computational complexity of $O(m.n^2)$, where m is the number of users in the training data and n is the number of attributes. Therefore, needing fewer features translates to significantly faster training times.

As an example, as outlined in figure 4.2, to predict gender from Facebook Likes using SVD it takes 500 features, when ECA is able to achieve that level of accuracy in 100 features. This, theoretically, can lead to a training time with ECA that is only 4% of the training time of the same model with SVD, at the same level of accuracy. This was a best-case scenario but in all cases, ECA is able to achieve better prediction accuracy at fewer features compared to other methods. We also observe that ECA can lead to more accurate results in a lot of instances, this is due to the fact that highly compressed yet unsupervised and noisy data makes it harder for decision trees to learn rules from, which translates to more complex decision trees which can often become unfeasible for many applications.

While ECA is very effective in big social data, especially data with high redundancy or high levels of noise, ECA is fairly ineffective in data that does not have any redundancy or is not noisy at all. This includes data that are pre-processed to only include relevant information (often medical datasets), or data that has redundancy removed from it using expert human knowledge. In a way, ECA is only useful when it is not feasible to use expert human knowledge to filter and prepare the data.

ECA is not a very fast algorithm, especially as it has to be computed for every class separately. Compared to SVD which is computed once and used on as many classes as possible, ECA adds a great amount of computation to the analysis. However, the most time consuming process in ECA is the computation of the correlation matrix, which is at $O(m.n^2)$ for a dataset of m users by n attributes, which needs to be computed only once as it is independent of the classes. Therefore, the next iterations are only at $O(n^2)$. This is still slower than SVD which does not require further analysis for new classes. However, as long as decision tree predictive models are used, this can still translate to a lower combined dimensionality reduction and model training time.

4.6 Summary

Unsupervised dimensionality reduction methods such as singular value decomposition usually aim to preserve as much variance in the data as possible by combining attributes into a smaller number of features, often referred to as components. They are more effective at reducing dimensions of datasets that mostly include relevant information and little noise, however the preserving of all variance non-discriminatingly makes them disadvantageous for very large, sparse datasets such as those often gathered from online social networks. Supervised feature selection methods remove noise, but also remove redundancy by erasing seemingly redundant variables. This also removes the useful information hidden inside seemingly redundant variables. This is a challenge for highly sparse datasets such as data from social networks because variables often are not perfectly redundant.

ECA can be defined as a supervised feature selection system that instead of removing redundant variables, it uses an unsupervised dimensionality reduction method to preserve most variance among redundant variables. This is in contrast with other attempts at combining supervised learning with unsupervised dimensionality reduction algorithms, cited in section 4.1, where they first use an unsupervised algorithm to reduce the dimensions, and then only preserve components that fit well with the class variable. This does little to remove noise from the dataset as the noise is often embedded into reduced variables with the signal and they are compressed into the same components. ECA prioritises elimination of noise and compresses data only after noise is removed.

Advantages of ECA can be summarised as a systems advantage where accurate predictions can happen with fewer features, and a prediction advantage where lack of noise makes predictive models more capable of providing accurate predictions from the cleaner data that they are presented with. I can summarise limits of ECA as being mostly useful for noisy and sparse datasets, its slower speed and its supervised nature, which necessitates repeating dimensionality reduction for every class variable. In chapter 5, ECA will be used to reduce dimensions of Facebook Likes for the prediction of five-factor personality traits.

Chapter 5

Predictive Models

This thesis focuses on assessment of individual differences using data from Facebook. In chapter 4, I introduced ECA, the method that is capable of reducing dimensions of the large data collected from Facebook, while removing noise and maintaining useful information that might be hidden among seemingly redundant attributes. In machine learning, to get from reduced dimensioned data to prediction of class variables, we need predictive models. In this chapter, I explain how accurate assessments of five-factor personality model are made from the reduced dimensioned data, and discuss technical details of my proof-of-concept personality prediction machine learning system. This chapter mostly focuses on technical aspects of the machine learning system, chapter 6 is where the psychometric properties of a personality test developed using machine learning methods introduced in this chapter and chapter 4 are investigated in terms of reliability, internal consistency and discriminant and external validity.

In section 5.1, I review the predictive models of existing attempts in literature at predicting personality from online social networks. Section 5.2 explains key differences between the model used in this projects and prior works. In section 5.3, I implement conventional methods of predicting five factor personality from Facebook data, and compare them with the method explained in this chapter. This is done in the form of ablative analysis where the effect of each model decision on the overall prediction power is investigated. I also compare this system with a deep learning methodology that was investigated but not picked as the preferred method. Finally, I summarise this chapter in section 5.5.

5.1 Conventional predictive models

All prior published work into predicting five-factor personality from Facebook data have used linear regression models as their predictive model [86, 117, 152, 109, 146, 15, 131, 30]. Linear regression models are easy to understand, usually perform well and do not require heavy optimisations, therefore they are often the go-to models for predicting desired information from big social data.

However, vanilla machine learning models used in prior research are not optimal for accurate assessment of psychological traits from Facebook data. There are several reasons for this. I categorise the reasons into three groups, and explain each in details in sections 5.1.1, 5.1.2 and 5.1.3, as well as introducing ways that my model will differ from them.

5.1.1 Direct personality prediction

In a traditional self-report personality test, several items are presented to the user. A sample item is provided in table 2.1 with the choices offered to the individual. Once the user has responded to all items, the test is scored by averaging all responses to compute scores for each personality trait. MyPersonality dataset includes both the computed personality scores, and the individual responses to each item. So instead of predicting the final personality score directly from the data, as prior research have done [86, 152], I use the item-level data to train the predictive models and use them to predict the answers to each item on the questionnaire for the test users, to construct answers to a personality questionnaire that the user would have provided if they had been given a self-report personality test.

This is advantageous for several reasons. First, responses to a lot of items measuring personality traits are easily inferable from a user's footprints on Facebook. For example, a person's voting preferences, or their interest towards arts or philosophy, are part of their five-factor personality assessment, and they are predictable from the pages that they like on Facebook. This allows the models to be trained for each specific item, rather than a trait as a whole and the model records a more detailed understanding of patterns in data relating to each personality trait compared to directly predicting an overall personality score.

Second, just as the reliability of self-report measurements of personality increase with the length of questionnaires, online prediction of personality also becomes more accurate as the more finer detailed information from the user are used during the training of the machine learning models. So instead of using a personality score as one reference point for training, responses of a user to a 20-item test for each personality trait are used for training. This allows the models to capture up to 20 times more information.

Finally, using individual questions enables a detailed understanding of what facets of each personality trait are predictable from online data and what facets are not. This knowledge helps to determine if a passive method of measurement of personality is suitable for a specific application. Direct prediction of final personality scores do not allow for this level of understanding. This is further investigated in chapter 7.

5.1.2 Linear regression models

Generally, linear regression models are useful for finding linear direct relationships between different observations, in this context, between liking certain pages on Facebook and being more likely score a certain way in a personality test. This is why all prior work use linear regression models [86, 117, 152, 109, 146, 15, 131, 30]. For example, A linear regression model for this task is the one presented by Kosinski et al. [86] where Facebook Likes are compressed into a smaller set of 100 dimensions using Singular Value Decomposition (SVD), then a linear regression model is trained on the data to find the right coefficients of a linear function that can directly map the SVD components onto the desired personality trait, for example, openness to experience. Linear regression models are often used for their versatility among a wide range of applications. However in this context, they present two key disadvantages.

First, personality is defined as patterns of behaviour, not patterns of digital leftovers of behaviour. Facebook Likes on the other hand, are not user behaviour, they are digital leftovers of user behaviour. Using a single linear function to map digital leftovers into a personality score does treat personality as a pattern of digital leftovers of behaviour rather than patterns of behaviour. Having liked a page on Facebook is *not* a behaviour, it is a digital leftover of a behaviour. The behaviour is, for example, getting news updates from a favourite celebrity or getting football scores of a favourite club. The existence of a Like in the user's data is what is left over from the behaviour, but the behaviour itself is truly latent. The data does not say why a user has liked a page on Facebook, only that they have liked it.

For example, Bob may like FC Barcelona's Facebook page to get the scores for football matches when there is a match as he does not have time to read sports news websites or watch the football matches. He checks his Facebook feed at nights and this is a great way to ensure that he receives the information when it is fresh, but without intruding his lifestyle. Jason likes the same page because he wants to show to his friends that he is a fan of the club, because he thinks it is part of being socially accepted, but really has no interest in the club's results or its news. Carlos on the other hand, likes the page because he wants to know exactly when tickets to the matches and parties go on sale, as he cannot miss those events. Three very different behaviours which could be indicators of very different types of personality, as Carlos is likely an extravert who enjoys parties and likes to watch sports games in person and socialise with friends, while Jason is likely a neurotic individual who is trying to present an ideal and more likeable image of himself on Facebook, and Bob is simply a conscientious person who likes to organise the way he receives news updates. However, the digital leftover for all three users is the same, they all liked FC Barcelona's Facebook page.

Treating the liking FC Barcelona's Facebook page as a direct predictor of personality will automatically make the assumption that people who like this page do it for the same reasons, because in the end of the training of the models, a constant coefficient will be assigned to this specific observation. This illustrates that we need *judgements of personality*, rather than a calculation of personality scores. A system capable of passive assessment of personality needs to be able to distinguish between different behaviours that might leave similar footprints.

Second, linear models are also not able to discriminate between users on the basis of their data. There is a rich literature about cross age, gender and cultural differences in personality. An individual's demographics are obtainable from their profiles (or can be predicted very accurately from their Facebook Likes, as demonstrated in section 4.3), and users with similar discriminating features can be grouped together and separate models can be trained for each of them. This is the basis of decision trees which solves both problems.

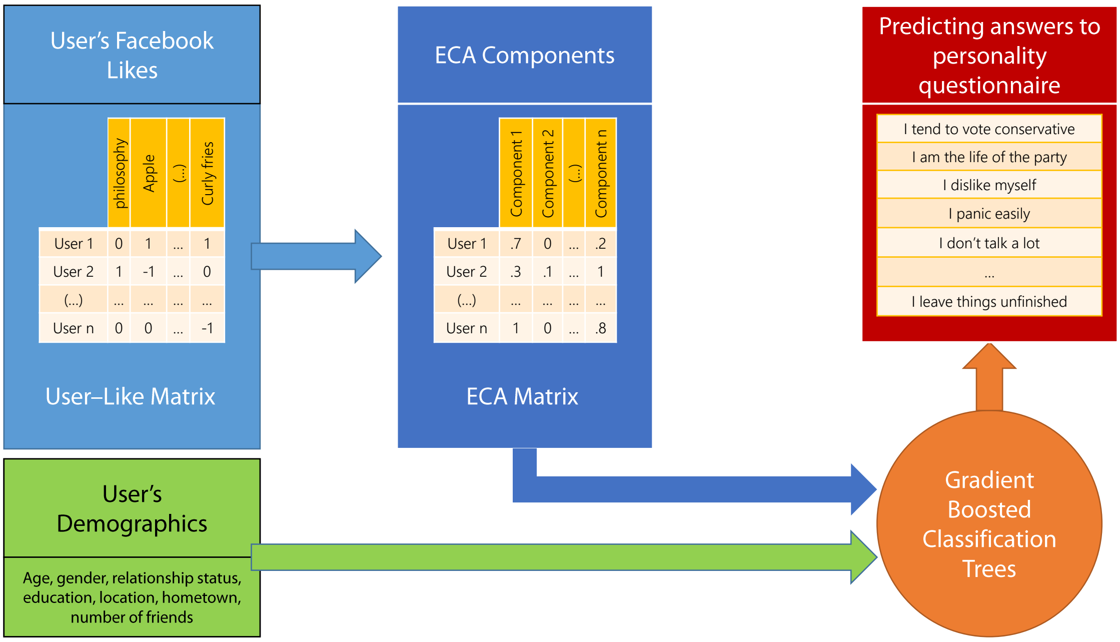
5.1.3 Pages that are not liked

In MyPersonality dataset, for each Facebook page, for each user there exists a binary value. That value shows whether the user has liked the page or not. This presents challenges as a *lack of like* does not necessarily mean a lack of interest, or a dislike. There are hundreds of thousands of pages on Facebook and it is not possible for any user to express interest towards all pages. Grouping all the *not liked* pages together means telling the model to treat them equally.

For example, a user might like the following pages: *Barack Obama*, *Hillary Clinton* and *Bernie Sanders*. The user has not liked the pages *Donald Trump*, *Ted Cruz* and *Michelle Obama*. Telling the model that the attitude of the user towards Donald Trump, Ted Cruz and Michelle Obama is the same, by classifying all three as *not liked*, is intuitively wrong. Instead, based on the pages that the user has liked, it is possible to rate other pages at potential likes, or potential *dislikes*. There is no ground truth to compare them against, however. Because Facebook does not allow users to express an expression of dislike towards pages. The role that potential dislikes can play in predictions is further investigated in section 5.2.1.

5.2 Predictive Model

In this section I describe how internal structure of my prediction model works. The step-by-step process is described in figure 5.1.


User's Facebook Likes
User–Like Matrix
ECA Components
ECA Matrix
Predicting answers to personality questionnaire
User's Demographics
Age, gender, relationship status, education, location, hometown, number of friends
Gradient Boosted Classification Trees

	philosophy	Apple	(...)	Curly fries
User 1	0	1	...	1
User 2	1	-1	...	0
(...)	...	...	...	...
User n	0	0	...	-1

	Component 1	Component 2	(...)	Component n
User 1	.7	0	...	.2
User 2	.3	.1	...	1
(...)	...	...	...	...
User n	1	0	...	.8

I tend to vote conservative
I am the life of the party
I dislike myself
I panic easily
I don't talk a lot
...
I leave things unfinished

Fig. 5.1 Overall structure of the prediction model

First a matrix of m users by n Likes is constructed. Then a stage of ECA, as described in chapter 4, is performed to reduce the dimensions and to remove unrelated and redundant information from the data. This is repeated 100 times for each of the 100 items on the MyPersonality's IPIP personality questionnaire. The output of ECA is used to train gradient boosted classifiers for each one of the 100 IPIP items, which are then used to predict the answers a user would have given to each item on a IPIP five-factor personality test.

IPIP Five-factor personality tests use the mean of the responses to items, keyed positively or negatively, to compute personality scores. While this scoring system works very well for self-report traditional psychometrics tests, it does not work well for predicted responses to personality items. It is because the level of accuracy of the predictions differ for different items. Some items are easier to predict than others, and some items are easier to predict for different users, based on their demographics. This is further investigated in chapter 7. As a result, different weights need to be used for different items based on the profile of the user. Figure 5.2 shows the structure of the method used to score the test and guide the users to the right regression model.

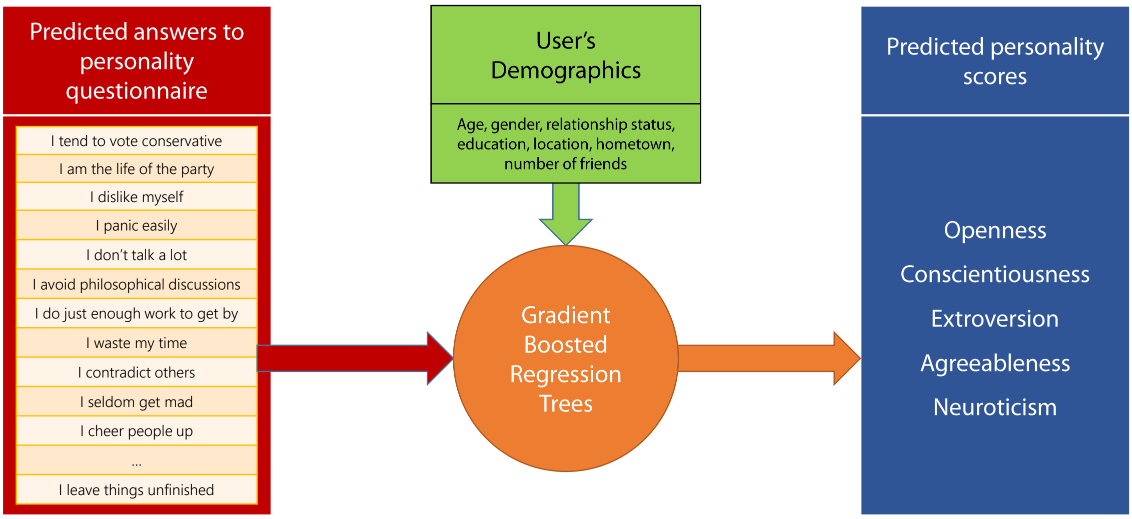

Predicted answers to personality questionnaire
I tend to vote conservative
I am the life of the party
I dislike myself
I panic easily
I don't talk a lot
I avoid philosophical discussions
I do just enough work to get by
I waste my time
I contradict others
I seldom get mad
I cheer people up
...
I leave things unfinished
User's Demographics
Age, gender, relationship status, education, location, hometown, number of friends
Gradient Boosted Regression Trees
Predicted personality scores
Openness
Conscientiousness
Extroversion
Agreeableness
Neuroticism

Fig. 5.2 Overall scoring structure of the prediction model

The decision trees are not only able to learn to assign different weights to different items during training, but they also learn to find patters within the demographics of users that explain differences in the model's predictive power for each item for various groups and scores each group according to their optimal items. The actual personality scores are calculated from the predicted responses to questionnaires, demographics are used only to train the decision trees.

The final output of the system is a single personality score, to be compared against self-reports for evaluation. This is investigated in chapter 6.

5.2.1 Dislikes

As explained in section 5.1.3, I do not treat all pages that are not liked in the same way, as Facebook pages are only liked when a user visits them and not liking a page does not necessarily mean a lack of interest. I developed a user-based collaborative filtering recommender system for Likes, and added the most unlikely recommendations as pages that the user would not have liked, and added them to the data as values of -1, as shown in figure 5.1. The rest of the possible pages are treated as unknown.

Admittedly, there is no way to evaluate if the so-called dislikes are actually pages that the user would dislike if Facebook had provided that option. They are instead, pages that the user is least likely to find useful had they been presented to them in a recommender system. They are, in a way, only here to avoid the pages that correlate negatively with the user's interests to be mistaken as pages that the user has no genuine interest towards. As an example, a user that likes multiple pages from Democratic Party politicians in the United States would get a -1 score for very conservative politicians, while they would get an unknown for, for example, football players in East Asia that they have expressed no interest towards.

Since there is no ground truth, it is very important to only add a certain number of dislikes to the user's data. Adding too much would increase the risk of adding noise to the data, since the dislikes cannot be evaluated independently. The number of disliked pages that are added for each user is fixed at 20 percent of the number of pages that they have liked. This number has been chosen via trial-and-error. A comparison of the effect of adding more or less than 20 percent is studied later, in section 5.3.1.

I used Apache Mahout [105] to develop this recommender system, further explanations of the framework is presented in appendix A. User-based recommender systems calculate the similarity between different users and recommend items most used by the most similar users. Instead of picking pages that are best fits for recommendation, pages that are worst fits are picked as so-called dislikes. This ensures that dislikes are only picked from pages that are contextually relevant to the user. I used the log likelihood ratio test to measure similarity and differences to account for surprise and coincidence.

5.2.2 Classification and regression models

Prior studies [86, 117, 152, 109, 146, 15, 131, 30] used liner regression models for predicting personality traits from digital data. In short, I did not use linear regression models for the following reasons: (a) personality is latent patterns of behaviour, rather than patterns of footprints of behaviour, which a model with direct link (linear or higher polynomial degrees) between the Facebook data and the personality scores incorrectly assumes; (b) linear regression models do not allow the use of dichotomous variables such as gender, location and cultural differences in a discriminatory way; and (c) they do not take to account that people may leave the same footprint for very different behavioural reasons which can be related to latent psychological traits.

Decision trees

Instead of linear regression, I use decision trees. It means that instead of training one function to predict the desired results (answer to each personality question, or a personality score) for all users, it uses various trained classification or regression models for various groups of users, and the rules of the tree, learned from the data, are used to guide the users to the right model.

This is intuitive: Humans take into account demographics when making psychological judgements from behaviour. It's only natural that an artificial intelligent judge of personality should do the same. For example, a 70-year-old who listens to classical music might just be average for his age, whereas a young person who listens to classical music might be considered to be rather open-minded and introverted, given his peer group. This pattern is actually visible in the data. Another example is that liking sports related pages is a strong indicator of extraversion only for under 27s whereas it is not indicative of personality traits for older people. Perhaps this is because young people actually play sports in teams whereas older people are sports fans who might not necessarily watch with friends.

Training of decision trees is the process of constructing the rules that can divide the users into different groups, and is done using the principle introduced by the C4.5 algorithm [118] which uses the same concept of *Information Gain* as described in chapter 4. The choice of *Information Grain* as the criteria to select features in ECA is coupled with using the same algorithm to train the decision trees. In other words, the features are selected and constructed to be used to train decision trees.

Gradient boosting

In machine learning, a *strong learner* is defined as *one* predictive model where its output correlates to a high degree with the class variable. A linear regression model or a large decision tree are examples of strong learners. A *weak learner* is a predictive model that correlates to a smaller degree with the class variable. Multiple weak learners can be added together to construct a strong learner. This method is called boosting and models utilising this method are called additive models. Robert E. Schapire [126] demonstrated the strength of using multiple weak learners in highly sparse datasets. Gradient boosting [48] combines the boosting technique with decision trees, meaning that decision trees are trained to be weak learners, and a combination of decision trees are used to construct a strong learner, which provides the predictions.

More formally, there are m weak learners, $h_m(x)$, where each learner is a decision tree of a fixed size. $F(x)$ is the final strong learner. γ is the weight that is assigned to each weak learner.

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \tag{5.1}$$

Where the model is calculated step-wise:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{5.2}$$

At each stage, a weak learner is chosen that minimises a loss function L. y is the class variable that we are trying to predict.

$$F_m(x) = F_{m-1}(x) + \arg\min \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) - h(x)) \tag{5.3}$$

The minimisation problem is solved numerically using the gradient steepest descent algorithm. The weight of each weak learner is calculated using the line search algorithm:

$$\gamma_m = \arg\min \sum_{i=0}^{n} L(y_i, F_{m-1}(x_i)) - \gamma \frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)} \tag{5.4}$$

The role of loss function is to measure the prediction error at each stage of the recursion. Huber's loss function [75] is used for regression tasks during the scoring. Prediction of the likely responses to personality questionnaires is done as a classification task, even though the responses are not independent classes. As shown in table 2.1 in a sample personality questionnaire, users have the option to choose from the range of very inaccurate, somewhat inaccurate, neither accurate nor inaccurate, somewhat accurate and very accurate. A typical classification loss function will treat all incorrectly predicted values in the same way, however the distance between possible choices should be taken into account. To achieve this I use the Hinge loss function, while the penalty becomes compounded as the distance between the predicted class and the label class increases.

Gradient boosting decision trees were implemented using the Scikit-learn framework [113]. Further explanation of the framework is provided in appendix A.

5.3 Ablative analysis

In this section I examine the effect of each model design decision in the overall predictive power of the model. This section is not the psychometric evaluation of the entire system, as that is covered in chapter 6. Here, the basis of the evaluation of the machine learning models is their pure predictive power in predicting five-factor personality scores. Predictive power of the models are measured by correlating their predictions with self-reports using Pearson's correlation. This helps us measure how models compare in the amount of variance of the desired variable that they are able to account for in their predictions. These findings are reported as enhancements towards the baseline, as the values of the correlations themselves are not the focus of this chapter. This is done to avoid confusion with chapters 6 and 7, where these correlations will be examined using psychometric standards of reliability and validity.

