Universidade do Minho
Escola de Engenharia
Departamento de Informática

Hugo Ribeiro

Applying Data Science and Machine Learning for Psycho-Demographic Profiling of Internet Users

November 2018

Universidade do Minho
Escola de Engenharia
Departamento de Informática

Hugo Ribeiro

Applying Data Science and Machine Learning for Psycho-Demographic Profiling of Internet Users

Master dissertation
Master Degree in Computer Science

Dissertation supervised by
Paulo Novais

November 2018

ACKNOWLEDGMENTS

In this initial section, I would like to thank all the persons that backed me up, that influenced me to work hard, to never give up during these hard and long five years since the journey began in 2013, and now concludes with this thesis.

I would like to start by thanking my parents, by allowing me to study for five years, supporting me financially and helping with everything I needed. My brother, whom is far away but still supported me and helped me every time something was not good throughout this process, as well.

To my friends, I would like to thank everyone that was with me since the beginning. Some of them deserve a honorable mention, they are, Luis Filipe Peixoto, Luís Araújo, José Sendão, Carlos Martins and José Pontes, whom were there for me when I needed the most, when I felt pressure, and kept me sane during some tough times.

To the ones I made during these five years, I have so much to thank for. Since teaching me new things, giving me insight on how to deal with problems and just by helping not giving up on this. Filipe Silva, Benjamim Fernandes, Rui Fernandes, Miguel Rodrigues, Nuno Fernandes, Pedro Lopes, Maria João Fernandes, Miguel Moreira, André Caetano, Rui Rua and Mario Abreu.

I am thankful to Carlos Pereira, João Faria, José Mário, Carlos Leite, Rui Gama and Nuno Guerra, as we did most of our works and projects together in some way during our graduation, for being good friends, good mentors and for making me company all this time.

I have also to thank to my friend from my hometown, who were always with me when I went to visit and never let me down, Fernando Oliveira, José Pedro Barbosa, Francisca Alves, Joel Barros, Francisco Barros, Francisco Velho, Diogo Carneiro, Nelson Barros, Ricardo Teixeira, Carlos Silva, Hélder Bruno Costa, Henrique Carvalho, Vitor Meireles, David Rodrigues, João Laranjeira and Bruno Vale.

Lastly, I am so thankful for the help and time disposed from my tutor, Marco Gomes, who guided me for the last year, taught me many things and helped building the path that culminates with the finishing of this project and I have to thank my supervisor, professor Paulo Novais for believing in me and giving me the pleasure to work with him.

For all of you, I have to say, thank you from the bottom of my heart, you are special.

DEDICATION

This thesis is dedicated to my parents, Paulino Ribeiro and Ana Paula Ribeiro, and to my brother, Rui Ribeiro.

ABSTRACT

There always have been a huge interest in working with public data from online social media users, with the exponential growth of social media usage, this interest and researches on the area keep increasing.

This thesis aims to address prediction and classification tasks on online social network data. The goal is to predict psycho-demographic - personality and demographic - traits by doing text emotion analysis on social networks as Twitter and Facebook. Our main motivation was to raise awareness to what can be done with users' social media or network information or usual behaviours on the web, such as from text analysis we can trace their personality, know their tastes, how they behave and so on, and to spread the emotion-text relation on social networks subject, because it only started to be studied recently and there's so much data and information to do it.

To perform these tasks mentioned above we carried an extensive review of literature of previous works to define the state-of-art of the project and to learn and identify work strategies. Almost all of the past researches, based their results on a vast sample of users and data, but because some frameworks and APIs were shutdown in recent years, such as MyPersonality from Facebook adding to some frameworks being paid for, resulted in a small sample of users' data to analyze in our thesis which can prejudice the results.

We start by gathering data from Twitter and Facebook with users consent. On Twitter we focused on tweets and retweets, on Facebook we focused on all of what the user typed by using the DataSelfie plugin that stored all that data on a server that can be retrieved later. Our next step was to find emotions on their text data with the help of a lexicon that categorized words by eight different emotions, two of them were put away because we focused only on the six major emotions - this is explained later - and we had to remove stopwords and apply stemming to all of the text and do a word-matching of every word of our data with every word from the lexicon. After this, we asked our participants to fulfill a "Big-Five" personality questionnaire and to provide us their age, so we added the Big-Five traits and age to each users individual dataset. We got their final versions, ready to apply machine-learning algorithms to

find correlations between emotions and personality or demographic attributes. We focused on practical and methodological aspects of the user attribute prediction task. We used many techniques and algorithms that we thought it were best fit for the data we had and for the goal that we had to achieve.

We gathered data in two datasets that we tested, one of them we called "Mixed Language Dataset", contains all text entries from each user, and the other "User Dataset", contains one entry per user after we analyze every text entry for all users in order to have a more general view on each one. For the first mentioned dataset we achieve best results with the decision trees algorithms, from 58% on the agreeableness trait, to 68% on the neuroticism trait. This dataset had a problem with the way data was spread, so it was impossible to predict age and gender with efficiency. As for the latter, regarding demographic characteristics all of the classifiers had a good classifying percentage, from K-nearest's 73% to Naive Bayes' 95%. The most solid classifier for personality traits was the one using the CART decision tree algorithm, it ranged from 50% on the openness trait to 76% on the agreeableness one. There were classifiers with terrible results, there were others that were a bit dull, and there were some that stood out as we stated above. We had a small sample, and that was a problem as it wasn't consistent or solid in terms of data value and that can change our results, we believe that our results would be way better if we applied the same mechanisms to a much bigger sample.

Concluding, we demonstrate how we can predict personality or demographic traits - BigFive traits, age or gender - from studying emotions in text. As stated above, we hope this thesis will alert people for what can be done with their online information, we only focus on psycho-demographic profiling, but there are many other things that can be done.

RESUMO

Sempre houve um enorme interesse em trabalhar com dados públicos dos utilizadores das redes sociais online, com o crescimento exponencial do uso das redes sociais, esse interesse e pesquisas na área continuam a crescer imenso.

Esta tese tem como objetivo abordar tarefas de previsão e classificação de dados de redes sociais online. O objetivo é prever traços psico-demográficos - de personalidade e demográficos - fazendo análises de emoções presentes no texto em redes sociais como Twitter e Facebook. A nossa principal motivação foi consciencializar os utilizadores sobre o que pode ser feito com as informações dos utilizadores ou com os seus comportamentos na web, por exemplo, com a análise de texto, podemos traçar a sua personalidade, conhecer os seus gostos, saber como eles se comportam e assim por diante, e para espalhar a relação texto-emoções nas redes sociais, porque só começou a ser estudado recentemente e há imensos dados e informações para isso.

Para realizar essas tarefas mencionadas acima, realizamos uma extensa revisão da literatura de trabalhos anteriores para definir o estado da arte do projeto, aprender e identificar estratégias de trabalho. Quase todas as pesquisas anteriores basearam os seus resultados numa vasta amostra de utilizadores e dados, mas como algumas frameworks e APIs foram encerradas nos últimos anos, como a MyPersonality do Facebook, adicionando a algumas frameworks que são pagas, o resultado foi que na nossa tese tivemos uma pequena amostra de dados de utilizadores para analisar o que pode prejudicar os resultados.

Começamos por recolher os dados do Twitter e do Facebook com o consentimento dos utilizadores. No Twitter, concentramo-nos nos tweets e retweets, no Facebook concentramo-nos em tudo o que o utilizador digitou usando o plugin DataSelfie que armazena todos os dados num servidor que podem ser recuperados mais tarde. O nosso passo seguinte foi encontrar emoções no texto digitado por cada utilizador com a ajuda de um léxico que categoriza palavras por oito emoções diferentes, duas dessas emoções foram descartadas, concentrando-nos apenas nas seis principais emoções - o processo é explicado mais tarde - e tivemos que remover as stopwords e aplicar stemming a todo o texto e fazer uma correspondência de cada palavra dos nossos dados com cada palavra do léxico. Depois disto, pedimos aos nossos participantes que preenchessem um questionário de personalidade "Big-Five" e nos dessem a conhecer a sua idade. Adicionamos as 5 características do "Big-Five" e a idade ao dataset individual de cada utilizador e obtivemos as suas versões finais, prontas para aplicar algoritmos de aprendizagem de máquina para encontrar correlações entre as emoções e personalidade ou atributos demográficos. Focamo-nos nos aspectos práticos e metodológicos da tarefa de predição e classificação de atributos do utilizador. Muitas técnicas e algoritmos foram utilizados, aqueles que consideramos mais adequados para os dados que tínhamos e o objetivo que tínhamos que alcançar.

Obtemos dados para dois datasets diferentes que testamos no final, um deles chamado de "Mixed Language Dataset", contém todas as entradas de texto de cada utilizador e o outro "User Dataset" contém uma entrada por utilizador após analisarmos todas as entradas de texto de todos eles para ter informação mais concisa geral sobre cada um. Para o primeiro conjunto de dados mencionado, os melhores resultados obtidos foram com os algoritmos de árvores de decisão, de 58% na característica de agreabilidade, para 68% na característica de neuroticismo. Este conjunto de dados tinha um problema com a forma como os dados estavam compostos no dataset, por isso foi impossível prever idade e género com eficiência. Quanto ao último dataset, em relação às características demográficas, todos os classificadores tiveram uma boa percentagem de classificação, de 73% de K-nearest para 95% com Naive Bayes. O classificador mais sólido para os traços de personalidade foi o que usou o algoritmo de árvore de decisão, CART, que varia apenas entre 50% no traço de "abertura a experiências" e 76% no de agreabilidade. Tivemos classificadores com resultados terríveis, houve outros que foram um pouco "aborrecidos", e houve alguns que se destacaram como afirmamos acima. A nossa amostra era consideravelmente pequena e isso foi um problema para nós, pois não era consistente ou sólido em termos de valores de dados e isso provavelmente alterou alguns dos nossos resultados, com uma amostra bem maior, mais profunda, acreditamos que aplicando os mesmos processos e mecanismos, teríamos resultados mais sólidos e mais consistentes.

Concluindo, demonstramos como é possível prever traços de personalidade ou demográficos - traços BigFive, idade ou género - a partir do estudo de emoções presentes em texto. Como foi dito acima, esperamos que esta tese permita que os utilizadores tenham mais consciência da importancia dos seus dados e do que conseguimos atingir com eles.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

The internet has become increasingly social in the last ten to fifteen years (Dabbish et al., 2012). It has also become one of the vast sources of information for users to access and retrieve data. As this happens, users should have the best access to what information they want and to what content they want to see or have in their browser.

With the increase of internet content and use, also the use of social networks and social media (SNs or SM) spiked, and they now represent a significant role in our daily life, however, though this extensive use of it seems inoffensive, sometimes it's not. For instance, there are a lot of web users, that use or not SNs or SMs, who don't know how to protect themselves over the information they leave behind as a trace, such as data from SNs or SMs, whether in the form of posts, tweets, likes, shares and more.

This is called *Digital Footprint* and represents something recurrent in the modern digital era. It happens when users leave, knowingly or not, unique traceable data from their actions, activities or communications, manifested on the internet or in digital devices.

This thesis aims to provide a perspective on data science, machine learning, predictive analysis, and digital identity. In our society, algorithms, big data and web content define our lives more and more. Therefore it is crucial - especially to those indifferent to this subject- to be aware of the power and influence their digital footprint has over each one of them.

A lot of people don't know the dangers of leaving personal information of any kind while on the internet, this is a concerning problem, once they feel they and their data are secured, but it can sometimes be traceable and used for a malicious end. With that referred, the main concern is to make it possible to do psycho-demographic profiling of internet users by applying data science and machine learning algorithms in their data. This will have a variety of potential results, such as, recommending activities to them taking into account their personality and tastes, strengthen security mechanisms by analyzing their habits on the internet, to help in the psychological area by aiding the study of users' personality by their online behaviour.

The method addressed to solve this problem is called *Data Science* using *Machine Learning* algorithms, which is the process of extracting knowledge from data, whether it is a huge or a small amount, as long as it's directed to a goal, or to answer a question.

The project will be developed following an action-research methodology, in which, faced with the presence of a given challenge, a solution hypothesis is stipulated. A compilation and organization of information relevant to the problem will be carried out, and a solution proposal for the problem will be designed.

The main problem will be contextualized and deeply defined in the following sections after addressing the investigation planning this thesis will follow.

1.1 CONTEXT AND PROBLEM DEFINITION

This thesis addresses a certain type of online information - user-generated content on social networks. Consider the case of social networks and social media, all of them are, mostly, tons of data from each user - what would *Facebook* be without any users? - so, social networks and social media users, as consumers of technological services and producers of data will be the main focus of this thesis as there is much personal and psychological user-generated information to scrape from them.

This data will be used to do profiling of individual users, and, by evaluating a series of combining aspects and data, we'll try to match each user with a specific psycho-demographic profile. The way we're going to approach this problem is by analyzing users' text data and try to identify the emotions presented in it. After identifying these emotions, we're going to try to find correlations between said emotions and psycho-demographics attributes, such as demographic characteristics, and personality traits through machine-learning classification and prediction algorithms that will be assessed later in this thesis.

1.2 MOTIVATION AND MAIN CONTRIBUTIONS

In this *world of technology*, as said before, many more people have easy access to the internet without knowing their hazards. One of the motivations of this thesis aims to assist, by reaching its goal of user profiling, to raise awareness to what can be done with their social media or network information or normal online behaviours.

The biggest motivation is to develop a system which can classify internet users in psycho-demographic profiles, that enables the possibility of having a series of results as explained in section 1.

Applying machine-learning for psycho-demographic profiling will have a significant impact on today's society, if done with a high accuracy rate, which will provide, in the future, better prepared Artificial Intelligence (AI) systems to deal and understand human behaviour. In other fields, to be able to help someone by recommending them some activity, such as a vacation plan or, to be able to protect users through the fortification of their online security by learning and analyzing their usual online behaviour and habits is a significant contribution and help to the dangers that exist on the internet. Bear in mind that there are a lot of challenges about trying to predict users personality and demographic profiles:

- Biased Data: Almost all of the researches on social media analytic rely on samples of social network data. Therefore, biases introduced by sampling and annotation procedures have to be addressed, and the representativeness and validity of the studies have to be verified (Tufekei, 204; Volkova, 2015).

- Users language: As it was addressed above the main focus passes by analyzing users text and language used. The problem here is that on social networks users, most of the time, use abbreviations or terms that are not incorrect grammatical norms, i.e. (gr8 = great, cya = see you, etc.). This is just noise on our datasets and can change the final results for worse.

- Changes in data content: Social media subjects and content changes quickly. With new trends, new subjects to discuss, statistic models and studies were done in the past may not correspond, to studies done sometime after (Dredze et al., 2010; Fromreid et al., 2014; Osborn et al., 2014).

- User behaviour and online time spent: Every user has different usage time habits. Some of them are constantly tweeting, or posting in social networks and have a lot of data to gather, and others may not do it for days or weeks. Models trained on data from active users don't usually generalize to the average users who have no or limited content (Sap et al., 2014; Vapnik et al., 1996a).

- Limited user sample: Many of the works done on this subject have samples of hundreds or even thousands of users, in our case, we have a great sample of tweets, but only 31 users, which will not produce the results a sample of thousands of users would do.

In this thesis, we will do experiments on different models and datasets to try to infer relations between emotions and psycho-demographic attributes (that will be explained later in the document).

1.3 CONCEPTUAL DEFINITIONS

This section will serve to review and define in full-detail the definitions of the main concepts that build this thesis.

1.3.1 *Digital Footprint*

This concept can be defined in various forms, per example, the Scientific Working Group on Digital Evidence, SWGDE (2000) stated that digital footprint could be described as "information of probative value stored or transmitted in digital form" from an individual user.

Another definition that came later in the decade was created in Madden et al. (2007) which referred that two different categories of the digital footprint can be distinguished: active and passive. An active digital footprint refers to personal data made accessible through the deliberate action of an individual, whereas a passive digital footprint is created by personal data made available online with no deliberate.

Although there are a lot of definitions and none of them is consensual, the chosen to be used as reference in this thesis is stated by Zezulka (2016), that refers both types (passive and active), and also refers the personal information that has a significant weight throughout the research. It follows as:

The concept of so-called *Digital Footprint* represents a phenomenon of the modern digital era. People who use digital services create, deliberately or unknowingly, a kind of digital imprint which contains sensitive personal information.

There's a famous saying that goes "there's no such thing as a free lunch" said by Friedman, contextualizing, this can be understood as the information that users give access to a service, whether online or offline, that generates knowledge about them or their personality, which, may or may not be used for a malicious end. In other words, *everything* we do while in the web has a price, whether it is our real-time location or the access we give to our intimate personal information, all of this increases the traces of data we leave behind, creating a bigger digital footprint of ourselves.

To access this information that will later be processed and analyzed we chose two of the most worldwide used social networks/media - *Facebook* and *Twitter* - which will later be *mined*. There are a lot of studies and researches on these social networks/media, which will be later addressed in section 1.6. The next subsections will describe each one of them succinctly.

Facebook

This is a popular, if not the most popular, social networking where users can post comments, share photographs and post links to news or other interesting content on the web, chat live, and watch a short-form video. You can even order food, do live streams, support charities on Facebook if that's what you want to do. Shared content can be made publicly accessible, or it can be shared only among a select group of friends or family, or with a single person. Per statista, a global statistic portal, in the second quarter of 2018, there were as much as 1400 million daily active users, making it the most used social network of the last few years.

Twitter

Twitter is a social networking website where users can post short messages, called tweets, to other users to see. It's members can see, retweet or reply to users' tweets on their feed by "following" them which is the corresponding to "add a friend" on Facebook. Twitter is considered by some to be a microblogging. People post there a lot of personal information and share a lot of emotions, which makes it great to perform some analysis on it. Compared to Facebook, per statista, it gets a bit behind regarding daily active users in 2018, with "only" 336 million users, which is still a considerable number of users per day.

1.3.2 *Psycho-demographic Profile*

As stated by Omelianenko (2017), the considered psycho-demographic profile comprises of psychometric scores based on the Big-Five model of personality and demographic scores such as age, gender or ethnicity.

Psycho-demographics can be defined as a quantitative methodology used to describe consumers/users on psychological attributes (Senise, 2007). It has been applied to the study of personality, values, opinions, attitudes, interests, and lifestyles (Wells, 1975). Psycho-demographics attributes are contrasted with demographic variables, and this is the main focus of this thesis, being able to find correlations between personality, demographic factors and emotions found on text, combining them into a "psycho-demographic profile" - used term in this paper.

