UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DSIC
DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

MASTER'S THESIS

Inferring the curiosity by using Facebook profile data

Inferencia de la curiosidad mediante el uso de datos de perfil de Facebook

Author:
Rebeca Janina DELCONTE FERREIRA DOS SANTOS

Supervisor:
Dr. Laura SEBASTIA

A thesis submitted in fulfillment of
the requirements for the Master's degree in
Engineering and Technology of Software Systems

in the

DSIC - Department of Information Systems and Computation
September 2017

Resum

En els últims anys, l'ús d'Internet i xarxes socials s'ha tornat quasi indispensable en la nostra societat. Malgrat el pensament general, els estudis han demostrat que el perfil en les xarxes socials mostra en gran manera la personalitat real de l'individu. Dit açò, seria valuós extraure algun coneixement de tal quantitat de dades i utilitzar-ho per a millorar el rendiment de les eines de personalització intel·ligent, com la recomanació de productes i serveis o altres tipus de màrqueting.

En aquest treball, hem optat per treballar amb les dades disponibles en el projecte anomenat myPersonality, que posa a disposició tant els perfils de Facebook com les respostes al qüestionari psicològic Big Five. El nostre enfocament es va centrar en l'anàlisi de dos trets específics de la personalitat humana, l'obertura a l'experiència i la extroversió, que estan estretament relacionats amb la curiositat. Hem desenvolupat models de regressió i de classificació per a predir aquests dos trets. Els resultats han demostrat una significància feble, a pesar que la extroversió va obtenir resultats lleugerament millors, indicant que algunes dades de Facebook poden expressar el grau de extroversió de l'usuari.

Paraules clau: Xarxes socials, curiositat, Facebook, myPersonality, Big Five

Resumen

En los últimos años, el uso de Internet y redes sociales se ha vuelto casi indispensable en nuestra sociedad. A pesar del pensamiento general, los estudios han demostrado que el perfil en las redes sociales muestra en gran medida la personalidad real del individuo. Dicho esto, sería valioso extraer algún conocimiento de tal cantidad de datos y utilizarlo para mejorar el rendimiento de las herramientas de personalización inteligente, como la recomendación de productos y servicios u otros tipos de marketing.

En este trabajo, hemos optado por trabajar con los datos disponibles en el proyecto llamado myPersonality, que pone a disposición tanto los perfiles de Facebook como las respuestas al cuestionario psicológico Big Five.

Nuestro enfoque se dió al análisis de dos rasgos específicos de la personalidad humana, la apertura a la experiencia y la extroversión, que están estrechamente relacionados con la curiosidad. Hemos desarrollado modelos de regresión y de clasificación para predecir estos dos rasgos. Los resultados han demostrado una significancia débil, a pesar de que la extroversión obtuvo resultados ligeramente mejores, indicando que algunos datos de Facebook pueden expresar el grado de extroversión del usuario.

Palabras clave: Redes sociales, curiosidad, Facebook, myPersonality, Big Five

Abstract

In the recent years, the use of internet and social networks has become almost indispensable in our society. Despite the general thinking, studies have shown that the profile in social networks shows a great extent of the actual personality of the individual. That said, it would be valuable to extract some knowledge from such amount of data and use it to enhance the performance of intelligent personalization tools such as recommendation of products and services or other types of marketing.

Seizing the great amount of data available under the project called myPersonality, which makes available both Facebook profiles and responses to the Big Five psychological questionnaire, we have chosen to work on this data.

We were interested in two specific traits of human personality, the openness to experience and the extroversion, which are closely related to the curiosity. We built regression and classification models to predict extroversion and openness of the users. Results shown weak significance, although extroversion obtained a slightly better result, indicating that some Facebook features can express the extroversion of the person.

Keywords: Social networks, curiosity, Facebook, myPersonality, Big Five

Acknowledgements

Many thanks to David Stillwell and Michal Kosinski for access and guidance regarding the data from myPersonality project data; to my supervisor Laura who was always present when I doubted; and to my lovely husband Alan who has always a motivation word for me, thank you all for supporting me.

Contents

List of Figures

List of Tables

Listings

Chapter 1

Introduction

1.1 Motivation

The human personality says how and why we act this or that way, and it can be used not only for us to better know ourselves, but also to aid improving a huge amount of technology to what we are exposed to. Especially because technology exists to make our lives better and easier. A psychometric test is one of the most common and easy ways people can find out their personality, and there are several different tests that can be taken, according to different authors and approaches. For instance, Big Five is one of those that allow to measure the personality.

Given the fact that nowadays the use of the internet, especially the social networks, have been constantly increasing and changing the way people interact and communicate with each other, it emerged the idea to turn the data available useful in ongoing projects also related to psychological factors.

Facebook is not used only for communication; it has been becoming a business tool. Some companies have Facebook pages to promote their products or services, while others make real business inside the platform, whether selling products (there is a resource called Facebook store), whether through advertisements, specific for each user thanks to the recommendation algorithms.

The personality is constantly expressed in our daily activities, in our social relations, but these footprints can also be found in the online world, where anything can be done, from ordering a pizza to taking a course. It can be measured by means of questionnaires, but it also could be measured by finding relations among those online footprints. Thus, the discovery of the personality of the user has been opening a new window in the research of personality and it has created a new field of investigation by computing researchers. For instance, Menk and Sebastiá, 2016 have found significant correlations between the curiosity (measured by means of the CEI-II questionnaire) and some features of Facebook profiles of the participants.

In order to perform similar research with a more diverse and big extent of users, we have chosen myPersonality data. The project called myPersonality (Kosinski et al., 2015) was a Facebook application developed in 2007, which contained a Big Five-questionnaire in order to measure the personality of the users on that social network, thus recording the tests and some data from their profile online (with their consent).

Sine is not that easy to obtain such amount of users and information, their idealizers made available for other researchers to use those data.

Since then, over 200 researchers have been working with myPersonality data, and we hope to answer some questions not clearly answered about relations between our personality and the usage of social networks, for instance Facebook, and how researchers could use this relation to better understand people's needs and habits. We have seen some researches talking about the relation between myPersonality project and Big Five, so we are interested in explore the relation between myPersonality and other questionnaires, such as CEI-II. This is one of the most accepted questionnaires to measure the human curiosity. Our main goal is to deeper analyze two of the five personality traits, openness and extroversion, which are closely correlated to curiosity; we believe that curiosity could be a relevant variable/factor to increment in recommendation systems in order to improve the results obtained in terms of satisfaction of the users.

1.2 Objectives / Goals

The main goal of this work is to generate a prediction model for the curiosity based on the features of Facebook. As secondary objectives, we want to individually analyze the eight lengths of the 20-100 IPIP questionnaire for Big Five and detect with which one we get the best correlation results, besides to identify the features more correlated with the curiosity.

1.3 Structure

The development of this thesis was divided into 5 chapters. First, the definition and goals of this study are presented; next, we make an overview on the main subjects addressed along this work, Chapter 3 is dedicated to the explanation of the model developed. In Chapter 4 we describe the results obtained and, finally, in Chapter 5, we discuss them and present future possible work.

Chapter 2

Related Works

In recent days, it is clear that a huge amount of data is available online. Media files, such as photos and videos, text files from e-mails and messages changed in social networks, and given that people are always connected

The amount and diversity of data extracted from the internet is immense and updated every moment, especially in the social networks.

According to IDC, 2014, the digital universe is doubling in size every two years, and will reach 44 Zettabytes in 2020, as we can see in fig. 2.1. Such growth has no way back, so why not to take advantage of this enormous data extent? An important role is to filter, analyze and then extract valuable information of that data

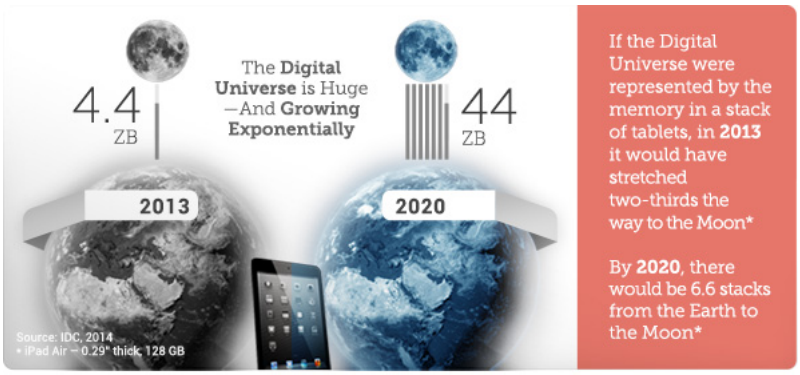

Fig. 2.1: EMC Digital Universe, 2014

One type of data that is clearly growing is regarding the use of social networks, especially Facebook. The website has launched in 2004 by Mark Zuckerberg inside Harvard's University ((Phillips, 2007)), originally known as "The Facebook". The initial intention was to use it as a network within the students, but after its popularity had grown that fast, it promptly extended to other universities. Until 2006, it was still restricted to universities' students, but had spread to other countries. Nowadays, it is the most known social network, achieving the figure of two billion users this year, according to Zuckerberg himself.

Given the widespread impact that this social network represents, it could be valuable to use that data to discover patterns in human behavior. But, do the online profiles represent the reality, or they are just a way to show the world and idealized version of themselves? Different of previous researchers that had this thinking, Back

et al., 2010 shows that people do not use social networks to promote an idealized or projected personality. It means that, by using Facebook data, it would be possible to obtain real psychological profiles without the need to filling forms, and use this result to improve the recommendation systems, among other things.

Talking about the human personality, it has not a single definition. (John, 2008) mention that a common way that psychologists refer to personality is as people's consistent patterns of feeling, cognition, and behaving. When studied, two broad areas can be focused: understanding individual differences in particular personality characteristics, such as sociability or irritability; and understanding how the various parts of a person come together as a whole (American Psychological Association, 2017). The measurement of the personality is often used to diagnose diseases and mental disorders, or to determine workplace suitability. But they can be used in other fields, as we will discuss later. Several authors have been studying ways to interpret and measure the characteristics of personality, so various approaches have been created over the years. And, although psychological researchers recognize that personality is also about biological factors, the role of culture, and needs, among other things, the most studied perspective involves the *trait* concept; the study of personality traits began with Allport and Allport, 1921, and refers to personality types or qualities, mostly explained by single adjectives as items. To measure it, there are psychometric tests that assess the personality by traits, as it would be too difficult to assess it by single responses. Thus, the personality traits are groups of adjectives that people use when describing themselves or others.

The most widely known, studied, and generally accepted - albeit not unanimous among all researchers (Block, 1995) - model of personality, according to Goldberg, 1993; Costa Jr and McCrae, 1992 is the Big Five model, a five-factor framework of the human personality described by a lexical method, that is, based on a linguistic analysis. Those five factors are personality dimensions that summarize several more specific facets that comprise a person's personality. Its statistical technique is through the application of questionnaires that are able to reduce a huge amount of information in a synthetic and relevant set. In other words, the "Big Five" is a reasonable representation of human personality (McCrae and Costa, 1987). The five factors are the following:

- Agreeableness: includes attributes such as trust, altruism, kindness, affection, and other prosocial behaviors. People who are high in agreeableness tend to be more cooperative while those low in this trait tend to be more competitive and even manipulative (Poropat, 2009).