5.3.1 Study 1: Analysis of the effect of dislikes

As explained in section 5.2.1, I used a user-based collaborative filtering recommender system to predict a set of pages that are the least useful options to be recommended to a user, here referred to as dislikes. These pages are added to the user's data as -1 to distinguish them from not-liked pages. In this section, I investigate the effect that adding these values to the dataset has to the predictive performance of the models.

Sample

25,000 American Facebook users from MyPersonality dataset have been chosen for this study. The users satisfy the inclusion criteria outlined in section 2.3. The reduction in the number of users from 98,515 users has been made only to speed up machine learning experiments. The decision to keep American users was made to keep the users contextually relevant to each other, as the same number of users from worldwide would have made the data less contextually rich.

Method

Predictive power is compared when no dislikes are added to the data, when dislikes equal to 10% of the user's overall liked pages are added, when 20% are added, when 50% are added, and when 100% are added. All models have been trained with their respective percentage of dislikes added. To avoid overfitting bias, all analysis have been 10-fold cross validated. The users have been split into 10 bins, 9 are used for training and 1 for testing, and the process is repeated until all bins have been tested. The outcomes of all folds are averaged to produce the final result.

Results

Table 5.1 shows the comparative predictive power of the model, with and without dislikes. The results have been normalised to the baseline.

Trait	*No dislike*	10%	20%	50%	100%
Openness	1	1.02	1.05	1.01	.97
Conscientiousness	1	1.01	1.04	1.04	.98
Extraversion	1	1.03	1.06	1.02	.94
Agreeableness	1	1.02	1.04	1.01	.96
Neuroticism	1	1.01	1.02	.98	.94
Mean	1	1.02	1.04	1.01	.96

Table 5.1 Ablative analysis of adding dislikes to dataset

Since there was no ground truth to evaluate dislikes, it initially came at a surprise that adding them are helpful. However, we should look at them as a way of making data cleaner, as it helps relevant but not-liked pages to not be confused with genuinely irrelevant pages. So dislikes can be defined as a group of contextually relevant pages that are not liked, and are not likely to be liked, as they are picked from the worst recommended pages by the recommender system.

Adding dislikes equal to 20 percent of the user's likes adds something to the predictions, however, adding more makes the predictions less accurate as the recommendations may start to be less accurate and relevant. This is most likely because genuinely irrelevant pages are added to relevant-but-not-liked pages at that point. Overall, they help increase the variance accounted for by the models by about 8%.

Comparatively among the five personality traits, dislikes add a similar amount of predictive power to the models, except in the case of Neuroticism where the amount of added predictive power is less compared to other traits. A detailed analysis of overall predictibility of various traits is performed in chapter 7.

5.3.2 Study 2: Comparison of predictive models

In this study I compare the predictive power of linear regression, a strong learner decision tree (without boosting), and gradient boosting decision trees.

The purpose of this study is to test whether the rationale presented in sections 5.1.1 and 5.1.2 about disadvantages of linear regression models for personality research will translate to practical results as well. It also tests whether the consensus [126] about the suitability of using a combination of weak learners compared of strong learners on sparse datasets will hold on Facebook data, which is highly sparse.

Sample

25,000 American Facebook users from MyPersonality dataset have been chosen for this study. The users satisfy the inclusion criteria outlined in section 2.3. The reduction in the number of users from 98,515 users has been made only to speed up machine learning experiments. The decision to keep American users was made to keep the users contextually relevant to each other, as the same number of users from worldwide would have made the data less contextually rich.

Method

I have implemented the three predictive models for the analysis of the effectiveness of using decision trees and gradient boosting compared to simple linear regression models. All other model decision are as described in figure 5.1. To avoid overfitting bias, all analysis have been 10-fold cross validated. The users have been split into 10 bins, 9 are used for training and 1 for testing, and the process is repeated until all bins have been tested. The outcomes of all folds are averaged to produce the final result.

Results

Table 5.2 shows normalised predictive power of different models compared to the baseline.

		Decision Trees	
Trait	*Baseline*	*Without boosting*	*With boosting*
Openness	1	1.30	1.79
Conscientiousness	1	1.34	1.89
Extraversion	1	1.43	1.65
Agreeableness	1	1.63	1.92
Neuroticism	1	1.23	1.63
Mean	1	1.39	1.78

Table 5.2 Comparison of different predictive models

I chose the baseline as the linear regression model as it is the most widely used model in literature. Using a strong learner decision tree as the predictive models improves upon the predictive powers of the linear regression model, because of the reasons outlined in section 5.1.2. A model that is able to discriminate between users based on rules learned from data can point the users to better optimised regression and classification functions based on their own data. Overall, for all five traits combined, using a strong learner decision tree can add about 93% to the variance accounted for.

Moving from one big decision tree to a gradient boosted method further increases the predictive power of the models. This is inline with the literature of using gradient boosted methods for sparse datasets [126]. A combination of many weak learners are able to capture finer details of the sparse data easier than a big learner can. Compared to a strong learner, we observe an increase in variance accounted for of about 64%. Compared to linear regression, this increases the variance accounted for by 217%.

Comparatively among the five personality traits, using a big learner decision tree (without boosting) improves upon the predictability of all five traits, most for agreeableness and extraversion and least for openness and neuroticism. A decision tree looks for patterns in the data to divide the users into smaller groups and uses separate regression models to calculate their personality scores, this can indicate that for agreeableness and extraversion, there exist more markers in social networking data that can split the users for the predictive models to train for separately compared to neuroticism or openness. However, a strong caveat here is that this can be a factor of the way the trees are trained. Different algorithms that use different criteria for construction of the decision tree might lead to different results. The predictability of personality traits is discussed in chapter 7.

Among the five personality traits, prediction of agreeableness and conscientiousness are best aided by the use of a combination of weak learners as decision trees. The least aided traits are neuroticism and extraversion. There is however, less variance among the added benefit of gradient boosted decision trees ($\sigma^2 = .014$) compared to a strong learner decision tree ($\sigma^2 = .019$). This can be attributed for their better ability to manage sparse datasets.

5.3.3 Study 3: Comparison with direct prediction

In this study I compare the prediction power of models when item-level data are predicted first, and the score is predicted from them, and when the five-factor personality traits are predicted directly. This is to provide a comparison between my approach and prior work in literature and to investigate if training the models with finer, more focused data is beneficial to the overall accuracy of the predictions.

Sample

25,000 American Facebook users from MyPersonality dataset have been chosen for this study. The users satisfy the inclusion criteria outlined in section 2.3. The reduction in the number of users from 98,515 users has been made only to speed up machine learning experiments. The decision to keep American users was made to keep the users contextually relevant to each other, as the same number of users from worldwide would have made the data less contextually rich.

Method

In this study I use all other ideas presented in my prediction models, including gradient boosted regression trees, to directly predict personality scores from ECA components. This gives insight as to how significant is the benefit of predicting responses to a personality items compared to directly predicting personality scores. To avoid overfitting bias, all analysis have been 10-fold cross validated. The users have been split into 10 bins, 9 are used for training and 1 for testing, and the process is repeated until all bins have been tested. The outcomes of all folds are averaged to produce the final result.

Results

Table 5.3 shows results of this analysis.

Trait	*Direct prediction*	*Predicting individual items*	*Increase in variance accounted for*
Openness	1	1.31	71%
Conscientiousness	1	1.65	172%
Extraversion	1	1.15	32%
Agreeableness	1	1.35	82%
Neuroticism	1	1.59	153%
Mean	1	1.41	99%

Table 5.3 Comparison of different learner models

Overall, using item-level data to train the models is very beneficial to the predictive models, as they are able to look for patterns in the data for specific items and then compute the personality scores accordingly. There is 99% increase in the variance accounted for all the traits.

Comparatively among the five personality traits, neuroticism and conscientiousness are most helped by using the item-level data. Neuroticism and conscientiousness are also traits that highly correlate with life outcomes and are easier to predict from behaviour. This indicates that using the item-level data to train the predictive models helps utilise footprints of behaviour on Facebook which are indicators of a user having various degrees of neuroticism or conscientiousness that would otherwise have been missed. Extraversion is least helped by the item-level data. The most likely explanation is that the greater predictor for extraversion in Facebook data is the size of the user's social network. It correlates well with most of the items from the item-level data as well as the personality score itself.

5.3.4 Study 4: Comparison with state-of-the-art

Current the best published accuracy for prediction of personality from Facebook data is reported by Youyou et al. [152]. That paper also uses MyPersonality dataset. The purpose of this study is to compare the level of correlation of predicted personality traits with self-report personality traits of my method and the best published.

Sample

98,515 worldwide Facebook users from MyPersonality dataset have been chosen for this study. The users satisfy the inclusion criteria outlined in section 2.3. There is a cross-section between the users in my study and Youyou's, however there are also differences. Youyou used users with more than 200 Facebook Likes in her study. The inclusion criteria for inclusion in my study is only 100 Facebook Likes. So the method in this thesis would work with a wider segment of the Facebook population.

Method

All ideas explained in this chapter are used in the predictions for this study, to compare them to current state-of-the-arts. This includes using item-level data for training, using gradient boosted decision trees and using dislikes. To avoid overfitting bias, all analysis have been 10-fold cross validated. The users have been split into 10 bins, 9 are used for training and 1 for testing, and the process is repeated until all bins have been tested. The outcomes of all folds are averaged to produce the final result.

Results

Table 5.4 shows the results of Youyou et al. [152] compared to the predictions made by the models developed in this chapter.

Trait	*Previous best [152]*	*This prediction*	*Increase in variance accounted for*
Openness	.65	.77	40%
Conscientiousness	.52	.68	71%
Extraversion	.55	.71	67%
Agreeableness	.56	.72	65%
Neuroticism	.49	.58	40%
Mean	.56	.70	52%

Table 5.4 Correlations with self-reports, comparative analysis with previous best

All the improvements proposed in this thesis combined add about 52% to the overall variance accounted for compared to the best published results in literature. This is for users with a lower limit of inclusion criteria (100 Likes compared to 200 in Youyou's study [152]). Conscientiousness, extraversion and agreeableness were most helped by the methods proposed in this thesis, while models for openness and neuroticism gained less predictive power over the best published outcomes. Comparative analysis of predictability of various traits is provided in chapter 7.

5.4 Alternative model: Deep learning

An alternative way to do personality predictions is to use deep learning. Deep learning models see only high-level data abstractions (such as a personality score) and construct a graph-based internal structure of a model automatically in a data-driven way [14]. This comes in contrast to models that have their internal structure designed by expert knowledge in the specific domain, such as the model introduced in this chapter.

In a traditional personality test, each item is used to assess a certain aspect of an individual's personality (e.g. political opinions and interest toward arts are aspects of Openness). I use an internal hierarchal model because a lot of aspects of personality, such as the people's political preferences or interest towards arts are very well represented in their Facebook Likes and digital footprints in general. I use the items of a personality questionnaire as our hidden layer as I will be able to do supervised learning of hidden layers from people's responses to each question during training. Theoretically, an automated deep learning algorithm should be able to learn similar hierarchal interactions mainly from the data, assuming that enough data exists.

There are various approaches to deep learning. The choice of the approach is often based on the data and task at hand. Perhaps the most widely used and publicised architecture is Convolutional Neural Networks (CNNs), widely used in computer vision [88, 90], but also for speech recognition as they are very effective at dealing with time-series data [92]. This is because they engineer higher-ordered features using convolution operations based on convolution theorem in mathematics, which relates the time and space domains together. An example of the use of convolution is the Fourier transform [19], where signals can be converted from the time domain into the frequency domain. Fourier transform is widely considered to be one of the most influential algorithms in history, and its variants were also the key algorithms in historical computer vision methodology [67].

Outside of computer vision, when it comes to highly sparse datasets, Unsupervised Pretrained Networks (UPNs) tend to be very helpful [39]. Engineering higher order features from highly sparse datasets as the pretraining stage is similar to treating the first set of hidden layers as principal components and the rest of the architecture is learned from there, which this part can be supervised. Deep Belief Networks (DBNs) [94] are UPNs which allow for a hierarchical layered architecture without allowing the layers to communicate, this is important as we prefer the engineered features to include less dependency and therefore, less redundancy. We also want independent predictors of personality in these methods. To put this theory to test, I developed a multi-layered Deep Belief Network, learned using Hinton's greedy algorithm [70] to compare with the model introduced in this chapter. To ensure this model is suitable, I used this deep learning architecture to predict demographics details of age, gender, relationship status and voting preferences from Facebook Likes, the same variables predicted in section 4.3 and found excellent predictive power.

After engineering features, the most optimised layered structure is the use of four layers with size of each layer is 50. Larger or deeper models begin to overfit at 100,000 users. Overfitting is observed when increasing the complexity of the structure results in increased correlations of the predicted values with the class variable of samples from the training subset, but correlation of predicted values with the class variable in the testing subset is decreased.

Table 5.5 shows the correlations of the predicted personality traits with self-reports, for both the deep learning method and the method introduced in this chapter. The same sample of 98,515 worldwide Facebook users from the MyPersonality dataset were used for both methods. The users satisfy the inclusion criteria outlined in section 2.3.

The deep learning model is implemented using the DeepLearning4J framework [31]. Further explanation of the framework is in appendix A.

Trait	*Deep learning*	*This prediction*	*Increase in variance accounted for*
Openness	.71	.77	18%
Conscientiousness	.60	.68	28%
Extraversion	.66	.71	16%
Agreeableness	.62	.72	35%
Neuroticism	.49	.58	40%
Mean	.62	.70	27%

Table 5.5 Comparative analysis with deep learning

In my experiments, while deep belief networks do outperform the previous best reported scores [152], reported in section 5.3.4. However even with close to 100,000 users, deep learning methods are not able to outperform models constructed by expert psychological knowledge. Therefore we can conclude that they require an even larger sample size to be well-trained and remove the need for expert psychological knowledge in model construction. Larger sample sizes are necessary to investigate if deep learning models are able to outperform expert models for prediction of personality traits from Facebook.

5.5 Summary

In this chapter I have introduced the predictive models of the machine learning system. There are several improvements to the liteature, they include using gradient boosted decision trees compared to linear regression models, using item-level data to train the models in order to be able to predict a user's potential responses to a questionnaire instead of predicting their personality scores directly, and by not mixing contextually relevant but not-liked pages with non-relevant pages to provide a cleaner data.

All the enhancements introduced in this chapter, combined with ECA, as introduced in chapter 4, make the machine learning system capable of doing judgements of personality that can demonstrate psychometric properties in a way that a self-report questionnaire does. This is investigated in chapter 6.

Chapter 6

A Passive Personality Test

In chapter 4, I introduced ECA, a method to effectively reduce the dimensions of the highly sparse data of online social networks while eliminating noise, reducing redundancies and keeping maximum variance among only relevant variables. In chapter 5, I explained how the predictive model works, and how its predictive power is increased by the enhancements that were introduced. In this chapter, I demonstrate that personality predictions made by a machine learning system using the methodology described chapters 4 and 5 can demonstrate strong psychometric properties of reliability, internal consistency, external validity and discriminant validity.

For simplicity, I refer to personality predictions of my method as *Machine Learning* predictions, or *ML-rated personality*. Personality scores computed by a traditional 100-item IPIP questionnaire of the same users is reported as *self-report personality*.

6.1 Sample

Data used for this analysis is collected by the MyPersonality project, as introduced in chapter 2. It consists of 98,515 volunteers (worldwide) who completed a 100-item IPIP personality questionnaire. It includes Facebook Likes and the following demographics: Age, gender, relationship status, education, location, hometown, and number of friends.

Inclusion criteria is explained in section 2.3. To summarise, all users in this study have a minimum of 100 Likes. This ensures that users who have very recently joined the network are not included. The mean number of Likes of the users used in this study is 194, which is fewer than the average number of Likes of Facebook users, as reported by Youyou et al. [152], so this method works for most users.

6.2 Cross-validation

I apply a 10-fold cross-validation to avoid overfitting bias in the results. The users are randomly divided into 10 subsets, 9 subsets are used for training and one subset for evaluation. This is repeated 10 times to make personality predictions of the whole dataset. Separation of the training and testing subsets is done at step zero, before construction of the models and before dimensionality reduction. This ensures that there is no information leak between the training and testing subsets.

6.3 Analysis of reliability and validity

In order to assess the psychometric properties of the ML-rated personality traits, I use four different studies: Self-ML agreement in order to measure the reliability of the test, split-half correlations to measure internal consistency, external validity to measure the predictive power of the ML-rated personality in terms of other life outcomes, and discriminant validity to ensure that the traits that are designed to be independent from each other, are indeed so.

6.3.1 Self-ML agreement

The self-ML agreement is the primary way to measure accuracy of the predictions, and it is the criteria by which the model was designed and trained. I use test-retest correlation after 1 year to measure reliability of the self-report questionnaire. Figure 6.1 demonstrates the correlations of ML-predicted personality with self-reports, and compares them to 1-year test-retest correlation of self-reports.

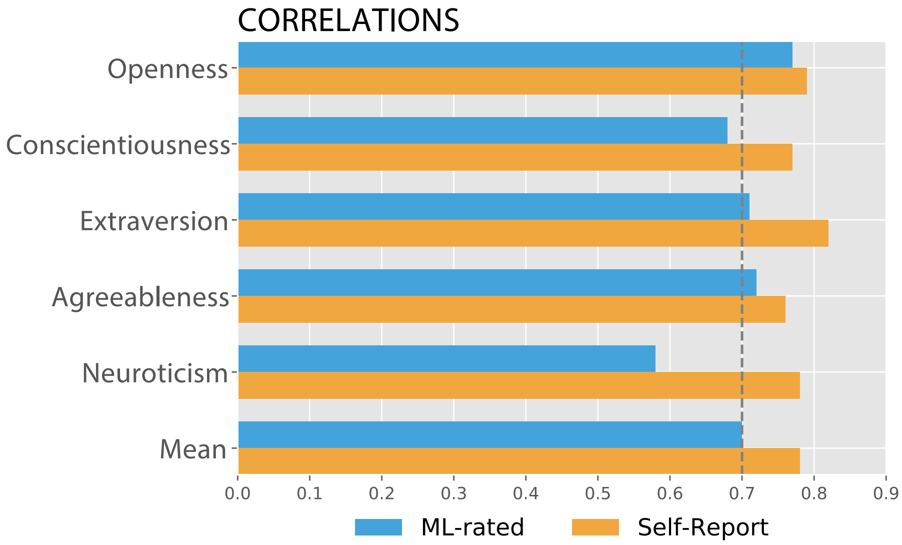


Fig. 6.1 Comparison of Self-ML agreement and 1 year test-retest correlation of self-reports. The results are corrected for attenuation. The mean were averaged using Fisher's r-to-z transformation.

Generally, Psychometric rules of thumb require a personality questionnaire to have a reliability between $r = .7$ and $r = .9$ in order to measure a trait reliably for an individual [124] for application in services tailored at individuals such as career advice, recruitment and financial services. Here we observe that ML-rated personality satisfies this condition and is only worse than 1-year test-retest correlations of a self-report test by a small margin. It should also be noted that Facebook data used in these predictions are not time stamped, and are accumulated over a long period of time, years or even more than a decade. However, the new Facebook Graph API includes time stamped data as well. So practically, only recent Likes can be used for prediction, or can be given a stronger weight compared to older data, to simulate a more recent personality test which should theoretically improve accuracy of the predictions.

Openness is the most predictable trait. This might be because of the way social networks are used as mediums of expression and discussion about politics, philosophy and arts. There are many pages on Facebook dedicated to these subjects and liking them is a great indicator of higher degrees of openness.

Contentiousness, extraversion and agreeableness are predictable to about the same degree. Neuroticism is the least predictable trait and the only trait with correlations of less than $r = .7$. This trend is unanimous in literature of predicting personality from Facebook Likes [86, 152] and might be because the options to leave evidence of low or higher levels of neuroticism are limited on Facebook Likes. Neuroticism however, is better apparent from linguistic cues [97] and Park et al. [109] was able to predict neuroticism at the same level of accuracy as other traits from Facebook status updates.

Further investigation of the predictability of the traits, as well as an analysis of facet-level and item-level predictability is performed in sections 7.1 and 7.2.

6.3.2 Split-half correlations

As a further estimate of reliability I use split-half correlations to measure the internal consistency of the ML-rated personality traits. I randomly split the Likes of each user into two subsets of equal size and use each one separately to predict their personality traits and correlated their results. The users in this pool had at least 300 Likes. Overall, we can observe in table 6.1 that we are moving into the range of the accuracy of traditional psychometrics tests on individual level where has not been possible before. Youyou et al. [152] reported a mean split-half correlation of $r = .62$, which is the best reported result in literature.

	Split-half correlations	
Trait	*ML-rated*	*Self-report*
Openness	.79	.87
Conscientiousness	.67	.89
Extraversion	.73	.91
Agreeableness	.71	.90
Neuroticism	.54	.88
Mean	.70	.89

Table 6.1 Split half correlations of ML-predicted and self-report personality. The mean was averaged using Fisher's r-to-z transformation.

Similar to self-ML agreements, openness shows a better scale reliability compared to the rest of the traits ($r = .79$). This can be explained by the nature of Facebook, as a large amount of what people do on Facebook is following news and politics, it is reasonable to expect that splitting a person's Likes into half would still involve pages that can lead the model into the right predictions. Facebook is also a big online forum for discussions and activity towards arts, another important indicator for openness.

Extraversion, conscientiousness and agreeableness have relatively similar split-half correlations (around $r = .70$), which is acceptable but lower than ideally desired. This demonstrates a higher dependency of the predictive models to the inclusion of pages that are highly important to the predictions, compared to openness. Neuroticism, the least predictable trait in terms of Self-ML agreements, also has the lowest rate of split-half correlations. This indicates that compared to other traits, Facebook pages that are indicative of lower or higher degrees of neuroticism are more unique, since the data that they contain is missed when the Likes are split into half.

These statements however, should come with certain caveats. The way the dimensions of the data are reduced, and the way the predictive models work can affect the variance among the split-half correlations of different traits. None of the prior work report the per-trait split-half correlation of the five-factors personality traits that are predicted from online social networks, therefore there is no point of outside comparison. However, changing the parameters of the model can help with this investigation.

Optimising the methods for better internal consistency

Low levels of split-half correlations indicate the possibility of existence of pages on Facebook that are highly beneficial to the prediction, and their removal from one of the halves lowers the overall split-half correlations. This can be a byproduct of the way the dimensions of the data are reduced, since that is when noise removal and handling of redundancy happens. In section 4.2.2, I explained how the parameters of the Entropic Component Allocation (ECA), the dimensionality reduction method introduced in chapter 4, need to be adjusted. In this section, I investigate if ECA can be adjusted in a way to improve internal consistency, and I also investigate if the variance among the internal consistency of different ML-rated traits changes as parameters of ECA are tuned.

To reduce dependency to single points in the data, $threshold_c$ needs to be decreased. This allows more variables to be be seen as redundant, which is counter-intuitive to the task, however since the model preserves variance among redundant variables, this reduces the chance of their dismissal. $threshold_d$ needs to be increased, to ensure that the added variance among the redundant variables are preserved. I reduced $threshhold_c$ by 20%, and increased $threshhold_d$ from 95% to 99%. This is an optimised value achieved from doing several experiments. This resulted a 7% increase in the number of features. Table 6.2 shows the split-half correlations of ML-rated predictions with the adjusted parameters.

	Split-half correlations	
Trait	*ML-rated*	*Self-report*
Openness	.79	.87
Conscientiousness	.70	.89
Extraversion	.74	.91
Agreeableness	.72	.90
Neuroticism	.64	.88
Mean	.72	.89

Table 6.2 Optimised split-half correlations of ML-predicted and self-report personality. The mean was averaged using Fisher's r-to-z transformation.