There are two well-known personality characterization models. One of them presented by (Jung, 1971; Myers and Myers, 1980), says that there are 16 types of personalities which are a combination of four criteria that was chosen to characterize people by their general attitude. They are the following:

- Extroverted (E) vs. Introverted (I)
- Sensing (S) vs. Intuition (N)
- Thinking (T) vs. Feeling (F)
- Judging (J) vs. Perceiving (P)

The combination of these four traces is said to characterize a person's personality, p.e, ISTJ stands for introverted, sensing, thinking, judging.

The other and the considered one for this work comprises of psychometric scores based on a five-factor model of personality and demographic scores such as Age, Gender. The Big-Five model is named after Goldberg's early researches (Goldberg, 1981), this is a title chosen not to reflect their intrinsic greatness but to emphasize that each of these factors is extremely broad. (John and Srivastava, 1999; Goldberg, 1990). The so called big-five traits are *openness, conscientiousness, extroversion, agreeableness, and neuroticism*.

This model was first referred to in the past century, where many researchers studied it throughout the years, as it will be explained in section 1.6.2. We chose this one because as mentioned above the psycho-demographic profile uses psychometric scores based on this model.

Big 5 Traits	
Trait	Description
Openness to Experience	Artistic, Curious, Imaginative, Insightful, Original, Wide Interests
Conscientiousness	Anxious, Self-pitying, Tense, Touchy, Unstable, Worrying
Extroversion	Active, Assertive, Energetic, Enthusiastic, Outgoing, Talkative
Agreeableness	Appreciative, Forgiving, Generous, Kind, Sympathetic
Neuroticism	Anxious, Self-pitying, Tense, Touchy, Unstable, Worrying

Table 1.: Big-Five traits characterized with adjectives

To give a detailed explanation of this personality model, we will comply an own characterization of each trait.

Openness is a general appreciation for art, emotion,(Jacob B. Hirsh et al., 2009) imagination, curiosity, and variety of experience. Usually, when compared to people who score low on openness, they are more creative and more aware of their feelings.

Concerning the other Big Five factors, openness to experience is weakly related to neuroticism and extroversion and is mostly unrelated to agreeableness and conscientiousness (Ones et al., 1996).

Conscientiousness is a tendency to display self-discipline, act dutifully, and work for better and more achievements (Conrad and Patry, 2012). It is related to the way in which people control, regulate, and direct their impulses. High scores on conscientiousness indicate a preference for planned rather than spontaneous behavior. Also, an high-score in this trait usually relates to being a good student and having a better academic course (Noftle and Robins, 2007).

Conscientiousness was found to correlate somewhat negatively with neuroticism and somewhat positively with agreeableness but had no discernible relation to the other factors (Ones et al., 1996).

Extroversion is being energetic, active, enthusiastic and so on. People with high-scores on extroversion enjoy interacting with people and are often perceived as full of energy. Those high in extroversion are likely to value achievement and stimulation, and unlikely to value tradition or conformity (Roccas et al., 2002).

People with low scores on extroversion, also known as introverts, have lower social engagement and energy levels than extroverts, they tend to seem quiet, low-key, deliberate, and less involved in the social world. Their lack of social involvement should not be interpreted as shyness or depression; the introvert needs less stimulation than an extrovert and prefers to be alone (Krišto, 2012; Zafar et al., 2017).

When analyzed concerning the other Big Five factors, extroversion correlated weakly and negatively with neuroticism and loosely positively related to openness to experience (Ones et al., 1996).

The **Aggreableness** trait reflects the desire to fulfill social obligations or follow established norms, or it may spring from a genuine concern for the welfare of others. Whatever the motivation, it is rarely accompanied by cruelty, ruthlessness, or selfishness. They are generally considerate, appreciative, forgiving, generous, kind and sympathetic. Agreeable people also have an optimistic view of human nature (Barrick and Mount, 1991; Ones et al., 1996).

Antagonists, as people with a low score on this trait are called, care more about themselves and don't feel the need to get along with other people. This can cause them to be suspicious, unkind, and uncooperative. Agreeableness correlates weakly with extroversion and is somewhat negatively related to neuroticism and somewhat positively associated with conscientiousness (Ones et al., 1996).

Neuroticism is the tendency to experience negative emotions, such as anger, anxiety, or depression (Jeronimus et al., 2014). Moreover, individuals high in neuroticism tend to have more negative self-esteem and general self-efficacy, as well as the individual locus of control (Judge et al., 2002).

At the other end of the scale, individuals who score low in neuroticism are less easily upset and are less emotionally reactive. They have more capacity to be calm and to stay emotionally stable.

Neuroticism was found to correlate somewhat negatively with agreeableness and conscientiousness, in addition to a weak, negative relationship with extroversion and openness to experience (Ones et al., 1996).

1.3.3 *Big Data*

Big Data appeared as a concept not so long ago, because technology evolved exponentially in the last decades with tons of data being generated per minute everywhere on the web and other sources. It describes the storage and analysis of large and complex data sets using a series of appropriate techniques such as machine learning and others

(Ward and Barker, 2013). Without using the term *big data*, Laney (2001) proposed a definition encompassing the three V's, and they were, Volume, Velocity, and Variety. This was stated to remark upon the increasing size of data, the growth rate at which it is produced, and the increasing range of formats and representations employed (Ward and Barker, 2013). More recently this model went up to four V's as Veracity was added to it Djicks (2013); Turner et al. (2012) and others since it includes questions of trust and uncertainty with regards to data and the outcome of the analysis of that data.

1.3.4 *Data Science*

This has received widespread attention in academic and industrial circles (Zhu and Y, 2015). All of the studies in this area has composed various definitions on what it is and means. One of the salient revelations of today, with the vast and growing amount of data, is that domain knowledge and analysis cannot be separated Waller and Fawcett (2013). As it can be seen below, some researchers address knowledge and data extraction or analysis in their data science definition.

Loukides (2013) stated that data science should enable the creation of data products rather than working only as a simple application with data. Later Provost and Fawcett (2013) affirmed that one of the central concepts of data science is extraction knowledge from data to solve business problems. One other popular definition was designated by Dhar (2013) where he referred to data science as the study of the generalizable extraction of knowledge from data.

As it can be concluded, this is an ambiguous term, mainly because it is relatively new in the technology area and there are a lot of studies and papers on the subject, each one with their view on the matter.

This dissertation will focus on Dhar's definition stated in 2013, is the one which represents better the focus and objectives presented here.

Data scientists need in-depth domain knowledge and a broad set of analytic skills. Developing a comprehensive set of analytic skills and an in-depth domain knowledge requires consistent investments of time. This is explained on figure 3 below. (Waller and Fawcett, 2013)

After this detailed introduction on the subject and the most important concepts, the following section will explain the motivation of this work, and why it was chosen.

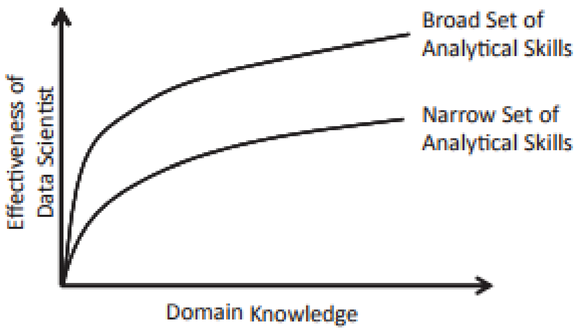


Figure 1.: Effectiveness, domain knowledge, and breadth of analytic skill set

1.4 OBJECTIVE QUESTION

This section will serve to provide an insight on what were the crucial points and questions we focused on to go through our work, and also what we thought were the steps we had to take to fulfill our objectives.

The two main questions this thesis aims to answer is "Is it possible to predict the psycho-demographic profiles by doing a constructive text emotion analysis of users data in social networks?" and "Is it possible to find emotions in the text to find correlations with users personalities?" The objectives that have to be full-filled for these questions, in a general way, are:

1. Define how to represent a psycho-demographic profile;
2. Design and codify the technological infrastructure needed to extract data, treat them and use them in the prediction process;
3. Test the developed system with data from real environments;
4. Analyze results and make assumptions and conclusion on what needs to be better, to optimize the system.

1.5 RESEARCH METHODOLOGY

This section will present an overview of our research methodology, detailing our strategy and the steps behind it to reach the results we wanted with high efficiency.

Similar research goals can be sought in entirely different ways depending on the accessibility and proximity of experts, synergies with ongoing research projects and so forth. Because of this research context, the chosen research strategy, represented in Figure 5, is based on the following activities:

1. Update the acquired knowledge by reviewing recent and state-of-the-art publications;
2. Design and develop the different parts of the proposed models enlarging the scope gradually in an iterative process;
3. Experiment on and evaluate the system;
4. Attend conferences and workshops to present partial results and to learn of existing state-of-the-art advancements;
5. Redesign the system with the feedback obtained from all the above means;
6. Develop and deploy the final system for context and behavioural analysis in real world-like scenarios to gather results;
7. Disseminate the obtained knowledge and experiences to the research community.

This research process is the action-research methodology composed of five different phases:

- Diagnosing: identifying the problem;
- Action planning: considering possible courses of action;
- Taking action: selecting a course of action;
- Evaluating: analyzing the consequences of the course of action;
- Specifying results: identifying general findings.

These phases were applied to all the outlined research activities with the aim of providing rigour, reflexive critiques, and constant challenges.

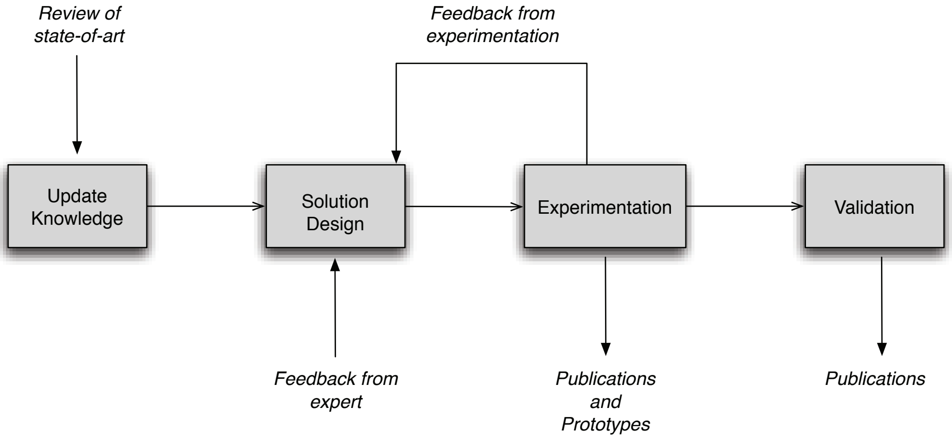


Figure 2.: Schematic view of the research process.

1.6 STATE OF ART

This section will present a complete review of related work on users' psycho-demographic profiling trough digital footprint whether by studying users' browsing history or their online social networks or media. To achieve this, an exhausting study and analysis on the area were done.

The significant growth of the internet in the last few decades generated a massive amount of data flow in itself. Nowadays it's easier to have internet access, every coffee shop, library, school or even public open spaces have free internet connection for everyone to use. Before smart-phones people rarely had access to the internet in their telephones, meaning they now spend even more time in the web and with more devices, generating that increased growth mentioned above. Last but not least, the social media "boom". Since Facebook was created in 2004, the number of people on the internet, exchanging, giving, or leaving data there took an exponential jump, and today Facebook has over 2 billion active users, as it can be quickly checked on Google, once it is in constant growth.

With this increase in technology, people are always online writing, liking or sharing information. The way they write and what they write will be of vital importance for our study once we will analyze it and try to find emotions or correlations between emotions found on text and psycho-demographic traits. This means that there is way more data than there was, ten years ago, but much less than there will be ten years from now. It also means that there are a lot of new studies and researches to be done, that will be addressed in the following section.

This is what matters in this thesis, to mine data from users on the web, mainly on social networks and social media to be able to extract knowledge and information about them, psychological and demographic wise.
The next subsections will address in detail some of the works and researches already done in the area as a review of literature. Every different topic or field of study will have it's separated section.

1.6.1 *Social Media and Networks*

Noticeably the internet has become one of the huge sources of information. Social media has facilitated easy communication and broadcasting of information, most typically, via commenting, sharing and publishing online content (Gauch et al., 2007). The relation of digital footprint and social networks or social media platforms has been studied for years and served to a lot of purposes for the research field. For instance, Twitter has been used in a large number of studies to predict demographic factors through many and diverse techniques and manners.

Twitter Researches

Twitter is one of the most studied online social networks because of it's users' profile structure and because of the tweets which are only text, which has a lot to offer to a machine learning process. Rao et al. (2010) focused on classifying users attributes such as age, gender, regional origin and political orientation using stacked-SMV based classification algorithms. Burger et al. (2011) did another research on this subject and trait using a wide variety of different classifier types, including *support vector machines, naive bayes, and balanced Winnow2*, based on caseful word unigram features from tweet texts to predict users' gender.

Other explored demographic characteristic is age, Sloan et al. (2015) used pattern-matching techniques to extract age and occupation-related information from tweets, in this tone Volkova et al. (2016) used logistic regression models learned using binary word unigram features to predict age and other psycho-demographic traits. Twitter served as a base for Coppersmith et al. (2014) research which quantified mental health signs in tweets, using LIWC (Linguistic Inquiry and Word Count), a tool developed and studied in Pennebaker et al. (2007), LMs (Language Models) and Pattern of Life Analytics to have multiple results on the subject.

Political references have a considerable weight in today's society, and it is also a research target, Conover et al. (2011b,a) applied SMVs and LSA (Latent Semantic Analysis) to predict political alignment on twitter.

Lastly, in the demographic category, Culotta et al. (2015) fit a regression model to predict demographics such as ethnicity and gender.

Psychology also has its studies and researches based on Twitter's users. Quercia et al. (2011) used an algorithm that generates model trees with linear models on the leaves using the M5' rules to predict personality traits. Also, Volkova et al. (2016), already mentioned above, was able to associate personality traits with Twitter profiles and users' profile information as well.

Golbeck et al. (2011) uses *Twitter* to predict users personality and fit it on the Big-Five model by analyzing users' profile and style of writing and present emotions on their text, by applying machine-learning algorithms in the *Weka* interface, having prediction results around 50-60%. This is pretty close to what we're trying to achieve and do on this study.

Facebook Researches

As for the world's largest online social network, Facebook, there is also a significant number of papers and researches about users information.

For example, Kosinski et al. (2013b) analyzed users likes on *Facebook* and after using a linear logistic regression model, they were able to predict psycho-demographic characteristics. Apart from likes, there are way more works done on *Facebook* profile information. Kosinski et al. (2012) extracted users' data from the Facebook application, MyPersonality, and show how multivariate regression allows prediction of the personality traits of an individual user given their Facebook profile. A little different but also related with personality traits, Schmit (2012) conducted a textual analysis on 16 different user profiles and asked each user to take a Myers-Briggs Type Indicator Test, there is a reference to this test in section 1.2.2, and then analyzed participants traits in relation with their profiles. In the same spirit, Gosling (2007) examined impressions based on 133 Facebook profiles, comparing them with how the targets see themselves and are seen by close acquaintances, targets, informants, and observers made their ratings on the Ten Item Personality Inventory (TIPI) (see, (Gosling et al., 2003)) which measures the Big Five personality dimensions.

Kosinski et al. (2013a) also studied demographic traits over Facebook profiles, they trained predictive models based on the publicly available myPersonality datasets containing users' Facebook likes, to predict and classify users age and gender.

Other Social Media

There are cases of data mining and analysis on other online social networks or media as there is on *Flickr*, where Onder et al. (2014) performed a polynomial regression analysis and was able to provide a representation of tourist numbers in Austria through users data. Later, Girardin et al. (2014), presented tourism statistics from the geolocation of users posts on this same network.
MySpace, once was a hot and trending social network, but since Facebook appeared it soon had a massive decline, it still counts with an enormous number of active users and so, *MySpace* was also used to extract information on people's personality traits by their digital footprint, see Chang et al. (2010) that used a Bayesian approach to estimating the distribution of ethnicity of a population given only their names. Some of these machine learning methods mentioned in this section will be explored in section 1.6.3.

1.6.2 *Psycho-Demographic Content*

In the *internet era* it's way easier to get more and correct information on demographics, whether by online surveys with millions of participants or because of social media and networks platforms where users give their personal information for free. This means that now there are a large number of studies done on the subject as mentioned above, but these next ones are just about demographics itself, not necessarily related to web content.

Referring to the psychological side of a psycho-demographic profile, this is a trait that defines a person, their behaviour, tastes, way of being, etc., and there have been made a lot of studies on how to divide people in to psychological categories in order to know more about the human mind, and human society itself. About psychological and demographic subjects, or as it's called in this paper, psycho-demographic matters, there is a lot of work and researches were already done, mainly in the past century. Because it was never consensual, there is a considerable number of studies on the Big-Five factor model of personality, see Tupes and Christal (1961); Goldberg (1998);

Bouchard (1994); Goldberg (1993a,b); Budaev (1998). In the 20th century Wells (1975) worked on demographics and researched the matter.

More recently there's been researches on how demographics correlate with lexical variation methods, as it was done by O'Connor et al. (2011) or Nguyen et al. (2011).

Most of the these mentioned researches didn't use internet or artificial intelligence to find behaviours or patterns in their experiments, although, recently Novais et al. (2015), were able to find correlations between mouse movement, during an online exam, and stress. Another research related with stress was done by Gomes et al. (2014b) who studied the effects of stress on negotiation behavior.

1.6.3 *Machine-Learning Techniques*

Machine learning is the base of this thesis; it also is a data analysis method that automatizes the analytic model development. By using iterative algorithms which learn from the given data, machine learning allows computers to find hidden insights without being specifically programmed to do so, as stated by Samuel (1959), he who first coined the term "machine learning".

There are two main machine learning algorithm groups, the supervised and unsupervised ones. Every instance of the datasets used by machine learning algorithms is represented using the same set of features. These can be continuous, categorical or binary. If instances are given with no labels, then it's called supervised, else if instances are unlabeled, it's called unsupervised (Kotsiantis, 2007).

Reinforcement Learning exists as well but is not used as much as the two mentioned above. It can be defined by the problem faced by an *agent* that has to learn correct behaviours through trial and error interactions with the environment surrounding it (Kaelbing et al., 1996).