- Conscientiousness: is a tendency to show self-discipline, act dutifully, and aim for achievement. In other words, it is related to planning, organization, and dependability (Barrick and Mount, 1991).

- Extraversion: regarding positive emotions, assertiveness, sociability, talkativeness, and also the tendency to seek stimulation in the interaction with others.

- Neuroticism (sometimes reversed and called Emotional Stability): describes vulnerability to unpleasant emotions like anger, anxiety, depression, or vulnerability. Neuroticism also refers to an individual's level of emotional stability and impulse control and is sometimes referred to as emotional stability (Barrick and Mount, 1991).

- Openness to experience: it is related to a person's curiosity of the person, creativity, and preference for novelty and variety. Some disagreement remains about how to interpret this factor, which is sometimes called intellect.

Fig. 2.2 shows the complete taxonomy of Big Five, where it is possible to find all the adjectives related to each trait. As we can observe, each trait of the Big Five model is broad and consist of a range of "synonyms" or more specific characteristics. Its structure, divided into 5 groups, was derived from statistical analysis of which traits tend to arise in the description of people about themselves or about other people. The underlying correlations, on the other hand, are probabilistic, and exceptions may occur.

Extraversion		Agreeableness		Conscientiousness		Neuroticism		Openness	
Low	High	Low	High	Low	High	Low	High	Low	High
−.83 Quiet	.85 Talkative	−.52 Fault-finding	.87 Sympathetic	−.58 Careless	.80 Organized	−.39 Stable	.73 Tense	−.74 Commonplace	.76 Wide interests
−.80 Reserved	.83 Assertive	−.48 Cold	.85 Kind	−.53 Disorderly	.80 Thorough	−.35 Calm	.72 Anxious	−.73 Narrow interests	.76 Imaginative
−.75 Shy	.82 Active	−.45 Unfriendly	.85 Appreciative	−.50 Frivolous	.78 Planful	−.21 Contented	.72 Nervous	−.67 Simple	.72 Intelligent
−.71 Silent	.82 Energetic	−.45 Quarrelsome	.84 Affectionate	−.49 Irresponsible	.78 Efficient		.71 Moody	−.55 Shallow	.73 Original
−.67 Withdrawn	.82 Outgoing	−.45 Hard-hearted	.84 Soft-hearted	−.40 Slipshot	.73 Responsible		.71 Worrying	−.47 Unintelligent	.68 Insightful
−.66 Retiring	.80 Outspoken	−.38 Unkind	.82 Warm	−.39 Undependable	.72 Reliable		.68 Touchy		.64 Curious
	.79 Dominant	−.33 Cruel	.81 Generous	−.37 Forgetful	.70 Dependable		.64 Fearful		.59 Sophisticated
	.73 Forceful	−.31 Stern	.78 Trusting		.68 Conscientious		.63 High-strung		.59 Artistic
	.73 Enthusiastic	−.28 Thankless	.77 Helpful		.66 Precise		.63 Self-pitying		.59 Clever
	.68 Show-off	−.24 Stingy	.77 Forgiving		.66 Practical		.60 Temperamental		.58 Inventive
	.68 Sociable		.74 Pleasant		.65 Deliberate		.59 Unstable		.56 Sharp-witted
	.64 Spunky		.73 Good-natured		.46 Painstaking		.58 Self-punishing		.55 Ingenious
	.64 Adventurous		.73 Friendly		.26 Cautious		.54 Despondent		.45 Witty
	.62 Noisy		.72 Cooperative				.51 Emotional		.45 Resourceful
	.58 Bossy		.67 Gentle						.37 Wise
			.66 Unselfish						
			.56 Praising						
			.51 Sensitive						

Note. Based on John (1990). These items were assigned to one Big Five domain by at least 90% of the judges and thus capture the most prototypical (or central) content of each Big Five domain. The factor loadings, shown here only for the expected factor, were obtained in a sample of 140 males and 140 females, each of whom had been described by 10 psychologists serving as observers during an assessment weekend at the Institute of Personality and Social Research at the University of California at Berkeley (see also John, 1989).

Fig. 2.2: Big Five Prototypes: Most Central Trait Adjectives Selected Consensually by Expert Judges and Their Factor Loadings in Personality Ratings by 10 Psychologists Serving as Observers

For instance, Srivastava, 2006 mentions that the traits talkativeness and assertiveness are both associated with Extraversion, but they do not go together by logical necessity: you could imagine somebody that is assertive but not talkative. Nevertheless, many studies indicate that people who are talkative are usually also assertive (and vice versa), which is why they go together under the broader Extraversion factor.

It is worth highlighting that some aspects of the personality are not subsumed into the Big Five model, since the term personality trait has a special meaning in personality psychology that is narrower than the everyday usage of the term.

Other personality traits such as motivations, emotions, attitudes, abilities, self-concepts, social roles, autobiographical memories, and life stories are just a few of the other "units" may have theoretical or empirical relationships with the Big Five

traits, but they are conceptually distinct. This way, even a very comprehensive profile of somebody's personality traits can only be considered a partial description of their personality (Srivastava, 2006).

Regarding the measurement approaches of the Big Five dimensions, we can mention the NEO-PI. The very first version of this measurement was published in 1978 by Paul Costa, Jr. and Robert McCrae, so called NEO-I, and included only three factors, Neuroticism, Extraversion and Openness. The other two traits were included in a revised version (Costa and McCrae, 1985) and renamed as NEO Personality Inventory (NEO-PI), where NEO became no longer an acronym, but part of its name. Revised again, and nowadays considered the most comprehensive inventory (Costa Jr and McCrae, 1992), the NEO-PI-R contains 240 items and permits the measurement of the Big Five domains and six specific facets within each dimension. Given its considerable length and time spent to complete it, a number of shorter questionnaires arose. We highlight the 100 trait descriptive adjectives (TDA), where Goldberg, 1992 have selected only adjectives that uniquely defined each trait; the NEO-Five Factor Inventory (NEO-FFI), composed of 60 items (Costa Jr and McCrae, 1992), which consists of items that loaded highly on one of the five factors; and the Big Five Inventory (BFI)(John and Srivastava, 1999), which includes 44 items and, instead of using adjectives in the questionnaires, it uses short phrases based on the trait adjectives; for instance, the "Openness" adjective became the BFI item "Is original, comes up with new ideas".

Besides those approaches, Goldberg, 1999 introduced the International Personality Item Pool (IPIP), under the reason that broad-bandwidth inventories, such as NEO-PI, are proprietary, thus they cannot be freely used by other researchers, in addition to not being constantly revised. Another issue is that most publishers do not allow the use of their copyrighted inventories on the internet (Goldberg et al., 2006). Hence, IPIP scales were designed to measure constructs similar to those in existing personality inventories. Their reliability lies in the correlation between the IPIP scale and the scale on which it was based. For instance, there is a 20 to 100-item IPIP proxy for NEO-PI-R domains (Five Factor Model), which is a widely accepted personality questionnaire. All scales that have been constructed from the items are gathered in the IPIP website[1], which is intended as an international effort to develop and continually refine a set of personality inventories, whose items are in the public domain, and whose scales can be used for both scientific and commercial purposes (IPIP, 2017).

Our personality can be expressed by deliberate attempts to make statements to others. However, other forms of expression may simply be individual's inadvertent actions. This is known as *behavioral residue* (Gosling et al., 2011). It has been found in personal websites (Marcus, Machilek, and Schütz, 2006) and in social networks such as Facebook (Kosinski, Stillwell, and Graepel, 2013). Given the rising amount of data available online, including very "intimate" information (preferences, tastes,

[1] http://ipip.ori.org/

purchase history, travel routes, amount of friends on social networks, political view, sexual orientation, religion etc.), the tendency is that such online behavioral residues will increase, which could aid in the task of inferring personality through this data. We can find several projects that use data available on social networks to infer personality implicitly to the user (without the need to fill in forms). For instance, Segalin et al., 2015 explored how self-assessed personality profiles can be inferred by looking at the Facebook profile pictures. They could demonstrate a positive and significant correlation between the profile pictures and Extroversion and Neuroticism, the two traits that obtained better scores.

A very interesting paper (Bachrach et al., 2012) have shown correlations between some traits of Personality measured by Big Five and some features from Facebook profiles. By using data from 180.000 users from the project MyPersonality, they have analyzed the following Facebook features: friends, groups, likes, photos, statuses and tags. Their results show some correlations, such as: positive correlation between Openness and number of users' likes; negative correlation between agreeableness and number of likes. Then, they have subset 5.000 individuals and have used their Facebook profiles to predict their personalty traits, using a multivariate linear regression with 10-fold cross-validation. The best accuracy for the prediction was found in Extraversion and Neuroticism traits, and the lower in Agreeableness.

Another study (Golbeck, Robles, and Turner, 2011) performed with 279 users from different countries took into account linguistic features (by applying analysis methods), in addition to personal information, number of friends and activities and preferences from Facebook. They had to complete a personality test (a 45-question version of the BFI) and then, they analyzed the correlations between those features and the five personality traits from Big Five, such as that conscientiousness is negatively correlated to swear words, but positively correlated with words surrounding social processes. After that, they have predicted the score of the personality traits by means of a regression analysis in Weka (M5'Rules and Gaussian Processes). Their results have shown strong correlations on Openness, Conscientiousness, Extraversion and Neuroticism.

In the same research line, also using the BFI, Gao et al., 2013 demonstrated that Big Five personality traits also can be extracted from other data sources and languages. By collecting status text from 1766 users of the Chinese micro-blog Sina (the most popular micro-blog in China, which is reported as having 300 million register users), having those users responded to the BFI questionnaire, the authors first performed a feature extraction of the statuses, and then, built a prediction model using Gaussian process, M5'Rules and Pace Regression. algorithms, finding significant correlations for Conscientiousness, Extroversion, and Openness to Experience.

Solinger et al., 2014 obtained positive prediction accuracy in all traits of Big Five, achieving 65% for Extraversion and Agreeableness, 55% for Neuroticism, 50% for Openness and 40% for Conscientiousness. They used data collected from Facebook like: profile bios, status updates, photos, and number of "Friends". To achieve these

results, they included additional cognitive psychology metrics in a multidisciplinary approach to increase Facebook personality prediction accuracies. This way, they demonstrated that the inclusion of additional personality dimensions, specifically the need for cognition and agreeableness, may be predicted with a high degree of accuracy.

Ortigosa, Carro, and Quiroga, 2014 have developed TP2010, a Facebook application, working with data of 20.988 users. Through it, the users filled in a questionnaire based on ZKPQ-50-cc, corresponding to the five traits of the Alternative Five Model. The authors were able to get the users' personality (extended with questions related to infrequency); then, they have tried to infer personality from interactions of the users on the social network. They performed a prediction with 3-class and 5-class model. The results show that the classifiers obtained a level of accuracy higher than 70% for all personality traits.