Overall, split-half correlations have been increased to $r = .72$. The only significant change is in neuroticism, where it has been increased from $r = .54$ to $r = .64$. This came at the expense of more features, which means slower training and testing, and a slight reduction in self-ML agreements, which now averages to $r = .69$. This demonstrates that being less rigorous in allowing variables into the models might help improve internal consistency, at the expense of added training time.

This demonstrates that optimisations in dimensionality reduction algorithms or predictive models can influence the internal consistency of the personality assessment. This is a very interesting finding which demonstrates that sole focus on enhancing correlations with self-reports might not be the best way to build more reliable tools of passive psychological assessment. This also presents a challenge in methods, as machine learning models, such as decision trees or linear regression models, utilise a loss function as part of their training process. The role of the loss function is to evaluate the predictive performance of the model during training. Loss functions almost always measure the error in prediction of the class variable, which the model aims to minimise. Therefore the training model is always guided to maximise correlations between the predicted outcome and the observed outcome, in this case between the predicted personality and self-reports. This shows that enhancing the internal consistency of passive psychometric tests might require creating entirely new loss functions dedicated to focusing on internal consistency. This is a very interesting area of further research into the methods of passive psychological assessment.

6.3.3 External validity

A third and perhaps most important measurement of validity of a test is its predictive power of external variables. Five-factor personality has been demonstrated to have a good ability to predict consequential life outcomes [106]. In this section, each of the five-factor personality traits are used to predict a series of external factors. These include: satisfaction with life, sensational interests, impulsivity, self-monitoring, Schwartz's values and depression. These data were collected during the MyPersonality project and were cross-referenced with the users from the inclusion criteria (section 2.3).

Figure 6.2 outlines the predictive power of ML-rated openness compared to self-report openness.

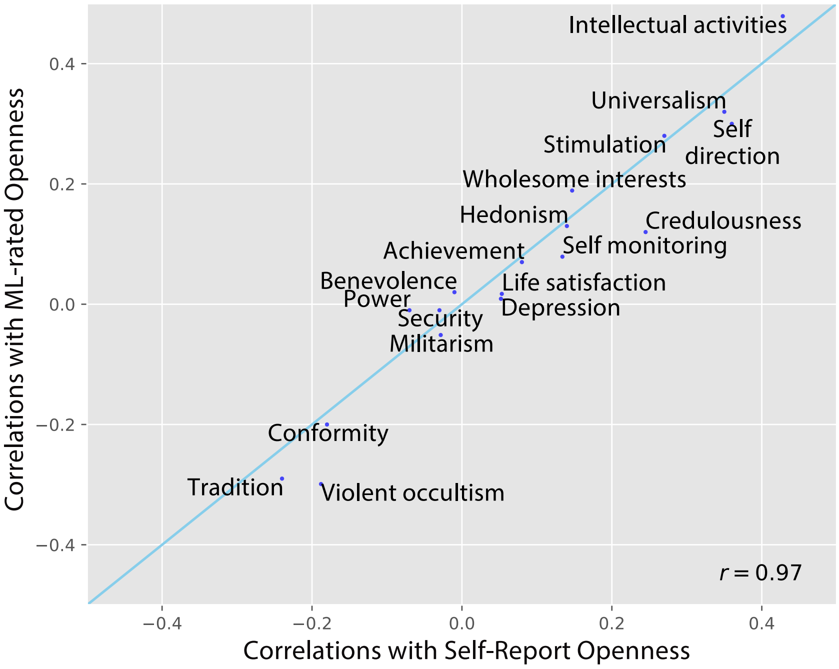


Fig. 6.2 Comparison of the predictive power of ML-rated openness and self-report openness. The correlation coefficient on the lower right side of the graph is the correlation between the predictive power of ML-rated and self-report openness, after applying the r-to-z transformation to the original correlation coefficients of the predictions.

There is excellent agreement between the accuracy of the predictions of external factors by ML-rated openness compared to self-report openness. Self-report openness is better at predicting violent occultism and credulousness while ML-rated openness is slightly better at predicting intellectual activities and power. However at an overall r = .97, ML-rated openness is as good as self-report openness in terms of external validity.

Figure 6.3 outlines the predictive power of ML-rated conscientiousness compared to self-report conscientiousness.

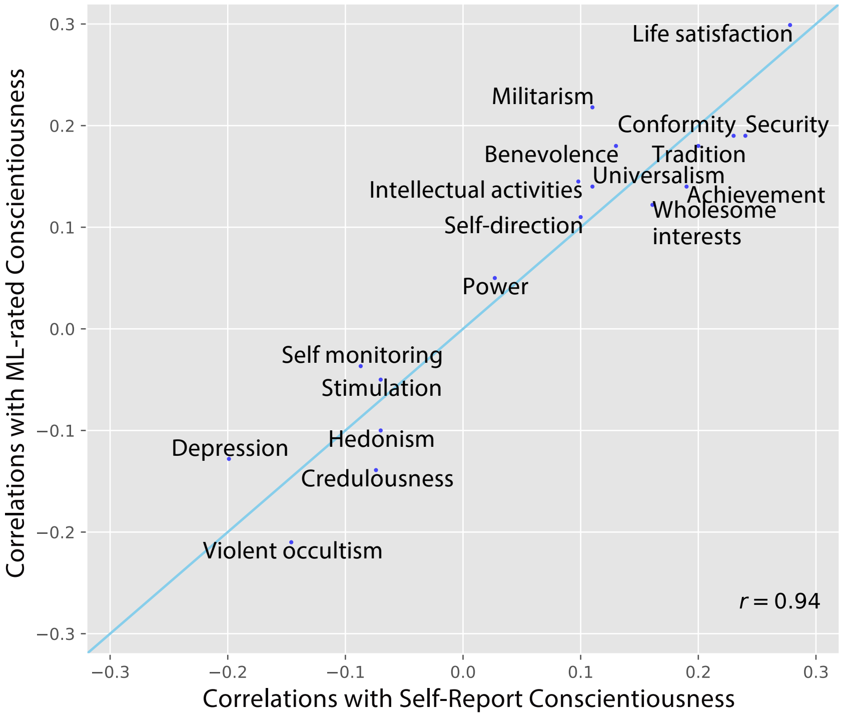


Fig. 6.3 Comparison of the predictive power of ML-rated conscientiousness and Self-Report conscientiousness. The correlation coefficients in the lower right side of each graph is the correlation between the predictive power of ML-rated and self-report conscientiousness, after applying the r-to-z transformation to the original correlation coefficients of the predictions.

Figure 6.4 outlines the predictive power of ML-rated extraversion compared to self-report extraversion.

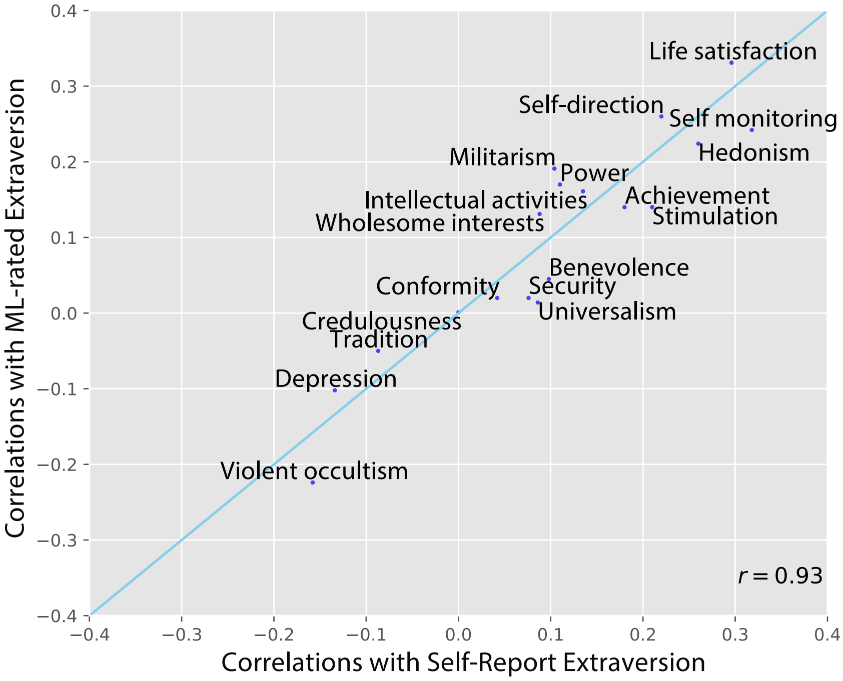


Fig. 6.4 Comparison of the predictive power of ML-rated extraversion and self-report extraversion. The correlation coefficients in the lower right side of each graph is the correlation between the predictive power of ML-rated and self-report extraversion, after applying the r-to-z transformation to the original correlation coefficients of the predictions.

There is excellent agreement between the accuracy of the predictions of external factors by ML-rated extraversion compared to self-report extraversion. For almost all external validity variables, they have a relatively similar predictive powers. At an overall $r = .93$, ML-rated extraversion is as good as self-report extraversion in terms of external validity.

Figure 6.5 outlines the predictive power of ML-rated agreeableness compared to self-report agreeableness.

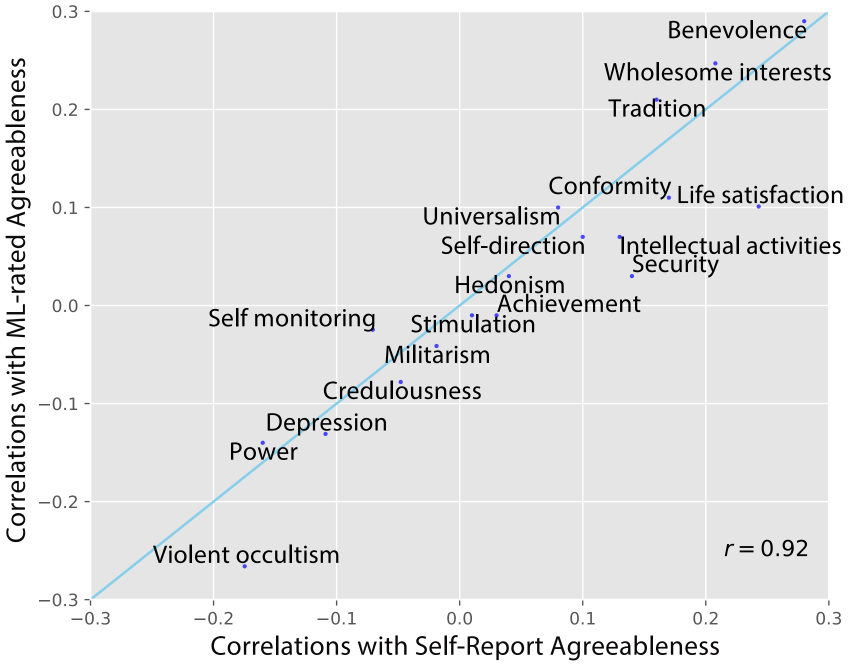


Fig. 6.5 Comparison of the predictive power of ML-rated agreeableness and self-report agreeableness. The correlation coefficients in the lower right side of each graph is the correlation between the predictive power of ML-rated and self-report Agreeableness, after applying the r-to-z transformation to the original correlation coefficients of the predictions.

There is excellent agreement between the accuracy of the predictions of external factors by ML-rated agreeableness compared to self-report agreeableness. Self-report agreeableness is better at predicting violent occultism, intellectual activities, life satisfaction and security. ML-rated agreeableness is slightly better at predicting self-monitoring, tradition and wholesome interests. However at an overall $r = .92$, ML-rated agreeableness is as good as self-report agreeableness in terms of external validity.

Section 6.3.1 showed that neuroticism was the least predictable trait. Figure 6.6 outlines the predictive power of ML-rated neuroticism compared to self-report neuroticism. For almost all external validity variables, they have a relatively similar predictive powers. At an overall $r = .98$, ML-rated neuroticism is as good as self-report neuroticism in terms of external validity.

This demonstrates that the slightly lower self-ML agreement and split-half correlations of neuroticism compared to other traits does not translate to a lower level of agreement with self-reports in terms of prediction of outcomes variables.

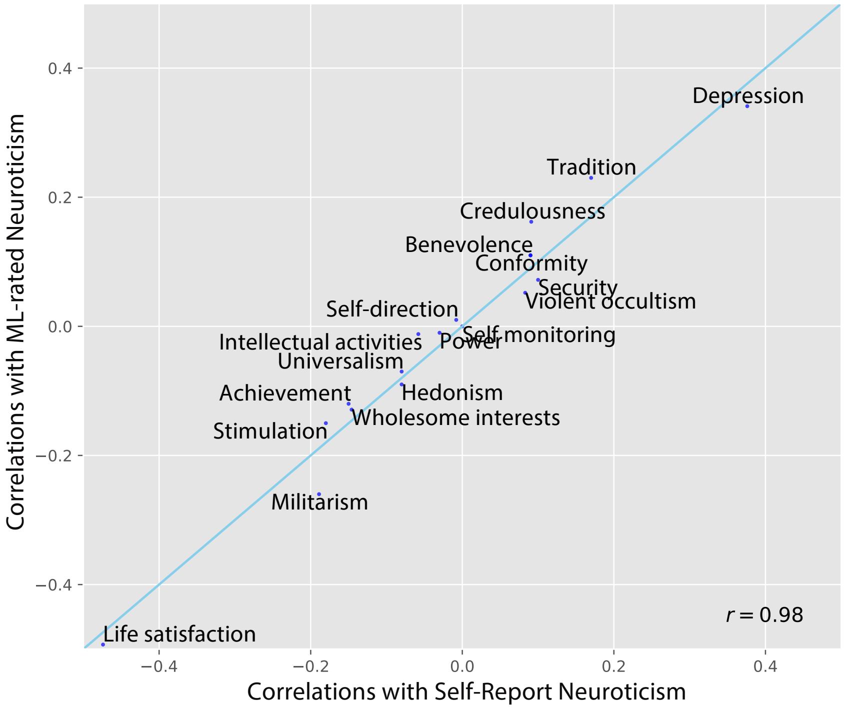


Fig. 6.6 Comparison of the predictive power of ML-rated neuroticism and self-report neuroticism. The correlation coefficients in the lower right side of each graph is the correlation between the predictive power of ML-rated and self-report neuroticism, after applying the r-to-z transformation to the original correlation coefficients of the predictions.

With *r* ranging from .92 to .98, in terms of external validity, ML-rated personality has a very similar predictive power compared to self-report personality. It is important to note that the correlation coefficients reports how close ML-rated personality and self-report personality are able to predict external factors, these are not correlations with the external factors themselves. Table 6.3 covers the actual values of correlations with external variables.

Per-trait comparisons of external validity between self-reports and predicted personality from Facebook Likes have not been reported in literature. However, Park et al. [109] reported the external validity of the predicted personality traits compared to the self-report personality, and found that their predictions of personality from language on social media had great agreement with the predictions of self-report personality in terms of predicting external variables. Table 3.5 outlines the level of agreement between the predictive power of predicted personality from language on Facebook and self-report personality in Park's study compared to the findings of this section.

Youyou et al. [152] reported a lower combined predictive power of their predicted personality traits compared to self-reports. Table 6.3 demonstrates the combined predictive power of all ML-rated personality scores, compared to self-report personality scores. The mean $r = .43$ for ML-rated personality demonstrates equal or better external validity compared to self-report personality of mean $r = .40$.

		Predictions by personality	
Group	*Item*	*ML-rated*	*Self-report*
Life Satisfaction		.47	.50
Sensational Interests	Wholesome interests	.33	.31
	Intellectual activities	.53	.52
	Violent occultism	.41	.28
	Militarism	.46	.26
	Credulousness	.39	.26
	Mean	.42	.33
Impulsivity		.48	.53
Self-monitoring		.44	.39
Schwartz's values	Conformity	.15	.17
	Tradition	.20	.14
	Benevolence	.17	.13
	Universalism	.27	.21
	Self-direction	.19	.14
	Stimulation	.22	.21
	Hedonism	.17	.12
	Achievement	.07	.11
	Power	.26	.27
	Security	.05	.12
	Mean	.18	.16
Depression		.41	.38
Mean		.43	.40

Table 6.3 Comparison of external validity of ML-rated personality with self-report personality. Predictions are linear regressions using the big five traits as predictors. The results are averaged using Fisher's r-to-z transformation and weighted by sample size.

6.3.4 Discriminant validity

Discriminant validity indicates how far a trait that is meant to be independent from other traits is indeed so. Table 6.4 shows how the self-report and ML-rated personality traits correlate with each other.

		Self-Reports					*ML-rated*				
		O	C	E	A	N	O	C	E	A	N
Self-Reports	O										
	C	.04									
	E	.14	.17								
	A	.04	.19	.17							
	N	-.06	-.30	-.34	-.34						
ML-rated	O	**.77**	-.08	-.09	-.02	-.01					
	C	-.09	**.68**	.14	.13	-.12	-.03				
	E	-.07	.11	**.71**	.11	-.15	.10	.21			
	A	-.03	.12	.14	**.72**	-.12	-.07	.20	.23		
	N	-.02	-.12	-.13	-.09	**.58**	-.05	-.22	-.24	-.31	

Table 6.4 Comparison of the discriminant validity of ML-rated personality with self-report personality. Values represent correlation (r). O, Openness; C, Conscientiousness; E, Extraversion; A, Agreeableness; N, Neuroticism;

ML-rated personality shows better discriminant validity in 6 out of 10 possible relationships between the personality traits. This includes the correlations of neuroticism with all four traits, and correlations of conscientiousness with all traits except agreeableness. The overall mean is calculated for self-report personality (mean $r = .18$) and ML-rated personality (mean $r = .17$), and shows that the ML-rated personality traits are equal or better at only measuring the trait that they intend to measure.

Park et al. [109] reported the discriminant validity of their predicted personality scores compared to self-reports, as reported in table 3.4. With mean discriminant correlations of $r = .28$ in various traits in Park's predictions, the ML-rated personality traits in this study are less dependant on each other compared to traits predicted from language features of status updates on Facebook.

6.4 Discussion

6.4.1 Reliability and validity

Using machine learning methods described in chapters 4 and 5, to predict five-factor personality from Facebook Likes (mean r = .70) improved upon the previously best published accuracy (mean r = .56) [152] - an increase in variance accounted for of 52%. Previous studies [86, 152, 109] only used participants from the United States, so the prediction in this thesis are on a more diverse sample from worldwide, with fewer average Likes per user. The increase in accuracy is valuable for practice since it differentiates between a prediction that is good enough for analysis at group level (e.g. research into the differences between groups or communities) and a prediction that is good enough at the individual level (accurate and reliable assessment of one's personality for a wide variety of applications). As well as being reliable, a psychometric test needs to be demonstrated to be valid in predicting external factors. Predictions of a range of outcomes showed that predicted personality is as externally valid as self-report tests (mean r for 19 outcomes = .43 compared to .40 for self-report). In terms of discriminant validity, ML-rated personality traits were demonstrated to be more as independent from other ML-rated traits as self-reports traits are independent from each other.

Overall, openness was the the most reliable trait to predict from Facebook, in terms of agreement with self-reports and internal consistency. This can be attributed to the fact that Facebook is a major medium for access to politics, news and arts, therefore users leave more footprints that share informations about their attitudes towards the subject, and as a result the prediction of openness becomes more reliable. This is in contrast to other mediums of personality judgement, for example career performance, where conscientiousness is the trait that is easier to predict. This is further investigated in chapter 7.

In terms of external validity, all ML-rated traits demonstrated a similar level of predictive power to self-reports in predicting external and outcome variables, with an agreement of r ranging from .92 to .98.

6.4.2 Bias

Normally in psychometric testing, item responses are scored the same way for all participants. In fact, psychometricians use differential item functioning to look for item that are biased against demographics such as gender and ethnicity [136]. The goal is to produce a test whose result depends only on the latent trait being measured, as opposed to the demographics of the test taker. Sometimes to control for bias, questionnaires use different norms for different groups, for example the Empathy Quotient Scale uses different norms for males and females [11]. The machine learning models introduced here remove bias earlier in the process, by using decision trees to intentionally look for differences in the meaning of Likes for different groups and producing different prediction models for those people. This corrects for bias using rules learned from the big data. Critically though, this method controls for the bias only as far as the gold standard test is unbiased, as the prediction model is trained to produce those scores. Therefore, predictions from digital footprints cannot replace high quality self-report questionnaires; instead, they make their application more practical by using existing digital data to simulate their use.

Psychological assessment by digital data overcomes several disadvantages of self-reports, such as mood and memory influences as data is recorded over a long period of time, self-enhancing becomes much more difficult as shaping digital data to present an enhanced version to the model is typically very difficult without deep understandings of how the prediction models work (unlike self-reports where the test-taker can simply lie). Reference-group effect is also removed as the prediction models do not take into account the social circle of the individual.

6.4.3 Application limitations

The reliance of this method on digital data can be a limiting factor in deployment, as not everyone uses online services, however in the modern world, it is difficult to avoid making digital footprints, as they include an individual's use of online social networks, their cellphone records, their browsing and search histories, their e-mails, and their purchases as recorded by credit or loyalty cards.

Facebook has recorded over 2 billion monthly *active* users in 2017, and 1.32 billion daily active users in June 2017 [42]. Twitter has reported 313 million monthly active users [139], Instagram has reported over 500 million monthly active users [76]. Other social networks such as Qzone, LinkedIn [1], Google+[2] and Tumblr[3] have reported hundreds of millions of

[1] https://www.linkedin.com

[2] https://plus.google.com

[3] https://www.tumblr.com

active users as well. While a specific social network might become more or less popular as fashion dictates, it is likely that businesses and society will collect increasingly large and detailed records of individuals' behaviour.

Another major limitation of the method is the need for training data for the specific social network, and for a large number of users. Expanding this functionality to a new social network will require new studies where users of the new social network need to perform personality tests and their data needs to be collected, only then we are able to train similar predictive models that can provide passive psychometrics. Progress of time is also a limitation, as data on social networks also changes with time. What is fashionable and popular at a certain time might not be so in a few years or decades. This translates to a constant need for data collection and cross-examination, which can be done as the service is being used but also might require completely new data collection when after a long time.

Further discussions on limitations of the passive approach, including ethical concerns are covered in chapter 9.

6.4.4 Convenience

The most important advantage of passive psychological assessment using machine learning is convenience for both participants and administrators. It allows a test to be pushed to millions of people quickly, and it requires no time commitment from the participants. Researchers who want to record an individual's personality for a study can only request access to their social networking data, rather than asking the individual to complete long questionnaires, and in practice, an employer can easily request access to social networking data of the job seekers and infer their personality instead of requiring them to participate in self-report tests, which itself is prone to self-enhancement bias.

6.5 Summary

In this chapter, I used the methodology introduced in chapters 4 and 5 to construct a proof-of-concept machine learning system capable of accurate personality predictions. The predicted personality scores were assessed in terms of correlations with self-reports, internal consistency, external validity and discriminant validity and were found to be very similar to self-reports in terms of their psychometric properties.

While this chapter mostly focused on investigating psychometric properties of the outcomes of the machine learning system, chapter 7 investigates why the predictions are the way they are. This includes analysis of predictability of various traits, predictability item-level data, and how personality scores relate to predictability of users on Facebook.

Chapter 7

Analysis of Predictability

In chapters 4 and 5, I introduced a dimensionality reduction method and predictive learning models that are capable of accurate prediction of personality of users from social networking data. In chapter 6, I investigated the psychometric properties of the outcomes of the machine learning system when used to predict five-factor personality in terms of reliability and validity.