In the next subsections, some of the most used machine learning algorithms and techniques, that were used - and will be used in our study - in some of the papers and studies cited in section 2.2 will be given a more detailed perspective.

Support Vector Machines

SVMs were created and first invented by Vapnik et al. (1992), and the current standard incarnation was proposed by Vapnik and Cortes (1995), in the time known by *Support*

Vector Networks. They were developed to solve the classification problem, but recently they have been extended to the domain of regression problems (Vapnik et al., 1996b).

Although it's always ambiguous, SVMs are typically used to describe classification with support vector methods; they can also be used for regression, Vapnik et al. (1996a) proposed Support Vector Regression as the correct term for it (Gunn, 1998).

Support Vector Classification: Classification is a fundamental issue in machine learning and data mining. In classification, the goal of a learning algorithm is to construct a classifier given a set of training examples with class labels. SVMs are based on the idea of finding a hyper-plane that best divides a dataset into two classes, as we can see in figure 5.

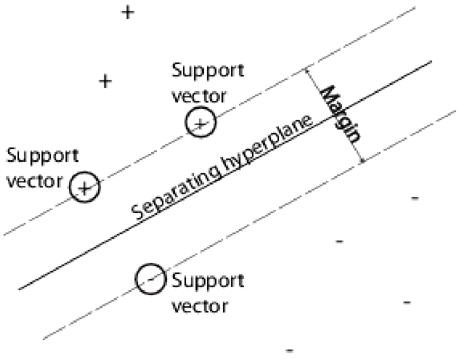


Figure 3.: hyperplane dividing the two classes - from mathworks website

Support vectors are the data points nearest to the hyperplane, the points of a dataset that, if removed, would create a different hyperplane in position terms. For a classification task with only two features (like the image above), you can think of a hyperplane as a line that linearly separates and classifies a set of data.

To be considered a good classifier, the data points need to lie the furthest possible to the hyperplane while being on the right side of the classification, thus giving more distance between the data points of different class labels of the classification, meaning it's a proper classification.

The margin that can be seen in the figure is the distance between the nearest data point from each class (Zhang, 2004). The objective is to choose a hyperplane with the most significant possible margin between the hyperplane and any data point in the training set, making possible the better classification of a new case.

Support Regression Machines: In regression problems, the goal is trying to predict continuous values as the output which differs from classification, where the output is binary whether as a category or class.

As in classification, SRMs contain all the main features which characterize a maximum margin algorithm: linear learning machine uses a non-linear function to map it into high dimensional kernel-induced feature space. As well as with the classification approach there is a motivation to optimize the generalization bounds given for regression. They rely on defining the loss function that ignores errors, which are situated within a certain distance of the true value, usually designated by ε-intensive loss function, as figure 6 shows. In a practical way, instead of attempting to classify new unseen variables x′ into one of two categories $y' = \pm 1$, it's now possible to predict a real-valued output for y′ so that the training data is of the form (Fletcher, 2009):

$$\{\mathbf{x}_i, y_i\} \text{ where } i = 1 \dots L,\ y_i \in \Re,\ \mathbf{x} \in \Re^D$$

$$y_i = \mathbf{w} \cdot \mathbf{x}_i + b$$

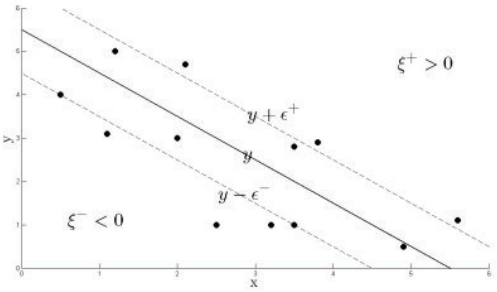


Figure 4.: Simple Regression with ε-insensitive tube

Using the *intensive* loss function ensures the existence of the global minimum and at the same time optimization of reliable generalization bound.

Bayesian Approach

Naive Bayes is one of the most efficient and effective inductive learning algorithms for machine learning and data mining (Zhang, 2004).

We can look at the Bayesian approach as a probability, except it has a prior *belief* that goes on to be updated following the Bayes rules. In machine learning the goal is to select the best hypothesis (h) given data (d). If it's a classification problem, the hypothesis (h) may be the class to assign for a new data instance (d) else, in a regression problem, the hypothesis (h) is a continuous value.

One of the easiest ways of selecting the most probable hypothesis given the data gathered is to use previous knowledge about the problem. Bayes' theorem provides a way that makes it possible to calculate the probability of a hypothesis given prior knowledge.

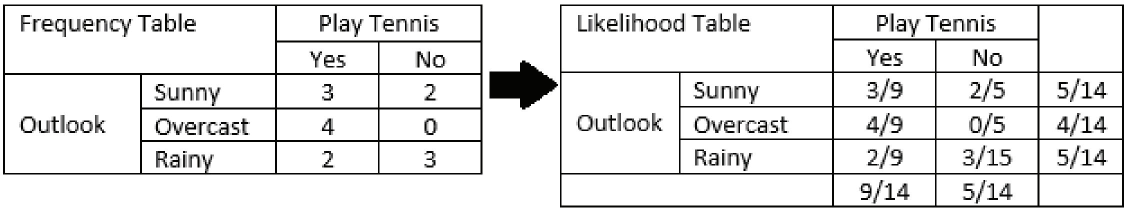
Bayes' Theorem is stated as:

$$P(h|d) = \frac{(P(d|h) * P(h)}{P(d)}$$

Where:

- $P(h|d)$ is the probability of hypothesis h given the data d. This is called the posterior probability.
- $P(d|h)$ is the probability of data d given that the hypothesis h was true.
- $P(h)$ is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h.
- $P(d)$ is the probability of the data (regardless of the hypothesis).

An example is illustrated in figure 7 where:



Frequency Table		Play Tennis	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

Likelihood Table		Play Tennis		
		Yes	No	
Outlook	Sunny	3/9	2/5	5/14
	Overcast	4/9	0/5	4/14
	Rainy	2/9	3/15	5/14
		9/14	5/14	

Figure 5.: Frequency and Likelihood Tables from an Bayes Approach example

$P(d|h) = P(Sunny|Yes) = 3/9 = 0.33$

$P(d) = P(Yes) = 9/14 = \ 0.64$

And the posterior probability that we want is: $P(h|d) = P(Yes|Sunny) = \frac{0.33 * 0.64}{0.36} = 0.60$

With this it is known that the probability to play when it's a sunny weather is 60% with the data from this example.

The tautological Bayesian Machine Learning algorithm is the Naive Bayes classifier, which is based on the so-called Bayesian theorem and is particularly suited when the dimension of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods.

Decision Trees

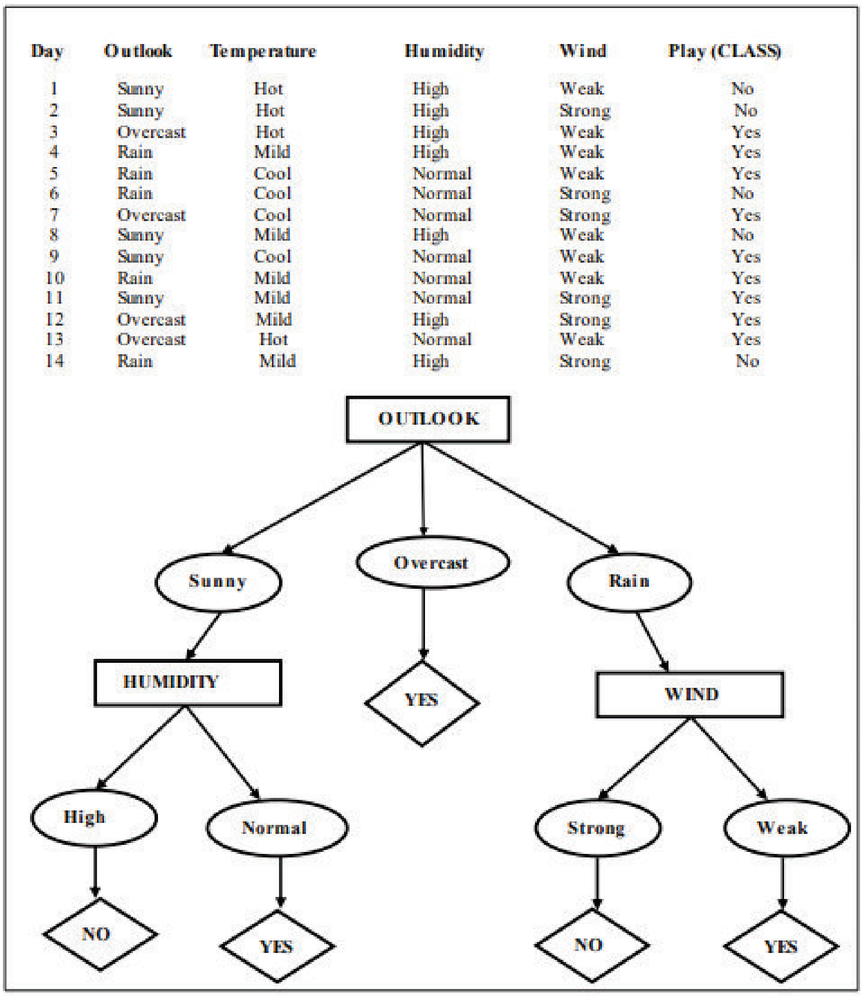
Decision trees(DT) are the most powerful approaches to knowledge discovery and data mining. It includes the technology of research large and complex bulk of data to discover useful patterns (Sharma et al., 2013).

A decision tree is a decision support system that uses a tree-like graph decision and their possible after-effect, including chance event results, resource costs, and utility. It is used to learn a classification function which concludes the value of a dependent attribute(variable) given the values of the independent(input) attributes(variables). This verifies a problem known as a supervised classification because the dependent attribute and the counting of classes(values) are given (Korting, 2014).

There are many DTs algorithms to use, and we used two of them in our research, J48, and CART because they were the only ones that fit our work better which will be explained in detail below.

C4.5 algorithm: C4.5 or as it's represented in Weka, J48, is an evolution of ID3, presented by Quinlan (1993). The C4.5 algorithm generates a decision tree for the given data by recursively splitting that data. The decision tree grows using depth-first strategy. The C4.5 algorithm considers all the possible tests that can split the data and selects a test that gives the best information gain (i.e. highest gain ratio). For each discrete attribute, one test is used to produce many outcomes as the number of distinct values of the attribute. For each continuous attribute, the data is sorted, and the entropy gain is calculated based on binary cuts on each distinct value in one scan of the sorted data (Singh and Gupta, 2014). This process is repeated for all continuous attributes.

The C4.5 algorithm allows pruning (technique in machine learning that reduces the size of decision trees by removing sections of the tree that provide little or no power to classify instances) of the resulting decision trees. This increases the error rates on



Day	Outlook	Temperature	Humidity	Wind	Play (CLASS)
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Figure 6.: Decision Tree example

the training data, but importantly, decreases the error rates on the unseen testing data. The C4.5 algorithm can also deal with numeric attributes, missing values, and noisy data.

C4.5 as advantages such as: (Singh and Gupta, 2014)

- can handle both continuous and discrete attributes.
- allows attribute values to be marked as ? for missing. Missing attribute values are simply not used in gain and entropy (measure of impurity in data) calculations.

- goes back through the tree once it's been created and attempts to remove branches that do not help by replacing them with leaf nodes.

And disadvantages:

- constructs empty branches which it makes the tree bigger and more complex.
- overfitting happens when algorithm model picks up data with uncommon characteristics.

CART algorithm: CART stands for classification and regression trees (Breiman et al., 1984). It is characterized by the fact that it constructs binary trees, namely each internal node has exactly two outgoing edges. The splits are selected using the twoing (splitting algorithm) criteria, and the obtained tree is pruned by cost-complexity pruning. When provided, it can consider *misclassification* costs in the tree induction. It also enables users to offer prior probability distribution and is able to generate regression trees. Regression trees are trees where their leaves predict a real number and not a class. In the case of regression, this algorithm looks for splits that minimize the prediction squared error (the least–squared deviation). The prediction in each leaf is based on the weighted mean for node (Gayatri and Dattetrya, 2017).

This algorithm also has advantages on others: (Singh and Gupta, 2014)

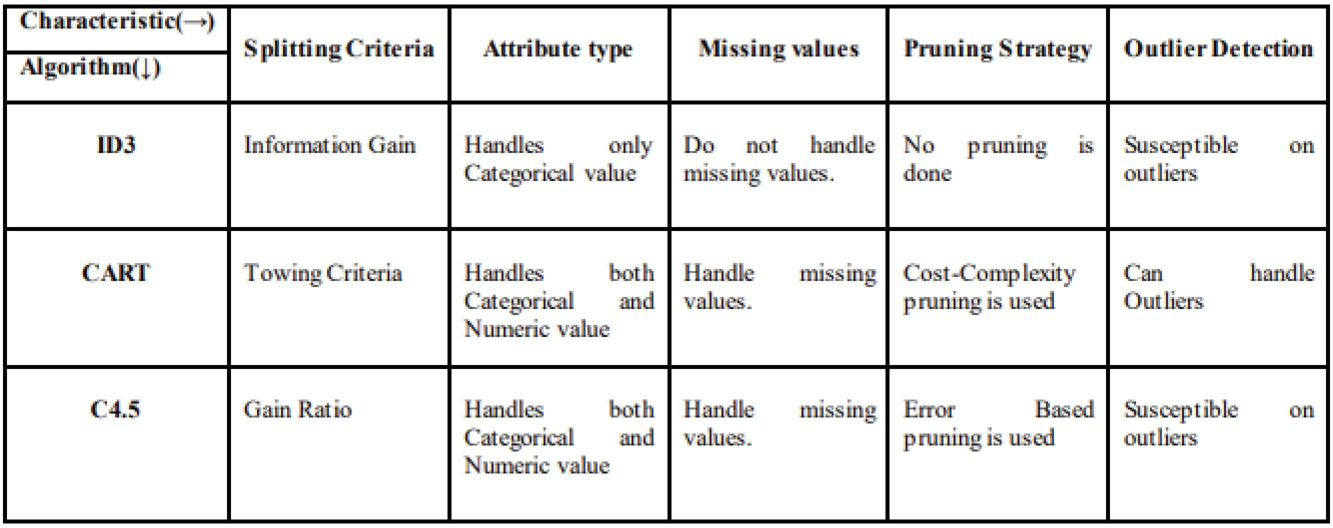
- can easily handle both numerical and categorical variables.
- will itself identify the most significant variables and eliminate non-significant ones
- can easily handle outliers.

And disadvantages:

- may have an unstable decision tree.
- splits only by one variable.

1.6.4 *Detecting Emotions*

Emotion is any conscious experience - each response triggered from parts of the brain to the body, and from parts of the brain to other parts of the brain, using both neural



Characteristic(→)
Algorithm(↓)
	Splitting Criteria	Attribute type	Missing values	Pruning Strategy	Outlier Detection

ID3	Information Gain	Handles only Categorical value	Do not handle missing values.	No pruning is done	Susceptible on outliers
CART	Towing Criteria	Handles both Categorical and Numeric value	Handle missing values.	Cost-Complexity pruning is used	Can handle Outliers
C4.5	Gain Ratio	Handles both Categorical and Numeric value	Handle missing values.	Error Based pruning is used	Susceptible on outliers

Figure 7.: Decision Tree Algorithms characteristics

and humoral routes, (Damassio, 1998) - characterized by intense mental activity and a certain degree of pleasure or displeasure it also is often intertwined with mood, temperament, and personality.

Analysis of emotions in a text can help determine the opinions and effective intent of writers, as well as their attitudes, evaluations, and inclinations concerning various topics (Aman and Szpakowicz, 2007). A text does communicate not only informative contents but also attitudinal information, including emotional states (Alm et al., 2005). There were always a lot of researches on the emotion and sentiments detection from text applied to various types of text, such as emails as Mohammad and Yang (2011) did to predict a person's gender or blogs as shown by Aman and Szpakowicz (2007), even news headlines were studied to find emotions in a text by Strapparava and Mihalcea (2008), also Gomes et al. (2014a) relied on social network data to identify how the parties' social networks correlate to their negotiation performance.

Although there is a large number of investigations on this subject, it wasn't until the last recent years that investigators started to turn their attention to detect emotion in text from online social networks, including Twitter and Facebook which are the focus of this thesis, (Roberts et al., 2012; Choudhury et al., 2012; Bollen et al., 2011; Wemb et al., 2012; Kim et al., 2012; Qadir and Riloff, 2013; Kumar et al., 2018; Mohammad and Kiritchenko, 2014).

On this thesis We will focus on the six significant emotions that were proposed by Ekman (1992); Lazarus and Smith (1991), anger, disgust, fear, joy, surprise and sadness, although our lexicon contains values for both anticipation and trust, we focused just on those six.

1.7 STRUCTURE OF THE DOCUMENT

This thesis is structured as follows:

- Chapter 1 initiates the thesis and introduces the concept of the research that was done.
- Chapter 2 guides us through the technologies we used to process data and apply algorithms to it to obtain our final results.
- Chapter 3 expresses our development regarding getting the data and how we handled and polished it to be prepared to be used by machine learning algorithms and techniques. It also explains the work method that this thesis follows.
- Chapter 4 gives a detailed view of our experiment setup, the techniques we used and how did we come up with the ideas to use each algorithm or machine learning process. In this chapter, our results on the classification and prediction of psycho-demographic profiles are exposed and discussed in detail.
- Chapter 5 as the name says is the conclusion of our paper, it presents the final thoughts and explanations about all the work and mostly about the results we achieved. It also provides an idea of the work that needs to be done to improve what we've done as a prospect for future work.

1.8 SUMMARY

Summarized, this chapter has the purpose of introducing the problem we have in hands as much as the objectives we have to fulfill. It also gives an overview of state of the art and the conceptual definitions of many of the terms that we're going to use throughout the paper.

The following chapter will introduce state of the art in this area and will give an overview of the review of literature done before the active development.

2

PROBLEM SOLUTION AND ENABLING TECHNIQUES

2.1 INTRODUCTION

During this chapter we will explain what technologies and methods were used to prepare data to be ready to be used by machine learning algorithms so we would be able to extract information and our final results from it; also we will explain the thought process that leads us to find a solution for our problem.