There are also works that predict the personality traits from Big Five with less known social networks, such as Foursquare. Chorley et al., 2013 created an online personality experiment examining the relationship between Big Five traits and the number and types of places visited by Foursquare users. The authors have found a positive correlation between Conscientiousness and the number of places visited, possibly due to the organized routine required to consistently check in at places. On the other hand, they also identified a negative correlation between Neuroticism and the number of places visited.

The variety of data sources and questionnaires used shows that, independently of the social network used, predictions of personality can be performed, and this has opening many opportunities to developers, researchers from Psychology and also from Computing.

Nevertheless, it is worth highlighting that, although the results of the predictions have been performed in an "implicit" way, that is, without the need of filling lengthy forms, the prediction of human personality is in an initial stage. As we can observe from the projects previously mentioned, the level of precision is not as high as the psychological tests. Yet, we believe that the improvement of the prediction models, the escalation of the amount and type of data available, apart from the internet of things, the enhancement of the predictions is a matter of time.

We are especially interested in measuring the curiosity of the users. We identified that this facet is highly correlated with extroversion and openness, according to previous studies (Kashdan et al., 2009). Curiosity is broadly defined as a desire for acquiring new knowledge and new sensory experience, and widely recognized as an important antecedent of exploration (Litman, Hutchins, and Russon, 2005; Spielberger and Starr, 1994). In the psychology domain, it is considered one of the fundamental strengths and personality traits studied by psychologists (Seligman, 2012; Berlyne, 1954; Reis and Judd, 2000).

The most reliable and practical questionnaire to measure the degree of curiosity of a person is the Curiosity and Exploration Inventory (CEI-II), consisting of only

10 items (Kashdan et al., 2009). When we compare the Big Five framework and the CEI-II, the curiosity is considered a lower-order, central facet of openness to experience Kashdan et al., 2009. Further, being in accordance with research Fredrickson, 1998 that defines curiosity as a discrete positive affect, we can observe the fig. 2.3, where the curiosity has a large positive correlation with openness (r = .51) and also with extroversion (r = .42), considered to be a reflection of positive affection and reward sensitivity. As another demonstration of links with psychological flexibility, curiosity was associated with more liberal political values (r = .26). To conclude, of the three studies, Kashdan et al., 2009 highlights the strongest correlate of the embracing sub-scale of the CEI-II was mindful awareness (low scores reflect greater mindful awareness) (r = .22).

	CEI-II-total	CEI-II-stretching	CEI-II-embracing
Mean (*SD*)	32.90 (7.48)	17.08 (3.88)	15.82 (4.37)
α	.85	.78	.75
Big Five			
Openness to experience	.51**	.50**	.43**
Conscientiousness	.20*	.31*	.07
Extraversion	.42**	.29*	.46**
Agreeableness	−.04	.03	−.09
Neuroticism	−.27*	−.30*	−.20*
Mindfulness	−.14	−.02	−.22*
Conservative political views	−.26*	−.28*	−.19*

Lower scores reflect greater mindfulness.
* $p < .05$.
** $p < .001$.

Fig. 2.3: Means, standard deviations, and Pearson correlations between Big Five and CEI-II

2.1 myPersonality

The project myPersonality[2] started in June 2007 as David's personal side project, between his undergraduate and postgraduate studies. Only after it became popular, he decided to consider its research possibilities. Nowadays, myPersonality data is being used by more than 200 researchers from 149 universities around the globe. To have the access granted to work with myPersonality data, it is needed to send them a request by e-mail, in which has to be explained the objective of the research, besides to mention which data one wants to have access. After approval, an account is created in the webpage of the project and the user can download the datasets.

The psychological profiles that myPersonality provides is vast and comprises of questionnaires and scales about satisfaction with life, music preferences or body consciousness, among others. However, we were especially interested in the Big

[2]http://mypersonality.org

Five personality scores, a 20 to 100-item IPIP proxy for Costa and McCrae's NEO-PI-R domains (Goldberg et al., 2006). With 8 length versions, the participants could either decide the length of the questionnaire they want to take, or they can take extra questions in blocks of 10 until they have finished all 100 items. Appendix A shows the psychometric test performed by the users.

Regarding the data from Facebook, there are available demographic details such as basic profile, race, location, education, political and religious views, but also "likes", status, photos and general activity from Facebook, besides Last.FM music listening data. The data that we were interested in will be described in Chapter 3.

Chapter 3

Methodology

The development methodology of this study was performed in the following steps:

- data collection from myPersonality project;
- development of a database and posterior pre-processing of data;
- correlation analysis and detection of the best predictors for extroversion and openness;
- generation of the prediction models;
- analysis of the models generated.

The following topics present the details of each step as well as the results obtained.

3.1 Data Source

As previously mentioned, the data used in this work were provided by the myPersonality project, which is derived by an homonym Facebook application that allowed users to take psychometric tests and let their profiles on that web be stored.

The access to the databases has to be granted by their founders, David Stillwell and Michal Kosinski; for that, it is needed to register as a collaborator, specifying which datasets will be used, the student bond which includes a supervisor, and also the main goal of the general project, besides accepting the terms of use.

The datasets available for use and analysis are related to psychological profiles, demographic data and the activity on Facebook of the users who took the personality tests. The access to some selected files was granted, more specifically to the following datasets:

- *demog.csv:* basic profile, containing personal (and anonymous) data such as age, gender, relationship status etc.;
- *fb_school.csv:* schools dictionary, with id and the name of the school;
- *user_education.csv:* diads between the user and the schools;

- *freq.csv:* Facebook activity, which contains number of likes, friends, among others;
- *big5_domains_item_level.csv:* results of the questionnaires Big5, specified by its length (20, 30, 40, 50, 60, 70, 80, 90 and 100 questions);
- *location.csv:* location of the users; and
- *hometown.csv:* their hometown.

After we had the access granted, each of the above mentioned datasets was downloaded in comma-separated values (*csv*) format to then make part of a database, which was built in MySQL 5.5, hosted in the cloud. The creation of this database was necessary to make possible the correlation and subsequent analysis among all the data. In figure 3.1 it is possible to see the relationship between the tables of the database.

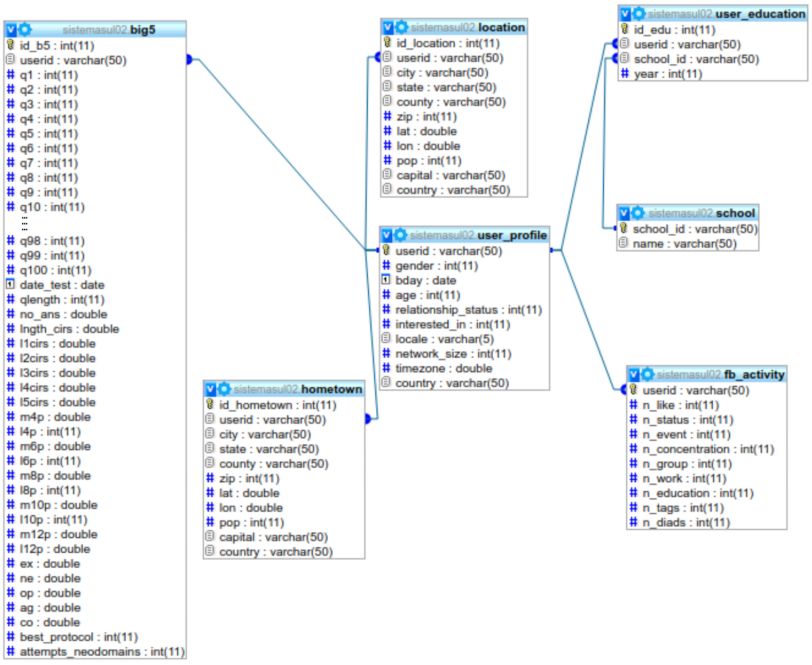

Fig. 3.1: Model of the entity-relationship diagram generated by MySQL

It is valuable to mention that some tables have more than one instance (tuple) per user. It can happen in the following tables: *big5*, since the user could take the personality test as many times as wanted; *location*; *hometown*; and *user_education*, since the user can have more than one study. That is why *userid* is not the Primary Key for those tables.

After the database was built, an SQL query was created in order to have it all together into a single *csv* file, for posterior analysis in different softwares. So, we selected the variables we thought it would be valuable for the analysis. A posteriori some of them could be discarded if they showed to be not relevant. As the goal of this work is to predict the curiosity of the users through their Facebook data,

we necessarily need information about at least their Facebook activity (fb_activity), results of their psychometric tests (big5) and, of course, general information about the users (user_profile). As we wanted only the best score of each user from *big5* table, the boolean column *best_protocol = 1*, indicating the best score for that user, was used as a filter when retrieving the results.

We performed previous queries in which we detected that several users had more than one hometown, location, or school; this way, when retrieving all this data altogether, their unique information (e.g. he psychometric test) appeared repeated. That is why we decided not to retrieve that information in the general query.

Then, we ran the SQL code (3.1) into the environment of the R studio tool, thus generating an R object that could be used in this environment and also easily exported in *csv* format, if needed. It have retrieved from our database 1.336.998 instances: this will be the size of the dataset we will use in this work.

Listing 3.1: SQL code to retrieve the wanted data from MySQL database

```
SELECT DISTINCT u.userid, u.gender, u.age, u.
    relationship_status, u.interested_in, u.network_size,
f.n_like, f.n_status, f.n_event, f.n_concentration, f.n_group
    , f.n_work, f.n_education, f.n_tags, f.n_diads,
b.q1, b.q3, b.q4, b.q7, b.q10, b.q14, b.q16, b.q18, b.q21, b.
    q23, b.q24, b.q29, b.q31, b.q33, b.q34, b.q39, b.q41, b.
    q43, b.q44, b.q49, b.q51, b.q53, b.q54, b.q59, b.q61, b.
    q63, b.q64, b.q69, b.q71, b.q73, b.q74, b.q79, b.q81, b.
    q83, b.q84, b.q89, b.q91, b.q93, b.q94, b.q99, b.ex, b.op,
     b.ne, b.ag, b.co, b.qlength, b.no_ans, b.best_protocol
FROM user_profile u
INNER JOIN fb_activity f
ON u.userid = f.userid
INNER JOIN big5 b
ON u.userid = b.userid
WHERE b.best_protocol = '1'
```

3.2 Preprocessing

After having performed the query to retrieve all the relevant data we want from the original dataset, we got 1.336.998 tuples as stated above. From the complete dataset, we have that 62.6% of the users are identified as female and the other 37.4%, as male. Additionally, 35.3% are single, and 29.9% did not mention their relationship status. Table 3.1 gives us an overview on the numeric columns of the dataset. The data taken from the psychometric tests are complete, which does not happen with the other variables. That is because only 30%-40% of users have given their consent

in sharing their Facebook profile data for research; in addition, not all users tell Facebook about their age, for instance (Kosinski et al., 2015).