Five factor personality is very effective in predicting behaviour and important life outcomes [106, 121], such as happiness, spirituality, physical health, peer, family and romantic relationships, occupational choice and performance, political and ideological values, and criminality. This is partly why this research started, to examine whether it is possible to go from footprints of behaviour (online in the work of this thesis) to accurate assessment of personality.

Being able to predict personality scores presents a unique opportunity to investigate if the five-factor personality traits can predict a new outcome that has not been studied before, which is online predictability.

In section 7.1, I examine the level of predictability for each of the personality traits and compare them to the literature. In this section I also investigate why neuroticism is the most difficult trait to predict, as observed in chapter 6, and draw a new study which further investigates a potential answer. In section 7.2, I perform new studies to investigate the per-item perfectibility for each of the 100-items of MyPersonality's personality test, from the IPIP inventory and discuss the results. In section 7.3, I investigate what role demographics and personality play in explaining how predictable a specific user is and perform relevant studies. Finally, section 7.4 summarises the findings.

7.1 Predictability of the five-factor model

When it comes to predicting behaviour and important life outcomes, different traits are better or worse at predicting different things. Conscientiousness is best able to predict academic [104] and work performance [12]. My own analysis in section 6.3.3 showed that both ML-rated and self-report neuroticism are best at predicting depression and life satisfaction. Overall, conscientiousness is the most predictable trait from behaviour. This brings about the expectation that the two traits should be easiest to predict here, as well. However, this was not observed in chapter 6. Table 7.1 shows the correlations between the ML-rated and self-report personality traits, as discussed in chapter 6.

Trait	*Self-ML agreement*
Openness	.77
Conscientiousness	.68
Extraversion	.71
Agreeableness	.72
Neuroticism	.58

Table 7.1 Correlation of ML-rated and self-report five-factor personality model

Openness is the easiest trait to predict and neuroticism is the most difficult trait to predict. This is not a method-dependent limitation, as this pattern exists in all predictions made in chapter 5 with different model decision. Table 7.2 shows the correlation between predicted and self-report personality traits in literature.

Study	*O*	*C*	*E*	*A*	*N*
Kosinski et al. [86]	.43	.29	.40	.30	.30
Youyou et al. [152]	.65	.52	.55	.56	.49
Park et al. [109]	.41	.26	.36	.41	.39
Golbeck et al. [50]	.65	.59	.55	.48	.53

Table 7.2 Correlation of ML-rated and self-report five-factor personality models in literature. O, Openness; C, Conscientiousness; E, Extraversion; A, Agreeableness; N, Neuroticism;

Across all studies, openness is the easiest trait to predict, which is inline with the results of this thesis as well. As explained in earlier chapters, this is most likely the result of the way Facebook is used as a medium for accessing news, politics and an online forum to access and discuss topics of arts. Kosinski et al. [86] reported conscientiousness as the most difficult trait to predict, with agreeableness and neuroticism being only slightly easier. Kosinski's study is also based on MyPersonality and used a lot of the same users as this thesis, but with a narrower inclusion criteria, as described in chapter 3. Similarly, another study which used MyPersonality is Youyou et al. [152]. It also reports neuroticism as the most difficult trait to predict.

Park et al. [109] uses MyPersonality, but instead of Facebook Likes, they use the user's status updates to train their models. They found conscientiousness to be the most difficult trait to predict. Golbeck's study also primarily uses language features on social media and it finds agreeableness to be most difficult to predict.

There is no other literature focusing on the predictability of personality traits from Facebook data. However, having access to personality scores and predicted personality, I can investigate if personality traits themselves correlate with predictability of users. This is what I investigate in the next study in section 7.1.1.

7.1.1 Study 1: Can personality explain predictability?

Five-factor personality traits are very informative about the behaviour and life outcome of the individuals, as outlined earlier in the chapter. Here I investigate if they can also explain predictability of users. This is to test the hypotheses outlined in section 2.1.1.

Previous literature suggested that individuals with higher scores on openness and extraversion use social networks such as Facebook more often and share more information about themselves [6, 56, 6, 149], which translates to leaving more footprints which in turn can make the predictive models more powerful. In this section, I investigate if openness and extraversion are indeed correlated with higher degrees of predictability. Conscientiousness is negatively correlated with the use of social networks [125, 149], but it is positively correlated with being organised and efficient online [127], therefore it is interesting to see how conscientiousness relates to predictability.

Agreeableness is not correlated with using social networks more or less often [123], but it is related to how users present themselves on social media. Highly agreeable individuals tend to present a more accurate picture of themselves on social media and [91] and are less concerned with self-presentation [127]. Neuroticism is correlated with seeking acceptance on social media [127] and presenting an ideal image rather than a true image of the self [91]. This can complicate personality judgements from the data, as data artificially tuned to look a different way can indeed confuse the predictive models.

Sample

The entire 98,515 that were tested in chapter 6 were involved in this study. They all satisfied in the inclusion criteria outlined in section 2.3.

Methods

Using methodology explained in chapter 5, the ML-predicted personality scores of all the users have been predicted from their respective cross-validation iteration and are compared with their self-report personality to calculate the prediction error. Self-report personality traits are then used in a linear regression model to predict the prediction error for each trait and goodness of fit is measured using R^2.

Results

Table 7.3 illustrates the predictive power of each personality trait in explaining the errors observed for each prediction of each personality trait.

Neuroticism proves to be the greatest predictor of prediction error in personality predictions. It is positively correlated with prediction error for all five personality traits. This means neurotic people are the most unpredictable in terms of their personality traits from online social networks. This makes sense as presenting yourself in an ideal rather than realistic way, a tendency that neurotic individuals have on Facebook [127], will add noise to the Facebook data and make it less useful for predictions. Ross et al. [123] reports that neuroticism is positively correlated with using the Facebook wall compared to other features, this can also contribute as in this thesis I am only using Facebook Likes for the predictions.

	Openness error	Conscientiousness error	Extraversion error	Agreeableness error	Neuroticism error
Openness	N	$.08^{(-)}$	$.09^{(-)}$	$.06^{(+)}$	N
Conscientiousness	$.09^{(-)}$	N	N	$.04^{(-)}$	N
Extraversion	$.02^{(-)}$	$.05^{(-)}$	N	$.02^{(-)}$	N
Agreeableness	N	N	N	N	N
Neuroticism	$.07^{(+)}$	$.12^{(+)}$	$.12^{(+)}$	$.03^{(+)}$	$.09^{(+)}$

Table 7.3 Results of analysis of predictability of users as a function of their personality trait in terms of R^2. All are statistically significant at $p < .01$, N means no statistical significance.

Openness is negatively correlated with prediction error of conscientiousness and extraversion. Therefore, the people who score high on an openness scale can have more accurate predictions for their conscientiousness and extraversion from Facebook data while their predictions for agreeableness is less accurate. This is mostly inline with expectations, as openness is correlated with an increased usage of social networks and users who use social networks more often will leave more footprints which means more data for predictive models. Agreeableness has no effect on the prediction accuracy of any traits. Extraverts are also slightly more likely to have accurate predictions of openness, conscientiousness and agreeableness.

Conscientious individuals use social networks less often, however according to these results, they are not less predictable. In fact conscientiousness is positively correlated with predictability of openness and agreeableness. This might be due to the way conscientious individuals use Facebook. Conscientious individuals are more careful when they use Facebook [91, 101] and are less interested in attention seeking and ideal self-presentation [127]. This can indicate that their footprints on Facebook are more authentic and informative of their private traits compared to individuals with lower conscientiousness.

The relationship between personality traits and predictability has not been studied in literature, therefore I can make no direct comparisons. However, to investigate if any of the above outcomes can be related the choice of predictive model rather than being indicative of aspects of personality theory, I repeat the above test with the use the a strong learner decision tree instead of gradient boosting. This results in less accurate predictions, however the patterns should be similar if the results are thought to be universal, the two methods were compared directly in section 5.3.2. Table 7.4 shows the results of this investigation.

	Openness error	Conscientiousness error	Extraversion error	Agreeableness error	Neuroticism error
Openness	N	$.09^{(-)}$	$.10^{(-)}$	N	N
Conscientiousness	$.07^{(-)}$	N	N	$.05^{(-)}$	N
Extraversion	N	$.07^{(-)}$	N	N	N
Agreeableness	N	N	N	N	N
Neuroticism	$.09^{(+)}$	$.11^{(+)}$	$.12^{(+)}$	$.04^{(+)}$	$.10^{(+)}$

Table 7.4 Results of analysis of predictability of users as a function of their personality trait, using a strong learner instead of a combination of weak learners, in terms of R^2. All are statistically significant at $p < .01$, N means no statistical significance.

We observe the same pattern for neuroticism, however extraversion is no longer correlated with prediction accuracy of openness or agreeableness. Openness is also no longer negatively correlated with prediction accuracy of agreeableness. Other patterns are similar to the findings of table 7.3. Therefore, we can conclude that the findings do translate across different predictive models.

In addition to being a statement about personality theory and five-factor's predictive power in life outcomes and external variables, this is also a statement about Facebook as a communication medium and a platform for social psychology research. These findings demonstrate that Facebook is a less effective medium for accurate psychological assessment of highly neurotic individuals. In other words, the footprints which they leave on Facebook are less useful. In contrast, Facebook is a more effective medium for individuals who score high on openness, conscientiousness and extraversion.

With neuroticism being the greatest predictor of unpredictability, it becomes an interesting question as to what extent the predictions for less neurotic individuals are more accurate than the results for the entire population.

7.1.2 Study 2: Low neuroticism and predictability

In this study, I investigate the reliability of the ML-rated personality for users with low neuroticism and compare that to the entire user-base. The purpose of this study is to see how much more accurately are we able to predict all five factor personalities for peoples who have lower than average neuroticism. Ideal self-presentation, in contrast to authentic self-presentation, is a pattern that is positively correlated with neuroticism, and this is a risk in personality prediction. A similar risk also exists in self-report questionnaires in terms of self-enhancement bias [89]. Self-enhancement is also positively correlated with neuroticism [110].

Sample

Out of the 98,515 users who we investigated in chapter 6, only the half in the lowest 50 percentiles of the neuroticism scale are included.

Methods

Using methodology explained in chapter 5, the ML-rated personality scores of the users has been predicted and compared with their self-report personality. To avoid overfitting bias, all analysis have been 10-fold cross validated. The users have been split into 10 bins, 9 are used for training and 1 for testing, and the process is repeated until all bins have been tested. The outcomes of all folds are averaged to produce the final result.

Results

Table 7.5 shows the results of the self-ML agreement of five-factor personality model, for all users and users with lower than average neuroticism.

	Self-ML agreement	
Trait	*All users*	*Low neuroticism users*
Openness	.77	.85
Conscientiousness	.68	.79
Extraversion	.71	.80
Agreeableness	.72	.77
Neuroticism	.58	.71

Table 7.5 Correlation of ML-rated and self-report five-factor personality model, for all users and users with lower than average neuroticism

Overall predictive accuracy for all traits has improved significantly. Neuroticism is still the most difficult trait to predict, even for individuals with lower than average neuroticism. However, for users with lower than average neuroticism, we observe a relatively similar levels of correlation between ML-rated and self-report openness, conscientiousness and extraversion.

Table 7.6 shows the results of the split-half correlations of ML-rated personality, for all users and users with lower than average neuroticism.

	Split-half correlations	
Trait	*All users*	*Low neuroticism users*
Openness	.79	.87
Conscientiousness	.67	.80
Extraversion	.73	.84
Agreeableness	.71	.82
Neuroticism	.54	.80

Table 7.6 Split-half correlations of ML-rated personality, for all users and users with lower than average neuroticism

These findings demonstrate that for users with lower than average neuroticism, the tests are more internally consistent for all five factors. These are more or less equal to the internal consistency of the self-reports, which demonstrates that for users who are less neurotic, we observe better psychometric properties. This confirms the idea that Facebook, as a medium for collection of social networking data, is not only less effective at capturing the sort of behaviour that might be indicative of neuroticism, but is also less effective at capturing all sorts of behaviour for users that are highly neurotic, due to their tendency to self-present ideally rather than authentically.

To investigate how effective Facebook is at collecting footprints that might be indicative of various aspects of personality traits, in the next section I investigate the predictability of a user's responses to each item on MyPersonality's five-factor personality test.

7.2 Per-item predictability

We observed that openness is the most predictable trait while neuroticism is the least predictable, while conscientiousness, extraversion and agreeableness are predictable to relatively the same degree. Recalling from chapter 5, personality predictions are done by first predicting a user's potential responses to a personality questionnaire. In the following studies, I analyse the accuracy of the predictions for each individual item.

7.2.1 Study 1: Predictability of openness

Openness to experience shows an appreciation for new experiences, adventures, curiosity and ideas. Social networks offer an avenue for expression of such interests. For example, users can show their interest towards arts, philosophy, political liberalism, or adventures in the form of Facebook Likes.

Sample

In this study I use 25,000 American Facebook users from the MyPersonality dataset who completed the personality questionnaire. The users satisfy the inclusion criteria outlined in section 2.3. The reduction in the number of users from 98,515 users has been made only to speed up machine learning experiments. The decision to keep American users was made to keep the users contextually relevant to each other, as the same number of users from worldwide would have made the data less contextually rich. The personality questionnaire for openness includes 20 items where the users have the choice of answering them by selecting whether the statement in the item applies to them is *Very inaccurate*, *Somewhat accurate*, *Neither accurate nor inaccurate*, *Somewhat accurate*, and *Very accurate*.

A discussion of how well MyPersonality's openness questionnaire actually captures the trait openness is presented in chapter 2.

Methods

I use ECA, as described in chapter 4 to reduce the dimensions and eliminate noise from the raw data, and use the learning model outlined in figure 5.1 and chapter 5 to predict the response of the users to each individual item. Accuracy is measured by the percentage of times that the model is able to predict the right choice for the item, known as precision.

Results

Table 7.7 shows all 20 personality items with respect to Openness, and how accurately they can be predicted.

The most predictable items on the openness scale are items about political opinions and voting preferences, and interests towards arts. This might be easily explained by the abundance of pages on Facebook relating to those topics. By liking or not liking those pages, the user records a clear footprint in their digital data that can be used to accurate predict their responses to those items in a personality questionnaire. This is the best explanation as to why we are able to predict openness more accurately than other traits.

Items about interest towards philosophy, literature and abstract ideas are less predictable than voting preferences or arts, but are still very predictable. This can be because there are pages decided to such activities, but they are less common and less widely-used than pages relating to arts and politics.

Rank	*Item*	*Prediction Accuracy*
1	Tend to vote for conservative political candidates	58%
2	Tend to vote for liberal political candidates	56%
3	Believe in the importance of art	49%
4	Do not enjoy going to art museums	45%
5	Do not like art	45%
6	Believe that too much tax money goes to support artists	44%
7	Have a rich vocabulary	43%
8	Am not interested in theoretical discussions	42%
9	Am not interested in abstract ideas	41%
10	Avoid philosophical discussions	40%
11	Do not like poetry	35%
12	Have difficulty understanding abstract ideas	35%
13	Rarely look for a deeper meaning in things	34%
14	Enjoy wild flights of fantasy	33%
15	Carry the conversation to a higher level	31%
16	Have a vivid imagination	30%
17	Get excited by new ideas	30%
18	Can say things beautifully	29%
19	Enjoy hearing new ideas	29%
20	Enjoy thinking about things	27%

Table 7.7 Analysis of predictability of Openness items. Precision.

Finally, items about attitude of the user towards hearing about things, thinking about things, their imagination and getting excited about ideas are the least predictable items, and this can be because capturing such interests on social networking websites is rather difficult. Note that data about user-user interactions on Facebook (status updates, likes, comments or sharing other status updates) have not been used in this thesis, which can theoretically improve prediction accuracy for items such as *I enjoy hearing new ideas*, or *I can say things beautifully*. Even in the least predictable items on the openness scale, the prediction accuracy is better than random (20%), which demonstrates that while intuitively a lot of those items might not seem inferable from typical activity on Facebook, the data collectively can determine responses towards those items, to some degree.

At the facet-level, items measuring the values and aesthetics facets are most predictable. This is to be expected as a lot of Facebook pages are about arts and politics. The items measuring the ideas facet are also very predictable, due to abundance of pages on Facebook about philosophy and literature. The items measuring the actions and fantasy facets are least predictable. This might be because the questionnaire was weighed less towards adventurousness. The relative unpredictability of the fantasy facet compared to other facts might suggest that Facebook is a less suitable and successful medium of sharing information on people's imaginations, compared to arts and politics.

7.2.2 Study 2: Predictability of conscientiousness

Conscientiousness shows a need for being efficient, disciplined, neat and thorough. While intuitively difficult to understand from digital data, collectively Facebook data provides significant insight into how a user might respond to items on a conscientiousness scale on a five-factor personality test.

Sample

In this study I use 25,000 American Facebook users from the MyPersonality dataset who completed the personality questionnaire. The users satisfy the inclusion criteria outlined in section 2.3. The reduction in the number of users from 98,515 users has been made only to speed up machine learning experiments. The decision to keep American users was made to keep the users contextually relevant to each other, as the same number of users from worldwide would have made the data less contextually rich. The personality questionnaire for conscientiousness includes 20 items where the users have the choice of answering them by selecting whether the statement in the item applies to them is *Very inaccurate*, *Somewhat accurate*, *Neither accurate nor inaccurate*, *Somewhat accurate*, and *Very accurate*.

A discussion of how well MyPersonality's conscientiousness questionnaire actually captures the trait conscientiousness is presented in chapter 2.

Methods

I use ECA, as described in chapter 4 to reduce the dimensions and eliminate noise from the raw data, and use the learning model outlined in figure 5.1 and chapter 5 to predict the response of the users to each individual item. Accuracy is measured by the percentage of times that the model is able to predict the right choice for the item, known as precision.

Results

Table 7.8 shows all 20 personality items with respect to Conscientiousness, and how accurately they can be predicted.

Rank	*Item*	*Prediction Accuracy*
1	Do just enough work to get by	42%
2	Waste my time	38%
3	Find it difficult to get down to work	37%
4	Mess things up	36%
5	Make a mess of things	35%
6	Get chores done right away	33%
7	Finish what I start	33%
8	Leave things unfinished	33%
9	Carry out my plans	32%
10	Shirk my duties	31%
11	Make plans and stick to them	31%
12	Do things according to a plan	30%
13	Don't put my mind on the task at hand	30%
14	Complete tasks successfully	29%
15	Follow through with my plans	29%
16	Am exacting in my work	29%
17	Need a push to get started	27%
18	Am always prepared	24%
19	Don't see things through	23%
20	Pay attention to details	21%

Table 7.8 Analysis of predictability of conscientiousness items. Precision.

Most predictable conscientiousness items are about general attitude towards work and use of time. This is intuitive, as there are many pages on Facebook dedicated with tips and guides about being efficient at work, time saving and organization. These pages are good indicators of a person being more conscientious.

Least predictable items are about a person's attitude towards attention, motivation and task completion. Especially as the lowest items tend to approach the level of accuracy of a random guess, this demonstrates that Facebook Likes are not able to capture relevant information about how well an individual pays attention to detail, or if they need a push to get started. On the other hand, being self-motivated and attention to details are big determinants of success at work, therefore behaviour that is captured from a person's work outcome is more likely to capture data that is useful for prediction of responses to these items compared to Facebook data. This might explain why conscientiousness is the greatest trait at predicting work-related performance, while it is not the easiest to predict from footprints on Facebook.

All facets of conscientiousness seem to be predictable to about the same level, as items representing their corresponding IPIP scale seem to be evenly distributed among the predictability rankings.

7.2.3 Study 3: Predictability of extraversion

Extraversion shows energetic behaviour, being talkative and outgoing. Footprints relative to partying, size of friendship network, and liking pages related to such activities should intuitively be patterns of such traits.

Sample

In this study I use 25,000 American Facebook users from the MyPersonality dataset who completed the personality questionnaire. The users satisfy the inclusion criteria outlined in section 2.3. The reduction in the number of users from 98,515 users has been made only to speed up machine learning experiments. The decision to keep American users was made to keep the users contextually relevant to each other, as the same number of users from worldwide would have made the data less contextually rich. The personality questionnaire for extraversion includes 20 items where the users have the choice of answering them by selecting whether the statement in the item applies to them is *Very inaccurate*, *Somewhat accurate*, *Neither accurate nor inaccurate*, *Somewhat accurate*, and *Very accurate*.

A discussion of how well MyPersonality's openness extraversion actually captures the trait extraversion is presented in chapter 2.

Methods

I use ECA, as described in chapter 4 to reduce the dimensions and eliminate noise from the raw data, and use the learning model outlined in figure 5.1 and chapter 5 to predict the response of the users to each individual item. Accuracy is measured by the percentage of times that the model is able to predict the right choice for the item, known as precision.

Results

Table 7.9 shows all 20 personality items with respect to extraversion, and how accurately they can be predicted.

The most predictable extraversion items are the ones about the attitude towards parties and social situations. This can be because people generally record their social activities on social networks, and more extraverted people use do it more often.

The least predictable extraversion items are about how a person feels. People usually don't leave footprints on social networking websites that involves insight into, for instance, *if they know how to captivate people*. As such, these items are least predictable. Similar to openness and conscientiousness, even the least predictable items are indeed much better than a random guess, which indicates that digital footprints on social networks, while intuitively might not seem rich enough to have insights insights into a user's feelings about social interactions, include latent insight that are used to make such predictions.

At the facet-level, most items measuring gregariousness are in the top half of the table, demonstrating that gregariousness is the most predictable facet of extraversion on Facebook. Other facets however, seem to be predictable to a similar degree.

Rank	*Item*	*Prediction Accuracy*
1	Am the life of the party	46%
2	Talk to a lot of different people at parties	44%
3	Am skilled in handling social situations	43%
4	Feel comfortable around people	39%
5	Don't talk a lot	38%
6	Would describe my experiences as somewhat dull	37%
7	Start conversations	36%
8	Find it difficult to approach others	36%
9	Make friends easily	35%
10	Retreat from others	34%
11	Keep in the background	34%
12	Am hard to get to know	33%
13	Keep others at a distance	33%
14	Avoid contact with others	33%
15	Cheer people up	32%
16	Have little to say	32%
17	Don't like to draw attention to myself	31%
18	Know how to captivate people	31%
19	Do not mind being the centre of attention	30%
20	Warm up quickly to others	30%

Table 7.9 Analysis of predictability of extraversion items. Precision.

7.2.4 Study 4: Predictability of agreeableness

Agreeableness is the tendency to be cooperative and loving with others rather than being suspicious and competitive. Facebook is not a collaborative work platform, although online collaborative work platforms do exist, and those services have pages on Facebook.

Sample

In this study I use 25,000 American Facebook users from the MyPersonality dataset who completed the personality questionnaire. The users satisfy the inclusion criteria outlined in section 2.3. The reduction in the number of users from 98,515 users has been made only to speed up machine learning experiments. The decision to keep American users was made to keep the users contextually relevant to each other, as the same number of users from worldwide would have made the data less contextually rich. The personality questionnaire for agreeableness includes 20 items where the users have the choice of answering them by selecting whether the statement in the item applies to them is *Very inaccurate*, *Somewhat accurate*, *Neither accurate nor inaccurate*, *Somewhat accurate*, and *Very accurate*.

A discussion of how well MyPersonality's agreeableness questionnaire actually captures the trait agreeableness is presented in chapter 2.

Methods

I use ECA, as described in chapter 4 to reduce the dimensions and eliminate noise from the raw data, and use the learning model outlined in figure 5.1 and chapter 5 to predict the response of the users to each individual item. Accuracy is measured by the percentage of times that the model is able to predict the right choice for the item, known as precision.