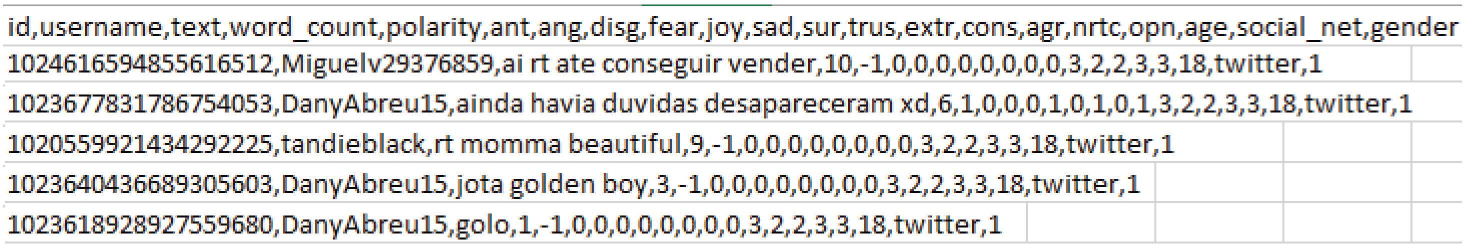
2.2 DATA

Our approach is relevant generally to the multi-class prediction on social media. We focused on predicting users personality scores and their demographic attributes, age, and gender, if possible. All users on our sample filled a personality questionnaire, which will be detailed later, and gave us their age. We gathered data from Twiter using Twint and from Facebook using Data Selfie. We were able to gather 31 users and about 46k text entries, the datasets that we build are detailed in the next section.

2.3 DATASETS AND ANNOTATION

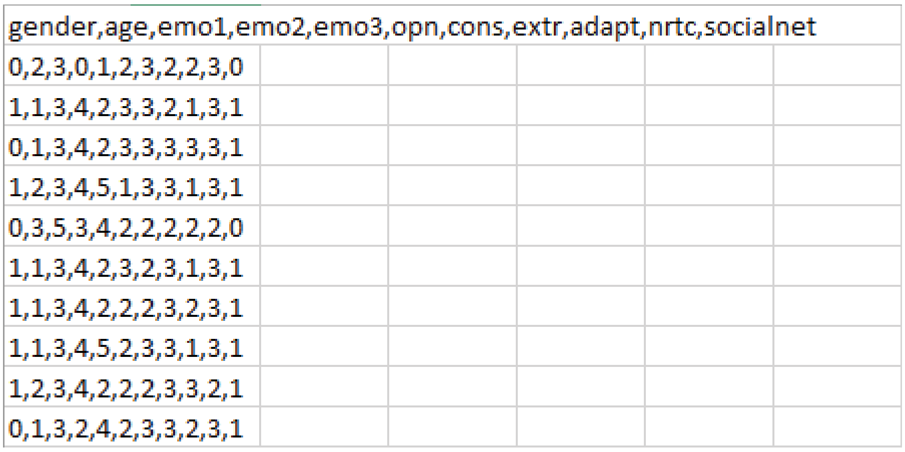
During this research and work we have built several datasets, but chose to only work with 2 of them, they can be assessed below as there is an explanation and a representation of both.

- Mixed Dataset - 31 users, 19 males, 12 females with ages from 15 to 32, with more than 48 thousand entries.
- Final Dataset - 31 users, 1 entry per user in a more concise and general fashion, with ages between 15 to 32, again 19 males and 12 females.



```
id,username,text,word_count,polarity,ant,ang,disg,fear,joy,sad,sur,trus,extr,cons,agr,nrtc,opn,age,social_net,gender
102461659485561651 2,Miguelv29376859,ai rt ate conseguir vender,10,-1,0,0,0,0,0,0,0,0,3,2,2,3,3,18,twitter,1
1023677831786754053,DanyAbreu15,ainda havia duvidas desapareceram xd,6,1,0,0,0,1,0,1,0,1,3,2,2,3,3,18,twitter,1
1020559921434292225,tandieblack,rt momma beautiful,9,-1,0,0,0,0,0,0,0,0,3,2,2,3,3,18,twitter,1
1023640436689305603,DanyAbreu15,jota golden boy,3,-1,0,0,0,0,0,0,0,0,3,2,2,3,3,18,twitter,1
1023618928927559680,DanyAbreu15,golo,1,-1,0,0,0,0,0,0,0,0,3,2,2,3,3,18,twitter,1
```

Figure 8.: Mixed Language Dataset Representation



```
gender,age,emo1,emo2,emo3,opn,cons,extr,adapt,nrtc,socialnet
0,2,3,0,1,2,3,2,2,3,0
1,1,3,4,2,3,3,2,1,3,1
0,1,3,4,2,3,3,3,3,3,1
1,2,3,4,5,1,3,3,1,3,1
0,3,5,3,4,2,2,2,2,2,0
1,1,3,4,2,3,2,3,1,3,1
1,1,3,4,2,2,2,3,2,3,1
1,1,3,4,5,2,3,3,1,3,1
1,2,3,4,2,2,2,3,3,2,1
0,1,3,2,4,2,3,3,2,3,1
```

Figure 9.: User Dataset Representation

Several approaches have been used, and several tries were made to annotate the data with user attributes. This is insanely tiring and endless because we never know what's the best way to do such a process. In the end, we got these two datasets, and we used them to do our experiments and tests.

2.4 WORK PROCESS

In this section, our work process will be deeply detailed in the form of a *Data Analysis Pipeline*.

2.4.1 *Data Analysis Pipeline*

We used a data analysis pipeline as a work process. A simple pipeline usually has inputs going through some processing steps chained together in some way to produce some output. Here, the same happens, but its purpose is for data analysis, see, there is data as input, which is going through some processes such as pre-processing, data checking, analysis, and others, later resulting in either a data product or set of decisions and their supports. As stated in NationalResearchCouncil (2015), Temple Lang described the data analysis pipeline, outlining the steps in one example of data analysis and exploration process, which had to suffer an alteration to be adapted to this thesis needs, they are the following:

1. Ask the objective question.
2. Refine the question, identify data, and understand data and meta-data, related to the answers desired to be obtained.
3. Data Scraping. Get raw data from users social networks or media profiles.
4. Clean and prepare the scraped data. There may be inconsistent or missing information.
5. Compute features that will serve as input to machine learning models.
6. Build learning models.
7. Model evaluation. Test the models made to see what's left to refine and optimize.
8. Perform diagnostics. This helps to understand how well the model fits the data and identifies anomalies and aspects for further study.
9. Analyze the percentage of correct prediction and decide whether it is conceivable to a good result or not.
10. Convey results.

2.5 ENABLING TECHNOLOGIES

Here, we explicit why our used technologies were chosen and what do they serve for in our scenery.

During the development of this thesis, we mainly used two technologies: *Machine Learning and Natural Language Processing*. By combining this two, we aim to get a particular information from a text source, i.e., our data. NLP can be defined as a field of computer science artificial intelligence and computational linguistics that studies the interactions between computers and human languages. It can be used to obtain syntax, semantic or discourse features. It enables us to extract features from the text (tokens, stemming and stopwords in our case), that makes it possible to apply Machine Learning techniques to obtain a predictive model that we can get information from. (**?**)

Next, we will give an insight into how these techniques can be used in python programming language.

2.5.1 *Pandas*

In Pandas' website, `https://pandas.pydata.org/`, it is defined as an open source, a BSD-licensed library providing high-performance, easy-to-use data structures, and data analysis tools. (see pandas.pydata.org) There are three data structures that can be used with pandas:

1. Series - 1-dimensional labeled homogeneous array, size immutable. It is similar to an array, list or dictionary generally used in Python.
2. DataFrames - general 2-dimension labeled, size-mutable, a tabular structure with potentially heterogeneously typed columns. A DataFrame is a container of Series. This is the data structure that we used on our work.
3. Panels - General 3-dimension labeled, size-mutable array and is a container of DataFrames.

We used pandas do read CSV files to a DataFrame structure, so it's easier to work with, per example:

- DataFrame object for data manipulation with integrated indexing.
- Reading and writing data in different formats.
- Intelligent label-based slicing, indexing and subsetting of large datasets
- Easy to work with dataset's columns, whether to insert or to remove them.
- High performing merging and joining of datasets.

2.5.2 *Scikit-Learn*

Scikit-Learn is stated in `https://scikit-learn.org/stable/` as an open source, BSD-licensed Python library as well, providing which provides simple and efficient tools for data mining and analysis. It has many built-in types of algorithms, such as machine learning, preprocessing, cross-validation and more. Scikit allows us to perform classification (identifying to which category an object belongs to), regression (predicting a continuous-valued attribute associated with an object), clustering (automatic grouping of similar objects into sets), dimensional reduction (reducing the number of random variables to consider), model selection (comparing, validating and choosing parameters and models) and preprocessing (feature extraction and normalization). We'll focus on classification, once that is what we have to do with our data, classify in classes users' age, gender or personality traits. Scikit's classification algorithms are plenty, and we used only NaiveBayes and SVC (trees we went with WEKA) because they were the best fit for our problem.

2.5.3 *Natural Language Toolkit*

NLTK is a platform for building Python programs to work with human language data. It provides easy-to-use interfaces and lexical resources, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, semantic reasoning and wrappers for industrial-strength NLP libraries as it is defined in `https://nltk.org`. Some things it provides are:

- tokenize and tag some text
- identify named entities
- display parse-trees

We used it to perform stemming, tokenization and to list the stopwords we would later remove.

2.5.4 *Weka*

Weka is defined in `https://cs.waikato.ac.nz./ml/weka` as a data mining software written in Java programming language. It consists of a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization. It also allows us to process big data and perform data learning. We used Weka to apply decision tree algorithms for data classification, and we used J48 (C4.5) and CART from the many listed there, that choice was explained in section 2.

2.6 SUMMARY

We get to know in this chapter, that we gathered data from both Facebook and Twitter social networks.

Also, that we chose a data analysis pipeline approach as our work process, so we had to follow a series of steps since the beginning of the thesis until the very end when we got our results.

To finalize the chapter, we gave a detailed synopsis on our used enabling technologies that when combined with the python language gave us all we needed to get our final results after a lot of pre-processing, which is explained in the following chapter.

3

EXPERIMENTAL SETUP IMPLEMENTATION

3.1 INTRODUCTION

In this chapter we are going to explain in full our setup implementation and explain how we got to our final data, ready to apply the algorithms and methods necessary to obtain our desired results.

3.2 EXPERIMENTAL SETUP

This section will describe our decisions throughout the research since data mining, to data already pre-processed to be ready to apply the chosen machine learning algorithms. For the data mining part, we will divide this sections into two subsections once we applied it in two different social networks with different techniques. The following subsection will address our choice for a word-emotion lexicon in which we would use to associate emotions presented in words with the text and information mined. Lastly, in the final subsection of this chapter, we will explain our choice for the personality questionnaire.

3.2.1 *Facebook*

For *Facebook* we went with using Data Selfie plugin on Chrome to *mine* data from *Facebook* users. This plugin would already predict personality traits by itself and their algorithms, but we did not want that. All that we ask the users that were part of the sample was the data file that could be exported from the plugin in JSON format. That file contained the time they logged in, and out of *Facebook*, what they looked at, what they commented or posted and even what they typed in personal messages - users were aware of that. See, *www.dataselfie.it*

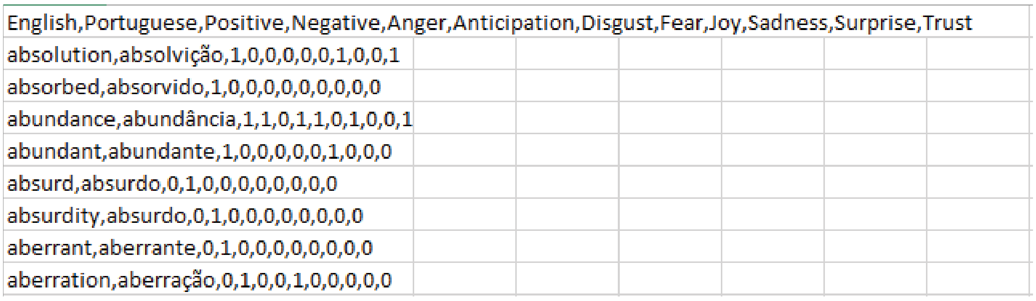
For *Facebook* our sample was considerably small, consisting only on five users and about 6 thousand dataset text entries.

3.2.2 *Twitter*

On *Twitter* we went on a different note and used a twitter scraping tool written in *Python*, know as *Twint*. This tool allows for scraping Tweets from Twitter profiles without using *Twitter's API*. It utilizes *Twitter's* search operators to let you scrape Tweets from specific users, scrape Tweets relating to certain topics, hashtags & trends, or sort out sensitive information from Tweets like e-mail and phone numbers. *Twint* also makes special queries to *Twitter* allowing you also to scrape a *Twitter* user's followers, Tweets a user has liked, and who they follow without any authentication, API, Selenium, or browser emulation. See, *https://github.com/twintproject/twint/tree/master/twint* In our case, what we did was extract all tweets or retweets from users that participated in our project (not replies once there can reply to private profiles, and we cannot get access to them). We had a sample of 27 users for *Twitter* and were able to retrieve more 42 thousand tweets or retweets to build our dataset.

3.2.3 *EmoLex*

EmoLex also known as ***NRC Word-Emotion Association Lexicon*** was at first a list of only English words and their associations with eight basic emotions (anger, fear, surprise, sadness, joy, and disgust) and two sentiments (negative and positive) but has since evolved to a multi-language lexicon. The annotations were manually done by crowd-sourcing.



```
English,Portuguese,Positive,Negative,Anger,Anticipation,Disgust,Fear,Joy,Sadness,Surprise,Trust
absolution,absolvição,1,0,0,0,0,0,1,0,0,1
absorbed,absorvido,1,0,0,0,0,0,0,0,0,0
abundance,abundância,1,1,0,1,1,0,1,0,0,1
abundant,abundante,1,0,0,0,0,0,1,0,0,0
absurd,absurdo,0,1,0,0,0,0,0,0,0,0
absurdity,absurdo,0,1,0,0,0,0,0,0,0,0
aberrant,aberrante,0,1,0,0,0,0,0,0,0,0
aberration,aberração,0,1,0,0,1,0,0,0,0,0
```

Figure 10.: Lexicon dataset representation

This lexicon was approved by the *Research Ethics Board* of the *National Research Council*. It contains over 30k word-entries which cover the most used words or terms used in the English language and assigns a set of emotions per word. After cutting all the non-used languages, we got the dataset that can be seen above in figure 12.

3.2.4 *Ten Item Personality Inventory*

To compare and classify our data, we needed our sample to take a simple questionnaire about their personality.

There were some options for it, such as *International Personality Item Pool* - IPIP - created by Wim Hofstee with the help of his colleagues and students at the University of Groningen in the Netherlands (Hendriks and Raad, 2002; Hendriks and Jolijn, 1997) or the *Big Five Inventory* - BFI - referenced by John and Srivastava (1999). We went with the *Ten Item Personality Test*, also known as TIPI introduced by Gosling et al. (2003), as the name suggests it consists only in ten questions to determine all big 5 traits. As it is more concise (and as stated in Gosling et al. (2003)), it reaches adequate convergence with the Big-Five measures. The questionnaire is shown in figure 13.

We created a set of classification procedures from both of the two types of data we have, Facebook and Twitter. Our goal is to explore the following prediction classes:

- Both types of the above data mixed (twitter and facebook text samples)
- All users with no text entries, instead we have the result of analyzing each one of them individually, where each user is one single dataset entry

For all of the above, we performed various machine-learning techniques that were explored by some of the authors mentioned in section 1.6. We applied 10 and 3-fold cross-validation and a train-test split on different classification tasks, such as Naive-Bayes, Decision Trees and SVM.

We used the Weka tool, for decision trees and the Scikit-Learn library for Python programming language, embedded algorithms for NaiveBayes and SVM.

Classification tasks assume the existence of an unknown function that maps predictor variables to a nominal target variable. This function can be defined as $Y = f(X1, X2, ..., Xp)$, where Y is the nominal target variable, $X1, X2, ..., Xp$ are features describing the items and $f()$ is the unknown function we want to approximate. In order to obtain an approximation (a model) of this unknown function we use a data set with examples of the function mapping (known as a training set), i.e. $D = \{\langle h_i, y_i \rangle\}_{i=1}^{n}$ (Muniz, 2009).

With all of our "working tools" introduced and explained in the next section, we will detail all of the pre-processing process of the research and how we got to the final datasets that we needed to apply machine-learning techniques and got results out of it.

3.3 PRE-PROCESSING

As stated above, the data that we receive from dataselfie comes in a JSON (JavaScript Object Notation) file which contains arrays of information that interest us such as:

- looked - content looked at in the form of text
- typed - content typed, whether a comment or post or message
- clicked - URL of the links clicked on
- timespent - start and end time of the logged session

From this four different "classes" we focused only on *looked* and *typed*, so we had to create a parser script in order to turn that *JSON* data into *CSV* entries to build users' datasets. Each post, message or comment made by the user equals one dataset entry there is no minimum or maximum of words per entry.

As for *Twitter*, using the *twint* script already returns a *CSV* file with all tweets and retweets and user information, such as tweet time-stamps, tweet ids, user nicknames and type of tweet (if it was a tweet made by the user or if it was a retweet). After the first part of pre-processing, the only features of the dataset were only the tweet text itself, user's id and user's nicknames.

Remember that both two of this types of the dataset are still raw and need to be pre-processed before we can apply any machine learning method to get knowledge about the data. After gathering the data for both datasets, we had to make it go through a pre-processing phase which is explained below.

3.3.1 *Finalizing Datasets*

To make them ready to apply the machine learning algorithms we created a script to add the next columns:

- word_count - the number of words presented in each tweet, post, etc.
- polarity - represents whether the text has positive, negative or neutral polarity (-1 for negative, 0 for neutral, 1 for positive)
- ang - whether the anguish emotion is presented in the text (0 for false, 1 for positive)
- disg - whether the disgust emotion is presented in the text (0 for false, 1 for positive)
- fear - whether the fear emotion is presented in the text (0 for false, 1 for positive)
- joy - whether the joy emotion is presented in the text (0 for false, 1 for positive)
- sad - whether the sadness emotion is presented in the text (0 for false, 1 for positive)
- sur - whether the surprise emotion is presented in the text (0 for false, 1 for positive)
- extr - the value of the extroversion trait after the original value from the TIPI questionnaire was discretized (range of 1-3)
- cons - the value of the conscientiousness trait after the original value from the TIPI questionnaire was discretized (range of 1-3)
- agr - value of the agreeableness trait after the original value from the TIPI questionnaire was discretized (range of 1-3)
- nrtc - value of the neuroticism trait after the original value from the TIPI questionnaire was discretized (range of 1-3)

- opn - value of the openness trait after the original value from the TIPI questionnaire was discretized (range of 1-3)
- age - participant's age
- gender - participant's gender (0 for male, 1 for female)
- social_net - if the participant's data came from *Facebook* or *Twitter*

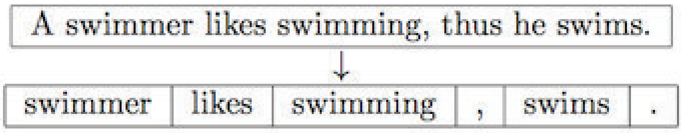
This origin the dataset that is portrayed above in figure 16. The following sections will explain some of the process applied to text and some attributes, such as stopwords removal, stemming, discretization and word-matching, this latter one to find polarity and emotion relations with words.