Table 3.1: Summary of the complete dataset

Label	n	Mean	SD	Min	Max
age	795,740	26.65	9.37	1	112
network_size	1,122,512	227.73	245.32	1	5008
n_like	180,461	218.83	404.94	1	4819
n_status	115,802	142.38	162.54	1	2450
n_event	13,216	27.86	83.51	1	2840
n_concentration	41,798	1.60	.96	1	12
n_group	159,600	32.37	45.06	1	812
n_work	580,491	1.09	.47	1	44
n_education	650,748	1.25	.69	1	19
n_tags	644,529	28.08	91.90	1	3081
n_diads	409,515	66.92	161.28	2	5057
ex	1,336,998	3.60	.80	1	5
op	1,336,998	3.82	.67	1	5

Then, some preprocessing analysis were performed with the dataset by means of the tool R. First of all, the columns related to binary data, such as gender, or related to classification, such as relationship status, were converted to type factor.

Next, we divided the dataset by length of questionnaires. This way, we obtained 9 smaller datasets, from *big5_20* (with 724.115 observations) to *big5_100* (with 254.052 observations). Then, we removed all users who have not completed the entire questionnaire; this was recorded during the realization of the psychometric test by the column *no_ans*, which indicates the number of questions that the user has not filled in. We did this for all the 9 datasets. We then removed this column and also the column *qlength*, which denotes the size of the questionnaire taken for each user and was used here only for sub-setting the datasets.

3.3 Correlation Analysis

Having all datasets separated by length of questionnaire, a correlation analysis was performed for each of them, in order to find the individual values of correlation for the different questionnaire lengths and, then, analyze which of them obtains the best score. Table 3.2 shows the result we obtained for the main numeric variables of the datasets.

The correlations highlighted in bold denote the best score for each feature, where it is possible to see that the dataset that obtained the best score is the one of 90 questions (big5_90). Thus, we are going to use this dataset to perform the subsequent analyses. Tables 3.3 depicts the summary of this dataset. Its correlation matrix (lower triangle) is shown in 3.4, where it is possible to observe the meaningful correlation between extroversion and openness, which reinforces the strong correlation between

Table 3.2: Correlation analysis between all lengths of questionnaires

	20		30		40		50		60		70		80		90		100	
	ex	op	ex	op	ex	op	ex	op	ex	op	ex	op	ex	op	ex	op	ex	op
age	-.05	-.02	-.05	-.01	-.03	.01	-.04	.02	-.04	-.02	-.02	-.01	-.02	.02	.00	.01	**.01**	**.04**
network_size	.16	.03	.17	.05	.17	.02	.19	.02	.20	.04	.19	.05	.20	.01	**.21**	**.05**	.19	.01
n_like	.03	.03	.06	.04	.05	.03	.04	.00	.01	-.01	**.09**	.04	.06	**.05**	-.03	-.06	.02	.02
n_status	.09	.05	.13	**.07**	.08	.04	.09	.05	.07	.05	.11	.03	**.19**	.01	.16	.04	.08	.04
n_event	.05	.04	.07	.00	.09	.03	.04	.10	-.01	**.11**	.16	-.01	-.09	.03	**.19**	.03	.07	.01
n_concentration	.02	.06	.02	.08	**.06**	.09	.02	.08	.05	.06	.04	.13	.02	.07	**.06**	**.14**	.03	.09
n_group	.05	.05	.07	**.07**	.04	.06	**.08**	.04	**.08**	.02	.07	.06	-.01	.05	.03	.01	.05	.04
n_work	.01	.03	.00	.03	.00	.03	.01	.03	-.02	.03	-.02	**.04**	.00	**.04**	**.03**	.02	.00	.03
n_education	-.01	.03	-.02	.03	**.01**	.02	.00	**.05**	-.02	.03	-.01	.04	.00	.04	.00	.04	-.02	.04
n_tags	.03	.01	.03	.01	.04	.01	.03	.01	.02	.02	.02	**.03**	.00	.01	**.05**	**.03**	.02	.01
n_diads	.02	.00	.00	.01	.04	.01	.02	.03	.01	.01	.02	.04	.04	-.01	**.06**	**.05**	.02	.01

them and the curiosity, as explained in Chapter 2. We also find moderate correlations between other variables, such as number of dyads and tags (r = .47); number of friends and groups (r = .43); number of work and education (r = .38).

Table 3.3: Summary of dataset big5_90

Labels	n	Mean	SD	Min	Max
age	3,240	26.36	8.92	15	110
network_size	4,659	215.41	208.64	1	3097
n_like	686	209.85	371.58	1	3349
n_status	448	141.92	144.18	1	959
n_event	56	35.73	79.33	1	553
n_concentration	171	1.51	.71	1	4.00
n_group	621	34.93	47.33	1	350
n_work	2,535	1.09	.49	1	11.00
n_education	2,736	1.22	.64	1	6.00
n_tags	2,472	30.75	102.16	1	1372
n_diads	1,667	63.47	143.20	2	1456
ex	5,434	3.52	.76	1.17	5.00
op	5,434	3.91	.52	1.06	5.00

We also performed a transformation in the dataset, changing the variables *ex* and *op* to a factor(ordinal) type. For this, we grouped them into 3 categories: slightly, for *ex* and *op* levels from 1 to 2.33; moderately, from 2.34 to 3.67; and extremely, from 3.68 to 5. Our goal is to develop a classification model besides the regression models, as it will be described in Section 4.1. The data distribution, that is, the extent of users by level of *ex* and *op* (Table 3.5) is not uniform; as we can see, the class *slightly* is under represented in our set, which will be discussed later. Table 3.6 shows side by side the correlation values obtained with both approaches, where the best scores are highlighted in bold. We found that the correlation values obtained are slightly better when we deal with *ex* and *op* as factor variables, although not that much different.

[1]R could not compute the correlation for 5 variables, which are here represented by NA.

Table 3.4: Pairwise correlation analysis of dataset big5_90. Significant correlations at p < 0.05 level are shown in bold

	age	network_size	n_like	n_status	n_event	n_concentration	n_group	n_work	n_education	n_tags	n_diads	ex	op
age	1												
network_size	**-.11**	1											
n_like	**-.21**	**.14**	1										
n_status	**-.16**	**.24**	**.23**	1									
n_event	.09	.25	-.13	**.37**	1								
n_concentration	.09	.04	-.02	.15	-.05	1							
n_group	**-.10**	**.43**	**.36**	**.27**	.15	-.05	1						
n_work	-.02	**.06**	-.06	.13	-.01	.05	.02	1					
n_education	.03	**.08**	**-.17**	.01	-.09	**.21**	-.06	**.38**	1				
n_tags	-.02	**.22**	.03	**.32**	**.40**	-.03	**.18**	**.20**	**.33**	1			
n_diads	.04	**.33**	-.07	**.41**	.09	.08	**.17**	**.22**	**.35**	**.47**	1		
ex	.00	**.21**	-.03	**.16**	.19	.06	.03	.03	.00	**.05**	**.06**	1	
op	.01	**.05**	-.06	.04	.03	.14	.01	.02	**.04**	.03	.05	**.23**	1

Table 3.5: Distribution of the levels for *ex* and *op*

	Slightly	Moderately	Extremely
ex	440	2543	2451
op	19	1782	3633

Table 3.6: Comparison between the correlation matrix for *ex* and *op* as factor and numeric variables, respectively

	Factor		Numeric	
	ex	op	ex	op
age	.00	-.01	.00	**.01**
network_size	**.23**	**.07**	.21	.05
n_like	**-.02**	NA[1]	-.03	**-.06**
n_status	.12	NA	**.16**	**.04**
n_event	**.37**	NA	.19	**.03**
n_concentration	-.02	NA	.06	.14
n_group	.02	NA	**.03**	**.01**
n_work	.03	**.05**	.03	.02
n_education	**.01**	.04	.00	.04
n_tags	.05	.02	.05	**.03**
n_diads	.05	**.06**	**.06**	.05
ex	1	**.25**	1	.23

3.4 PCA - Principal Components Analysis

In order to better understand how is the data structured and simplify the model generation, we performed a PCA analysis for the dataset big5_90 with the help of FactoMineR package, taking *ex* and *op* as supplementary variables (as these are the ones we want to predict), and also normalizing the variables to have standard deviation equals to 1. Table 3.2 shows its summary. Fig. 3.3 depicts a matrix with eigenvalues. Fig. 3.4 depicts the correlation between variables and PCs.

```
Call:
PCA(X = big5_pca, scale.unit = TRUE, quanti.sup = 12:13, graph = TRUE)


Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7   Dim.8   Dim.9  Dim.10  Dim.11
Variance               3.188   1.319   1.157   0.962   0.930   0.746   0.676   0.583   0.553   0.455   0.431
% of var.             28.980  11.993  10.514   8.749   8.453   6.783   6.146   5.300   5.023   4.140   3.918
Cumulative % of var.  28.980  40.973  51.487  60.236  68.689  75.472  81.618  86.918  91.941  96.082 100.000

Individuals (the 10 first)
                    Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
1                |  1.746 | -0.936  0.005  0.287 | -1.214  0.021  0.483 | -0.479  0.004  0.075 |
2                |  1.634 | -1.023  0.006  0.392 | -1.049  0.015  0.412 | -0.165  0.000  0.010 |
3                |  1.847 | -1.161  0.008  0.395 | -1.141  0.018  0.381 | -0.429  0.003  0.054 |
4                |  1.755 | -1.157  0.008  0.435 | -1.102  0.017  0.394 | -0.341  0.002  0.038 |
5                |  4.897 |  2.893  0.048  0.349 |  1.127  0.018  0.053 | -0.197  0.001  0.002 |
6                |  8.173 |  4.681  0.127  0.328 |  2.798  0.109  0.117 |  0.747  0.009  0.008 |
7                |  1.290 | -0.485  0.001  0.141 |  0.523  0.004  0.164 |  0.042  0.000  0.001 |
8                |  1.528 | -0.022  0.000  0.000 |  0.152  0.000  0.010 | -0.206  0.001  0.018 |
9                |  5.135 |  3.636  0.076  0.501 |  2.112  0.062  0.169 | -0.194  0.001  0.001 |
10               |  5.751 |  2.598  0.039  0.204 | -2.504  0.087  0.190 |  0.894  0.013  0.024 |

Variables (the 10 first)
                   Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
age             |  0.287  2.588  0.083 |  0.577 25.248  0.333 |  0.172  2.567  0.030 |
network_size    |  0.360  4.061  0.129 |  0.083  0.519  0.007 |  0.269  6.258  0.072 |
n_like          |  0.486  7.418  0.236 | -0.329  8.214  0.108 |  0.585 29.585  0.342 |
n_status        |  0.689 14.912  0.475 | -0.180  2.468  0.033 | -0.123  1.307  0.015 |
n_event         |  0.315  3.118  0.099 | -0.255  4.931  0.065 | -0.600 31.127  0.360 |
n_concentration |  0.585 10.728  0.342 |  0.304  6.997  0.092 | -0.245  5.191  0.060 |
n_group         |  0.672 14.164  0.452 | -0.293  6.488  0.086 |  0.411 14.633  0.169 |
n_work          |  0.344  3.705  0.118 |  0.599 27.226  0.359 | -0.026  0.059  0.001 |
n_education     |  0.642 12.925  0.412 |  0.388 11.416  0.151 |  0.028  0.069  0.001 |
n_tags          |  0.666 13.926  0.444 | -0.234  4.133  0.055 | -0.242  5.071  0.059 |

Supplementary continuous variables
                   Dim.1   cos2    Dim.2   cos2    Dim.3   cos2
ex              |  0.039  0.001 | -0.004  0.000 |  0.004  0.000 |
op              |  0.054  0.003 |  0.007  0.000 | -0.003  0.000 |
```

Fig. 3.2: Summary of the PCA analysis

Fig. 3.5 displays the first two principal components, which explain approximately 40% of the variance in our dataset. The two supplementary variables are in blue. Fig. 3.6 shows only the five variables that have the highest contribution to explain the data.