Results

Table 7.10 shows all 20 personality items with respect to Agreeableness, and how accurately they can be predicted.

The most predictable items are about a user's actions towards others, such as insulting people, getting back at them, having a sharp tongue and working for personal gain. We can theoretically assume that such behaviour will leave footprints on mediums of communication such as Facebook, as there are pages on Facebook dedicated to such activities. Intuitively understanding of why some items are least predictable is more difficult. In general, these items are about beliefs, such as believing if others have good intentions, or they are about the sort of behaviour that Facebook provides no instrument to capture, such as holding a grudge.

At the facet level, the item on the modesty facet and most items on the cooperation facets are easier to predict. However, the item about holding the grudge also belongs to the cooperation facet however, it is one of the least predictable items. This is probably because insulting people, getting back at them and having a sharp tongue leaves footprints on Facebook as they are actions, however holding a grudge is a mental process, which leaves very little footprints on social networks.

Rank	*Item*	*Prediction Accuracy*
1	Believe that I am better than others	42%
2	Insult people	39%
3	Get back at others	37%
4	Am out for my own personal gain	35%
5	Have a sharp tongue	34%
6	Trust what people say	32%
7	Make people feel at ease	32%
8	Have a good word for everyone	31 %
9	Contradict others	31 %
10	Respect others	30%
11	Sympathise with others feelings	30%
12	Am concerned about others	30%
13	Suspect hidden motives in others	29 %
14	Cut others to pieces	28%
15	Treat all people equally	28 %
16	Believe that others have good intentions	27%
17	Hold a grudge	27 %
18	Make demands on others	25 %
19	Accept people as they are	24%
20	Am easy to satisfy	22%

Table 7.10 Analysis of predictability of agreeableness items. Precision.

7.2.5 Study 5: Predictability of neuroticism

Neuroticism, or emotional stability in its reverse form, is the tendency to experience negative emotions such as depression, anger, anxiety and moodiness.

Sample

In this study I use 25,000 American Facebook users from the MyPersonality dataset who completed the personality questionnaire. The users satisfy the inclusion criteria outlined in section 2.3. The reduction in the number of users from 98,515 users has been made only to speed up machine learning experiments. The decision to keep American users was made to keep the users contextually relevant to each other, as the same number of users from worldwide would have made the data less contextually rich. The personality questionnaire for neuroticism includes 20 items where the users have the choice of answering them by selecting whether the statement in the item applies to them is *Very inaccurate*, *Somewhat accurate*, *Neither accurate nor inaccurate*, *Somewhat accurate*, and *Very accurate*.

A discussion of how well MyPersonality's neuroticism questionnaire actually captures the trait neuroticism is presented in chapter 2.

Methods

I use ECA, as described in chapter 4 to reduce the dimensions and eliminate noise from the raw data, and use the learning model outlined in figure 5.1 and chapter 5 to predict the response of the users to each individual item. Accuracy is measured by the percentage of times that the model is able to predict the right choice for the item, known as precision.

Results

Table 7.11 shows all 20 personality items with respect to Neuroticism, and how accurately they can be predicted.

The predictability of items measuring the depression facet varies, items about mood swings or feeling down are more predictable whereas items about feeling blue are the least predictable. It is conceivable that mood swings are more likely to leave footprints compared to feeling blue, explaining the difference in predictability. Items on other facets are more difficult to predict compared to depression, with the item about not being bothered by things to be completely unpredictable, as its predictability rate is at the level of random chance.

As shown earlier, neuroticism is the least predictable trait as a whole, and its items are also least predictable. The items mostly refer to how an individual feels, rather than what they do. Looking through the items, it is not difficult to understand why Facebook is not a great medium for collecting footprints that might indicate various degrees of neuroticism, as these items are very seldom represented in the form of Facebook Likes.

Rank	*Item*	*Prediction Accuracy*
1	Panic easily	40%
2	Dislike myself	39%
3	Have frequent mood swings	36%
4	Am often down in the dumps	36%
5	Am filled with doubts about things	35 %
6	Remain calm under pressure	34%
7	Feel comfortable with myself	33%
8	Worry about things	33 %
9	Am very pleased with myself	33 %
10	Rarely get irritated	32%
11	Feel threatened easily	32 %
12	Am not easily frustrated	32%
13	Fear for the worst	32 %
14	Am relaxed most of the time	32 %
15	Seldom get mad	31%
16	Get stressed out easily	30%
17	Often feel blue	29 %
18	Rarely lose my composure	28%
19	Seldom feel blue	28 %
19	Am not easily bothered by things	19%

Table 7.11 Analysis of predictability of neuroticism items. Precision.

The studies by Park et al. [109] and Golbeck et al. [50] were able to predict neuroticism less inaccurately, with respect to their overall level of accuracy of their own predictions, compared to this thesis, the study by Youyou et al. [152] and the study by Kosinski et al. [86]. Interestingly, the first two studies use linguistic features from Facebook status updates to predict personality scores whereas this thesis and the two latter studies use Facebook Likes. This might indicate that status updates might include more relevant information useful for prediction of neuroticism than other traits, as we have demonstrated that Facebook Likes have more relevant information for prediction of other traits compared to neuroticism. In fact there is evidence for this in literature. Mairesse et al. [97] tried to predict the five-factor models from non-social anonymous corpora of texts and found neuroticism to be the second-easiest trait to predict, after openness. It is however unclear whether adding data from status updates to the current models in this thesis would enhance the model's predictive power for prediction of neuroticism, as the highest reported correlation of predicted neuroticism with self-report neuroticism is achieved in this thesis.

7.2.6 Discussion

Various social networks exist for various types of communications and relationships. The types of behaviour common on a social network determines which personality traits or which facets of each trait are more predictable using digital footprints stored on the social network. Facebook is not only a social network focusing on interpersonal relationships, but it is also a major source for communication of news. A survey by the Pew Research Center [1] in 2017 showed that 67% of Americans get at least some of their news from social media and Facebook is the most widely used social media for news [40]. This behaviour leaves footprints about a person's values, and items representing the values facet of openness are the most predictable in the entire item pool.

Facebook is also the world's largest social network for personal relationships. Friends catch up with each other on Facebook or plan events together. This is the sort of behaviour that leaves footprints, and items representing the gregariousness facet of extraversion are also the most predictable items among the extraversion items but also among the entire item pool.

Overall, the hypothesis that actions-based items are more predictable than items inquiring about thoughts and beliefs seems to be a visible pattern from the observations. This makes intuitive sense as well, as actions are more likely to leave footprints than thoughts or beliefs and the basis of the predictions in this thesis is the footprints left on social networks.

7.3 Demographics and predictability

Prior sections in this chapter investigated whether prediction accuracy correlates with personality traits, and found out that neuroticism is negatively correlated with the accuracy of the predictions for all five personality traits. In this section, I investigate if demographics are correlated with predictability.

[1] http://www.pewresearch.org

Sample

In this study I use 25,000 American Facebook users from the MyPersonality dataset who completed the personality questionnaire. The users satisfy the inclusion criteria outlined in section 2.3. The reduction in the number of users from 98,515 users has been made only to speed up machine learning experiments. The decision to keep American users was made to keep the users contextually relevant to each other, as the same number of users from world-wide would have made the data less contextually rich.

Methods

Using methodology explained in chapter 5, the ML-rated personality of the users has been predicted and compared with their self-report personality. The error in personality predictions is root mean squared among all five personality traits. Demographics data are used in linear regression and goodness of fit is measured using R^2.

Results

- *Age* is positively correlated with predictability at $p < .005$ and $R^2 = .16$. Therefore, older users are more predictable than younger users on Facebook. This might be explained by the fact that social networks are a more important part of the lives of younger people, therefore there can be a stronger incentive to do focus on self-presentation.
- *Gender* is not correlated with predictability. Men and women are equally predictable. $p > .05$.
- *Being in a relationship* is slightly negatively correlated with predictability. Note that Facebook data is accumulated over time and could include data from the time when the relationship status of users could have been different. $p < .01$ and $R^2 = .09$.
- *Having a college degree* is negatively correlated with predictability of personality traits. $p < .01$ and $R^2 = .10$.
- *Location* is not correlated with predictability. This includes both current location and hometown. $p > .05$.

In short, older people are easier to predict while people in relationships and people with college degrees are more difficult to predict.

7.4 Summary

Personality is a great predictor of behaviour and life outcomes. By observing behaviour and life outcomes, it is possible to make judgements about personality. This thesis focuses on making personality judgements by observing footprints of a user's behaviour on Facebook. This chapter focused on the reasons behind the variance in predictive accuracy among the five-factor personality traits. Openness is the most predictable trait while neuroticism is the least predictable trait. In fact, neuroticism is the largest predictor of predictability itself, users who are highly neurotic are more unpredictable with it comes to the prediction of any of their five personality traits. This seems to be because of the tendency of highly neurotic individuals to present an ideal rather than realistic image of themselves on Facebook.

To investigate it further, I assessed the reliability of the ML-rated personality scores while excluding users who had higher than average neuroticism, and found out stronger reliability across the measurement of all five personality traits in terms of correlations with self-reports and internal consistency. This indicates that the more neurotic people are, the less useful their Facebook data is for prediction of their personality, and overall, for psychology research. This informs us of a limitation in Facebook's role as a platform for social psychology research.

Personality is not the only predictor of behaviour and life outcome. The methodology outlined in chapters 4 and 5 is very capable of predicting personality and demonstrates psychometric properties similar to a 100-item self-report personality test. The question arises whether the same methodology can be employed to design other passive psychometric tests, using Facebook data. I investigate this question in chapter 8.

Chapter 8

New Passive Psychometric Tests

In chapter 6, I developed a proof-of-concept passive personality test that infers the user's five-factor personality scores only from digital data. In this chapter, I extend the same principle to three other psychometric tests, in order to examine the extent to which other psychometric tests can be made passive only from social networking data and to investigate if the methods of making passive psychological assessment that I introduced in earlier chapters are indeed universal, or whether they only apply to personality traits.

Users spend a lot of times on online social networks, this includes times that a user feels well, as well as times that the user does not. Therefore it is a reasonable assumption that the social network captures information about a user's well-being. Satisfaction with Life scale [33] is a way to measure an individual's overall life satisfaction. It correlates well with personality and is often used as a life outcome variable. In fact, I used life satisfaction among the outcome variables to test the ML-rated personality scores against the self-report personality scores, as a way to measure the external validity of the machine learning system's personality predictions. Section 8.1 develops a passive psychometric test for the Satisfaction with Life scale, and I assess the reliability of the test by measuring the correlations of the predicted satisfaction with life score with self-reports and by measuring the split-half correlations to test its internal consistency and comparing it to reported literature. I examine the validity of the predicted satisfaction with life scores by correlating and comparing the predicted scores and self-reports with the big-5 personality traits, to measure its external validity.

Another way to examine well-being is by investigating the users for depression. Section 8.2 develops a passive psychometric test for the CES-D Scale [119]. Depression was also used as an outcome variable in investigating the external validity of the passive personality test, in chapter 6. The psychometric properties of the passive depression test are assessed in a similar way to the passive satisfaction with life test.

As a function of their size, online social networks allow the users to virtually perform any of a limitless choice of actions. Therefore even in the online world, an individual exhibits a degree of self-control towards their actions. The Self-monitoring Scale measures self-observation and control in social situations [130]. It is interesting to investigate if the same trait can be predicted using an individual's digital footprints on Facebook. Section 8.3 focuses on the development of a passive self-monitoring test. The psychometric properties of the passive test for the self-monitoring scale are assessed in a similar way to the prior tests for satisfaction with life and depression.

8.1 Satisfaction with Life

Diener et al. [33] developed the Satisfaction with Life Scale, SWL, in 1985. It measures global life satisfaction. It's a test that demonstrates high internal consistency, and correlates well with other subjective measurements of well-being. The satisfaction with life scale provides five items, outlined in table 8.1.

	Satisfaction with Life items
1	In most ways, my life is close to ideal
2	The conditions of my life are excellent
3	I am satisfied with my life
4	So far I have gotten the important things I want in life
5	If I could live my life over, I would change almost nothing.

Table 8.1 Items in the Satisfaction with Life Scale.

Responders have the option to respond to the item in 7 scales: *Strongly disagree*, *disagree*, *somewhat disagree*, *neither agree nor disagree*, *somewhat agree*, *agree*, and *strongly agree*.

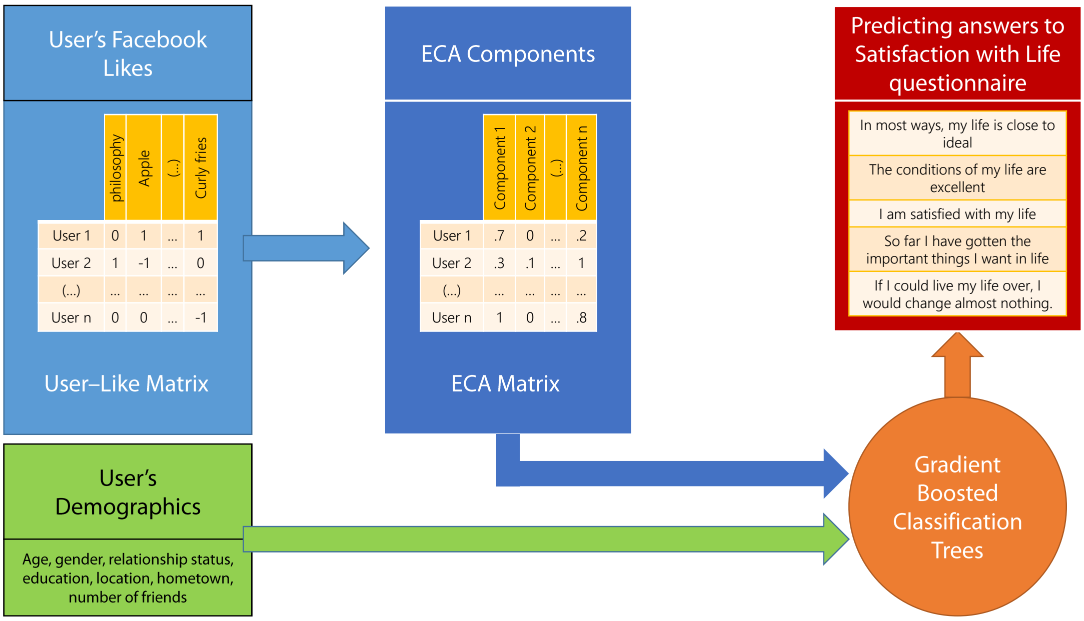
Kosinski et al. [86] is the only study in literature that predicted satisfaction with life using Facebook data, and showed a correlation of $r = .17$ between the predicted and self-report satisfaction with life scores. Their study uses MyPersonality dataset as well, with a narrower inclusion criteria as only users in the United States were used. While useful, this level of agreement is not good enough for a psychometric test.

8.1.1 Sample

MyPersonality provided the users with the option to opt into taking the five-item Satisfaction with Life test. Data from 43,419 worldwide users were used. These users had the same inclusion criteria as explained in section 2.3.

8.1.2 Method

Figure 8.1 illustrates the overall structure of the predictive model used for making the satisfaction with life assessment. The model uses the same configuration as the model developed for predicting five-factor personality. In fact, the only change is that it is now predicting items on a satisfaction with life scale, instead of items on a personality test.


User's Facebook Likes
User–Like Matrix
ECA Components
ECA Matrix
Predicting answers to Satisfaction with Life questionnaire
User's Demographics
Age, gender, relationship status, education, location, hometown, number of friends
Gradient Boosted Classification Trees

	philosophy	Apple	(...)	Curly fries
User 1	0	1	...	1
User 2	1	-1	...	0
(...)	...	...	...	...
User n	0	0	...	-1

	Component 1	Component 2	(...)	Component n
User 1	.7	0	...	.2
User 2	.3	.1	...	1
(...)	...	...	...	...
User n	1	0	...	.8

In most ways, my life is close to ideal

The conditions of my life are excellent

I am satisfied with my life

So far I have gotten the important things I want in life

If I could live my life over, I would change almost nothing.

Fig. 8.1 Overall structure of the prediction model for Satisfaction with Life scale

The scoring mechanism follows a similar structure as the scoring mechanism for personality assessments. The scoring structure is illustrated in figure 8.2.

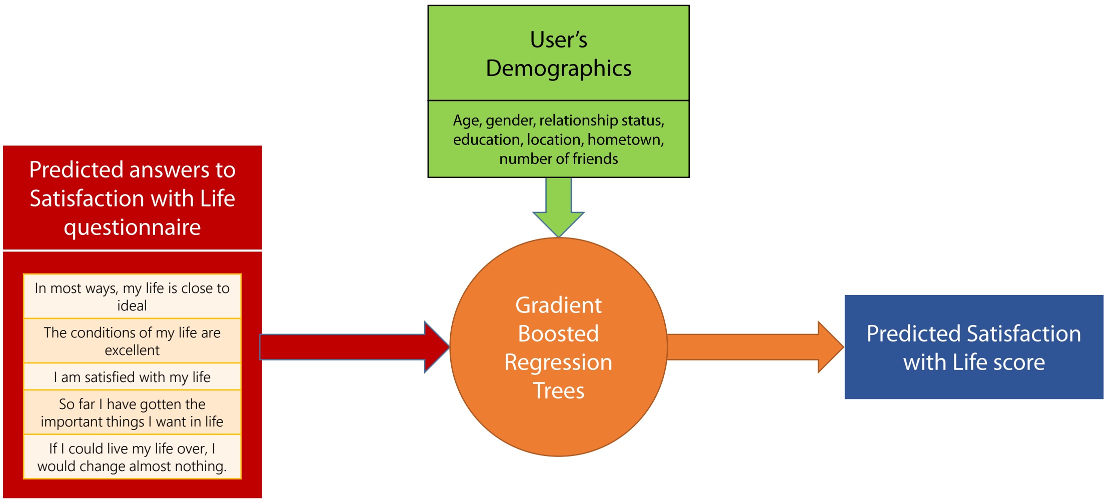

User's Demographics
Age, gender, relationship status, education, location, hometown, number of friends
Predicted answers to Satisfaction with Life questionnaire
In most ways, my life is close to ideal
The conditions of my life are excellent
I am satisfied with my life
So far I have gotten the important things I want in life
If I could live my life over, I would change almost nothing.
Gradient Boosted Regression Trees
Predicted Satisfaction with Life score

Fig. 8.2 Scoring structure of the prediction model for Satisfaction with Life scale

Similar to the personality test, the demographics are used only to train decision trees and guide the users to the right regression models, and are not used to compute the final satisfaction with life scores.

8.1.3 Results

The psychometric properties of the test are assessed by measuring Self-ML agreements, internal consistency and external validity. I applied a 10-fold cross-validation to avoid overfitting bias in the results. The users are randomly divided into 10 subsets, 9 subsets are used for training and one subset for evaluation. This is repeated 10 times to make satisfaction with life predictions of the entire dataset.

Self-ML agreement

Table 8.2 shows the level of agreement between the predicted satisfaction with life scores and self-reports.

	Self-ML correlation
Satisfaction with Life	.71

Table 8.2 Reliability of the Satisfaction with Life predictions.

The temporal reliability of the satisfaction with life scale has been studied, and short term test-retest correlations of satisfaction with life scale range from .6 to .8 have been observed in literature [111]. With a self-ML agreement, at $r = .71$, the predictions demonstrate psychometric properties, within the range reported by the literature.

Internal consistency

Similar to five-factor personality traits, I use split-half correlations to measure the internal consistency of the predicted scores. For each individual user, the Likes are split into two random subsets, the subsets are used to predict Satisfaction with Life scores. The scores predicted from each subset of the user's likes are correlated with each other. The literature reports Cronbach's alpha range of .79 to .89 for the Satisfaction with Life scale [111].

	Split-half correlations
Satisfaction with Life	.77

Table 8.3 Internal consistency of the Satisfaction with Life predictions.

With split-half correlations of $r = .77$, the satisfaction with life score predictions made from Facebook data have acceptable internal consistency.

External validity

In chapter 6, I used self-report satisfaction with life as an outcome variable, to assess the external validity of the ML-rated personality traits in comparison with self-report personality trait. The reversal of that process should also work, by using self-report personality traits as outcome variables to assess the external validity of ML-rated satisfaction with life in comparison to self-report satisfaction with life.

Table 8.4 shows the results of the external validity analysis of the ML-rated satisfaction with life scale and its comparison to the self-report satisfaction with life scale.

	Satisfaction with Life	
Traits	*ML-rated*	*Self-report*
Openness	.04	.05
Conscientiousness	.34	.30
Extraversion	.19	.25
Agreeableness	.20	.24
Neuroticism	-.44	-.47

Table 8.4 Correlation between self-report personality and ML-rated and self-report SWL, external validity. All are statistically significant at $p < .01$.

At at overall agreement of $r = .99$, the external validity of the ML-rated satisfaction with life is the same as the self-report satisfaction with life.

Discussion

The ML-rated satisfaction with life scores are as reliable and as internally consistent as self-reports and provide the same degree of external validity in terms of correlation with external outcomes.

Tests with lower temporal reliability usually measure variables that are prone to change in a person's life. It is important to note that the Facebook data used here are not time stamped, and they are a collection of data that have been gathered over a long period of time. Facebook was founded in 2004 and MyPersonality collected data until 2012 and as a result, data of a single user can be collected at any point in a timespan of over 8 years. Therefore, similar to the prediction of relationship status in section 4.3, this test measures satisfaction with life during the entirety of a user's active history on the social network, instead of the immediate moment of data collection. This limitation can be overcome by using time-stamped data from social networks. Facebook's graph API has only recently allowed for this feature.

Our ability to provide passive assessment of satisfaction with life using only Facebook data demonstrates that Facebook, as a platform for social interactions, indeed is able to capture information relevant to the user's subjective well-being. In recent years, Facebook has allowed users to express their feelings by adding a section dedicated to feelings and activities to a user's timeline. The users can use this feature to express a feeling of excitement, sadness, depression, happiness, love and anxiety. However, this is a newer feature and the data used for this study did not include them. The fact that accurate assessment of satisfaction with life is possible only from Facebook Likes shows that the way users choose the pages they like gives us insight into their well-being.

8.2 Depression

The CES-D Scale is a widely-used self-report scale for depression [119]. The test consists of 20 items, as outlined in table 8.5.

Users are given the following options for each item: *Rarely or none of the time (less than 1 day per week)*, *some or a little of the time (1-2 days per week)*, *occasionally or a moderate amount of the time (3-4 days per week)*, and *most or all of the time (5-7 days per week)*.

There is literature to connect the use of online social networks to depression. Pantic et al. [108] reported that the time spent on online social networks is related to the risk of depression in young adults. However, there are also studies that do not confirm these results. Jelenchick et al. [77] found no supporting evidence to form a relationship between the use of online social networks and depression. Choudhury et al. [30] found that the language used on social networks can be used as a predictor for depression. No prior study has used Facebook Likes to assess depression. This is the goal of this study.