3.3.2 *Stopwords*

Stopwords are a set of most commonly used words in any language. The reason we removed stopwords from our data was to focus on the important words instead. For example, in the context of a search engine, if your search query is "how to remove stopwords from the text?", the search engine will try to find web pages that contain the terms "how", "to" "remove", "stopwords", "from", "text", so this causes the search engine to find a lot more pages that contain the terms "how", "to" and "from" than pages that contain information actually removing stopwords from text because the terms "how", "to" and "from" are so commonly used in the English language. So, if we disregard these three terms, the search engine can focus on retrieving pages that contain the keywords: "remove" "stopwords" "retrieval" "text" – which would bring up many more pages that are really of interest without the ones that don't matter to us. In our case, it's more because it will be more efficient and fast to the word-matching phase, once all or most of these words have no polarity or emotion value attached to them, we don't need to do the test on them, and it saves a lot of memory and time. To remove them from our data, we used the stopwords module from "natural language toolkit" library in python.

3.3.3 *Stemming*

For grammatical reasons, all texts use different forms of a word, such as take, takes, and taking. Moreover, there are "word families" which are related words with similar



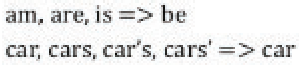
A swimmer likes swimming, thus he swims.

↓

swimmer	likes	swimming	,	swims	.

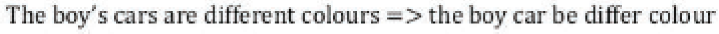
Figure 11.: Stopword removal example

meaning, such as, interact, interaction and interactive. What stemming aims to achieve is to reduce inflectional forms and sometimes derived related forms of a word to a common base form. For instance:



am, are, is => be
car, cars, car's, cars' => car

Turns into:

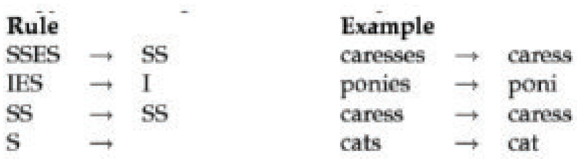


The boy's cars are different colours => the boy car be differ colour

Figure 12.: Text stemming representation

To our work stemming is even more important what's stated above, we applied stemming in our lexicon and in all of our words, see, if the word "car" was in the lexicon, and in our text the word "cars" appeared, the function that searches for the words that are in our data in the lexicon would return "false" and we wouldn't be able to get access to the polarity or emotion attached to the word, so stemming was really necessary. We used, *SnowBall Stemmer* from the "natural language toolkit" library in python, it is called SnowBall instead of Porter stemmer because it's creator, Porter, build a programming language with this name for creating new stemming algorithms (see, nltk.org), but follows the same rules of the original stemmer, those rules can be seen in the figure below:

It would be perfect to see how the results would differ with *Lemmatization* instead of Stemming, but there is no support for the Portuguese language which is the language of the major part of our data.



Rule			Example		
SSES	→	SS	caresses	→	caress
IES	→	I	ponies	→	poni
SS	→	SS	caress	→	caress
S	→		cats	→	cat

Figure 13.: Snowball/Porter stemmer rules representation

3.3.4 *Discretization*

Discrete values have important roles in data mining and knowledge discovery. They are about intervals of numbers which are more concise to represent and specify, easier to use and comprehend as they are closer to a knowledge-level representation than continuous values (Liu et al., 2002). Many studies show induction tasks can benefit from discretization: rules with discrete values usually are shorter and more understandable and discretization can lead to improved predictive accuracy. Rajalakshmi et al. (2016) states that this mechanism reduces the number of continuous features values, which brings smaller demands on system's storage. It also makes the learning process more accurate and faster. As for the *Big Five* traits, were values ranging from 1 to 7, and could be floats as in 1.5, 2.5 and so on. As we need to classify and predict each trait, being that the goal to achieve after having all data processed, we had to perform discretization on each trait, so if the value was between 1 and 3 (included) it is now represented by 1 (low), if it was in range of 3 to 5(included) it is represented by 2 (mid), finally if it was a value between 5 and 7, it is represented by 3 (high), by doing this we diminished our spectrum from [1, 1.5, 2, 2.5..7] to [1,2,3], which is a considerable upgrade in terms of a classification model. There are a lot of methods to perform this phenomenon, but we used Weka interface to do it in our required attributes.

1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7

Table 2.: BigFive values undiscretized

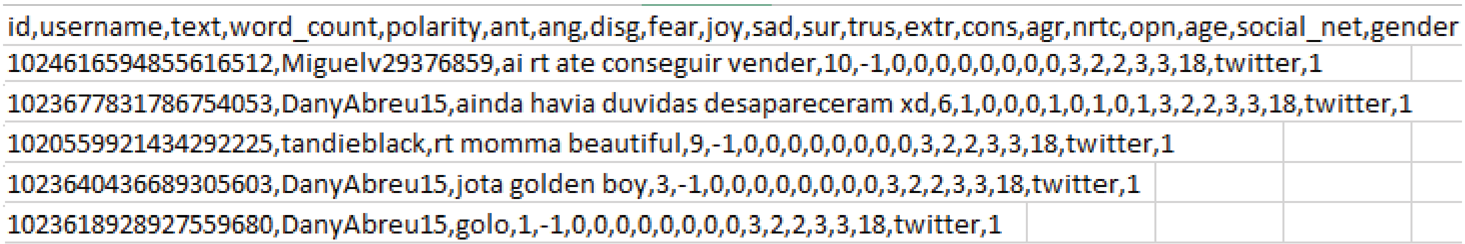
⇓

1	2	3

Table 3.: BigFive values discretized

3.3.5 *Word-Matching*

In order to know the polarity and which emotions are present in all our text data, we did a word-matching for every entry of the final datasets with all the words that are part of the lexicon, if the word were in the lexicon and had emotions attributed to it, the number one would be appended to that emotion column instead of zero, as well as the polarity attribute. This word-matching process was done after having already applied both the stopword removal process and after having stemmed all the words, of our data text and also from the lexicon. All of this originated the dataset that is portrayed (not in its entirety) in the figure below:



```
id,username,text,word_count,polarity,ant,ang,disg,fear,joy,sad,sur,trus,extr,cons,agr,nrtc,opn,age,social_net,gender
1024616594855616512,Miguelv29376859,ai rt ate conseguir vender,10,-1,0,0,0,0,0,0,0,0,3,2,2,3,3,18,twitter,1
1023677831786754053,DanyAbreu15,ainda havia duvidas desapareceram xd,6,1,0,0,0,1,0,1,0,1,3,2,2,3,3,18,twitter,1
1020559921434292225,tandieblack,rt momma beautiful,9,-1,0,0,0,0,0,0,0,0,3,2,2,3,3,18,twitter,1
1023640436689305603,DanyAbreu15,jota golden boy,3,-1,0,0,0,0,0,0,0,0,3,2,2,3,3,18,twitter,1
1023618928927559680,DanyAbreu15,golo,1,-1,0,0,0,0,0,0,0,0,3,2,2,3,3,18,twitter,1
```

Figure 14.: Post Pre-Processed Dataset Representation

Finally, our dataset (represented above) is ready, and we can now apply machine-learning algorithms in those more than 48k entries divided by 32 participants to get our results.

3.4 SUMMARY

This project's work project will have as support a data analysis pipeline which follows a series of steps with the purpose of data analysis, where data goes through some processes such as pre-processing, data checking, analysis, and others, producing either a data product or set of decisions and their supports. We opted to use data only from two of the major social networks - Facebook and Twitter - and utilized EmoLex as an emotion lexicon to have access to emotions attached to words so we can perform a text-emotion analysis on our data. About the personality questionnaire performed by all the participants, we went with TIPI - a 10 item questionnaire - which provided us the Big-Five traits for all the sample. With the decisions all made, our collected data suffered pre-processing to be ready to be studied and for us to apply machine-learning algorithms, The following chapter will explain all of our thought process on our chosen algorithms and the results that we achieved.

4

CASE STUDIES / EXPERIMENTS

4.1 INTRODUCTION

In this chapter, we will carry an experimental evaluation to obtain the final results. There are going to be some sets of experiments, once we have to apply the machine-learning algorithms to our two datasets. Our concerns when getting results where the ability to predict with efficiency the traits we mentioned throughout the last chapters, and to evaluate said predictions with evaluation methods that will be introduced in the next section. We will also present our results with a detailed explanation during the discussion section.

4.2 MODEL EVALUATION

The goal of this section is to explain the metrics that are used to evaluates a general classification task, and they are: (Powers, 2011)

- Precision: the number of true positives (the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (sum of true positives and false positives, which are items incorrectly labeled as belonging to the class).

$$precision = \frac{tp}{tp + fp}$$

- Recall: defined as the number of true positives divided by the total number of elements that belong to the positive class (sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been).

$$recall = \frac{tp}{tp + fn}$$

- F1-Score: measure of the classification accuracy. It is the harmonic mean of precision and recall.

$$f1 = 2.\frac{precision.recall}{precision + recall}$$

4.3 TRAIN-TEST SET VS K-FOLD CROSS VALIDATION

We trained our classification models in two different *environments*, firstly, and to what we call our base test, we split our data into two subsets, training, and testing, with the training subset containing 80% of our data, leaving the remaining 20% for the test. And secondly, we trained our classification model with a 10-fold cross validation method applied to our data, both of them are explained below, respectively.
In our base test, we divide data into two subsets as mentioned above:

- **Training Subset:** subset to train the classification model.
- **Testing Subset:** subset to test the classification model.

This method can be tricky because it's easy to achieve overfitting, which is the phenomenon that occurs when our training set is overtrained with our data, this happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model. The problem is that these concepts do not apply to new data and negatively impact the model's ability to generalize, so we have to be careful not to overtrain our model. But there's the other side of the coin here, we also need to look out not to make our model under fitted, meaning that our model is too simple – informed by too few features, regularized too much or just because it doesn't have enough data to train the model – which makes it inflexible in learning from the dataset. Underfitting refers to a model that can neither model the training data nor generalize to new data. We went with a training set of 80% of our total data, leaving 20% for our test set, which usually, including, in this case, doesn't cause overfitting nor underfitting.
After assuring that none of the two phenomena mentioned above occurs, we have to make sure that our test set meets the following three conditions:

- Is large enough to yield statistically meaningful results.
- Is representative of the data set as a whole. In other words, don't pick a test set with different characteristics than the training set.

After having all of this check, we train the model with our training subset and then test its efficiency by classifying the data presented in our test set.

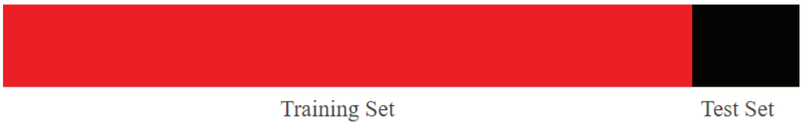


Figure 15.: Representation of sliced data from one dataset into a training set and a testing set

The goal of the other method we used is to help knowing how the machine learning model would generalize to an independent data set. This technique is used to estimate how accurate the predictions your model will give in practice. When working on a machine learning problem, we will be given two type of data sets — known data (training data set) and unknown data (test data set) as we referred above. By using this method, we are testing our machine learning model in the "training" phase to check for overfitting and to get an idea about how our machine learning model generalizes (i.e., an unknown dataset, for instance from a real problem). In one round of cross-validation, you will have to divide your original training dataset into two parts:

- Cross Validation Training Set
- Cross Validation Testing Set

We will train our model on the cross-validation training set and cross-validation testing the model's predictions on the test set. That gives us the accuracy of our machine learning model's predictions when compared to the model's predictions on the test set and the actual labels of the data points in the testing set. To reduce the variance, it performs multiple rounds of cross-validation by using different CV training and testing sets. The results from all the rounds are averaged to estimate the accuracy of the machine learning model to derive a more accurate estimate of the model prediction performance. (Seni and Elder, 2010). We used the K-Fold cross-validation, more specifically, the 10-fold, meaning that we divide the testing set in 10 equal parts, each containing different data and the model performs ten rounds of cross-validation. K-Fold is performed as the following steps indicate:

1. Partition the original training data set into k equal subsets. Each subset is called a fold. Let the folds be named as $f1, f2, \ldots, fk$
2. For $i = 1$ to $i = k$:
   a. Keep the fold f_i as validation set and keep all the remaining $k - 1$ folds in the

Cross-validation training set.
b. Train the machine learning model using the cross-validation training set and calculate the accuracy of your model by validating the predicted results against the validation set.

3. Estimate the accuracy of your machine learning model by averaging the accuracy derived in all the *k* cases of cross-validation.

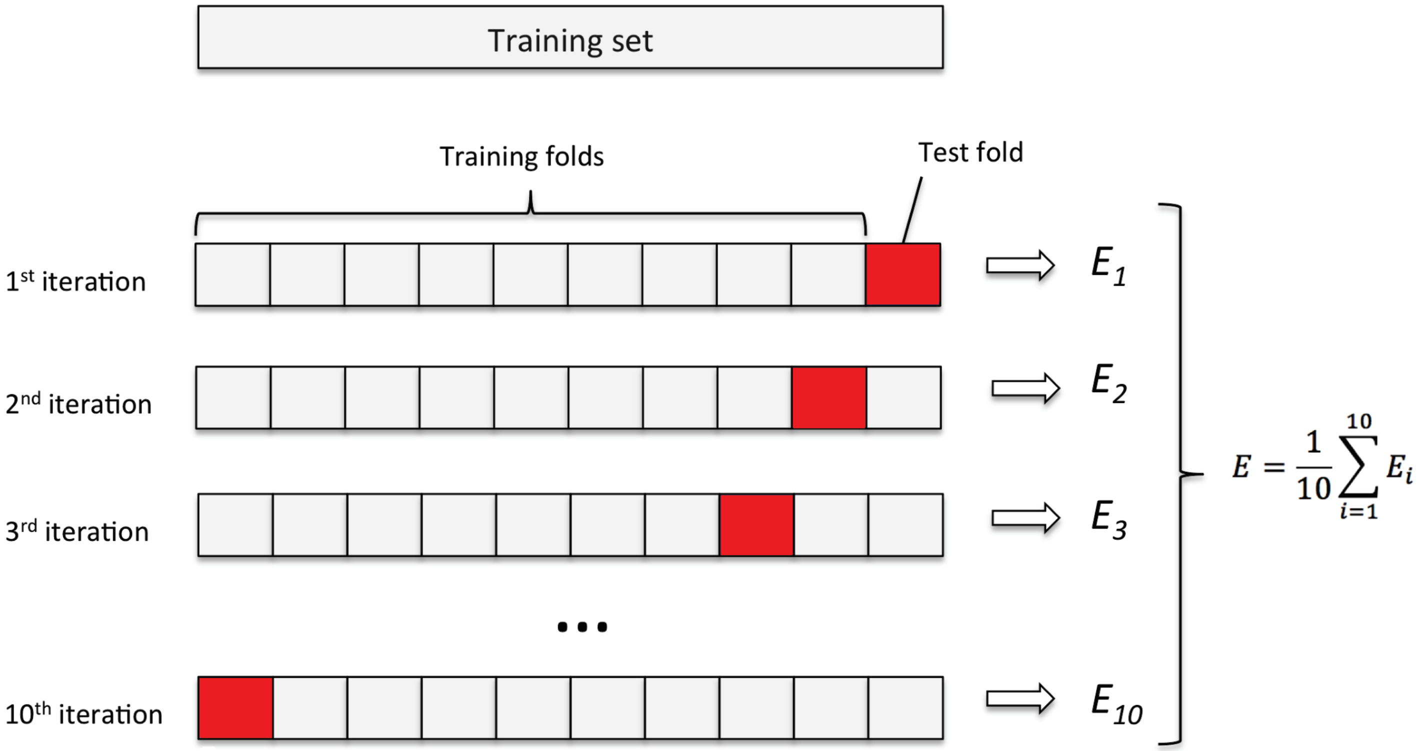


Figure 16.: Representation of how a 10-fold cross validation is applied on data

4.4 OUR SAMPLE

In this section we will give an overview of our sample, we will focus on the differences and resemblances between facebook and twitter users, male vs. female users, and will also present the relations between our users' personality traits with their most used emotions in their text samples.

4.4.1 *Facebook vs Twitter*

One of our curiosities was to see if users showed different emotions in their text on Facebook or Twitter by comparing them both. Our Facebook user's data is super small, we only have five users on our sample, but we did find some correlations between them. Our process was to find the three most frequent emotions presented in each user's written text.