The Proportion of Variance Explained (PVE) helps us to identify the variables that most influence the variance of each Principal Component. PVE is a positive quantity and the cumulative sum of PVEs will be 100%. A scree plot (Fig. 3.7), which displays the PVE on the vertical axis and the number of principal components found, could help deciding how many principal components should be chosen.

```
> pca.big5$eig
        eigenvalue percentage of variance cumulative percentage of variance
comp 1   3.1877540              28.979581                           28.97958
comp 2   1.3192429              11.993117                           40.97270
comp 3   1.1565594              10.514176                           51.48687
comp 4   0.9624050               8.749137                           60.23601
comp 5   0.9298113               8.452830                           68.68884
comp 6   0.7461293               6.782993                           75.47183
comp 7   0.6760869               6.146245                           81.61808
comp 8   0.5830316               5.300287                           86.91837
comp 9   0.5525288               5.022989                           91.94136
comp 10  0.4554484               4.140440                           96.08179
comp 11  0.4310026               3.918205                          100.00000
>
```

Fig. 3.3: matrix with eigenvalues

```
> pca.big5$var$coord
                    Dim.1       Dim.2       Dim.3         Dim.4       Dim.5
age             0.2872371  0.57713329  0.17230953  0.1577399786  0.46649340
network_size    0.3597807  0.08274182  0.26902770  0.7340259067 -0.39846980
n_like          0.4862633 -0.32918419  0.58495307 -0.1947680222  0.25527904
n_status        0.6894642 -0.18045312 -0.12296943 -0.1192525652  0.12997011
n_event         0.3152875 -0.25505005 -0.60000473  0.2921446031  0.43475066
n_concentration 0.5847890  0.30381064 -0.24503514 -0.4068387986 -0.29439036
n_group         0.6719455 -0.29256135  0.41139061 -0.0001397613  0.08895079
n_work          0.3436440  0.59931732 -0.02615184  0.1380974860  0.22815279
n_education     0.6418901  0.38807084  0.02830507 -0.2296524732 -0.21826854
n_tags          0.6662783 -0.23351157 -0.24217629  0.0260790824  0.02821658
n_diads         0.6301252 -0.17645624 -0.21857898  0.1523671559 -0.29543236
>
```

Fig. 3.4: Correlation between variables and PCs

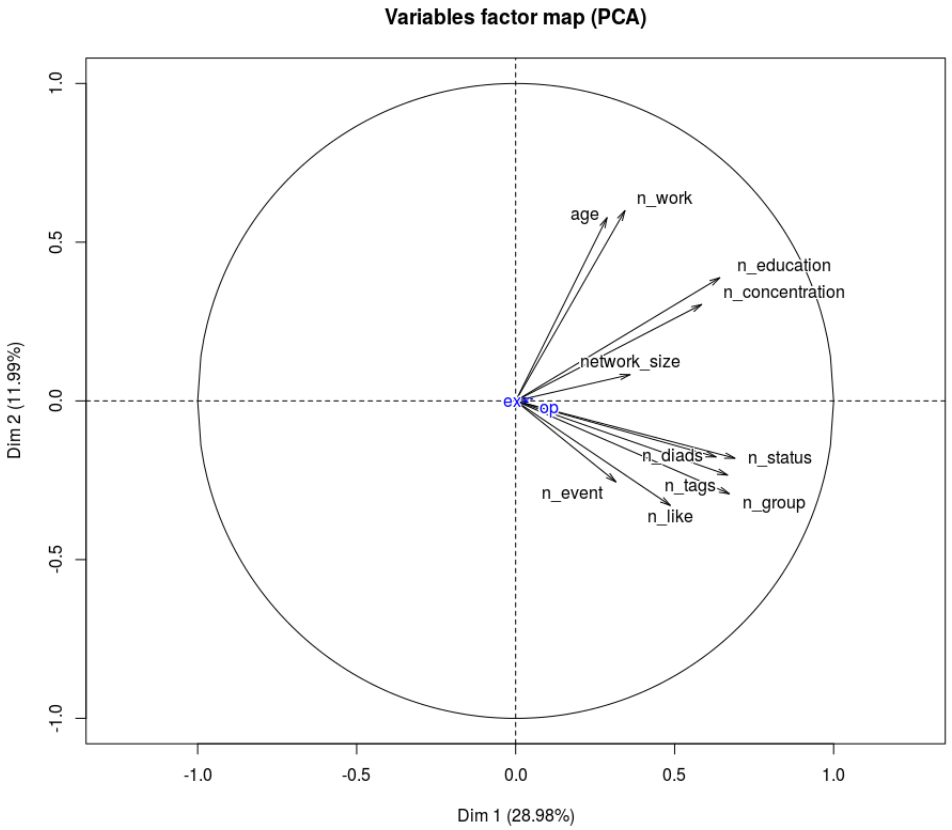

Fig. 3.5: Variables factor map

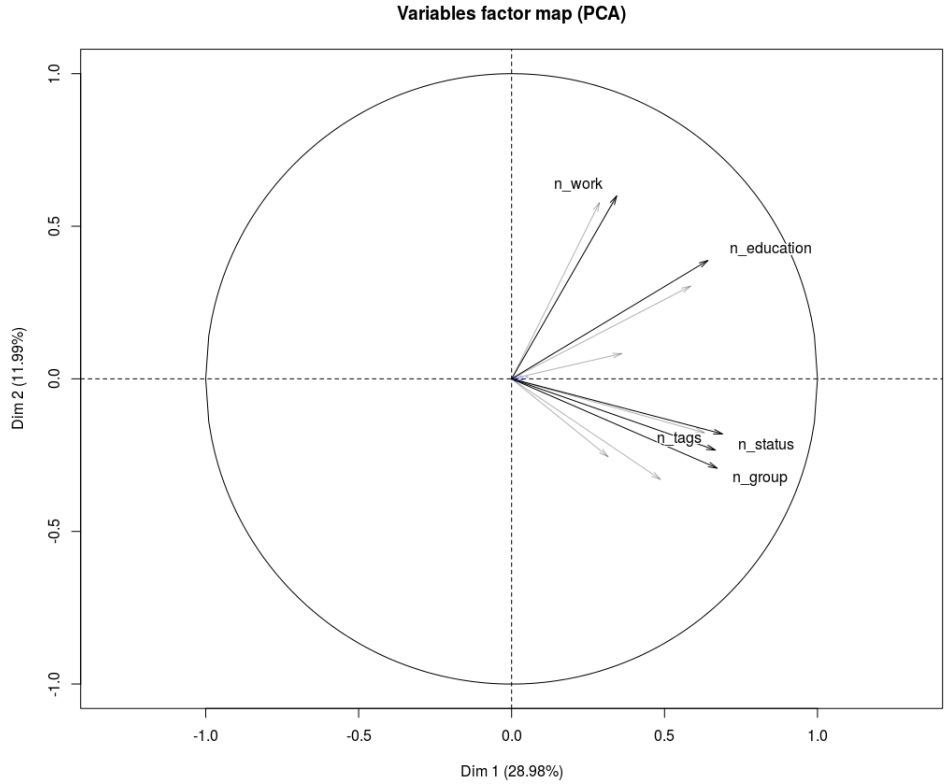

Fig. 3.6: Variables factor map depicting the 5 most important ones

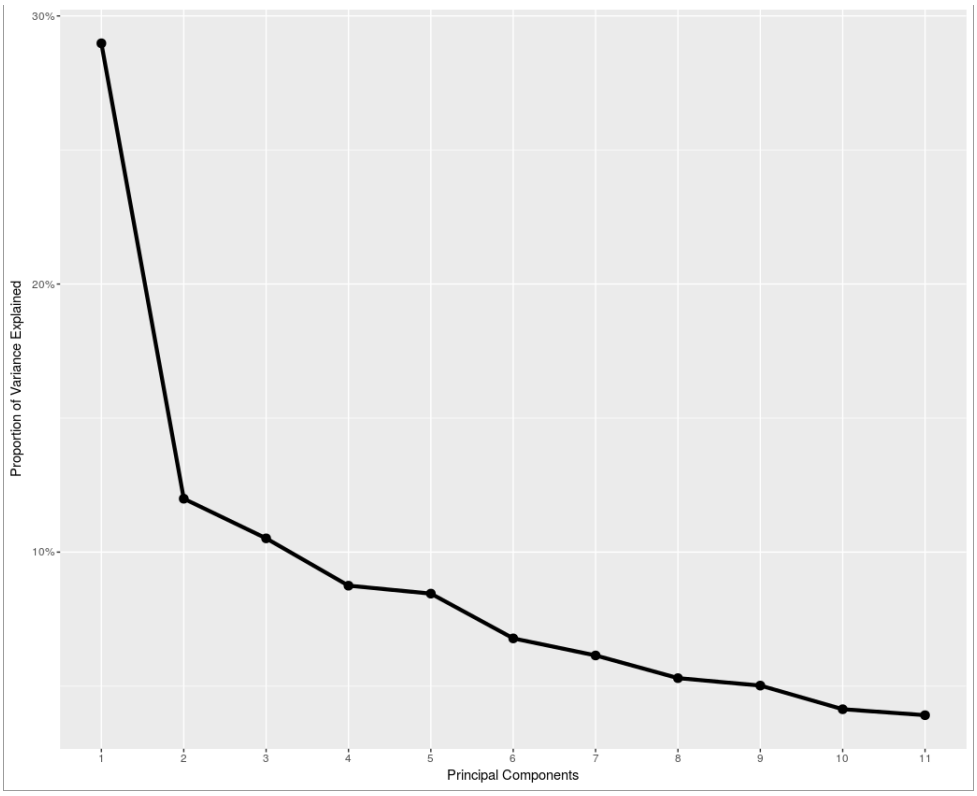

Fig. 3.7: Scree plot for Principal Components and Variance

3.5 Correlation for Population

We were also interested in finding if there is any correlation between the size of the city of the user (whether hometown or current location) and their degrees of extroversion and openness. To do so, we performed a specific query in order to get this information from the DB. As preprocessing, we removed NA values, besides the cases in which the population was 0. It can be explained by the fact that not always the users type a valid or correct city name. We also removed duplicated cities for the same user.