8.2.1 Sample

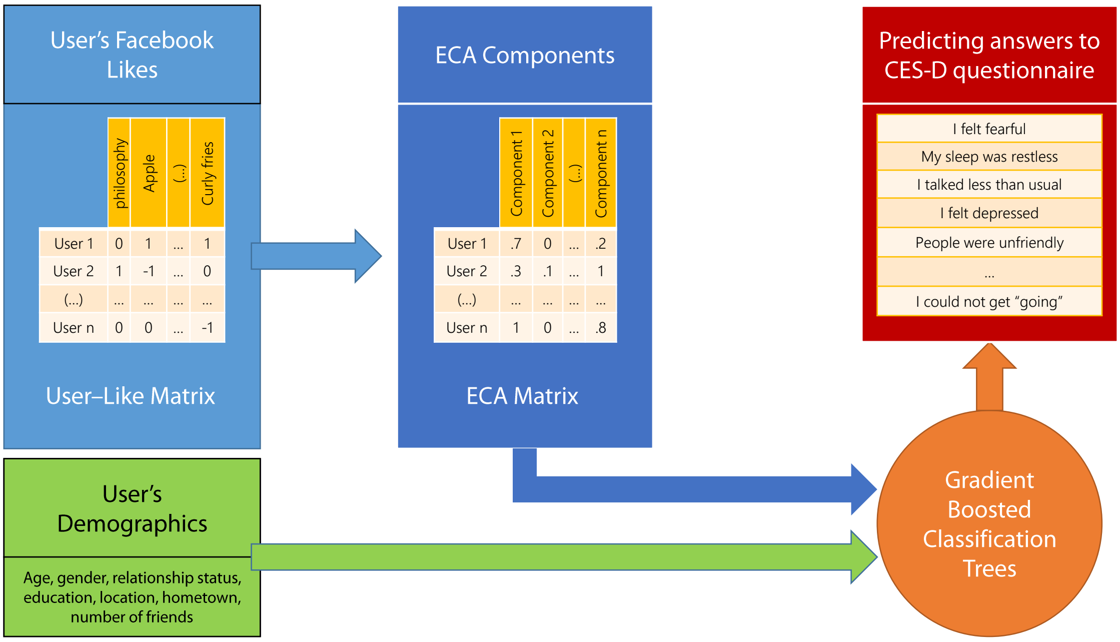
MyPersonality provided the users with the option to opt into taking the 20-item CES-D test. Data from 3,290 users were used. These users had the same inclusion criteria as explained in section 2.3.

	CES-D scale items
1	I was bothered by things that usually don’t bother me
2	I did not feel like eating; my appetite was poor
3	I felt that I could not shake off the blues even with help from my family and friends
4	I felt that I was just as good as other people
5	I had trouble keeping my mind on what I was doing
6	I felt depressed
7	I felt that everything I did was an effort
8	I felt hopeful about the future
9	I thought my life had been a failure
10	I felt fearful
11	My sleep was restless
12	I was happy
13	I talked less than usual
14	I felt lonely
15	People were unfriendly
16	I enjoyed life
17	I had crying spells
18	I felt sad
19	I felt that people disliked me
20	I could not get “going”

Table 8.5 Items in the CES-D scale for depression.

8.2.2 Method

Figure 8.3 illustrates the overall structure of the predictive model used for making the assessment. The model uses the same configuration as the model developed for predicting five-factor personality.



	philosophy	Apple	(...)	Curly fries
User 1	0	1	...	1
User 2	1	-1	...	0
(...)	...	...	...	...
User n	0	0	...	-1

	Component 1	Component 2	(...)	Component n
User 1	.7	0	...	.2
User 2	.3	.1	...	1
(...)	...	...	...	...
User n	1	0	...	.8

Fig. 8.3 Overall structure of the prediction model for CES-D scale

The scoring mechanism follows a similar structure as the scoring mechanism for personality assessments. The scoring structure is illustrated in figure 8.4.

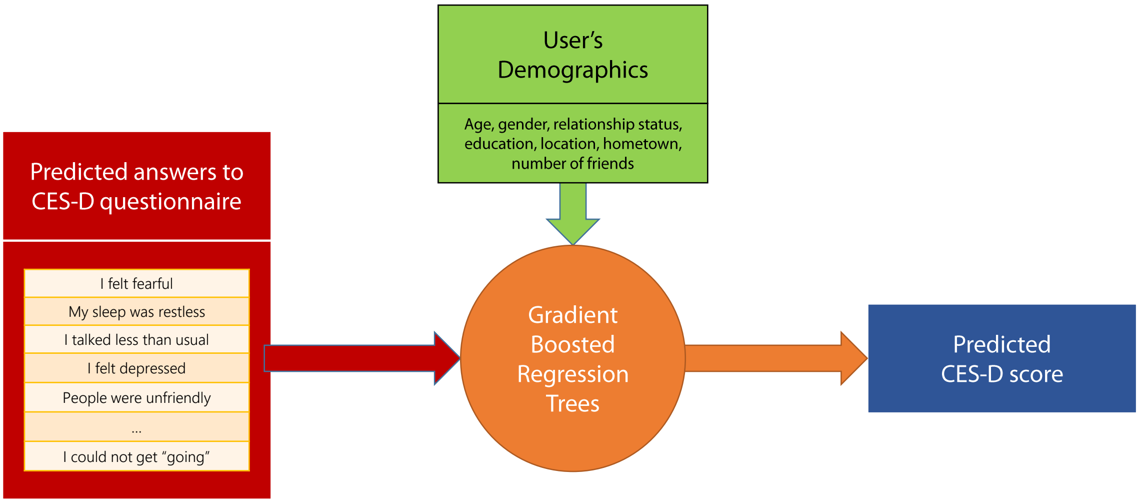


Fig. 8.4 Scoring structure of the prediction model for CES-D scale

The demographics are only used to train decision trees and guide the users to the right regression models, and are not used to compute the final CES-D score.

8.2.3 Results

The psychometric properties of the test are assessed by measuring Self-ML agreements, internal consistency and external validity. I apply a 10-fold cross-validation to avoid overfitting bias in the results. The users are randomly divided into 10 subsets, 9 subsets are used for training and one subset for evaluation. This is repeated 10 times to make depression predictions of the whole dataset.

Self-ML agreement

	Self-ML correlation
CES-D	.63

Table 8.6 Reliability of CES-D predictions.

The temporal reliability of the CES-D scale has been adequate, as the test-retest correlations at short intervals have been reported as $r = .69$ at four weeks, and $r = .59$ at eight weeks. So the reliability is within the range reported in literature. This demonstrates that the CES-D score predictions made from Facebook data are almost as accurate as retaking the test within a 4-8 week time frame. Note that these predictions are only trained on 3,290 users and while predictions accuracy is adequate, having access to more users would improve the predictions.

Internal consistency

Split-half correlations are used to measure internal consistency, in a similar way to previous tests. For each individual user, the Likes are split into two random subsets, the subsets are used to predict CES-D scores. The predicted scores are then correlated with each other. The literature reports the Cronbach's alpha and split-half correlations of about .85 in general populations [119].

	Split-half correlations
CES-D	.82

Table 8.7 Internal consistency of CES-D predictions.

With split-half correlations of $r = .82$, CES-D score predictions made from Facebook data have excellent internal consistency.

External validity

In chapter 6, I used self-report depression scores as an outcome variable, to assess the external validity of the ML-rated personality traits in comparison with self-report personality trait. The reversal of that process should also work, by using self-report personality traits as outcome variables to assess the external validity of ML-rated depression in comparison to self-report depression.

Table 8.8 shows the results of the external validity analysis of the ML-rated depression and its comparison to the self-report depression.

	CES-D	
Traits	*ML-rated*	*Self-report*
Openness	.03	.05
Conscientiousness	-.14	-.20
Extraversion	-.15	-.13
Agreeableness	-.05	-.11
Neuroticism	.42	.38

Table 8.8 Correlation between self-report personality and ML-rated and self-report CES-D depression scale, external validity. All are statistically significant at $p < .01$.

ML-rated depression was not as good as self-report depression in predicting agreeableness and conscientiousness while it was better at predicting neuroticism compared to self-report depression. At at overall agreement of $r = .98$, the external validity of the ML-rated depression is the same as the self-report depression.

Discussion

The ML-rated depression scores are as reliable, as internally consistent and provide the same degree of power in prediction of external outcomes as self-report depression scores.

The depression test was trained on a smaller dataset compared to the satisfaction with life test (3,290 for depression compared to 43,419 for satisfaction with life). It may explain the slightly weaker psychometric properties, however the results were still within the range reported in literature.

Similar to the satisfaction with life test, our ability to provide passive assessment of depression using only Facebook data demonstrates that Facebook, as a platform for social interactions, indeed is able to capture information relevant to the user's subjective well-being.

8.3 Self-monitoring

Snyder's Self-monitoring scale is a measurement of self-observation and control in social situations [130]. It consists of 25 items as outlined in table 8.9. The user rates the questions as *true* or *false*.

Self-monitoring has been shown to connect to fitness, behaviour in romantic relationships, and it correlates with the five-factor personality traits, as shown in chapter 6, as it was used as an outcome variable in investigating the external validity of the passive personality test. In this section I investigate if data left in the form of Facebook Likes can help with assessment of an individual's self-monitoring.

8.3.1 Sample

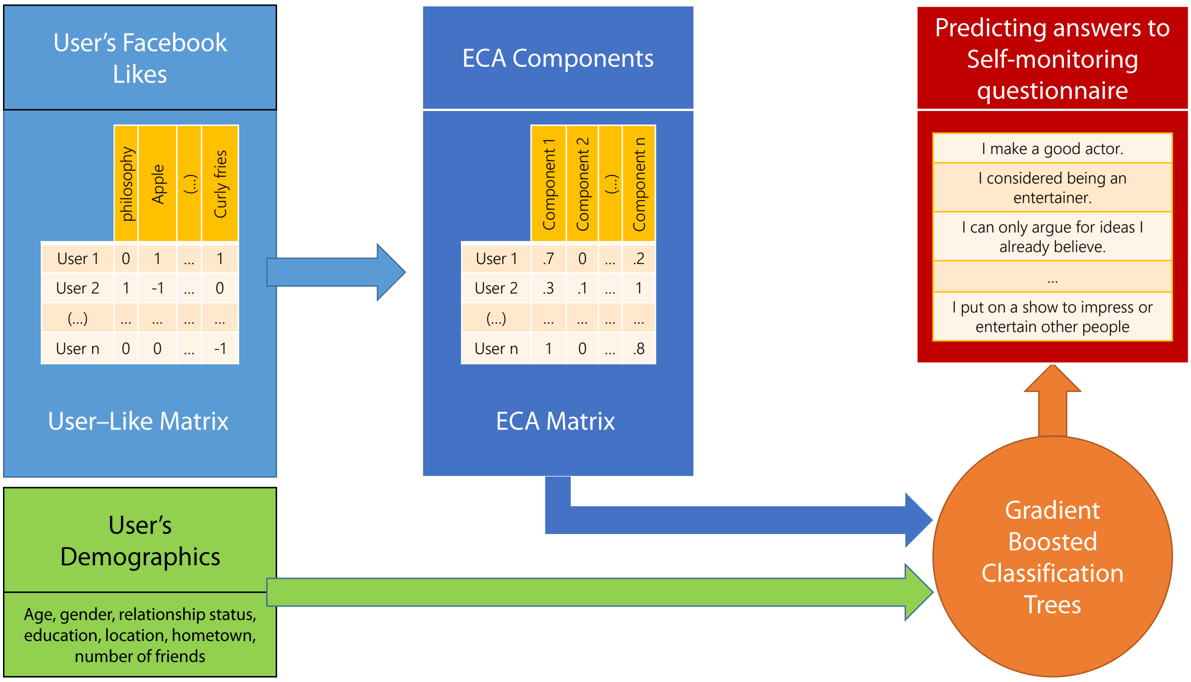
MyPersonality provided the users with the option to opt into taking the 20-item CES-D test. Data from 18,731 users were used. These users had the same inclusion criteria as explained in section 2.3.

	Snyder's Self-monitoring items
1	I find it hard to imitate the behaviour of other people.
2	My behaviour is usually an expression of my true inner feelings, attitudes, and beliefs.
3	At parties and social gatherings, I do not attempt to do or say things that others will like.
4	I can only argue for ideas which I already believe.
5	I can make impromptu speeches even on topics about which I have almost no information.
6	I guess I put on a show to impress or entertain people.
7	When I am uncertain how to act in a social situation, I look to the behaviour of others for cues.
8	I would probably make a good actor.
9	I rarely seek the advice of my friends to choose movies, books, or music.
10	I sometimes appear to others to be experiencing deeper emotions than I actually am.
11	I laugh more when I watch a comedy with others than when alone.
12	In groups of people, I am rarely the centre of attention.
13	In different situations and with different people, I often act like very different persons.
14	I am not particularly good at making other people like me.
15	Even if I am not enjoying myself, I often pretend to be having a good time.
16	I'm not always the person I appear to be.
17	I would not change my opinions (or the way I do things) in order to please someone else or win their favour.
18	I have considered being an entertainer.
19	In order to get along and be liked, I tend to be what people expect me to be rather than anything else.
20	I have never been good at games like charades or improvisational acting.
21	I have trouble changing my behaviour to suit different people and different situations.
22	At a party, I let others keep the jokes and stories going.
23	I feel a bit awkward in company and do not show up quite as well as I should.
24	I can look anyone in the eye and tell a lie with a straight face (if for a right end).
25	I may deceive people by being friendly when I really dislike them.

Table 8.9 Items in the Snyder's Self-monitoring scale.

8.3.2 Method

Figure 8.5 illustrates the overall structure of the predictive model used for making the assessment. The model uses the same configuration as the model developed for predicting five-factor personality.


User's Facebook Likes
User–Like Matrix
ECA Components
ECA Matrix
Predicting answers to Self-monitoring questionnaire
User's Demographics
Age, gender, relationship status, education, location, hometown, number of friends
Gradient Boosted Classification Trees

	philosophy	Apple	(...)	Curly fries
User 1	0	1	...	1
User 2	1	-1	...	0
(...)	...	...	...	...
User n	0	0	...	-1

	Component 1	Component 2	(...)	Component n
User 1	.7	0	...	.2
User 2	.3	.1	...	1
(...)	...	...	...	...
User n	1	0	...	.8

I make a good actor.
I considered being an entertainer.
I can only argue for ideas I already believe.
...
I put on a show to impress or entertain other people

Fig. 8.5 Overall structure of the prediction model for Snyder's Self-monitoring scale

The scoring mechanism follows a similar structure as the scoring mechanism for personality assessments. The scoring structure is illustrated in figure 8.6.

The demographics are only used to train decision trees and guide the users to the right regression models, and are not used to compute the final Self-monitoring score.

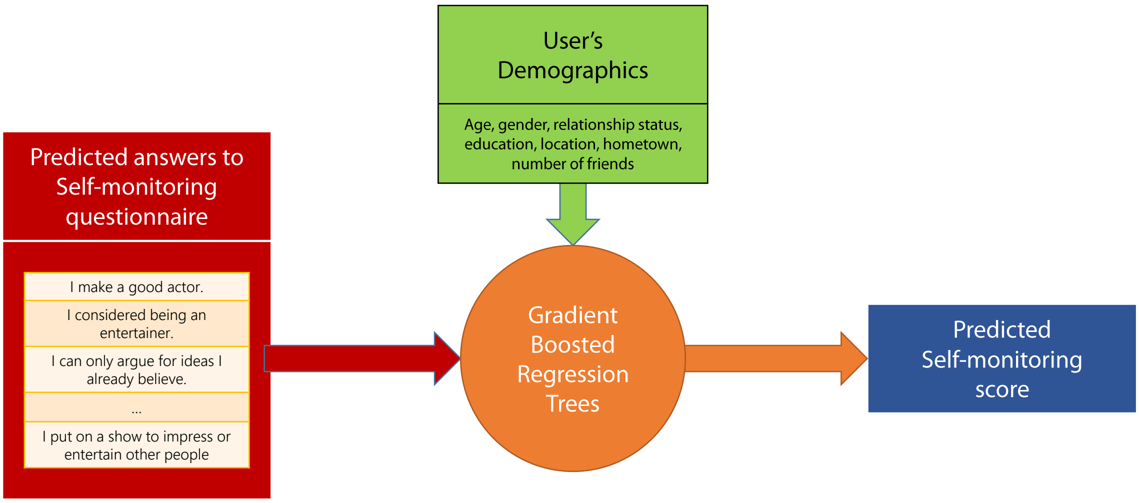

User's Demographics
Age, gender, relationship status, education, location, hometown, number of friends
Predicted answers to Self-monitoring questionnaire
I make a good actor.
I considered being an entertainer.
I can only argue for ideas I already believe.
...
I put on a show to impress or entertain other people
Gradient Boosted Regression Trees
Predicted Self-monitoring score

Fig. 8.6 Scoring structure of the prediction model for Snyder's Self-monitoring scale

8.3.3 Results

The psychometric properties of the test are assessed by measuring Self-ML agreements, internal consistency and external validity. I apply a 10-fold cross-validation to avoid overfitting bias in the results. The users are randomly divided into 10 subsets, 9 subsets are used for training and one subset for evaluation. This is repeated 10 times to make self-monitoring predictions of the whole dataset.

Self-ML agreement

	Self-ML correlation
CES-D	.67

Table 8.10 Reliability of Snyder's self-monitoring predictions.

Snyder reported excellent test-retest reliability of $r = .83$ [130]. The self-ML agreement of $r = .67$ is an adequate correlation with self-reports.

Internal consistency

Split-half correlations are used to measure internal consistency, in a similar way to previous tests. For each individual user, the Likes are split into two random subsets, the subsets are used to predict Self-monitoring scores. The predictions are then correlated with each other.

	Split-half correlations
Self-monitoring	.70

Table 8.11 Internal consistency of Self-monitoring predictions.

With split-half correlations of $r = .70$, Self-monitoring score predictions made from Facebook data have acceptable internal consistency.

External validity

In chapter 6, I used self-report self-monitoring scores as an outcome variable, to assess the external validity of the ML-rated personality traits in comparison with self-report personality trait. The reversal of that process should also work, by using self-report personality traits as outcome variables to assess the external validity of ML-rated self-monitoring in comparison to self-report self-monitoring.

Table 8.12 shows the results of the external validity analysis of the ML-rated self-monitoring and its comparison to the self-report self-monitoring.

	Self-monitoring	
Traits	*ML-rated*	*Self-report*
Openness	.10	.13
Conscientiousness	-.10	-.09
Extraversion	.29	.32
Agreeableness	-.10	-.07
Neuroticism	*N*	*N*

Table 8.12 Correlation between self-report personality and ML-rated and self-report Self-monitoring scale, external validity. All are statistically significant at $p < .01$. N means no statistical significance.

At at overall agreement of $r = .99$, the external validity of the ML-rated depression is the same as the self-report depression.

Discussion

The ML-rated depression scores are as reliable, as internally consistent and provide the same degree of power in prediction of external outcomes as self-report depression scores.

Our ability to provide passive assessment of self-monitoring using only Facebook data demonstrates that Facebook, as a platform for social interactions, indeed is able to capture information relevant to the user's ability to self-control.

8.4 Discussion

The same model structure and methodology used to make personality assessment from Facebook data was successfully used to make assessments of satisfaction with life, depression and self-monitoring scales. This demonstrates that psychological assessment without administering questionnaires is not only limited to five-factor personality, and can be very accurate for a wide variety of psychometric tests. The fact that the same models, without any further optimisation or adjustments, were able to design new passive psychometric tests for new scales is evidence for the universality of the methods developed in chapters 4 and 5 and in their efficacy to be used as basis of passive psychometric tests in the future as well. As a caveat, the methods are designed to maximise the predictive power when it comes to online social networking data. Developing passive psychometrics from other sources will likely require optimisations and adjustments.

Findings of this chapter are further evidence that Facebook, as a social interactions platform, is capable of being a platform for social psychology research as well. The footprints left on Facebook are not only useful for prediction of demographics and personality, but also well-being and self-observation and control in social situations. Critically though, whether data from a specific social network is suitable for accurate psychological assessment of certain traits remains relatively unknowable in advance, as intuitive understanding of what lies in the data is not often possible given the variety of ways the users use social networks and complexity of the prediction models. Results in this chapter however demonstrates that traits that have a high predictive power in a person's behaviour in social settings, such as personality or self-monitoring, usually leave their clues in the digital footprints left by a user on social networks, and machine learning methods are able to pick up the clues and construct an accurate psychological profile for the user. This introduces concerns for privacy of the users, as this fact is often not well-known. The general discussion about the findings of this thesis, as well as the discussion about its implication on privacy of users on social networks are covered in chapter 9.

Chapter 9

General Discussion

The aim of this work was to demonstrate that reliable, accurate and valid psychological assessments are possible without any direct input from the participant. This has only been made possible with the growth in popularity of online social networking websites such as Facebook, Twitter, Instagram and Qzone in the past decade where users from all over the world spend a considerable amount of time every single day. Digital footprints are records of an individual's behaviour, and by looking at the records of behaviour we can make accurate judgements about their private psychological traits.

There has not been prior work in literature that demonstrates psychometric properties for predicted personality scores from online social networks in terms of reliability, internal consistency and external validity. Youyou et al. [152] published the most accurate predictions in literature, with correlations between self-report and predicted personality of a mean $r = .56$ across the five-factor personality model. This falls short of psychometric standards. They measured internal consistency using split-half correlations of $r = .62$, which is a reasonable level of internal consistency however, it still falls short of the internal consistency of self-report personality tests, at about mean $r = .89$. Finally, their predicted personality had less predictive power to predict external variables compared to self-reports.

Overcoming these limitations required methodological advancements. Traditional dimensionality reduction methods, widely used for compressing the very large dimensions of data from social networks, have a trade-off between removing noise and preserving useful variance. This is due to the unsupervised nature of dimensionality reduction methods that preserve variance, such as Principal Component Analysis [73], Latent Semantic Indexing with Singular Value Decomposition [32], Latent Drichlete Allocation [16, 61] and Non-negative

Matrix factorization [93], and non-prioritisation of preserving variance in supervised feature selection methods, such as filter [95] and wrapper [82] based feature selection algorithms. Attempts to overcome this trade-off [153, 13, 8, 7, 132] have been mostly focused on adding a later supervised feature selection stage, after reducing dimensions using an unsupervised algorithm. This does little to remove noise from the data, as the initial unsupervised stage convolutes the signal and noise, making it embedded deep within the components in the new variable space. This is why such methods have not become status-quo in online social network research.

I overcome this challenge by first selecting features that are not deemed to be noise without any focus on redundancies, and instead of picking one parameter to represent a series of seemingly redundant variables, I then use an unsupervised dimensionality reduction algorithm to preserve all variance in the series of redundant variables. This ensures that relevant data are not missed during the dismissal of seemingly redundant variables. Since all attributes containing noise have already been removed, the unsupervised learning stage does not embed noise among the reduced dimensioned data, unlike prior attempts explained earlier. The inner workings and evaluation of this method are covered in chapter 4.

Once the data's dimensions are reduced, we need to train predictive models. The use of linear regression models is almost unanimous in literature [86, 117, 152, 109, 146, 15, 131, 30]. They are picked often because they are simple to understand and implement, they perform reasonably well and they do not require a lot of tuning and optimisations. Their use is however, disadvantageous for several reasons: (a) personality is latent patterns of behaviour, rather than patterns of footprints of behaviour, which a direct regression model (linear or higher polynomial degrees) incorrectly assumes; (b) linear regression models do not allow the use dichotomous variables such as gender, location and cultural differences in a discriminatory way; and (c) they do not take to account that people may leave the same footprint for very different behavioural reasons which can be related to latent psychological traits. To overcome these challenges, I used decision trees which are able to use markers in our datasets to discriminate between the users, and develop separate predictive models for groups of users based on the rules learned from the data. I also made the data cleaner by separating contextually relevant Facebook pages that the user has not liked from completely irrelevant pages. Finally, instead of predicting personality traits directly from Facebook data, I predict the user's potential responses to each item on a personality questionnaire, this allows the predictive models to look for much finer patterns in the data that can help predictions in the end. Chapter 5 details the description of the predictive models, and section 5.3 provides benchmarks to determine the contribution that each enhancement to the models have made to the overall predictions.