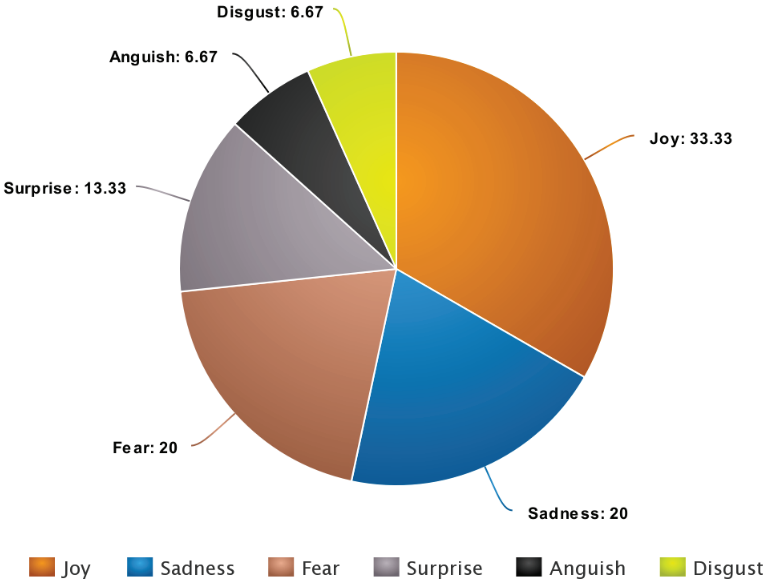


Figure 17.: Representation of facebook user's emotions

Analyzing the charts, we can see some resemblances between Facebook and Twitter users, both of them are covered by 33,3% of the emotion *Joy* which means that all of our samples had it as a part of their three highest emotions presented on their text. In both cases, *anguish* and *disgust* have the lowest percentage of representation on an user's top-3. Also, the three most represented emotions are the same three on both cases, first *joy*, followed by *sadness*, and *fear*, these latter two, on the Facebook chart are tied with 20% each but are still two of the most three represented emotions.
As for the differences, we can see that on Twitter, *sadness* and *fear* have higher percentages than on Facebook, 12 and 6%, respectively. The rest of the emotions have higher percentages on Facebook than on Twitter, meaning, *surprise, anguish, and disgust* for around 7, 5 and 5%, respectively as well.

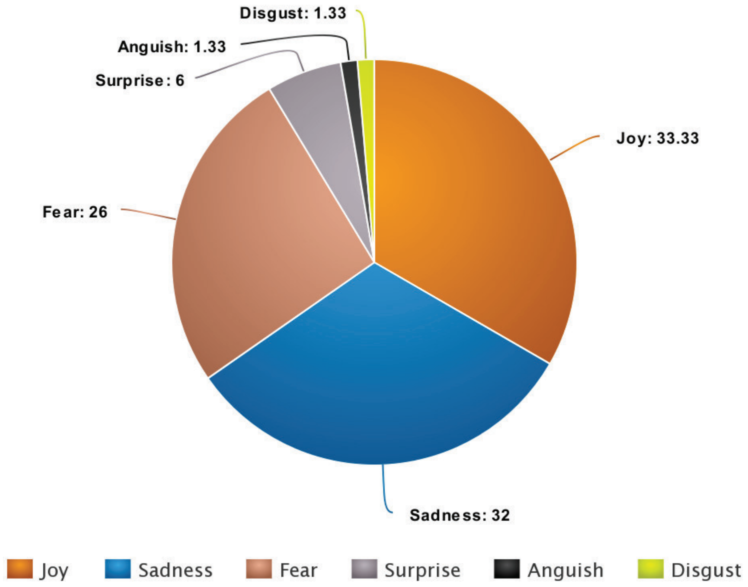


Figure 18.: Representation of twitter user's emotions

4.4.2 *Male vs Female*

As for relations and differences between both genders we found some interesting ones in our sample. Remember that as it was mentioned at the beginning of this thesis, we have 12 females and 19 males representing our sample of 31 users. We analyzed both their personality traits and preponderant emotions presented in all their data, one by one, and we got to some interesting conclusions. Male, have the same percentage of *joy* and *sadness* represented on their three most preponderant emotions ranking and both of them lead the rest with a percentage of 29.8, the lowest is *anguish* with 1.8%. Female on the other hand, have one emotion that tops all other, that is *sadness* with 33.3% followed by *joy* with 30.6%. The exciting thing about this group is the fact that *anguish* doesn't have a part in any of the top-3 emotion counters. The pie-charts confirming this information are shown in the support material chapter.
Referring to the personality traits male and female users differ in almost every single trait. We discretized the trait's values in three classes as we mentioned earlier, low, medium and high, and we analyzed their scores based on these. Analyzing trait by trait we get:
Interestingly enough, they diverge in more than half of all categories. Let's see: In all traits except *agreableness*, although the lower percentage fits in the *low* class in both of

	Male			Female		
	Low	Med	High	Low	Med	High
Openness	10,5	52,6	36,8	8,3	41,7	50,0
Conscientiousness	0,0	52,6	47,4	8,3	25,0	66,7
Extroversion	0,0	78,9	21,1	8,3	25,0	66,7
Agreableness	5,2	73,7	21,1	75,0	16,7	8,3
Neuroticism	0,0	52,6	47,4	0,0	33,3	66,7

Table 4.: Male and female personality traits comparison in percentages

them, there is a higher percentage of male users that fit in the *medium* than in the *high* class which is the exact opposite on the female side. Meaning that on these four traits, the class with the higher percentage for males is the *medium* one, and for females is the *high* one and they always "agree" when it comes to the *low* class, which has **always** the lowest percentage.
The most exciting trait to analyze here is *agreableness*. On the male side, it follows the other traits, *low* has the lowest percentage, *medium* has the highest and the last class stays in between. But here's when it gets amusing, on the female side, **seventy five** percent of our sample fits on the *low* class, comparing this value with other from the same class is absurd, the highest percentage on this class after this one is 10,5% represented by the male's openness trait. It is also the highest percentage number on any female trait and class which still stays a bit short of the male's *medium* class extroversion 78,9% which is the highest percentage acquiesced by any trait in any class. Bar charts with the numbers of users per class will be added to the support material chapter to corroborate these numbers. This serves to give an overview of what to expect from classifying and predicting our sample which is explained in the next section.

4.4.3 *Emotion and Psycho-Demographic traits relation*

During this research and study of our datasets, we found some relations between emotions and psycho-demographic traits. We will do an independent analysis of how the personality traits relate to the users' emotions followed by an analysis of how both of the demographic traits chosen by us relate with the emotions as well. The analysis has as base the three most expressed emotions by users in their text, and they are: joy which is present in all of the users top-3 expressed emotions, followed by sadness which is appears in 87% of users' top-3, and fear that occupies 71% of the spots. Of course, we will only relate the bigger and most interesting ones.

Relation Between Emotions and Personality Traits: After a deep analysis we found some interesting correlations between users expressed emotions and their personalities:

- **Joy:** We've discovered that 58% of the user's who have this emotion represent in their three most preponderant emotions, scored *high* (belong to the high class of that trait, after discretization) on both conscientiousness and neuroticism. Also, 92% of them, have a classification of *low* in agreeableness.
- **Sadness:** 59% of users who contain sadness in their three most frequent emotions also scored *high* on both conscientiousness and neuroticism. Again, 92% of them scored *low* on the agreeableness trait.
- **Fear:** This relates very well with joy as well, there are fewer users with fear as one of their three most preponderant emotions but, again, 58% of them score *high* on conscientiousness and neuroticism. This particular emotion still relates to a lower classification of aggrebeleeness with 59% of the users scoring *low* on that trait.

The results seem very similar between these three emotions and how they relate with the personality traits, this probably happens because of the reduced size of our sample, or maybe, it's just how it is. Concluding, users who express joy, sadness and fear a lot, are more likely to have a higher score in conscientiousness and neuroticism, and a lower score on agreeableness.

Relation Between Personality and Demographic traits: As we already compared male and female by their expressed emotions, here we will focus on finding relations between personality and demographic traits.

Remember we discretized the age attribute in three classes, $[15-20]$, $[21-25]$, $[26-31]$, the first one contains 74% of our sample, the second one contains only 16%, and the latter includes 10%. As for users whom belong to the first class, 52% of them scored *medium* on openness, extroversion and agreeableness and *high* on neuroticism, 48% of them scored *medium and high* on conscientiousness. Sixty percent of the ones that take part in the second class, scored *medium* openness, conscientiousness, and extroversion, and forty percent scored *medium and high* on neuroticism and *low and*

high on agreeableness. For the ones present on the latter class, 67% of them scored *medium* in all of the five traits.
As for gender relations with emotions, there are a lot of differences:

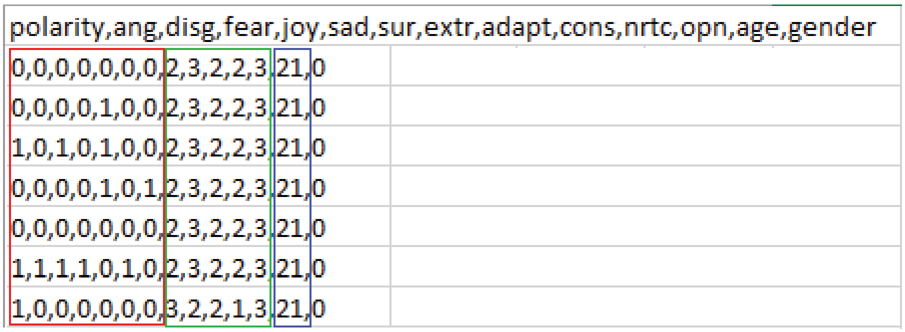
- **Openness:** 53% of males scored *medium* while only 42% scored that class. 50% of females scored *high* against the male's 37% on this trait. While only 37
- **Conscientiousness:** Here we have resemblances with both male and female's scoring higher on the *high* class, with 53 and 67% respectively.
- **Extroversion:** They diverge again here, when male's score 79% on *medium* against the 63% females score on *high*. A pretty big difference.
- **agreeableness:** Another huge difference can be acquiesced on this trait. Female's score 75% *low* and male's score 63% *medium*.
- **Neuroticism:** On the last trait they agree again, with the highest numbers appearing on the *high* class with 53% for the males and 67% for females.

We can conclude that no matter the age, users tend to score their personality traits on the *medium and high* classes except for neuroticism, since 40% users with age comprehended between 21 and 25 scored *low* on that trait. Gender-wise, we can deduce that men tend to be more neutral on three of the five traits, openness, extroversion, and agreeableness with a high percentage of them scoring *medium* on them, as long as women tend to be more open, extroverted and way less agreeable than men. As for the other traits, they are similar, both men and women score *high* on neuroticism and conscientiousness

4.5 RESULTS AND DISCUSSION

The goal of this section is to show our achieved results in two of the for mentioned datasets, mixed language, and user dataset. The other two will be displayed in the Support Material chapter because they're not part of the main focus of this work.dispersed in different subsections to individualize each dataset and to be able to find correlations in the results between the different datasets we're studying. In a dream world our sample would be rounding thousands of users instead of dozens, with this said, we had one objective and one line of thought from the beginning of our work and that was to use a dataset containing one entry per user, with the 3 most preponderant emotions presented in all of their tweets or Facebook text, their age, gender, and their personality traits, that would be the best way to get results in our research and it was the one that made more sense. We still did that, but with only 32 users, the results are not very trustworthy. So, in order to get some more results, we used a dataset containing all text from tweets and Facebook features from each user, that could work as well, but keep in mind that each entry for each user will repeat the columns with the personality traits, this can bias the results and it's not that much useful to predict outcomes. In this case, we had to make more columns, instead of the 3 most preponderant we had to put all emotions with values 0 (if not present), and 1 (if present) in thousand of small tweets and small portions of text, most of it will be 0, and that's not good to perform analysis and prediction based on that data, still, we did our best and got some expected results, considering our small sample.

Warning: On the "mixed-language dataset" as there are so much null values throughout all the emotion columns, it is impossible to achieve good classification numbers for personality traits without using demographic attributes as features, and vice-versa. We examined and analyzed this dataset knowing it would not be the best approach to achieve our goals of predicting both personality and demographic traits just from text emotion because of all the nulls that were mentioned above, but as we have such a small sample in our "user dataset" we decided to go for it and use emotions and demographic traits to predict personality traits and emotions and personality traits to predict demographic traits. As for the demographic traits (age and gender), we couldn't achieve real results on this dataset. The cause for this is that this dataset is way overfitted for those two attributes, because we have 48k tweets divided by only 32 users, even if we lowered the training sample, the model would always perform 100% prediction accuracy on these traits, we needed a way much broader sample in order to be able to predict these values with efficiency. As the personality traits are presented in every entry, and they are part of the features for the classification, even after the discretization of the age attribute in three classes, to make sure that there were no unique values as there were before discretizing - we had unique age numbers such as 15, 16, 28, 29, 31 -, this continues to happen. So the values presented below are not correct in any way. Meaning, that if we have a user with 2k tweets - dataset entries - it's age will always be associated with its personality traits for 2k entries, in our dataset these 2k entries represent almost 4% of our global dataset data, so if we're trying to make predictions from data presented on our dataset but not trained, the probability of the user which we're trying to predict the demographic attributes to belong also to the training set of the dataset is too high to be able to classify it correctly, so it means that it is a way overfitted model for this attributes.



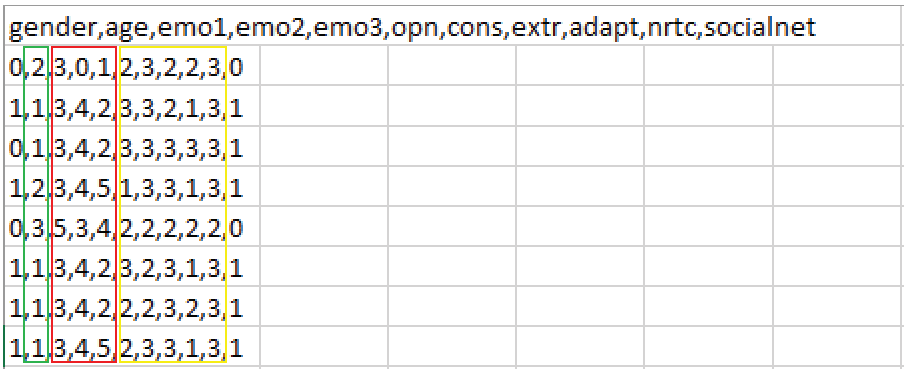
```
polarity,ang,disg,fear,joy,sad,sur,extr,adapt,cons,nrtc,opn,age,gender
0,0,0,0,0,0,0,2,3,2,2,3,21,0
0,0,0,0,1,0,0,2,3,2,2,3,21,0
1,0,1,0,1,0,0,2,3,2,2,3,21,0
0,0,0,0,1,0,1,2,3,2,2,3,21,0
0,0,0,0,0,0,0,2,3,2,2,3,21,0
1,1,1,1,0,1,0,2,3,2,2,3,21,0
1,0,0,0,0,0,0,3,2,2,1,3,21,0
```

Figure 19.: Representation of the dataset attribute's flaws. Red: All the null data on the emotion columns. Green and Blue: The repetition of the user's personality traits (already discretized) and user's age throughout the dataset.

For the "user dataset", as it is so small - 32 entries - , we had to divide it differently, we went with 66% for training and 34% for testing, because if we went higher on the training set, it would overfit the model. We wouldn't even perform analysis in this dataset because of our sample size, but we believe that with a more significant sample, of thousands of users, this would be the correct dataset to analyze, as we refine the emotion values, there are no null values, there are only the values of the top-3 emotions expressed by each user, there is much more consistency on data and it would be possible to predict demographic attributes without having the personality values, as it would be possible to predict personality class values without having the demographic values, because of it's consistency. As the sample is so small, the achieved results are not the best or the ones that we hoped.

4.5.1 *Self Performance*

Here we're going to show our classification reports on the *Mixed Language Dataset* based on the classifier's precision, recall, and f-1 score. This first table contains the results for each classifier performing on an eighty-twenty training/testing subsets, and we will analyze their results and then perform a general analysis of all of them. We will first analyze the personality traits results and only after it we will examine the demographic ones. Our focus here will be the F1 score which measures the "real" classification accuracy, with it being the harmonic mean of the first two



```
gender,age,emo1,emo2,emo3,opn,cons,extr,adapt,nrtc,socialnet
0,2,3,0,1,2,3,2,2,3,0
1,1,3,4,2,3,3,2,1,3,1
0,1,3,4,2,3,3,3,3,3,1
1,2,3,4,5,1,3,3,1,3,1
0,3,5,3,4,2,2,2,2,2,0
1,1,3,4,2,3,2,3,1,3,1
1,1,3,4,2,2,2,3,2,3,1
1,1,3,4,5,2,3,3,1,3,1
```

Figure 20.: Representation of the dataset advantages. Red: No null cells (0 to 5 represent the 6 available emotions if presented in the top-3 preponderant emotions). Green: Age, already discretized, but with a bigger sample could be non discretized in classes and it doesn't repeats itself for multiple entries because every entry represents a different user. Yellow: No repetition on the personality traits throughout the dataset because, once again, each entry represents a different user.

metrics. We can have a precision metric with a high score, but it still can have a lot of misclassifications what would provoke the recall metric to score pretty low, and we don't want that. So, we focus on the parameter that gives us the model's accuracy and hopes to score well.

Precision Score on the Mixed Language dataset

Classification Algorithm	Extr	Agr	Cons	Neur	Opn	Age	Gender
	P R F1	P R F1	P R F1	P R F1	P R F1	P R F1	P R F1
Naive Bayes	0.35 0.45 0.37	0.64 0.71 0.64	0.58 0.39 0.40	0.57 0.61 0.58	0.63 0.64 0.62	1.0 1.0 1.0	1.0 1.0 1.0
J48	0.60 0.60 0.59	0.58 0.59 0.58	0.63 0.63 0.65	0.68 0.70 0.68	0.64 0.65 0.64	1.0 1.0 1.0	1.0 1.0 1.0
CART	0.60 0.60 0.59	0.58 0.59 0.58	0.63 0.63 0.65	0.66 0.70 0.64	0.64 0.65 0.64	1.0 1.0 1.0	1.0 1.0 1.0
SVC	0.58 0.58 0.58	0.65 0.65 0.65	0.60 0.63 0.61	0.69 0.71 0.65	0.63 0.64 0.63	1.0 1.0 1.0	1.0 1.0 1.0
K-Neighbors	0.51 0.52 0.50	0.58 0.59 0.58	0.58 0.59 0.58	0.64 0.66 0.64	0.62 0.56 0.56	1.0 1.0 1.0	1.0 1.0 1.0
Average	0.53 0.55 0.53	0.61 0.63 0.61	0.60 0.57 0.58	0.66 0.68 0.65	0.63 0.64 0.62	1.0 1.0 1.0	1.0 1.0 1.0

Table 5.: Self Performance (80-20) - Mixed Language

In this dataset, we opted by dividing the dataset in 80% for training and 20% for testing, to avoid underfitting and overfitting, which we achieved.

Naive Bayes: For this classifier we can see that it is the worst classifier for personality traits, averaging the worst scores between the 5 of them. It scores best on the agreeableness trait with a 64% accuracy on the classifications, actually on this trait it gets the second best result, behind SVC, mas it has the worst results on the extroversion and conscientiousness traits with 37% - this being the worst score on any trait and algorithm - and 40% accuracy in each, respectively, following by being the second worst on the openness trait with a respectful 62%.