After the preprocessing, we binded this information with their degrees of extroversion and openness (from the previous complete dataset). Then, we got a data frame with 266.697 entries for hometown, and 198.320 entries for location. The results below (Tables 3.7 and 3.8) show that there is not a significant correlation between the size of the city of the user and his degree of extroversion and openness.

Table 3.7: Correlation between *ex*, *op*, and population of hometown

	p.home	ex	op
p.home	1		
ex	.01	1	
op	.03	.16	1

Table 3.8: Correlation between *ex*, *op*, and population of location

	p.loc	ex	op
p.loc	1		
ex	.00	1	
op	.00	.16	1

After those preliminary analysis, it was time to generate the models based on the correlations and previous analysis performed. We separately generated a linear regression for extroversion, then another one for openness.

Chapter 4

Models, Results and Discussion

4.1 Regression Model

In order to understand the causal relationship between the variables and perform further prediction of the levels of *ex* and *op*, we built some regression models, multiple and multivariate. We separate our dataset into train (75%) and test (25%) sets.

```
Call:
lm(formula = ex ~ age + network_size + n_like + n_status + n_event +
    n_concentration + n_group + n_work + n_education + n_tags +
    n_diads, data = big5_90_models)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3447 -0.5185  0.0733  0.5646  1.6522

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.45e+00   1.81e-02  190.66  < 2e-16 ***
age             -1.73e-03   7.27e-04   -2.38    0.017 *
network_size     6.91e-04   5.14e-05   13.43  < 2e-16 ***
n_like          -1.01e-04   7.99e-05   -1.26    0.207
n_status         8.96e-05   2.16e-04    0.42    0.678
n_event         -1.08e-03   1.22e-03   -0.88    0.377
n_concentration  4.54e-02   4.14e-02    1.10    0.273
n_group         -1.47e-03   6.68e-04   -2.20    0.028 *
n_work           2.69e-02   1.68e-02    1.60    0.109
n_education     -6.65e-02   1.64e-02   -4.05  5.3e-05 ***
n_tags           3.79e-04   1.72e-04    2.20    0.028 *
n_diads          2.24e-04   1.41e-04    1.58    0.114
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.742 on 5422 degrees of freedom
Multiple R-squared:  0.0394,    Adjusted R-squared:  0.0374
F-statistic: 20.2 on 11 and 5422 DF,  p-value: <2e-16
```

Fig. 4.1: Multiple regression for extroversion

Figs. 4.1 and 4.2 show the summary of the multiple linear regression model. First of all, it is possible to see that for extroversion there are more significant p-values ($p < .05$), for the predictor variables age, network_size, n_group, n_education, and n_tags. For openness, the predictor variables network_size and n_education contain significant p-values. The p-values of the model in general ($< 2 \times 10^{-16}$ for extroversion and .000458 for openness) would lead us to consider this model as statistically significant.

```
Call:
lm(formula = op ~ age + network_size + n_like + n_status + n_event +
    n_concentration + n_group + n_work + n_education + n_tags +
    n_diads, data = big5_90_models)

Residuals:
   Min     1Q Median     3Q    Max
-2.807 -0.327  0.023  0.379  1.121

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.88e+00   1.26e-02  308.04  < 2e-16 ***
age             -2.94e-04   5.07e-04   -0.58  0.56159
network_size     8.11e-05   3.58e-05    2.26  0.02372 *
n_like          -4.70e-05   5.57e-05   -0.84  0.39852
n_status         1.07e-04   1.50e-04    0.71  0.47706
n_event          3.51e-04   8.52e-04    0.41  0.67974
n_concentration -3.98e-04   2.89e-02   -0.01  0.98901
n_group          1.67e-04   4.65e-04    0.36  0.72035
n_work          -1.48e-02   1.17e-02   -1.27  0.20548
n_education      3.90e-02   1.15e-02    3.41  0.00065 ***
n_tags          -5.79e-05   1.20e-04   -0.48  0.62930
n_diads          1.13e-04   9.86e-05    1.14  0.25257
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.517 on 5422 degrees of freedom
Multiple R-squared:  0.00613,   Adjusted R-squared:  0.00411
F-statistic: 3.04 on 11 and 5422 DF,  p-value: 0.000458
```

Fig. 4.2: Multiple regression for openness

We then used the test set to predict the values for *ex* and *op* in order to validate our model. The results show a correlation accuracy between the actual and the predicted values of .16 for extroversion and .04 for openness. Thus, this model would need to be improved to be used for predict the levels of ex and op of the users at a significance level.

4.2 Principal Component Regression

In order to compare different models, we also performed a PCR - Principal Component Regression, a regression technique based on the principal component analysis. In order to corroborate our findings of the best dataset to use in the model generation, we performed the PCR on the biggest dataset (big5_20) and in the dataset which obtained the best correlation results as already related previously (big5_90). We set it to compute a ten-fold cross-validation error. As it can be seen in the Figs. 4.3 and 4.4, the last one obtained slightly better results if we look at the percentage of variance explained of each variable, thus justifying its use in the further steps.

```
> pcr_20 <- pcr(big5_20_pcr$ex~., data = big5_20_pcr, scale = TRUE, validation = "CV")
> summary(pcr_20)
Data:   X dimension: 878 12
        Y dimension: 878 1
Fit method: svdpc
Number of components considered: 12

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 comps  11 comps  12 comps
CV          0.8261   0.8077   0.8085   0.8051   0.8033   0.7992   0.7976   0.7980   0.7991   0.7999    0.8006    0.8012    0.8020
adjCV       0.8261   0.8076   0.8084   0.8050   0.8033   0.7984   0.7972   0.7976   0.7987   0.7994    0.7999    0.8005    0.8013

TRAINING: % variance explained
                1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 comps  11 comps  12 comps
X                 22.59   35.515   46.462   54.997   62.962   70.502   77.605   83.959   89.563    94.479    98.705   100.000
big5_20_pcr$ex     4.63    4.744    5.527    6.094    7.787    7.999    8.075    8.179    8.242     8.641     8.757     8.757
```

Fig. 4.3: Results of the PCR for extroversion with the dataset big5_20

```
> summary(pcr_ex)
Data:   X dimension: 5434 12
        Y dimension: 5434 1
Fit method: svdpc
Number of components considered: 12

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 comps  11 comps  12 comps
CV          0.7563   0.7558   0.7559   0.7562   0.7303   0.7265   0.7259   0.7255   0.7256   0.7259    0.7256    0.7237    0.7242
adjCV       0.7563   0.7558   0.7559   0.7563   0.7300   0.7263   0.7258   0.7254   0.7255   0.7258    0.7255    0.7234    0.7241

TRAINING: % variance explained
                  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 comps  11 comps  12 comps
X                 26.5993  37.5945  47.2331   55.702   63.707   71.321   77.531    83.16   88.014    92.618    96.408   100.000
big5_90_models$ex   0.1953   0.1954   0.1955    7.065    7.986    7.998    8.246     8.26    8.268     8.334     8.868     8.903
```

Fig. 4.4: Results of the PCR for extroversion with the dataset big5_90

Now using the dataset big5_90, the graph of the RMSEP (Fig. 4.5) shows that we do not get an error lower than 72%. Nevertheless, from the fourth component, the error significantly decreases, and the summary shows that with 5 components it is possible to explain 63.7% of all the variance, so we decided to select 5 components to build our model.

Listing 4.1: PCR code for extroversion

```
smp_size <- floor(0.75 * nrow(big5_90_models))
set.seed(3)
train_ind <- sample(seq_len(nrow(big5_90_models)), size = smp
    _size)
```

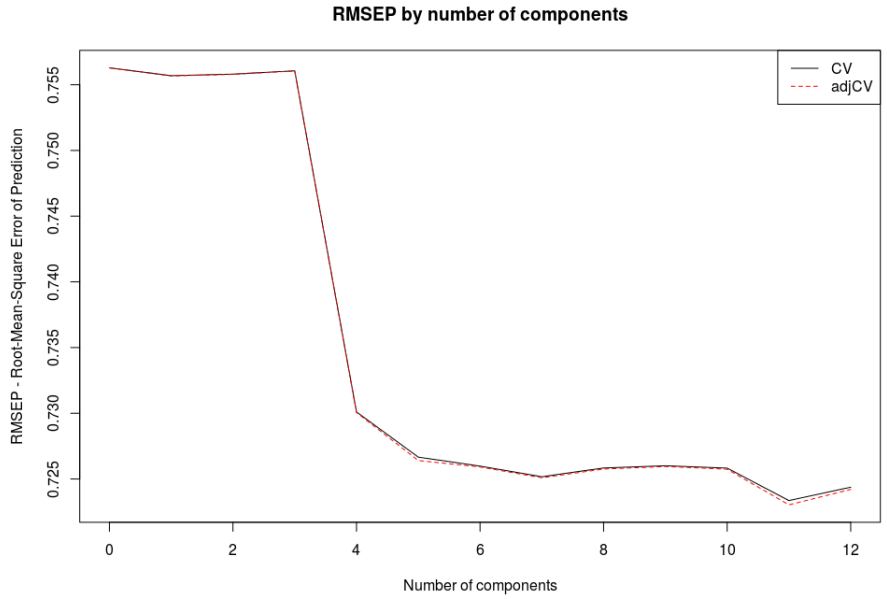

Fig. 4.5: Root-Mean-Square Error of Prediction by number of components

```
cols <- c(1:11, 13)
train <- big5_90_models[train_ind, ]
y_test_ex <- big5_90_models[-train_ind, 12]
# the other variables of the dataset
test_ex <- big5_90_models[-train_ind, cols]
set.seed(1)
model_pcr_ex <- pcr(train$ex~., data = train, scale = TRUE,
    validation = "CV")
pcr_pred_ex <- predict(model_pcr_ex, test_ex, ncomp = 5)
#MSE
mean((pcr_pred_ex - y_test_ex)^2)
#0.596
actxpred_ex <- cbind(y_test_ex, pcr_pred_ex)
cor_accur_pcrex <- cor(actxpred_ex)
# 0.0164
```

Listing 4.2: PCR code for openness

```
model_pcr_op <- pcr(train$op~., data = train, scale = TRUE,
    validation = "CV")
pcr_pred_op <- predict(model_pcr_op, test_op, ncomp = 5)
mean((pcr_pred_op - y_test_op)^2)
#0.275
actxpred_op <- cbind(y_test_op, pcr_pred_op)
cor_accur_pcrop <- cor(actxpred_op)
#0.135
```

First of all, we divided the dataset into training (75%) and test (25%) sets. Then, we performed again the PCR on the training set and evaluate its test set performance, defining the amount of components we wanted to use: 5. The MSE found was .60 for extroversion and .28 para openness. Then, we put together in a new dataframe the predicted and the actual values for both, and the correlation analysis found for extroversion and openness were .01 and .14, respectively.