Being able to provide accurate predictions (in terms of correlations with the observation) is not the only factor in determining whether a psychometric test is valid. The results of a new test not only need to correlate well with the results of existing tests, but they also need to be internally consistent. Meaning that a personality judgement made using a portion of user's data on Facebook need to correlate well with judgements made with a different portion of the data. This is to ensure that there is no single point in data that is mostly responsible for the predictions. Personality predictions produced in chapter 6 were shown to have a good level of internal consistency, at $r = .70$. In terms of external validity, predicted personality scores had the same level of predictive power as self-reports in predicting behavioural and life outcomes such as values, sensational interests, life satisfaction and depression. This is very important as in practice, outputs of a personality test need to be able to predict individual outcomes, otherwise the test would be of little value. In terms of correlations with self-reports, internal consistency and external validity, the predicted personality in this thesis outperform the best published results in literature, as described in section 5.3.4. Finally, to determine the level of independence of traits from each other, a discriminant validity analysis was performed and the predicted personality traits were found to be as independent from each other, as self-report personality traits are from each other.

9.1 Personality

To recall from chapters 6 and 7, we found openness to be the most predictable trait from Facebook data, and neuroticism to be the least predictable trait. This might seem at odds with general expectations, as conscientiousness and neuroticism are considered to be easiest to predict from behaviour. This is an interesting observation because Facebook, as a social communication medium and a social psychology research platform, is shown to be not as good at capturing records of behaviour that might indicate various degrees of neuroticism as it is for openness. To further investigate this issue, I investigated how the prediction accuracy differs for users with different personality traits and found neuroticism to be the greatest predictor of predictability of users on social media. The users who score higher on the neuroticism scale are the most difficult users to predict, and users who score lower on the neuroticism scale are the easiest users to predict. This observation indicates that the footprints that highly neuritic people leave on social networks are not as easy to decipher as the footprints that people with lower levels of neuroticism leave. This might be due to the tendency of neurotic individuals to present an ideal self rather than an authentic and realistic one on social media [127]. Similar to self-enhancement bias in a traditional self-report questionnaire [89], also correlated with neuroticism [110], self-enhancement on social networks adds bias to the data because the user no longer likes pages genuinely, and instead is trying to create an ideal image of themselves to present to the world.

It may also be a possibility that we need more sophisticated machine learning models to be able to construct personality profiles for users with higher levels of neuroticism. Indeed gradient boosted decision trees were able to capture neuroticism a lot better than linear regression models (accounting for an extra 165% of variance). The deep learning model introduced in section 5.4 also found neuroticism to be the least predictable trait. This pattern is also reported in the two other studies in literature that predicted personality scores from Facebook Likes [86, 152]. Therefore, it seems to be universal pattern. It might indeed be possible that more sophisticated machine learning models, probably trained with larger datasets, will be able to decipher footprints left by highly neurotic users better and provide more accurate results compared to the ones reported in this thesis, however such models are also likely to be able to do the same for all other users and traits as well, and it is expected that the same trend will be observed.

This observation can give us new insight into the way neurotic people behave on social networks. Ross et al. [123] reports that neuroticism is positively correlated with using the Facebook wall, that is where users post their status updates. This is inline with other studies [109, 50], discussed earlier in chapter 7, that found neuroticism to be among the easier to predict traits from status updates. This is inline with literature that neuroticism is easier to predict from text-based sources [97].

I believe the reason neuroticism itself is the least predictable trait is due to the nature of Facebook and the use of Facebook Likes as the primary factor for predictions in this thesis. As investigated in section 7.2.5, questions on a self-report neuroticism scale do not connect well with the options users have on Facebook in the form of pages. For example, the user will not like a page on Facebook that can indicate whether a user has frequent mood swings, or whether they worry about things, or whether they feel threatened easily. This is in contrast with the most predictable trait, openness, where there are numerous pages about politics, arts or philosophy. This is also the case for extraversion, as information about a person's attitude towards social situations, parties or friendships is captured by a social network that primarily focuses on such interactions. Extraversion is also the second easiest trait to predict, after openness.

While neuroticism was the least predictable trait from Facebook data, it should still be noted that in terms of the predictive power of the predicted neuroticism compared to self-report neuroticism when it comes to external variables, there was an agreement of $r = .98$, therefore predicted values of neuroticism are still very useful for research and practice.

9.2 Methods

Linear regression models, used almost unanimously in literature [86, 117, 152, 109, 146, 15, 131, 30], were shown to be less than fully effective at deciphering patterns in Facebook data that more sophisticated decision trees were able to do. This observation means that generating separate classification and regression models for subgroups of users enhances the predictive power of the models compared to using a single model for everyone. Therefore, this is evidence that not only entirely different behaviours by the users can leave identical footprints on social networks, but that they also leave markers in other parts of their data that can be used to make the right judgement about what those footprints mean, whether the markers are part of on their demographics or parts of other attributes in their Facebook data. This is the beauty of finding these markers in a data-driven way instead of manually separating users into separate groups based on our suspicions and intuitions. While our suspicions might be correct and our intuitive grouping of users might enhance predictive power, we can argue that if our intuitions are correct then there is evidence for them in the data and a data-driven method can learn the same things, assuming our data is large enough to account for all or most of the variances that users might have.

This is the promise of deep learning. That expert domain knowledge is not required in construction of the models, as deep neural networks are able to learn them only from the data. This idea has been successfully applied to the field of computer vision [88, 90]. Now there is a trend that tends to dismiss the idea that domain knowledge is useful when applying machine learning to a new research area, claiming a deep neural network will be able learn the same wealth of knowledge from the data itself. While this claim might actually be true, the real question is how large should a sample size be so that hidden layered structures learned automatically from the data become more knowledgeable than the wealth of prior research and literature in the domain. This is a question that needs addressing in various fields. In my experiments with deep learning, discussed in section 5.4, I was able to outperform the previous best published predictions. Deep learning could indeed have been the method of choice for the central work in this thesis, however I also observed that a model incorporating domain knowledge, while still benefiting from rules learned from the data, was the best performing model in terms of its predictive power. So the conclusion is that with a sample size of nearly 100,000 users from Facebook, domain knowledge is still beneficial to the predictions compared to the completely data-driven approaches. This might change at 150,000 or 200,000 users, or at 1,000,000 users. That investigation requires a larger sample size.

The most noticeable shortcoming of the predicted personality compared to self-report personality was the internal consistency of neuroticism, at split-half correlations of $r = .54$. Further optimisations and adjustments to the methods brought it up to $r = .64$ which is a considerable gain. However more importantly this demonstrates that methodology makes a noticeable different to internal consistency. This raises the question of how can we optimally train our machine learning models to optimise their predictions for higher levels of internal consistency, given that the loss functions of all training algorithms centre around tests that measure fitness of predictions with the class variable. This might indicate that we need major methodology advancements in terms of machine learning training algorithms for passive psychological assessments before we are able to completely meet the scale reliability of high-quality self-report questionnaires. I believe the development of a loss function coupled with gradient boosting which is capable of optimising the training process to produce highly internally consistent predictions is the research area that needs most attention in terms of methodology for future progress.

9.3 Limitations

The greatest limitation of this approach is the need for high quality training data. It is impossible to move to a new social network and immediately begin to predict personality scores for their users. We need a large project, such as MyPersonality, to collect high quality training data from tens of thousands of users. This means a large investment in time and resources. It took years for MyPersonality to collect this amount of information. In contrast, we can give a traditional self-report questionnaire to individuals immediately and provide psychological assessment. It takes more of an individual user's time compared to passive assessment, but the project overall might take less time. In terms of the quality of the predictions, a passive psychometric test is only as good as the gold-standard test that was used for its training, as it is aiming to reproduce those outcomes. Therefore, if the original test suffers from a lack of reliability or suffers from biases, the passive test will suffer as well.

The lack of need for users to provide any active participation in their psychological assessment is a major benefit of this method because it saves time for both the test giver and the test taker, but it can also be a limitation. In section 7.2, we observed that items on a personality questionnaire that are about feelings are more difficult to predict because Facebook Likes are not a great way to capture these thoughts when a user is using the social network. A direct response from the user might be useful in these instances. Therefore, passive psychometrics will never replace self-reports, instead it is an extra tool that we can use to capture the traits easier and faster, while still having very good reliability and validity for a wide range of applications.

Another major limitation for this work is privacy concerns and ethics. This research demonstrated that footprints of users on social networks is a very rich source of information, a source that we can use to construct very accurate psychological and demographic profiles for the user. A lot of users are not comfortable to hand over these profiles of themselves to others, especially when they learn of their true predictive power. In contrast, a self-report questionnaire is very focused and only captures the data a user directly sees and consents to. The user can stop a self-report test at any time, ignore any question or ask for clarification. When a user is filling a satisfaction with life questionnaire, they know they are not sharing information about their demographics or personality traits, however in a passive psychometrics system the same data is used for all assessments. Therefore, a test labelled as a satisfaction with life test can indeed also examine personality, depression, demographics, even sexual orientation or history of substance use [86], all from the same Facebook data. Section 9.4 focuses on the privacy aspect of passive psychometrics and research on social networks in general.

9.4 Privacy

Social networks are created to facilitate transfer of information between individuals. As a side-effect of facilitating the medium of information transfer, the social network itself becomes a third party that sees all transferred information. It is often been viewed as the cost of the service, as social networks are almost always free to the end-user. The users are paying for the service with their data, and the data is used to facilitate targeted advertisement to the individual. This equilibrium often works well with satisfaction of all involved parties.

Growingly though, studies are demonstrating that data stored on online social networks, emails, even on mobile phones are rich sources of private information that the user might not be aware of, or can intuitively understand. This thesis demonstrated that accurate assessment of a person's private individual traits such as personality and self-monitoring abilities, and their mental health status such as satisfaction with life and depression is possible only from data that is often even *publicly* accessible on their Facebook profiles. I also accurately predicted an individual's age, gender and relationship status from only their Facebook Likes. Kosinski et al. [86] demonstrated that sexual orientation, relationship status of parents, drug and alcohol use habits, religion, smoking habits and intelligence can be predicted from only Facebook data.

People are often uncomfortable when they are asked private information, and they expect their personal and mental boundaries to be respected. Not all people would willingly disclose information about their sexual orientation, history with substances, and certainly not a lot of people willingly fill long psychometric questionnaires. Most users do not understand that their profiles on online social networks are equally rich sources of private information and often people are unaware of their privacy settings, and their data is constantly being shared with various other websites and applications. While this presents interesting research and product design opportunities, it also confirms the urgency of attempts to grapple with users' privacy on online social networks as it is a significant concern.

Public awareness about privacy and security has been increasing in recent years. It is an effective tool to not only shape policies of governments and corporations, but also encourage individuals to become more cognizant of the reality of private data in an online world. The case of Edward Snowden [60, 10] is a great example of the power of increase in awareness. The social and political awareness about government surveillance has been significantly increased and not only violations of privacy are much better reported, but also individuals are more aware of their rights, and corporations have made adjustments to their policies to allow users more control towards their privacy.

9.4.1 Privacy concerns in passive psychometrics

As discussed in section 9.3, in contrast to self-report questionnaires, the user can have very little control or knowledge about what is actually being tested in a passive psychometric test. In this thesis I used the same data, Facebook Likes, to predict age, gender, relationship status, voting preferences, five-factor personality, satisfaction with life, depression and self-monitoring. A nefarious website claiming to provide prediction of relationship status from Facebook Likes can indeed end up using the same data for prediction of an entire set of demographics and psychological traits to sell to potential advertisers and the user has no way of ever knowing about it.

Technologically, there are ways to address this problem. For example, pushing the analysis to the client-side (web browser or mobile phone) instead of server can help give users assurances that they are in control of what is being tested and what is being communicated with the servers. This is not currently common however, as client computer hardware are usually less powerful than computing servers, and analysis are usually resource intensive and energy consuming, and most of the clients use mobile devices (laptops, tablets or mobile phones) which are battery powered. However, this could be an interesting trade-off between privacy and utility.

There are non-nefarious incentives for service providers to want to continue having access to the data of their users. As long as the users continue sharing their data with the service provider, predictive models can continue to be updated to provide even more accurate results, and adapt to changes in the data, as new content is uploaded daily and historical content sometimes become inactive and irrelevant. A predictive model that is not adapted to changes in its social network will lose its predictive power as the trends and fashionable content often changes quickly on these websites.

Transparency and consent are the key factors in overcoming the privacy concerns of passive psychometrics. The users need to know how their data is being used, what is being kept by the server and how the server will continue using these records in the future. They still have no control, as they cannot stop in the middle of the test, but they should become aware of the risks so that they can make informed decisions.

9.4.2 Privacy-preserving online services

It is often assumed that online social networks having unrestricted access to user data is the only way. It might not necessarily be so. It was once assumed that financial services and banking systems having unrestricted access to an individual's finances, and controlling the ledger centrally is the only possible way, however advancements in cryptocurrencies such as BitCoin[1] [103] have demonstrated that there are no technological limitations against decentralised financial services and ledgers.

I believe a similar concept can be applied to data stored on social networks. Currently, the data for each user is stored unencrypted on the data centres used by online social networks. The social network uses a database to store which other users have permission to access the data, based on user friendships, followerships and privacy settings. The social network itself sees and controls all of the data. This control can be given back to the user without sacrificing functionality of a social network. This can be achieved using the concept of broadcast encryption [44]. It is the method of broadcasting an encrypted message in a way that it can be decrypted by any of a group of decryption keys. This concept is used in Advanced Access Control System [17], the standard which manages the control access to Blu-Ray disks and other digital content. Each Blu-Ray player has a unique private decryption key hardcoded into their microprocessor. Each Blu-Ray disk is encrypted in a way that it can be decrypted by any of the millions of decryption keys stored in current and future Blu-Ray players.

This principle, applied to social networks, will look like:

- Each user on the network has their own private and public encryption and decryption key pair. The private keys are stored on local devices.
- User Bob's data are encrypted with his own public key, and are stored on social network servers. They can only be decrypted by Bob's own private key.
- Bob befriends User Alice. Data of Bob are now multicast encrypted and stored on the server so that they can be decrypted with Bob or Alice's private keys.
- Bob befriends new users, the data continues to be multicast encrypted so they are decryptable by all who are supposed to do so.
- Bob unfriends Alice. Bob's data gets re-encrypted which no longer allows for decryption with Alice's private key.

[1] https://www.bitcoin.org/

This ensures that the data is not seen, at any stage, by the online social network itself.

This principle applied in its most restrictive form, however, removes the ability of the online social network to provide any targeted advertisement. This significantly hurts their ability to generate revenue. Privacy is not about maximising restrictions to access to data, it is about being in control of access. The user can set which parts of the data they are willing to share with the social network for the purpose of advertisement, or service personalisation, and those parts of the user data can be multi-cast encrypted to allow the social network to decrypt them as well.

9.4.3 Privacy-preserving machine learning

There has been a recent shift towards adding privacy preservation into data mining and machine learning [1].

It is generally expected that big social data are anonymised. It involves changing the data in a non-reversible way, so that the identity of the people involved in the study are not recognizable. The MyPersonality dataset is anonymised. Users are identified with user IDs that do not correspond to their IDs on Facebook. No identifiable information such as names, pictures or email addresses are used in this thesis.

Randomisation methods [1] add noise to the data to remove identifiability from certain attributes. The noise itself has a high enough variance that makes recovery of original attributes impossible, but their distribution can be recovered. Randomisation methods are simple and data-independent, and they can be performed at the time of collection to ensure privacy. The major drawback is that the optimal probability distribution of noise added to the data, while independent of the data itself, is dependant on the type of learning model used to analyse the data [156]. Various privacy-preserving learning models, such as a Naive Bayes classifier [141] and a decision tree algorithm [142] have been developed.

These methods are not perfect and adversarial attacks on anonymisation and randomisation methods are possible. Some hashing functions are considered broken, which include widely used functions such as MD5 [148] or SHA-1 [147]. Various randomisation noise distributions have been shown to be broken too, as fairly-accurate estimation of original values has been demonstrated to be possible [74].

Datasets involved with medical data have always been concerned with privacy of users. *The Datafly System* [133] in 1997 was an automated system designed to remove identifiable information from medical data at the time of access. Information such as name, social security numbers and address were removed, and other identifiable features such as age and gender were made ambiguous by adding noise. The same approach can be used to data from social networks, for the purposes of research.

Lack of universal and widely subscribed guidelines on anonymising and randomising data is a major concern with needs to be addressed with urgency.

9.5 Role of passive psychometrics

Earlier in this chapter, I discussed the limitations of passive psychometrics. The question remains what role do I see for the future of passive psychometrics. In general, passive psychometrics is about making psychological measurements easier, faster, cheaper to scale and less inconvenient for the test-taker without sacrificing reliability and validity. This means where big social data are available and platform allows it, whether in research and practice, psychometric tests such as those of personality, self-monitoring or well-being can be completely automated without the need for active participation of the test-taker or administrator.

In social psychology and personality research, passive psychometric allows for much faster data collection compared to administering self-reports. As an example, a researcher who intends to study the effect of big-5 personality traits on consumer behaviour, an area of interest in research [81], can collect purchase history data for a large number of users only by asking them to allow connection to their Amazon[2] and Facebook accounts. Then the researcher can extract consumer behaviour data from the Amazon data and use a similar methodology as this thesis to predict big-5 personality traits, and perform their analysis. In contrast, a traditional way to perform this research would have been to bring users in for interviews (or invite them to an online questionnaire), inquire about their purchase history and consumer behaviour and administer a personality test. The passive method can allow for not only faster research production, but also allows the behaviour to be examined on a much larger scale, without the financial and time investments necessary to do a large self-report study.

Outside of research, there are benefits for passive personality assessment as well. An interesting area of potential application is in online recommender systems. Currently, the most prevalent problem in recommender systems for movies, music, books or general products is the *cold start* problem (also known as *new user* problem). This is when a user joins a new

[2] http://www.amazon.com

network and the network is not aware of the any of the user's preferences. For example, when a user join a movie subscription service such as Netflix [3], Netflix is unable to provide any cold start personalised recommendations. Once the user starts using the service or starts to rate movies, recommendations become possible.

Several studies have demonstrated that five-factor personality information enhances the performance of recommender systems, especially with the cold start problem [137, 36, 43]. However, these have not been very practical solutions since accurate assessment of a new user's personality required administering of questionnaires, which are not optimal in terms of user experience on a website that they have just joined. However, with passive assessment, the user only has to log into the service with their social networking credentials and agree to share the data with the new website that they have joined, and their personality can be assessed instantly, and new relevant personalised recommendations can be provided to the user.

Passive personality predictions can also provide a similar advantage to online shopping. Bosnjak et al. [18] reported the role that five-factor personality plays in online shopping behaviour. By accurately predicting personality when a user starts using a new online store, the store can provide recommendations or customise the experience, which benefits both the user and the store.

9.6 Summary

The most interesting idea that stems from this work is that reliable and valid passive psychometric tests are possible, and digital footprints on online social networking websites can be the data for such predictions. This develops a new direction in psychometrics where new tests can be developed that are accurate enough to be used at individual level in both research and practice, but are not time consuming and slow for both the test taker and administrator.

Passive psychometric is not without its disadvantages. It requires high quality training data which is difficult and costly to acquire, its reliability and validity are limited to the reliability and validity of the gold-standard test that it is trained to reproduce, and there are added privacy and ethics concerns compared to self-report questionnaires.

[3] http://www.netflix.com

Methodological contributions in this work demonstrate that analysis of big social data can benefit from methods designed and tuned specifically for big social data compared to reusing broader methods used in multiple disciplines. Especially with the introduction of ECA in chapter 4, I demonstrated that preparation of data for the predictive models is very important and vanilla feature selection or dimensionality eduction algorithms either do not capture enough useful information, or embed too much noise in the data which either overwhelms the predictive models which reduces their predictive power or require much more sophisticated models to decipher the useful information from them that translates to additional need for computing resources and time. Moving on from linear regression models onto models that can group users separately based on the rules learned from the data is beneficial for prediction of traits in social media.

Chapter 10 summarises on contributions of this work to the fields of personality psychology, psychometrics and machine learning.

Chapter 10

Conclusions

The present thesis looked at user data stored on Facebook as the sole source of information for assessment of five-factor personality traits. It used machine learning to provide the predictions. Facebook was found to be an accurate source of information for assessment of five-factor personality. Predicted personality traits from Facebook were found to be reliable, internally consistent and equally as good as self-report personality in terms of the prediction of external variables. Openness and extraversion were the most predictable traits from Facebook, which is expected because social networks are mediums of communication that enable social interactions, access to news and recent events. In contrast, neuroticism was the least predictable trait, which may suggest that Facebook pages are less powerful at capturing behaviours indicator of neuroticism compared to other traits.

This work extends on the past work on the five-factor personality by investigating the relationship between personality and predictability on social networks. Individuals with higher scores on openness and extraversion were easier to predict. Previous research also suggested that openness and extraversion are correlated with an increased use of Facebook, which translates to leaving more footprints for the predictive models to use and have and lower levels of motivation for inauthentic self-presentation. Both these findings are beneficial for being easier to predict. Highly neurotic users were the most difficult users to predict, this supports prior findings that neurotic individuals have a higher interest in ideal rather than authentic self-presentations, therefore adding bias to their Facebook data.

To further investigate data stored on Facebook as a source of reliable and valid passive psychological assessment, passive tests for self-monitoring, depression and satisfaction with life were also developed. These scales were chosen because they often relate to well-being and behaviour in social situations, which should theoretically influence the way a user conducts themselves on social networks. These are also often outcome variables used in assessment of external validity of other tests. Predicted scores for satisfaction with life, depression and self-monitoring were shown to be as reliable, internally consistent and externally valid as self-report scores. This provides further evidence that Facebook data is psychologically rich and accurate passive psychological assessments across a wide range of scales is possible.

This thesis extends the past work on prediction of personality traits from Facebook by producing a methodology capable of more accurate predictions, by examining psychometric properties of personality predictions and by demonstrating methodology's versatility and universality by not only being able to create a passive personality test, but also self-monitoring, depression and satisfaction with life tests.

These findings suggest that focusing on social networks as a source of data for constructing psychological profiles is an interesting and fruitful area of research, but also presents challenges in terms of ethics, privacy and issues relating to consent. A major limitation of this method is the need for high-quality self-reports and its corresponding social network data to be used for training.

In terms of machine learning, this thesis expands the research into methods to reduce dimensions of highly sparse very large datasets by demonstrating that supervised learning in dimensionality reduction does not need to come at the expense of poor forms of redundancy removal and preserving of variance can indeed be a method of removing redundancies for highly sparse datasets, while avoiding the embedding of noise deep in the reduced dimensioned data. This translates to cleaner and more relevant data for the predictive models.

This thesis added to the growing evidence that a combination of weak learners are better at dealing with highly sparse datasets compared to single strong learners. It also demonstrated that in a choice between automated learning of internal hidden layers (deep learning), and the use of expert human knowledge and literature to construct the model structure, the expert models can still outperform deep learning at about 100,000 users. However, the trained deep learning model did outperform the best published result in literature. Therefore, it is conceivable that with larger sample sizes, deep learning will be able to outperform the models constructed in this thesis.

This work also demonstrated that internal consistency of passive personality assessment is not necessarily correlated with the level of agreement between predicted personality and self-reports and it is possible to sacrifice a small degree of correlations between predicted and self-report values in order to gain a larger degree of internal consistency. This demonstrates the need for new training algorithms for machine learning models which can tune their loss function to measure internal consistency of the model's predictions during training, rather than training solely to maximise goodness of fits measurements between their predictions and self-reports. This can be the next area for further improving the methodology of passive psychometrics .