J48 and CART: Both of these decision tree algorithms average almost the same results in every metric. The only significant difference is on the neuroticism trait where J48 scored 4% higher on the F1 metric than the CART algorithm, with 68% over the last 64%. Both share the best accuracy score on three of the five traits, extroversion with 59%, openness with 65% and openness with 63%. J48 stands alone at the top of the accuracy scores on the neuroticism trait with 68%, which is the best result for any trait in any of the five classifiers.

SVC: This algorithm scores the best result on the agreeableness trait with a good score of 65%, also scores the second best with the neuroticism trait at 65%, while scoring the third best score, behind J48 and CART, on the extroversion, , and openness with percentages of 58, 61 and 63, respectively.

K-Neighbors: The best score achieved by this algorithm is 65% on neuroticism, the same result as SVC on that trait. As for the rest of the traits, it scores pretty low on all of them, with 50% for extroversion, 58% on agreeableness and conscientiousness, and a low 56% on the openness trait.

As said above these values are not correct in any way. The way to achieve some real results as if we were to remove the personality traits and use only polarity and emotions presented in the user's text, we did that, and we got results rounding the 62% for gender and 78.5% for age, this last one would still be "rigged" as there are so few different age numbers that the model still overfits for this attribute, and we can only blame our sample size on that.

Precision Score on the User dataset

Naive Bayes: Once again this classifier achieves the worst overall results, except for the age and gender attributes where it has an accuracy of 73% and 95% which is very good. As for the personality traits, the best accuracy it has is on the extroversion trait with 52%

Classification Algorithm	Extr	Agr	Cons	Neur	Opn	Age	Gender
	P R F1	P R F1	P R F1	P R F1	P R F1	P R F1	P R F1
Naive Bayes	0.70 0.47 0.52	0.36 0.36 0.34	0.43 0.40 0.37	0.54 0.49 0.49	0.34 0.36 0.33	0.70 0.72 0.73	0.96 0.95 0.95
J48	0.86 0.82 0.81	0.55 0.55 0.59	0.42 0.36 0.38	0.55 0.55 0.56	0.27 0.27 0.27	0.27 0.27 0.25	0.82 0.82 0.82
CART	0.55 0.55 0.55	0.73 0.73 0.76	0.55 0.55 0.55	0.55 0.55 0.56	0.46 0.46 0.50	0.64 0.64 0.65	0.82 0.82 0.82
SVC	0.67 0.64 0.60	0.54 0.57 0.52	0.29 0.49 0.37	0.38 0.47 0.39	0.21 0.31 0.23	0.52 0.71 0.60	0.90 0.86 0.85
K-Nearest	0.61 0.55 0.48	0.57 0.62 0.57	0.64 0.47 0.47	0.44 0.49 0.44	0.52 0.47 0.43	0.88 0.86 0.87	0.67 0.80 0.73
Average	0.55 0.61 0.59	0.55 0.57 0.61	0.46 0.45 0.43	0.49 0.51 0.49	0.36 0.37 0.35	0.60 0.64 0.62	0.83 0.85 0.83

Table 6.: Self Performance (66-34) - Users

J48: This decision tree algorithm scores pretty high on the first trait, extroversion, and gender as well, achieving 81, 82% respectively, which are pretty high values. Due to the sample being small, it also has attributes that this algorithm scores poorly, having the worst results of all of them on openness and age with miserable 27 and 25%.

CART: Has the best accuracy for agreeableness, conscientiousness, and neuroticism with 76, 55 and 56% respectively. It is, overall, the best classifier for this dataset, it's worst accuracy is with the openness trait, scoring only 50%.

SVC: This algorithm scores the second best result on extroversion and gender with 60 and 85% but also scores the worst results for accuracy on openness and neuroticism with the low accuracy of 23 and 39 %.

K-Neighbors: Surprisingly the second best algorithm in general, scores the best accuracy on the age attribute with a solid 87% and the second best results on extroversion, conscientiousness, and openness with an accuracy of 48%, 47% and 43%. Although it is the second best overall, it still has poor results as we can deduce.

The best classifier is without a doubt, CART, except for the openness attribute it has pretty good classifying accuracy throughout the rest of the attributes. The rest of them are pretty weak, but as explained earlier we did what we could with what we had, and we wished for a more significant sample because the results would be, for sure, way better and more reliable.

4.5.2 *K-Fold Cross Validation*

Here we tried to optimize our results with the cross-validation method, and we're going to analyze each table and then compare the results from above.

Classification Algorithm	Extr	Agr	Cons	Neur	Opn	Age	Gender
	P R F1	P R F1	P R F1	P R F1	P R F1	P R F1	P R F1
Naive Bayes	0.36 0.45 0.37	0.64 0.64 0.64	0.56 0.38 0.39	0.58 0.61 0.58	0.63 0.65 0.63	1.0 1.0 1.0	1.0 1.0 1.0
J48	0.59 0.59 0.59	0.67 0.67 0.66	0.63 0.63 0.65	0.65 0.70 0.64	0.64 0.65 0.64	1.0 1.0 1.0	0.97 0.97 0.93
CART	0.59 0.60 0.59	0.67 0.67 0.66	0.62 0.63 0.61	0.65 0.70 0.64	0.64 0.65 0.64	1.0 1.0 1.0	0.97 0.97 0.93
SVC	0.54 0.55 0.55	0.67 0.67 0.66	0.60 0.63 0.62	0.67 0.70 0.64	0.64 0.65 0.64	1.0 1.0 1.0	0.92 0.90 0.89
K-Neighbors	0.55 0.55 0.54	0.60 0.59 0.59	0.64 0.53 0.54	0.63 0.65 0.55	0.56 0.58 0.55	1.0 1.0 1.0	1.0 1.0 1.0
Average	0.53 0.55 0.53	0.65 0.65 0.64	0.61 0.56 0.56	0.64 0.67 0.61	0.62 0.64 0.62	1.0 1.0 1.0	0.99 0.99 0.99

Table 7.: Ten-Fold Cross Validation - Mixed Language

We used a 10-fold cross validation metric, but our results were pretty much the same as when we went with the training and testing subsets mentioned above. The average scores for every attribute are the same or with a variance of one to three percent. The best classifier, in general, was the J48 decision tree algorithm as well.

The interesting thing with the 10-fold cross validation was that the classification of the demographic attribute, gender, was not 100% in every classifier as it was earlier, it decreased do 93% with the decision trees algorithms and to 89% with SVC, this could mean that these three algorithms were able to combat and reduce the overfitting presented on the analysis explored above, they still got some very high values, but surely are more close to be real values than before.

Classification Algorithm	Extr	Agr	Cons	Neur	Opn	Age	Gender
	P R F1	P R F1	P R F1	P R F1	P R F1	P R F1	P R F1
Naive Bayes	0.78 0.55 0.52	0.26 0.32 0.26	0.58 0.59 0.56	0.54 0.55 0.54	0.22 0.29 0.24	0.74 0.81 0.77	0.97 0.97 0.97
J48	0.65 0.65 0.64	0.65 0.65 0.69	0.45 0.40 0.41	0.43 0.48 0.44	0.26 0.29 0.27	0.65 0.65 0.64	0.71 0.71 0.71
CART	0.61 0.61 0.60	0.71 0.71 0.76	0.58 0.58 0.58	0.33 0.55 0.41	0.42 0.42 0.41	0.74 0.74 0.74	0.67 0.68 0.66
SVC	0.63 0.61 0.58	0.59 0.61 0.55	0.30 0.42 0.35	0.40 0.52 0.42	0.43 0.42 0.37	0.55 0.75 0.64	0.92 0.90 0.89
K-Neighbors	0.57 0.55 0.52	0.60 0.65 0.60	0.35 0.35 0.35	0.45 0.58 0.50	0.26 0.35 0.28	0.57 0.75 0.65	0.92 0.90 0.89
Average	0.52 0.59 0.57	0.56 0.59 0.57	0.45 0.47 0.45	0.43 0.54 0.45	0.32 0.35 0.31	0.65 0.74 0.69	0.84 0.83 0.82

Table 8.: Three-Fold Cross Validation - Users

We opted for a 3-fold cross validation instead of a 10-fold, because of the sample size, with a 10 fold there would be basically nothing to train or test, and even with a 3-fold we just hoped it would help to deal with under and overfitting if that was the case.

Naive Bayes: Has the best score for gender classification accuracy with 97%, has average scores for extroversion, conscientiousness, and neuroticism, all rounding 50-55%, and has the worst scores for agreeableness and openness as it did on the other experiment.

J48: This one improves it's accuracy scores on three attributes, and worsens on another three, with openness staying the same with 27%. It has the best score for extroversion with 64% and the second best for agreeableness with 69%.

CART: With the 3-fold, it still has the best result for agreeableness with the same 76% accuracy score, and it is also the best classifier of conscientiousness and openness with 58 and 41%. Achieves the second-best results in extroversion and gender as well. Compared with the self-performance, it improves on three attributes, ties on one, and worsens on three as well.

SVC: This algorithm didn't react well with the 3-fold approach, it worsened every accuracy score except agreeableness, in which it improved by 3% to 55. It has some of the lowest classifications scores in conscientiousness, neuroticism, and openness with 35, 42 and 37% classification accuracy in each.

K-Neighbors: Continued to be one of the worst classification algorithms for this dataset, although it increased 16% on gender classification accuracy for 89%, it still has pretty bad and average scores for the rest of the attributes.

As we can conclude, performing the 3-fold cross validation method didn't improve our results, on average it decreased. CART algorithm didn't perform much worse, but still has attributes with some low accuracies.

Some of these classification percentages shine a light on the potential results that could have been achieved on this dataset, with a bigger and broader sample.

4.6 SUMMARY

In this final section of the sixth chapter, we will present an overview of the content of this chapter in a more summarized way. We started by detailing the way we were gonna evaluate our models and classification results on the "model evaluation" section. We opted to use three of the most used metrics to perform a proper evaluation, and they are, precision, recall, and f1-score. Following that, there's an explanation on what methods we used to train our datasets, and we went with the necessary train-test split and with k-fold cross validation for both of our datasets hoping to achieve different and better results with the latter method. There is a profound detailing of our datasets' sample. We compared and studied it by social network (comparing Twitter with Facebook users) and by gender (comparing male with female users) on what their average emotions and expressed emotions were, and we found some interesting relations between the lot. We end that section by showing relations between users' expressed emotions and their personality and demographic traits to find correlations between them. We end the chapter with the results and discussion section, and we show our results for two datasets, each one with two sets of results, one for each method applied to split the dataset. Our conclusion from our results is that the best algorithm classifier and most trustworthy is J48, followed by CART, which are both decision trees algorithms, and the worst is Naive Bayes and K-Neighbors clustering algorithm.

5

CONCLUSION

5.1 INTRODUCTION

In this chapter, we will make our conclusions on the work we made so far, and we will give an overview on what needs to be done on future work, and our expectations on that matter. Conclusions and future work.

5.2 CONCLUSIONS

The objective questions proposed at the beginning of this thesis were answered throughout the work and by our results. It is possible, with a medium-high percentage, to predict psycho-demographic profiles from a text emotion analysis on users social network data, as we can see from the results obtained by the J48 algorithm on the "Mixed Language" dataset with percentages of 59, 64, 65, 64 and 64 throughout all of the five personality traits, extroversion, agreeableness, conscientiousness, neuroticism, and openness, respectively. And by analyzing the "User" dataset we can see some really good percentages on the demographic attributes on all the classifying algorithms, with the worst score being 66% from CART and the best being 97% from naive Bayes on the gender attribute, and for the age attribute, 64% was the lowest score on J48 and Naive Bayes and the best scores being 74 and 77% with Naive Bayes and CART algorithms.
As for the second question, "Is it possible to predict emotions from text in order to find correlations with users personalities?" the answer is also positive, there was a high correlation from some emotions, as joy, sadness, and fear, with some personality traits, like conscientiousness, agreeableness, neuroticism having higher correlation percentages.

We faced some problems that are mentioning below, and also found that there's still so much more to do in this area that we'll address as future work to be done.

5.3 DISCUSSION

Here we will discuss the results we got and also address some of the encountered problems throughout the way.
Regarding the results, we concluded that we fulfilled our goal. They might not be the best results (problem addressed below), or the results we expected, but with the data that we could gather we think that it's more than enough to prove that there are correlations between text emotions and psycho-demographic profiling, with a more solid sample with higher chances of finding correlations and more relations between attributes and classes we think that the results would be way higher and more consistent. Still, there were some good results, rounding high sixties and low seventies in the *mixed language dataset* for the personality traits, and again some eighties, seventies, and high sixties mostly on demographic traits, but also in some personality characteristics on the *user dataset*.
The significant problems we faced were the lack of content on our data, not having a sample composed, minimum, by a couple of hundred of users prejudiced and might have rigged our results or some of them at least. We had problems in both datasets because we had to change some of our original thoughts on how to address our primary goal. For the "User" dataset, our problem is that we only have a dataset with 32 entries, the results are a bit dispersed, from the classifier to classifier, and sometimes, even in the same classifier but for different traits. On the "Mixed Language" user, our problem is with the demographic traits, and it's already addressed in chapter six. We made this dataset and used to try to get our proposed goals achieved, but this was an emergency approach as our number of users was concerning low, so we had some overfitting problems as to predict demographic attributes. We would like to have used deep learning and neural networks for our result, but we would have to have a broader, broader sample. So that was unfortunate as well.

5.4 PROSPECT FOR FUTURE WORK

For future work, the goal is definitely to improve our results, that can be done by just having a solid, deep sample of users because that will provide more information for our classifiers to work with. As stated above we want to be able to apply neural networks, as we think it might be the best solution, with more data we could think of applying deep learning techniques but that was just a waste of time with what we had.

We have an interest in continuing to work in this area of data mining, and psycho-demographic profiling. This has much potential once done correctly, one of the goals for the future is to relate these psycho-demographic profiles and these expressed emotions with behaviours, tastes and so on. There is much to learn and to research on psychology by analyzing users online behaviours and expressions. We didn't have the means for it as we mentioned in the section above, but initially we thought about making some kind of software implementation, that after performing the user's profiling, it would relate it with a particular group of people, in order help the user with suggestions of some sort, per example, let's say a person with a high score on the extroversion trait is a person that likes a particular type of video-games or activities, that we can know by analyzing their online data and text, if a user had a similar profile the software would suggest video-games for that specific profile type, like creating some "online suggester" with this information, although there is much to improve and there's still a long way to go.

REFERENCES

C. Alm, D. Roth, and R. Sproat. *Emotions from text: machine learning for text-based emotion prediction*. Human Language Technology Conference, 2005.

S. Aman and S. Szpakowicz. *Identifying Expressions of Emotion in Text*. 2007.

M. Barrick and M. Mount. *The Big-Five Personality Dimensions And Job Performance: A meta-analysis*. Personnel Psychology - Journal, 1991.

J. Bollen, H. Mau, and A. Pepe. *Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena*. Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 2011.

T. Bouchard. *Genes, Environment and Personality*. Science, 1994.

L. Breiman, J. Friedman, C. Stone, and R. Olshen. *Classification and Regression Trees*. 1984.

S. Budaev. *Sex diferences in the Big Five personality factors: Testing an evolutionary hypothesis*. Personality and Individual Differences, 1998.

J. Burger, J. Henderson, and G. Zarrella. *Discriminating Gender on Twitter*. Proceeding EMNLP '11 Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011.

J. Chang, Itamar Rosenn, Lars Backstrom, and Cameron Marlow. *ePluribus: Ethnicity on Social Networks*. Association for the Advancement of Artificial Intelligence, 2010.

M. Choudhury, M. Gamon, and S. Counts. *Happy, Nervous or Surprised? Classification of Human Affective States in Social Media*. Association for the Advancement of Artificial Intelli gence, 2012.

M. Conover, J. Rakiewicz, B. Francisco, A. Goçalves, A. Flammini, and D. Menezer. *Political Polarization on Twitter*. International Conference on Web and Social Media, 2011a.

M. Conover, J. Rakiewicz, A. Goçalves, A. Flammini, and D. Menezer. *Predicting the Political Alignment of Twitter Users*. IEEE, 2011b.

N. Conrad and M. Patry. *Conscientiousness and Academic Performance: A Mediational Analysis*. International Journal for the Scholarship of Teaching and Learning, 2012.

G. Coppersmith, M. Dredze, and C. Harman. *Quantifying Mental Health Signals in Twitter*. ACL, 2014.

A. Culotta, N. Ravi, and J. Cutler. *Predicting the Demographics of Twitter Users from Website Traffic Data*. 2015. ISBN Proceeding AAAI'15 Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence.

L. Dabbish, Colleen Stuart, Jason Tsay, and Jim Herbsleb. *Social Coding in GitHub: Transparency and Collaboration in an Open Software Repository*. ACM 2012 conference on Computer Supported Cooperative Work, 2012.

A. Damassio. *Emotion in the perspective of an integrated nervous system*. Brain Research Reviews, 1998.

V. Dhar. *Data Science and Prediction*. Magazine Communications of the ACM, 2013.

J. Djicks. *Oracle: Big Data for the Enterprise*. Oracle Enterprise Research Architecture White Paper, 2013.

M. Dredze, T. Oates, and C. Piatko. *We're Not in Kansas Anymore: Detecting Domain Changes in Streams*. 2010. ISBN Conference on Empirical Methods in Natural Language Processing.

P. Ekman. *An argument for basic emotions. Cognition Emotion*. Routledge, 1992.

T. Fletcher. *Support Vector Machines Explained*. 2009.

H. Fromreid, D. Hovi, and A. Soogard. *Crowdsourcing and annotating NER for Twitter drif*. 2014. ISBN European language resources distribution agency.

S. Gauch, Mirco Speretta, Aravind Chandramouli1, and Alessandro Micarelli. *User Profiles for Personalized Information Access*. Springer, Berlin, Heidelberg, 2007.

V. Gayatri and J. Dattetrya. *Applying Classification Technique using DID3 Algorithm to improve Decision Support System under Uncertain Situations*. International Journal for Research in Engineering Application Management, 2017.

F. Girardin, Josep Blat, Francesco Calabrese, Filippo Dal Fiore, and Carlo Ratti. *Digital Footprinting: Uncovering Tourists with User- Generated Content*. Journal of Travel Research, 2014.

J. Golbeck, C. Robles, M. Edmondson, and K. Turner. *Predicting Personality from Twitter*. IEEE International Conference on Social Computing, 2011.