4.3 Logistic Regression

The transformed dataset where the users are classified into slightly, moderately or extremely extroverted and open allowed us to perform a classification technique. With the help of the tool Weka, we performed a logistic regression (LR) configured with a 10-fold cross-validation. The performance of the models generated was evaluated by the analysis of correctly classified instances, Kappa statistic, and F-measure, shown in Table 4.3.

The coefficients (Tables 4.1 and 4.2) are weights that are applied to each variable, and indicate the probability that a new instance belongs to that class (the last class does not appear). As we can see, all the probabilities of the instances to belong to the classes moderately or extremely are very weak or inexistent. Table 4.4 shows the confusion matrix of the LR for extroversion and openness, respectively.

Table 4.1: Coefficients of the LR for extroversion

Variable	moderately	extremely
age	-.01	-.01
network_size	.00	.00
n_like	.00	.00
n_status	.00	.00
n_event	-.01	-.01
n_concentration	-.48	.02
n_group	.00	.00
n_work	3.69	3.82
n_education	.49	.52
n_tags	.01	.01

Table 4.2: Coefficients of the LR for openness

Variable	moderately	extremely
age	.01	.01
network_size	.00	.00
n_like	.00	.00
n_status	.00	.00
n_event	.01	.02
n_concentration	-1.17	-.83
n_group	-.01	-.01
n_work	-.02	.08
n_education	.26	.12
n_tags	.00	.00

As mentioned in 3.3, the dataset has not equally distributed classes for slightly, moderately and extremely extroversion and openness. This could directly influence the results of any model generated, so, albeit it would result in a small dataset, we generated another set with 440 instances for each of the three levels of extroversion, and only 19 instances for each in the case of openness, thus having 2 new

datasets with 1320 and 57, respectively. We wanted to check how the logistic regression would behave, however the results were not significant, so we decided to omit it in this work and discarded it.

Table 4.3: Logistic Regression (LR) results for extroversion and openness

	Correctly Classified Instances	Kappa Statistic	Average F-measure
LR for Extroversion	51.18%	0.09	0.47
LR for Openness	66.78%	-0.001	0.54

Table 4.4: LR confusion matrix for Extroversion (a) and Openness (b)

(a)

a	b	c	classified as
1914	627	2	a = moderately
1582	867	2	b = extremely
379	61	0	c = slightly

(b)

a	b	c	classified as
3629	4	0	a = moderately
1782	0	0	b = extremely
19	0	0	c = slightly

4.4 Comparison

As we saw in the previous chapter, we performed regression and classification analysis for extroversion and openness. When we look at the results obtained, the multiple regression models allow us to check the predictor variables that are more significant in each of them, where we did found some interesting similarities. For instance, the variables network_size is significant for both *ex* and *op*. However, the accuracy correlation for the regression models and the kappa for the logistic regression did not obtain satisfactory results.

In the Principal Component Analysis, we can see that the correlation accuracy for the model generated is not significant; it could be that the dataset is not distributed in some way that allows the combinations of variables into components to then have some applicable model.

Regarding the logistic regression performed through Weka, this classification analysis also showed not to be valuable for the dataset we are working on. Here, it is important to highlight that the considerable amount of NA values can have contributed to the low values found for accuracy; and when those values were omitted, the amount of the dataset was not considerably big to guarantee the confidence of the models

Similarly to the results found in Kosinski et al., 2014 with the same data source, extroversion obtained a better result than openness, which means that the Facebook profile of a person could better explain how extroverted the person is, but not his openness to new experiences. In addition, to predict *ex* and *op* we followed the same steps described in Menk and Sebastiá, 2016 to predict the curiosity, however we did

not achieve the same results, perhaps due to the lack of data, such as the places visited

4.5 Limitations

The first and most important limitation we found during the development of this work was the huge amount of null values. The users who participated on myPersonality project were offered to answer the 0-100 IPIP proxy questionnaire to measure the five traits of their personality. In exchange, they were asked to disclose the access and storage of their Facebook profiles, but it was not mandatory. As a result, around 40% of the more than 7 million participants have shared their online profiles only. It led us to find several NA values for their Facebook features, which have possibly influenced the weak results on the models we generated.

The data related to the level of education of the users may not reflect the reality. The variable *n_education* is a result of the sum of the schools that the user has inserted into their Facebook account. It is literally the number of education institutions listed on the individual's profile. For instance, John[1] might put his high school, undergraduate university, and Master university; in this case, he has 3 for n_education. Mary[1] might put her primary, secondary and undergraduate school. She also has 3 for n_education, however John has a postgraduate level, while she has a graduate level. That means it is not possible to determine the level of education (undergraduate, postgraduate etc.) of the person through this variable.

Another work motivated us to investigate about ways to predict the curiosity of the user (Menk and Sebastiá, 2016). One of the strong correlations they found was regarding the places visited, where the more places visited, the more curious tend is a person. We were also interested in check this correlation in our work, but we were not able to do so, because the project myPersonality does not have any information about the places the person visited.

[1]John and Mary are assumed names

Chapter 5

Conclusion and Future Work

5.1 Conclusions

Online data can tell a lot about ourselves. In recent days, the prediction of personality has been performed increasingly faster, easier and more accurately. Faster because the lengthy forms are no longer required; easier because the predictions can be performed by only sharing data from social networks; and more accurately due to the increasing extent of intimate data available on social networks.

In order to contribute in this scenario, this work focused on analyzing two of the Big Five traits of human personality, extroversion and openness to experience, which are strongly correlated to the curiosity. We aimed to predict those traits based on the Facebook profile of the users, by using Facebook and psychometric data from myPersonality project.

We first looked at the current digital world, which is in quickly expansion. Such huge amount of information available online opens a new challenge and also an opportunity in making sense of that data, and somehow use it for our well-being, though also for economic purposes.

Thus, its use for mapping personality's people is a reality. We reviewed the ways personality can be measured and, as we were interested in the curiosity, we wanted to be able to predict it through a different perspective: the traits extroversion and openness to experience, two traits of Big Five, since studies have shown strong correlations between them and the curiosity.

The project myPersonality was the way we find to develop our idea; more than 7 million people took an online questionnaire to measure their big five traits, and part of those users shared their Facebook profile information for research purposes.

So, we first investigate the correlations between extroversion and openness and the Facebook features, such as number of friends, number of education etc., besides to perform a principal component analysis to help the understanding the composition of the data. Then, we built some prediction models, by means of classification and regression techniques, with the help of the tools R and Weka, respectively. Although both models did not show significant results, the multiple regression allowed us to identify some significant variables, such as the number of friends (network_size) and the number of education. Similarly to other project that analyzed the

five traits, the models we generated shown a weak correlation and would need to be improved in order to predict extroversion and openness accurately and confidently.

5.2 Future Work

As future work, regarding the issue about the education level, mentioned in 4.5, the information related to the education could be deeper analyzed in order to classify the users into levels of education, which could be used for further predictions.

Another study that could be performed is regarding the correlation existent between curiosity, measured by the CEI-II questionnaire and extroversion and openness, measured by the Big5 questionnaires (IPIP proxy for Costa and McCrae's NEO-PI-R domains in the case of myPersonality project). It could be developed a formula in order to convert the two variables *ex* and *op* into a single variable, curiosity, taking into account the weighted correlation between each of them and the curiosity. It would be interesting to analyze the correlations with this only variable, but also it could ease the comparison between other questionnaires that measure this human facet, and the comparison between other projects that also aim to predict the curiosity.

People may have several profiles online, whether on Facebook or other social networks, such as Linkedin, Twitter etc. Future projects could inspect the possibility to get different profiles and merge them into a single online-user profile, which could improve future prediction models.

In this thesis, we focused on two of the five traits of personality. Future works could include the other traits, in order to search for further correlations. It also could be investigated other models of personality instead of the BigFive, and establish correlations between them.

Appendix A

Big Five Questionnaire (100 items) used in myPersonality

The Big Five Personality Questionnaire

Below, there are phrases describing people's behaviors. Please use the rating scale to describe how accurately each statement describes you. Describe yourself as you generally are now, not as you wish to be in the future. Rate yourself as you honestly see yourself, in relation to other people you know of the same sex as you are, and roughly your same age. If you are unsure of which response to choose (e.g. you act one way in a certain situation. and another way in a different situation), choose the response which feels most "natural" to you.

So that you can describe yourself in an honest manner, your answers to individual questions cannot be seen by others, only the overall calculation of your personality traits.

Phrase:	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
I am the life of the party	*1*	*2*	*3*	*4*	*5*
I leave things unfinished	*1*	*2*	*3*	*4*	*5*
I rarely look for a deeper meaning in things	*1*	*2*	*3*	*4*	*5*
I cut others to pieces	*1*	*2*	*3*	*4*	*5*
I am not easily frustrated	*1*	*2*	*3*	*4*	*5*
I keep in the background	*1*	*2*	*3*	*4*	*5*
I am always prepared	*1*	*2*	*3*	*4*	*5*
I have difficulty understanding abstract ideas	*1*	*2*	*3*	*4*	*5*
I suspect hidden motives in others	*1*	*2*	*3*	*4*	*5*
I seldom feel blue	*1*	*2*	*3*	*4*	*5*
I talk to a lot of different people at parties	*1*	*2*	*3*	*4*	*5*
I pay attention to details	*1*	*2*	*3*	*4*	*5*
I avoid philosophical discussions	*1*	*2*	*3*	*4*	*5*
I make people feel at ease	*1*	*2*	*3*	*4*	*5*
I am very pleased with myself	*1*	*2*	*3*	*4*	*5*
I know how to captivate people	*1*	*2*	*3*	*4*	*5*
I waste my time	*1*	*2*	*3*	*4*	*5*