References

[1] Agrawal, R. and Srikant, R. (2000). Privacy-preserving data mining. In *ACM Sigmod Record*, volume 29, pages 439–450. ACM.

[2] Aha, D. W. and Bankert, R. L. (1996). A comparative evaluation of sequential feature selection algorithms. In *Learning from data*, pages 199–206. Springer.

[3] Allport, F. H. and Allport, G. W. (1921). Personality traits: Their classification and measurement. *The Journal of Abnormal Psychology and Social Psychology*, 16(1):6.

[4] Allport, G. W. (1937). Personality: A psychological interpretation.

[5] American Association for the Advancement of Science (1999). Ethical and legal aspects of human subject research on the internet.

[6] Amichai-Hamburger, Y. and Vinitzky, G. (2010). Social network use and personality. *Computers in human behavior*, 26(6):1289–1295.

[7] Bai, B., Weston, J., Grangier, D., Collobert, R., Sadamasa, K., Qi, Y., Chapelle, O., and Weinberger, K. (2009). Supervised semantic indexing. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 187–196. ACM.

[8] Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2012). Prediction by supervised principal components. *Journal of the American Statistical Association*.

[9] Balabanović, M. and Shoham, Y. (1997). Fab: content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–72.

[10] Ball, J., Borger, J., and Greenwald, G. (2013). Revealed: how US and UK spy agencies defeat internet privacy and security. *Guardian*.
https://www.theguardian.com/world/2013/sep/05/nsa-gchq-encryption-codes-security
Accessed: September 2016.

[11] Baron-Cohen, S. and Wheelwright, S. (2004). The empathy quotient: an investigation of adults with asperger syndrome or high functioning autism, and normal sex differences. *Journal of autism and developmental disorders*, 34(2):163–175.

[12] Barrick, M. R. and Mount, M. K. (1991). The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology*, 44(1):1–26.

[13] Barshan, E., Ghodsi, A., Azimifar, Z., and Jahromi, M. Z. (2011). Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition*, 44(7):1357–1371.

[14] Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and trends® in Machine Learning*, 2(1):1–127.

[15] Bi, B., Shokouhi, M., Kosinski, M., and Graepel, T. (2013). Inferring the demographics of search users: Social data meets search queries. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 131–140, New York, NY, USA. ACM.

[16] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[17] Book, B.-r. D. P.-r. (2006). Advanced Access Content System (AACS).

[18] Bosnjak, M., Galesic, M., and Tuten, T. (2007). Personality determinants of online shopping: Explaining online purchase intentions using a hierarchical approach. *Journal of Business Research*, 60(6):597–605.

[19] Bracewell, R. N. and Bracewell, R. N. (1986). *The Fourier transform and its applications*, volume 31999. McGraw-Hill New York.

[20] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[21] Brown, J. D. (1986). Evaluations of self and others: Self-enhancement biases in social judgments. *Social cognition*, 4(4):353.

[22] Canli, T., Zhao, Z., Desmond, J. E., Kang, E., Gross, J., and Gabrieli, J. D. (2001). An fmri study of personality influences on brain reactivity to emotional stimuli. *Behavioral neuroscience*, 115(1):33.

[23] Caruana, R. and Freitag, D. (1994). Greedy attribute selection. In *In Proceedings of the Eleventh International Conference on Machine Learning*, pages 28–36. Morgan Kaufmann.

[24] Cattell, R. B. (1946). Description and measurement of personality.

[25] Correa, T., Hinsley, A. W., and De Zuniga, H. G. (2010). Who interacts on the web?: The intersection of users' personality and social media use. *Computers in Human Behavior*, 26(2):247–253.

[26] Costa, P. and McCrae, R. (1989). Neo Five-Factor Inventory (NEO-FFI). *Odessa, FL: Psychological Assessment Resources*.

[27] Costa, P. T. and McCrae, R. R. (1988). Personality in adulthood: a six-year longitudinal study of self-reports and spouse ratings on the neo personality inventory. *Journal of personality and social psychology*, 54(5):853.

[28] Costa, P. T. and McCrae, R. R. (1992). Four ways five factors are basic. *Personality and individual differences*, 13(6):653–665.

[29] Costa, P. T. and McCrae, R. R. (2008). The revised neo personality inventory (NEO-PI-R). *The SAGE handbook of personality theory and assessment*, 2:179–198.

[30] de Montjoye, Y.-A., Quoidbach, J., Robic, F., and Pentland, A. S. (2013). Predicting personality using novel mobile phone-based metrics. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 48–55. Springer.

[31] Deeplearning4j Development Team (2017). Deeplearning4j: Open-source distributed deep learning for the jvm, apache software foundation license 2.0. http://deeplearning4j.org. Accessed: August, 2017.

[32] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.

[33] Diener, E., Emmons, R. A., Larsen, R. J., and Griffin, S. (1985). The satisfaction with life scale. *Journal of personality assessment*, 49(1):71–75.

[34] Driscoll, J. C. and Kraay, A. C. (1998). Consistent covariance matrix estimation with spatially dependent panel data. *Review of economics and statistics*, 80(4):549–560.

[35] Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.*, 19(1):61–74.

[36] Elahi, M., Braunhofer, M., Ricci, F., and Tkalcic, M. (2013). Personality-based active learning for collaborative filtering recommender systems. In *Congress of the Italian Association for Artificial Intelligence*, pages 360–371. Springer.

[37] Elwood, D. L. and Griffin, H. R. (1972). Individual intelligence testing without the examiner: Reliability of an automated method. *Journal of Consulting and Clinical Psychology*, 38(1):9.

[38] Embretson, S. E. and Reise, S. P. (2013). *Item response theory*. Psychology Press.

[39] Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., and Bengio, S. (2010). Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11(Feb):625–660.

[40] ew Research Center (2017). News use across social media platforms 2017. http://www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017/. Accessed: August, 2017.

[41] Facebook Inc. (2015). Facenook terms of service. https://www.facebook.com/legal/terms. Accessed: August, 2017.

[42] Facebook Inc. (2016). Facebook press page. http://newsroom.fb.com/company-info/. Accessed: August, 2017.

[43] Fernández-Tobías, I., Braunhofer, M., Elahi, M., Ricci, F., and Cantador, I. (2016). Alleviating the new user problem in collaborative filtering by exploiting personality information. *User Modeling and User-Adapted Interaction*, 26(2):221–255.

[44] Fiat, A. and Naor, M. (1994). Broadcast encryption. In *Proceedings of the 13th Annual International Cryptology Conference on Advances in Cryptology*, CRYPTO ’93, pages 480–491, London, UK, UK. Springer-Verlag.

[45] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.

[46] Fowler, R. D. (1985). Landmarks in computer-assisted psychological assessment. *Journal of Consulting and Clinical Psychology*, 53(6):748.

[47] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

[48] Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378.

[49] Global Web Index Limited (2017). Daily time spent on social networks rises to over 2 hours. http://blog.globalwebindex.net/chart-of-the-day/daily-time-spent-on-social-networks/. Accessed: August, 2017.

[50] Golbeck, J., Robles, C., and Turner, K. (2011). Predicting personality with social media. In *CHI'11 extended abstracts on human factors in computing systems*, pages 253–262. ACM.

[51] Goldberg, L. R. (1990). An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology*, 59(6):1216.

[52] Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological Assessment*, 4(1):26.

[53] Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American psychologist*, 48(1):26.

[54] Goldberg, L. R. et al. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality psychology in Europe*, 7(1):7–28.

[55] Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., and Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in personality*, 40(1):84–96.

[56] Gosling, S. D., Augustine, A. A., Vazire, S., Holtzman, N., and Gaddis, S. (2011). Manifestations of personality in online social networks: Self-reported facebook-related behaviors and observable profile information. *Cyberpsychology, Behavior, and Social Networking*, 14(9):483–488.

[57] Gosling, S. D., Rentfrow, P. J., and Swann, W. B. (2003a). A very brief measure of the Big-Five personality domains. *Journal of Research in personality*, 37(6):504–528.

[58] Gosling, S. D., Rentfrow, P. J., and Swann, W. B. (2003b). A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37:504–528.

[59] Greene, R. L. (2000). *The MMPI-2: An interpretive manual*. Allyn & Bacon.

[60] Greenwald, G. (2013). NSA collecting phone records of millions of Verizon customers daily. *Guardian*.
https://www.theguardian.com/world/2013/jun/06/nsa-phone-records-verizon-court-order
Accessed: September 2016.

[61] Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235.

[62] Halko, N., Martinsson, P.-G., and Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288.

[63] Halko, N. P. (2012). *Randomized methods for computing low-rank approximations of matrices*. PhD thesis, University of Colorado.

[64] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

[65] Hall, M. A. (2000). Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 359–366, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[66] Hall, M. A. and Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *IEEE transactions on knowledge and data engineering*, 15(6):1437–1447.

[67] Haralick, R. M., Shanmugam, K., et al. (1973). Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621.

[68] Harper, F. M. and Konstan, J. A. (2015). The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4):19:1–19:19.

[69] Higgins, D. M., Peterson, J. B., Pihl, R. O., and Lee, A. G. (2007). Prefrontal cognitive ability, intelligence, big five personality, and the prediction of advanced academic and workplace performance. *Journal of Personality and Social Psychology*, 93(2):298.

[70] Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554.

[71] Holmes, M., Gray, A., and Isbell, C. (2007). Fast SVD for large-scale matrices. In *Workshop on Efficient Machine Learning at NIPS*, volume 58, pages 249–252.

[72] Holmes, M. P., Isbell, J., Lee, C., and Gray, A. G. (2009). QUIC-SVD: Fast SVD using cosine trees. In *Advances in Neural Information Processing Systems*, pages 673–680.

[73] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

[74] Huang, Z., Du, W., and Chen, B. (2005). Deriving private information from randomized data. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 37–48. ACM.

[75] Huber, P. J. et al. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

[76] Instagram Inc. (2016). Instagram press page. https://www.instagram.com/press/. Accessed: August, 2016.

[77] Jelenchick, L. A., Eickhoff, J. C., and Moreno, M. A. (2013). "facebook depression?" social networking site use and depression in older adolescents. *Journal of Adolescent Health*, 52(1):128–130.

[78] John, O. P. and Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999):102–138.

[79] Judge, T. A., Heller, D., and Mount, M. K. (2002). Five-factor model of personality and job satisfaction: a meta-analysis. *Journal of Applied Psychology*, 87(3):530.

[80] Karegowda, A. G., Manjunath, A., and Jayaram, M. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*, 2(2):271–277.

[81] Kassarjian, H. H. (1971). Personality and consumer behavior: A review. *Journal of marketing Research*, pages 409–418.

[82] Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324.

[83] Koller, D. and Sahami, M. (1996). Toward optimal feature selection. In Saitta, L., editor, *Proceedings of the Thirteenth International Conference on Machine Learning (ICML)*, pages 284–292. Morgan Kaufmann Publishers.

[84] Kosinski, M. (2014). *Measurement and prediction of individual and group differences in the digital environment*. PhD thesis, University of Cambridge.

[85] Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., and Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, 70(6):543.

[86] Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805.

[87] Kraut, R., , O., Banaji, M., , B., , C., and , C. (2004). Psychological research online: Opportunities and challenges. pages 105–117.

[88] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

[89] Kwan, V. S., John, O. P., Kenny, D. A., Bond, M. H., and Robins, R. W. (2004). Reconceptualizing individual differences in self-enhancement bias: an interpersonal approach. *Psychological review*, 111(1):94.

[90] Lawrence, S., Giles, C. L., Tsoi, A. C., and Back, A. D. (1997). Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks*, 8(1):98–113.

[91] Leary, M. R. and Allen, A. B. (2011). Self-presentational persona: simultaneous management of multiple impressions. *Journal of personality and social psychology*, 101(5):1033.

[92] LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995.

[93] Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

[94] Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th annual international conference on machine learning*, pages 609–616. ACM.

[95] Liu, H. and Setiono, R. (1996). A Probabilistic Approach to Feature Selection - A Filter Solution. In *International Conference on Machine Learning*, pages 319–327.

[96] Lushene, R. E., O'Neil Jr, H. F., and Dunn, T. (1974). Equivalent validity of a completely computerized mmpi. *Journal of Personality Assessment*, 38(4):353–361.

[97] Mairesse, F., Walker, M. A., Mehl, M. R., and Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500.

[98] Mcauliffe, J. D. and Blei, D. M. (2008). Supervised topic models. In *Advances in neural information processing systems*, pages 121–128.

[99] McCrae, R. R. and Costa Jr, P. T. (1997). Personality trait structure as a human universal. *American psychologist*, 52(5):509.

[100] Merton, R. K. (1968). *Social theory and social structure*. Simon and Schuster.

[101] Moore, K. and McElroy, J. C. (2012). The influence of personality on facebook usage, wall postings, and regret. *Computers in Human Behavior*, 28(1):267–274.

[102] Murdock Jr, B. B. (1962). The serial position effect of free recall. *Journal of experimental psychology*, 64(5):482.

[103] Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system.

[104] Noftle, E. E. and Robins, R. W. (2007). Personality predictors of academic outcomes: big five correlates of gpa and sat scores. *Journal of personality and social psychology*, 93(1):116.

[105] Owen, S., Anil, R., Dunning, T., and Friedman, E. (2011). *Mahout in Action*. Manning Publications Co., Greenwich, CT, USA.

[106] Ozer, D. J. and Benet-Martinez, V. (2006). Personality and the prediction of consequential outcomes. *Annu. Rev. Psychol.*, 57:401–421.

[107] Paninski, L. (2003). Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253.

[108] Pantic, I., Damjanovic, A., Todorovic, J., Topalovic, D., Bojovic-Jovic, D., Ristic, S., and Pantic, S. (2012). Association between online social networking and depression in high school students: behavioral physiology viewpoint. *Psychiatria Danubina*, 24(1.):90–93.

[109] Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., and Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6):934.

[110] Paulhus, D. L. (1998). Interpersonal and intrapsychic adaptiveness of trait self-enhancement: A mixed blessing? *Journal of personality and social psychology*, 74(5):1197.

[111] Pavot, W. and Diener, E. (1993). Review of the satisfaction with life scale. *Psychological assessment*, 5(2):164.

[112] Pearson, J. S., Rome, H. P., Swenson, W. M., Mataya, P., and Brannick, T. L. (1965). Development of a computer system for scoring and interpretation of minnesota multiphasic personality inventories in a medical clinic. *Annals of the New York Academy of Sciences*, 126(1):684–695.

[113] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[114] Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238.

[115] Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

[116] Phares, E. J. and Chaplin, W. F. (1997). *Introduction to Personality*. Longman.

[117] Quercia, D., Kosinski, M., Stillwell, D., and Crowcroft, J. (2011). Our twitter profiles, our selves: Personality with twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 180–185.

[118] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[119] Radloff, L. S. (1977). The ces-d scale a self-report depression scale for research in the general population. *Applied psychological measurement*, 1(3):385–401.

[120] Rammstedt, B. and John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of research in Personality*, 41(1):203–212.

[121] Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., and Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, 2(4):313–345.

[122] Robins, R. W., Trzesniewski, K. H., Tracy, J. L., Gosling, S. D., and Potter, J. (2002). Global self-esteem across the life span. *Psychology and aging*, 17(3):423.

[123] Ross, C., Orr, E. S., Sisic, M., Arseneault, J. M., Simmering, M. G., and Orr, R. R. (2009). Personality and motivations associated with facebook use. *Computers in human behavior*, 25(2):578–586.

[124] Rust, J., Golombok, S., Kosinski, M., and Stillwell, D. (2014). *Modern psychometrics: The science of psychological assessment*. Routledge.

[125] Ryan, T. and Xenos, S. (2011). Who uses facebook? an investigation into the relationship between the big five, shyness, narcissism, loneliness, and facebook usage. *Computers in Human Behavior*, 27(5):1658–1664.

[126] Schapire, R. E. (1990). The strength of weak learnability. *Machine learning*, 5(2):197–227.

[127] Seidman, G. (2013). Self-presentation and belonging on facebook: How personality influences social media use and motivations. *Personality and Individual Differences*, 54(3):402–407.

[128] Shannon, C. E. (1951). Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64.

[129] Sheldon, W. H., Stevens, S. S., and Tucker, W. B. (1940). *The varieties of human physique: An introduction to constitutional psychology*, volume 1. Harper.

[130] Snyder, M. (1974). Self-monitoring of expressive behavior. *Journal of personality and social psychology*, 30(4):526.

[131] Staiano, J., Lepri, B., Aharony, N., Pianesi, F., Sebe, N., and Pentland, A. (2012). Friends don't lie: inferring personality traits from social network structure. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, pages 321–330. ACM.

[132] Sun, J.-T., Chen, Z., Zeng, H.-J., Lu, Y.-C., Shi, C.-Y., and Ma, W.-Y. (2004). Supervised latent semantic indexing for document categorization. In *Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on*, pages 535–538. IEEE.

[133] Sweeney, L. (1997). Guaranteeing anonymity when sharing medical data, the datafly system. In *Proceedings of the AMIA Annual Fall Symposium, Journal of the Aerican Medical Informatics Association*, page 51. American Medical Informatics Association.

[134] The British Psychological Society (2014). Code of human research ethics.

[135] The British Psychological Society (2017). Ethics guidelines for internet-mediated research.

[136] Thissen, D., Steinberg, L., and Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models.

[137] Tkalcic, M., Kunaver, M., Tasic, J., and Košir, A. (2009). Personality based user similarity measure for a collaborative recommender system. In *Proceedings of the 5th Workshop on Emotion in Human-Computer Interaction-Real world challenges*, pages 30–37.

[138] Tupes, E. C. and Christal, R. E. (1992). Recurrent personality factors based on trait ratings. *Journal of personality*, 60(2):225–251.

[139] Twitter Inc. (2016). Twitter press page. https://about.twitter.com/company. Accessed: August, 2016.

[140] Twitter Inc. (2017). Twitter terms of service. https://twitter.com/en/tos/. Accessed: August, 2017.

[141] Vaidya, J., Kantarcıoğlu, M., and Clifton, C. (2008). Privacy-preserving naive bayes classification. *The VLDB Journal—The International Journal on Very Large Data Bases*, 17(4):879–898.

[142] Verykios, V. S., Bertino, E., Fovino, I. N., Provenza, L. P., Saygin, Y., and Theodoridis, Y. (2004). State-of-the-art in privacy preserving data mining. *ACM Sigmod Record*, 33(1):50–57.

[143] Vul, E., Harris, C., Winkielman, P., and Pashler, H. (2009). Puzzlingly high correlations in fmri studies of emotion, personality, and social cognition. *Perspectives on psychological science*, 4(3):274–290.

[144] Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., and Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge.

[145] Wald, R., Khoshgoftaar, T., and Sumner, C. (2012). Machine prediction of personality from facebook profiles. In *Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on*, pages 109–115. IEEE.

[146] Wang, N., Kosinski, M., Stillwell, D., and Rust, J. (2014). Can well-being be measured using facebook status updates? validation of facebook's gross national happiness index. *Social Indicators Research*, 115(1):483–491.

[147] Wang, X., Yin, Y. L., and Yu, H. (2005). Finding collisions in the full SHA-1. In *Annual International Cryptology Conference*, pages 17–36. Springer.

[148] Wang, X. and Yu, H. (2005). How to break MD5 and other hash functions. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 19–35. Springer.

[149] Wilson, K., Fornasier, S., and White, K. M. (2010). Psychological predictors of young adults' use of social networking sites. *Cyberpsychology, behavior, and social networking*, 13(2):173–177.

[150] Wojcik, S. P., Hovasapian, A., Graham, J., Motyl, M., and Ditto, P. H. (2015). Conservatives report, but liberals display, greater happiness. *Science*, 347(6227):1243–1246.

[151] Yang, Y. and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *ICML*, volume 97, pages 412–420.

[152] Youyou, W., Kosinski, M., and Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040.

[153] Yu, S., Yu, K., Tresp, V., Kriegel, H.-P., and Wu, M. (2006). Supervised probabilistic principal component analysis. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 464–473. ACM.

[154] Zhang, Z. and Wu, L. (2003). Optimal low-rank approximation to a correlation matrix. *Linear algebra and its applications*, 364:161–187.

[155] Zhou, T. and Tao, D. (2011). Godec: Randomized low-rank & sparse matrix decomposition in noisy case. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 33–40.

[156] Zhu, Y. and Liu, L. (2004). Optimal randomization for privacy preserving data mining. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 761–766. ACM.

Appendix A

Frameworks

Appendix A lists various frameworks that are used as part of this thesis.

A.1 Apache Mahout

Apache Mahout[1] [105] is an open-source Java[2] and Scala[3] library and toolkit for implementing scalable machine learning algorithms, developed and maintained by the Apache Software Foundation[4].

A.1.1 Singular Value Decomposition

Mahout implements a randomised SVD computation using the probabilistic approximate matrix decomposition reported by Halko et al. [62, 63].

A.1.2 Recommender systems

Mahout implements scalable user-based collaborative filtering algorithms [9]. They work by finding similarity between users and make recommendations to user A based on what the most similar users to user A has liked, or disliked, that user A has not expressed any

[1] https://mahout.apache.org

[2] https://www.java.com

[3] http://www.scala-lang.org

[4] https://www.apache.org

interest towards. The most important aspect of a user-based recommender system is the way similarity between users is judged. Ted Dunning's Log Likelihood Ratio (LLR) [35] accounts for surprise and coincide performs really well in order to measure similarity between users in a binary dataset. Table A.1 demonstrates the simple principle behind LLR. This is the same principle used in finding similarities between documents.

	`User A`	`NOT User A`
`User B`	X_{AB}	$X_{\overline{A}B}$
`NOT User B`	$X_{A\overline{B}}$	$X_{\overline{AB}}$

Table A.1 Similarity and difference between two users on Facebook

X_{AB} denotes the number of pages on Facebook that both `User A` and `User B` have liked. $X_{\overline{A}B}$ denotes the number of pages that `User A` has not liked, but `User B` has liked. $X_{A\overline{B}}$ denotes the number of pages that `User A` has liked, but `User B` has not liked. Finally, $X_{\overline{AB}}$ denotes the number of pages that neither `User A` nor `User B` have liked. *LLR* measures the amount of independence observed using entropy method developed by Shanon [128].

A.2 Weka

Weka[5] [64] is an open-source Java framework that packages a lot of widely used machine learning algorithms, and maintained by the Machine Learning Group at the University of Waikato.

A.2.1 Wrapper subset selection

Weka implements the wrapper subset selection method introduced by Kohavi et al. [82]. It also implements the Greedy search algorithm with forward stepwise selection and backwards stepwise elimination [23].

[5] http://cs.waikato.ac.nz/ml/weka/index.html

A.2.2 J48

J48 is an open-source implementation of the C4.5 algorithm [118] for training decision trees in Weka.

A.3 Scikit-learn

Scikit-learn [6] [113] is an open-source machine learning library for Python [7].

A.3.1 Random forest

Scikit-learn implements Boreiman's random forest algorithm for classifications [20]. Random forest is an ensemble model (uses a combination of weak learners compared to a single strong learner, as described in section 5.2.2).

A.3.2 Gradient boosting

Scikit-learn implements Friedman's approximation algorithm for training of gradient boosted decision trees [48, 47].

A.4 DeepLearning4J

DeepLearning4J [8] [31] is an open-source and scalable deep learning library for Java and Scala.

A.4.1 Deep belief networks

A deep belief network is a multi-layer neural network with hidden layers. DeepLearning4J implements Hinton's greedy algorithm for supervised training of deep belief networks [70].

[6] http://www.scikit-learn.org

[7] https://www.python.org

[8] http://www.deeplearning4j.org/