Phrase:	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
I do not like art	*1*	*2*	*3*	*4*	*5*
I have a sharp tongue	*1*	*2*	*3*	*4*	*5*
I often feel blue	*1*	*2*	*3*	*4*	*5*
I make friends easily	*1*	*2*	*3*	*4*	*5*
I complete tasks successfully	*1*	*2*	*3*	*4*	*5*
I can say things beautifully	*1*	*2*	*3*	*4*	*5*
I treat all people equally	*1*	*2*	*3*	*4*	*5*
I rarely get irritated	*1*	*2*	*3*	*4*	*5*
I find it difficult to approach others	*1*	*2*	*3*	*4*	*5*
I do just enough work to get by	*1*	*2*	*3*	*4*	*5*
I enjoy thinking about things	*1*	*2*	*3*	*4*	*5*
I respect others	*1*	*2*	*3*	*4*	*5*
I dislike myself	*1*	*2*	*3*	*4*	*5*
I warm up quickly to others	*1*	*2*	*3*	*4*	*5*
I do things according to a plan	*1*	*2*	*3*	*4*	*5*
I am not interested in theoretical discussions	*1*	*2*	*3*	*4*	*5*
I am easy to satisfy	*1*	*2*	*3*	*4*	*5*
I feel comfortable with myself	*1*	*2*	*3*	*4*	*5*
I start conversations	*1*	*2*	*3*	*4*	*5*
I shirk my duties	*1*	*2*	*3*	*4*	*5*
I tend to vote for liberal political candidates	*1*	*2*	*3*	*4*	*5*
I get back at others	*1*	*2*	*3*	*4*	*5*
I am filled with doubts about things	*1*	*2*	*3*	*4*	*5*
I would describe my experiences as somewhat dull	*1*	*2*	*3*	*4*	*5*
I make a mess of things	*1*	*2*	*3*	*4*	*5*
I have a vivid imagination	*1*	*2*	*3*	*4*	*5*
I contradict others	*1*	*2*	*3*	*4*	*5*
I rarely lose my composure	*1*	*2*	*3*	*4*	*5*
I have little to say	*1*	*2*	*3*	*4*	*5*
I don't put my mind on the task at hand	*1*	*2*	*3*	*4*	*5*
I have a rich vocabulary	*1*	*2*	*3*	*4*	*5*
I accept people as they are	*1*	*2*	*3*	*4*	*5*
I am often down in the dumps	*1*	*2*	*3*	*4*	*5*
I retreat from others	*1*	*2*	*3*	*4*	*5*
I carry out my plans	*1*	*2*	*3*	*4*	*5*
I believe that too much tax money goes to support artists	*1*	*2*	*3*	*4*	*5*
I hold a grudge	*1*	*2*	*3*	*4*	*5*
I am relaxed most of the time	*1*	*2*	*3*	*4*	*5*
I don't talk a lot	*1*	*2*	*3*	*4*	*5*
I follow through with my plans	*1*	*2*	*3*	*4*	*5*
I do not like poetry	*1*	*2*	*3*	*4*	*5*
I insult people	*1*	*2*	*3*	*4*	*5*
I fear for the worst	*1*	*2*	*3*	*4*	*5*
I cheer people up	*1*	*2*	*3*	*4*	*5*
I get chores done right away	*1*	*2*	*3*	*4*	*5*
I enjoy hearing new ideas	*1*	*2*	*3*	*4*	*5*
I make demands on others	*1*	*2*	*3*	*4*	*5*

Phrase:	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
I panic easily	*1*	*2*	*3*	*4*	*5*
I keep others at a distance	*1*	*2*	*3*	*4*	*5*
I find it difficult to get down to work	*1*	*2*	*3*	*4*	*5*
I tend to vote for conservative political candidates	*1*	*2*	*3*	*4*	*5*
I believe that I am better than others	*1*	*2*	*3*	*4*	*5*
I feel threatened easily	*1*	*2*	*3*	*4*	*5*
I am hard to get to know	*1*	*2*	*3*	*4*	*5*
I need a push to get started	*1*	*2*	*3*	*4*	*5*
I do not enjoy going to art museums	*1*	*2*	*3*	*4*	*5*
I trust what people say	*1*	*2*	*3*	*4*	*5*
I am not easily bothered by things	*1*	*2*	*3*	*4*	*5*
I don't mind being the center of attention	*1*	*2*	*3*	*4*	*5*
I don't see things through	*1*	*2*	*3*	*4*	*5*
I am not interested in abstract ideas	*1*	*2*	*3*	*4*	*5*
I have a good word for everyone	*1*	*2*	*3*	*4*	*5*
I get stressed out easily	*1*	*2*	*3*	*4*	*5*
I don't like to draw attention to myself	*1*	*2*	*3*	*4*	*5*
I make plans and stick to them	*1*	*2*	*3*	*4*	*5*
I enjoy wild flights of fantasy	*1*	*2*	*3*	*4*	*5*
I am concerned about others	*1*	*2*	*3*	*4*	*5*
I have frequent mood swings	*1*	*2*	*3*	*4*	*5*
I am skilled in handling social situations	*1*	*2*	*3*	*4*	*5*
I finish what I start	*1*	*2*	*3*	*4*	*5*
I get excited by new ideas	*1*	*2*	*3*	*4*	*5*
I am out for my own personal gain	*1*	*2*	*3*	*4*	*5*
I seldom get mad	*1*	*2*	*3*	*4*	*5*
I feel comfortable around people	*1*	*2*	*3*	*4*	*5*
I am exacting in my work	*1*	*2*	*3*	*4*	*5*
I believe in the importance of art	*1*	*2*	*3*	*4*	*5*
I believe that others have good intentions	*1*	*2*	*3*	*4*	*5*
I remain calm under pressure	*1*	*2*	*3*	*4*	*5*
I avoid contact with others	*1*	*2*	*3*	*4*	*5*
I mess things up	*1*	*2*	*3*	*4*	*5*
I carry the conversation to a higher level	*1*	*2*	*3*	*4*	*5*
I sympathize with others' feelings	*1*	*2*	*3*	*4*	*5*
I worry about things	*1*	*2*	*3*	*4*	*5*

Bibliography

Allport, Floyd H and Gordon W Allport (1921). "Personality Traits: Their Classification and Measurement." In: *The Journal of Abnormal Psychology and Social Psychology* 16.1, p. 6.

American Psychological Association, APA (2017). *Personality*.

Bachrach, Yoram et al. (2012). "Personality and patterns of Facebook usage". In: *Proceedings of the 4th Annual ACM Web Science Conference*. ACM, pp. 24–32.

Back, Mitja D et al. (2010). "Facebook profiles reflect actual personality, not self-idealization". In: *Psychological science* 21.3, pp. 372–374.

Barrick, Murray R and Michael K Mount (1991). "The big five personality dimensions and job performance: a meta-analysis". In: *Personnel psychology* 44.1, pp. 1–26.

Berlyne, Daniel Ellis (1954). "A theory of human curiosity". In: *British Journal of Psychology* 45.3, pp. 180–191.

Block, Jack (1995). "A contrarian view of the five-factor approach to personality description." In: *Psychological bulletin* 117.2, p. 187.

Chorley, Martin J et al. (2013). "Visiting patterns and personality of foursquare users". In: *Cloud and Green Computing (CGC), 2013 Third International Conference on*. IEEE, pp. 271–276.

Costa, PT and RR McCrae (1985). "The NEO-PI personality inventory". In: *Odessa, FL: Psychological Assessment Resources*.

Costa Jr, Paul T and Robert R McCrae (1992). "Neo personality inventory–revised (neo-pi-r) and neo five-factor inventory (neo-ffi) professional manual". In: *Odessa, FL: Psychological Assessment Resources*.

Fredrickson, Barbara L (1998). "What good are positive emotions?" In: *Review of general psychology* 2.3, p. 300.

Gao, Rui et al. (2013). "Improving user profile with personality traits predicted from social media content". In: *Proceedings of the 7th ACM conference on recommender systems*. ACM, pp. 355–358.

Golbeck, Jennifer, Cristina Robles, and Karen Turner (2011). "Predicting personality with social media". In: *CHI'11 extended abstracts on human factors in computing systems*. ACM, pp. 253–262.

Goldberg, Lewis R (1992). "The development of markers for the Big-Five factor structure." In: *Psychological assessment* 4.1, p. 26.

— (1993). "The structure of phenotypic personality traits." In: *American psychologist* 48.1, p. 26.

Goldberg, Lewis R et al. (1999). "A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models". In: *Personality psychology in Europe* 7.1, pp. 7–28.

Goldberg, Lewis R et al. (2006). "The international personality item pool and the future of public-domain personality measures". In: *Journal of Research in personality* 40.1, pp. 84–96.

Gosling, Samuel D et al. (2011). "Manifestations of personality in online social networks: Self-reported Facebook-related behaviors and observable profile information". In: *Cyberpsychology, Behavior, and Social Networking* 14.9, pp. 483–488.

IDC, EMC Digital Universe with Research & Analysis by (2014). *Data Growth, Business Opportunities, and the IT Imperatives*.

IPIP (2017). *International Personality Item Pool*.

John, Oliver P and Sanjay Srivastava (1999). "The Big Five trait taxonomy: History, measurement, and theoretical perspectives". In: *Handbook of personality: Theory and research* 2.1999, pp. 102–138.

John Oliver P., Robins Richard W. Pervin Lawrence A. (2008). *Handbook of personality: Theory and research*. 3rd ed. Pearson. ISBN: 978-1-59385-836-0.

Kashdan, Todd B et al. (2009). "The curiosity and exploration inventory-II: Development, factor structure, and psychometrics". In: *Journal of research in personality* 43.6, pp. 987–998.

Kosinski, Michal, David Stillwell, and Thore Graepel (2013). "Private traits and attributes are predictable from digital records of human behavior". In: *Proceedings of the National Academy of Sciences* 110.15, pp. 5802–5805.

Kosinski, Michal et al. (2014). "Manifestations of user personality in website choice and behaviour on online social networks". In: *Machine learning* 95.3, pp. 357–380.

Kosinski, Michal et al. (2015). "Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines." In: *American Psychologist* 70.6, p. 543.

Litman, Jordan, Tiffany Hutchins, and Ryan Russon (2005). "Epistemic curiosity, feeling-of-knowing, and exploratory behaviour". In: *Cognition & Emotion* 19.4, pp. 559–582.

Marcus, Bernd, Franz Machilek, and Astrid Schütz (2006). "Personality in cyberspace: personal Web sites as media for personality expressions and impressions." In: *Journal of personality and social psychology* 90.6, p. 1014.

McCrae, Robert R and Paul T Costa (1987). "Validation of the five-factor model of personality across instruments and observers". In: *Journal of personality and social psychology* 52.1, p. 81.

Menk, Alan and Laura Sebastiá (2016). "Predicting the Human Curiosity from Users' Profiles on Facebook". In: *Proceedings of the 4th Spanish Conference on Information Retrieval*. ACM, p. 13.

Ortigosa, Alvaro, Rosa M Carro, and José Ignacio Quiroga (2014). "Predicting user personality by mining social interactions in Facebook". In: *Journal of computer and System Sciences* 80.1, pp. 57–71.

Phillips, Sarah (2007). *A brief history of Facebook*.

Poropat, Arthur E (2009). "A meta-analysis of the five-factor model of personality and academic performance". In: *Psychological bulletin* 135.2, p. 322.

Reis, Harry T and Charles M Judd (2000). *Handbook of research methods in social and personality psychology*. Cambridge University Press.

Segalin, Cristina et al. (2015). "What your Facebook Profile Picture Reveals about your Personality : A Feature-based Approach". In: *ACM - Association for Computing Machinery*.

Seligman, Martin EP (2012). *Positive psychology in practice*. John Wiley & Sons.

Solinger, Carrie et al. (2014). "Beyond Facebook Personality Prediction". In: *International Conference on Social Computing and Social Media*. Springer, pp. 486–493.

Spielberger, Charles D and Laura M Starr (1994). "Curiosity and exploratory behavior". In: *Motivation: Theory and research*, pp. 221–243.

Srivastava, Sanjay (2006). "Measuring the Big Five personality factors". In: *Retrieved October* 11, p. 2009.