L. Goldberg. *Language and Individual Differences: The Search for Universals in Personality Lexicons*. APA science Vols. Studying lives through time: Personality and development, 1981.

L. Goldberg. *An Alternative "Description of Personality": The Big-Five Factor Structure*. Journal of Personality and Social Psychologs, 1990.

L. Goldberg. *The Structure of Personality Traits: Vertical and Horizontal Aspects*. APA science Vols. Studying lives through time: Personality and development, 1993a.

L. Goldberg. *The Structure of Phenotypic Personality Traits*. American Psychologist, 1993b.

L. Goldberg. *Demographic Variables and Personality: The Effects of Gender, Age, Education and Ethnic/Racial Status on Self-Descriptions of Personality Attributes*. Personality and Individual Differences, 1998.

M. Gomes, J. Alfonso-Cendón, P Marqués-Sánchez, D. Carneiro, and P. Novais. *Improving Conflict Support Environments with Information Regarding Social Relationships*. Springer International Publishing Switzerland, 2014a.

M. Gomes, T. Oliveira, D. Carneiro, P. Novais, and J. Neves. *Studying The Effects of Stress on Negotiation Behavior*. 2014b.

S Gosling. *Personality Impressions Based on Facebook Profiles*. ICWSM, 2007.

S. D. Gosling, P. J. Rentfrow, and W. Swann. *A Very Brief Measure of the Big Five Personality Domains*. Journal of Research in Personality, 2003.

S. Gunn. *Support-Vector Networks Support Vector Machines for Classification and Regression*. University Of Southampton, Technical Report, 1998.

Hendriks and A. Jolijn. *The construction of the five-factor personality inventory (FFPI*. University of Groningen, 1997.

Hendriks and H. Raad. *The Five-Factor Personality Inventory: assessing the Big Five by means of brief and concrete statements*. 2002.

J. Jacob B. Hirsh, C. DeYoung, and J. Peterson. *Metatraits of the Big Five Differentially Predict Engagement and Restraint of Behavior*. Journal of Personality, 2009.

B. Jeronimus, H. Riesse, R. Sanderman, and J. Ormel. *Mutual Reinforcement Between Neuroticism and Life Experiences: A Five-Wave, 16-Year Study to Test Reciprocal Causation*. Journal of Personality and Social Psychology, 2014.

Oliver P. John and S. Srivastava. *The Big-Five Trait Taxonomy:History, Measurement, and Theoretical Perspectives*. University of California at Berkeley, 1999.

T. Judge, J. Bono, and C. Thoresen. *Are Measures of Self-Esteem, Neuroticism, Locus of Control, and Generalized Self-Efficacy Indicators of a Common Core Construct?* Journal of Personality and Social Psychology, 2002.

C. Jung. *Psychology Types*. Princeton University Press, 1971.

L. Kaelbing, M. Liitman, and A. Moore. *Reinforcement Learning: A Survey*. Journal of Artificial Intelligence Research, 1996.

S Kim, J. Bak, and A. Oh. *Do You Feel What I Feel? Social Aspects of Emotions in Twitter Conversations*. Association for the Advancement of Artificial Intelligence, 2012.

S.K. Korting. *C4.5 algorithm and Multivariate Decision Trees*. National Institute for Space Research, 2014.

M. Kosinski, Yoram Bachrach, Thore Graepel, Pushmeet Kohli, and David Stillwell. *Personality and Patterns of Facebook Usage*. Web Science, 2012.

M. Kosinski, Bin Bi, Milad Shokouhi, and Thore Graepel. *Inferring the demographics of search users: Social data meets search queries*. World Wide Web Conference Committee, 2013a.

M. Kosinski, D. Stillwell, and T. Graepel. *Private traits and attributes are predictable from digital records of human behavior*. Proceedings of the National Academy of Sciences of the United States of America, 2013b.

S. Kotsiantis. *Supervised Machine Learning: A Review of Classification Techniques*. Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in

Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies, 2007.

S. Krišto. *The Relationship between Extroversion/Introversion, Perceptual Learning Styles and Success in English as a Foreign Language*. 2012.

R. Kumar, S. Ramanihan, and H. Albalooshi. *EmotionX-SmartDubai*$_N$*LP* : *DetectingUserEmotionsInSocialMediaText.ProceedingsoftheSixthInternationalWorkshoponNaturalLang*

D. Laney. *3d data management: Controlling data volume, velocity and variety*. Meta Group, 2001.

R. Lazarus and C. Smith. *Emotion and Adaptation*. Oxford, Oxford University Press, 1991.

H. Liu, F. Hussain, C. Tan, and M. Dash. *Discretization: An Enabling Technique*. Kluwer Academic Publishers, 2002.

M Loukides. *WHAT IS DATA SCIENCE?* O'Reilly Media, Inc., 2013.

M Madden, S. Fox, A. Smith, and J. Vitak. *Online identity management and search in the age of transparency*. Pew Internet American Life Project, 2007.

S. Mohammad and S. Kiritchenko. *Using Hashtags to Capture Fine Emotion Categories from Tweets*. Semantic Analysis in Social Media, Computational Intelligence, 2014.

S. Mohammad and T. Yang. *Tracking Sentiment in Mail: How Genders Differ on Emotional Axes*. 2011.

N. Muniz. *Prediction and Ranking of Highly Popular Web Content*. Faculdade Ciencias Universidade do Porto, 2009.

I. Myers and P. Myers. *Gifts Differing: Understanding Personality Type*. CPP Books, 1980.

NationalResearchCouncil. *3 Principles for Working with Big Data*. The National Academic Press, 2015.

D. Nguyen, A. Smith, and C. Rosé. *Author Age Prediction from Text Using Linear Regression*. Proceeding LaTeCH '11 Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, 2011.

E. Noftle and R. Robins. *Personality Predictors of Academic Outcomes: Big Five Correlates of GPA and SAT Scores*. Journal of Personality and Social Psychology, 2007.

P. Novais, D. Carneiro, J. Pego, N. Sousa, and J Neves. *Using Mouse Dynamics to Assess Stress During Online Exams*. International Conference on Hybrid Artificial Intelligence Systems, 2015.

B. O'Connor, J. Eisenstein, E. Xing, and A. Smith. *A Mixture Model of Demographic Lexical Variation*. JMLR: Workshop and Conference Proceedings, 2011.

I. Omelianenko. *Applying Deep Machine Learning for psycho-demographic profiling of Internet users using O.C.E.A.N. model of personality*. International Journal for Research in Engineering Application Management, 2017.

I. Onder, W. Koerbitz, and A. Hubmann-Haidvogel. *Tracing Tourists by Their Digital Footprints: The Case of Austria*. Journal of Travel Research, 2014.

S. Ones, C. Viswesvaran, and D. Reiss. *Role of social desirability in personality testing for personnel selection: The red herring*. Journal of Applied Psychology, 1996.

M. Osborn, A. Lall, and B. Durme. *Exponential Reservoir Sampling for Streaming Language Model*. 2014. ISBN Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.

J. Pennebaker, C. Chung, Irelandm M., A. Gonzales, , and R. Booth. *The Development and Psychometric Properties of LIWC2007*. LIWC.net, Austin, Texas, 2007.

W. Powers. *Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness Correlation*. Bionfo Publications, 2011.

F Provost and T. Fawcett. *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O'Reilly Media, Inc., 2013.

A Qadir and E. Riloff. *Bootstrapped learning of emotion hashtags hashtags4you*. Proceedings of WASSA, 2013.

D. Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. *Our Twitter Profiles, Our Selves: Predicting Personality with Twitter*. IEEE, 2011.

R. Quinlan. *C4.5: programs for machine learning*. 1993.

A. Rajalakshmi, R. Vinodhini, and F. Bibi. *Data Discretization Technique Using WEKA Tool*. IJCSET, 2016.

D. Rao, D. Yarowski, A. Shreevats, and M. Gupta. *Classifying Latent User Attributes in Twitter*. Proceeding SMUC '10 Proceedings of the 2nd international workshop on Search and mining user-generated contents, 2010.

K. Roberts, Roach A., J. Johnson, J. Guthrie, and M. Harabagiu. *Empatweet: Annotating and detecting emotions on Twitter.* Proceedings of LREC, 2012.

S. Roccas, L. Sagiv, S. Schwartz, and A. Knafo-Noam. *The Big Five Personality Factors and Personal Values*. Personality and Social Psychology, 2002.

A. Samuel. *Some Studies in Machine Learning Using the Game of Checkers*. IBM journal, 1959.

M. Sap, G Park, J. Eichstaedt, M. Kern, M. Stillwell, M. Kosinski, L. Ungar, and H. Schwartz. *Developing Age and Gender Predictive Lexica over Social Media*. 2014. ISBN Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing.

L Schmit. *Personality and its Effects on Facebook and Self-Disclosure*. Produced in Katherine Curtis's Spring 2012 ENC1102, 2012.

G. Seni and J. Elder. *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions*. Morgan Claypool, 2010.

J. Senise. *Who Is Your Next Customer?* Booz Allen Hamilton Inc, 2007.

G. Sharma, R. Bhargava, and M. Mathuria. *Decision Tree Analysis on J48 Algorithm for Data MiningDr*. International Journal of Advanced Research in Computer Science and Software Engineering, 2013.

S. Singh and P. Gupta. *COMPARATIVE STUDY ID3, CART AND C4.5 DECISION TREE ALGORITHM: A SURVEY*. International Journal of Advanced Information Science and Technology, 2014.

L. Sloan, J. Morgan, P. Burnap, and M. Williams. *Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data*. PLoS ONE 10(3): e0115545. https://doi.org/10.1371/journal.pone.0115545, 2015.

C. Strapparava and R. Mihalcea. *Learning to Identify Emotions in Text*. 2008.

SWGDE. *Digital Evidence: Standards and Principles*. 2000.

Z. Tufekei. *Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls*. 204. ISBN Proceedings of the 8th International AAAI Conference on Weblogs and Social Media.

E. Tupes and R. Christal. *Recurrent Personality Factors Based on Trait Ratings*. Journal of Personality, 1961.

D. Turner, M. Schroeck, and R. Shockley. *Analytics: The real-world use of big data*. IBM Institute for Business Value, 2012.

V. Vapnik and C. Cortes. *Support-Vector Networks*. Kluwer Academic Publishers, 1995.

V. Vapnik, I. Guyon, and B. Boser. *A Training Algorithm for Optimal Margin Classiers*. Proceeding COLT '92 Proceedings of the fifth annual workshop on Computational learning theory, 1992.

V. Vapnik, H. Drucker, C. Burges, Kaufman. L, and A. Smola. *Support Vector Regression Machines*. Proceeding NIPS'96 Proceedings of the 9th International Conference on Neural Information Processing Systems, 1996a.

V. Vapnik, S. Golowich, and A. Smola. *Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing*. Proceeding NIPS'96 Proceedings of the 9th International Conference on Neural Information Processing Systems, 1996b.

G. Volkova, Y. Bachrach, and B. Durme. *Mining User Interests to Predict Perceveid Psycho-Demographic Traits on Twitter*. Big Data Computing Service and Applications (BigDataService), 2016 IEEE Second International Conference, 2016.

S. Volkova. *Predicting Demographics and Affect in Social Networks*. 2015. ISBN The Johns Hopkins University.

M. Waller and S. Fawcett. *Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management*. Journal of Business Logistics, 2013.

J. Ward and A. Barker. *Undefined By Data: A Survey of Big Data Definitions*. 2013.

S. D. Wells. *Psychographics: A Critical Review*. Journal of Marketing Research, 1975.

W Wemb, C. Lu, K. Thirunarayan, and A. Sheth. *Harnessing Twitter 'Big Data' for Automatic Emotion Identification*. Proceedings of SocialCom, 2012.

S. Zafar, A. Khan, and K Meenakshi. *Extraversion-Introversion Tendencies and their Relationship with ESL Proficiency: A Study of Chinese Students in Vellore, India*. Pertanika Journals, 2017.

O. Zezulka. *The Digital Footprint and Principles of Personality Protection in the European Union*. Charles University, 2016.

H. Zhang. *The Optimality of Naive Bayes*. FLAIRS2004 conference, 2004.

Z Zhu and Xiong. Y. *Defining Data Sciences*. 2015.

A

SUPPORT MATERIAL

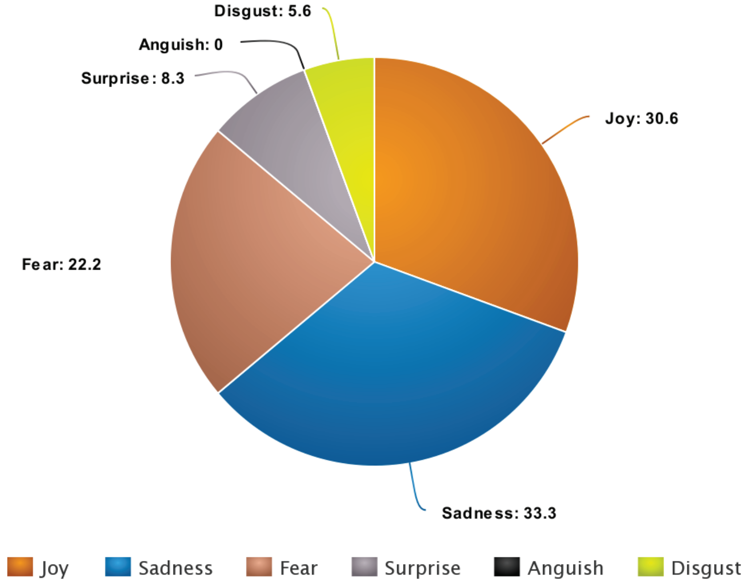


Figure 21.: Representation of female user's emotions

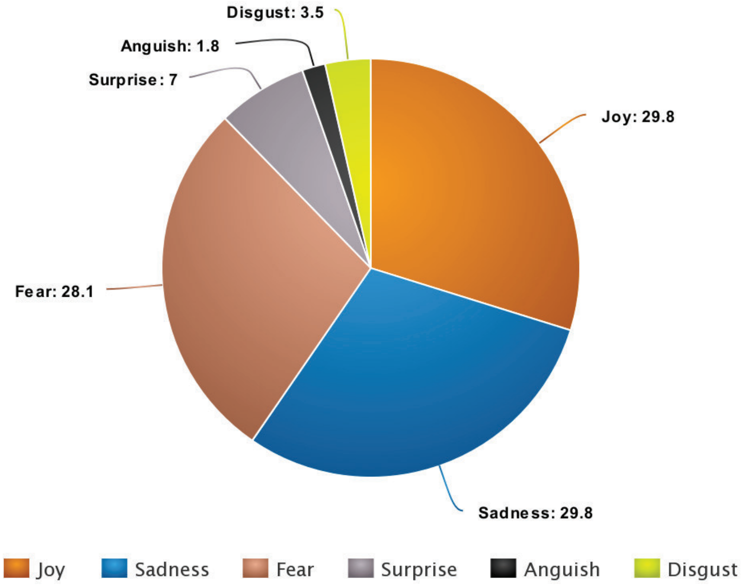


Figure 22.: Representation of male user's emotions

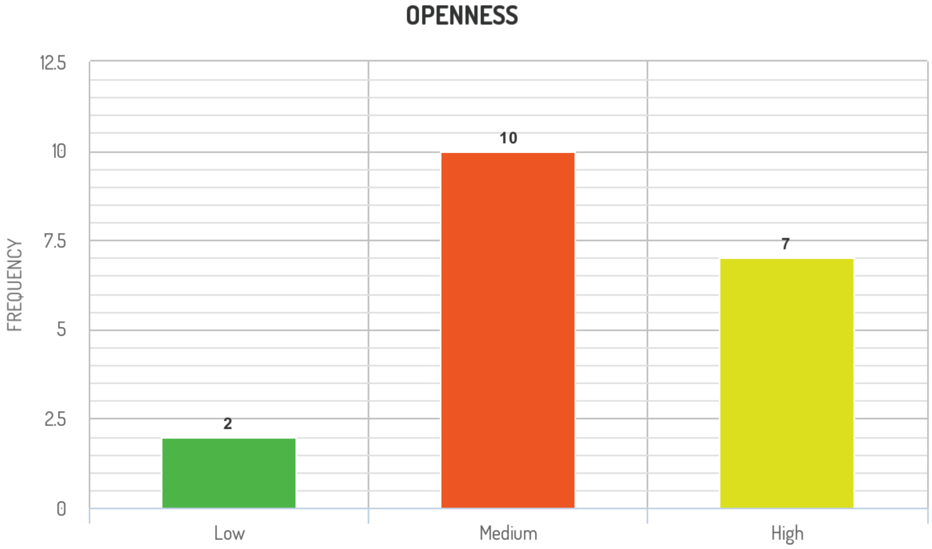


Figure 23.: Representation of male user's openness trait frequency

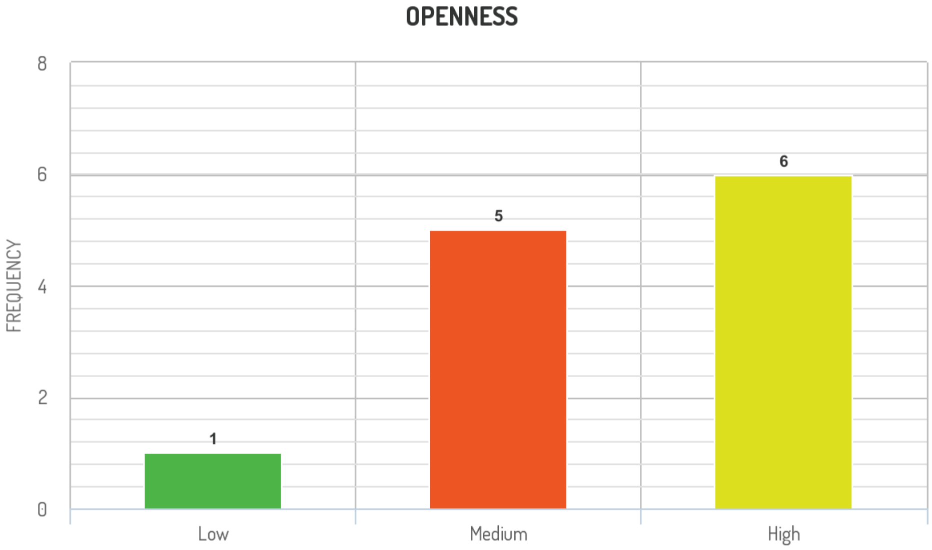


Figure 24.: Representation of female user's openness trait frequency

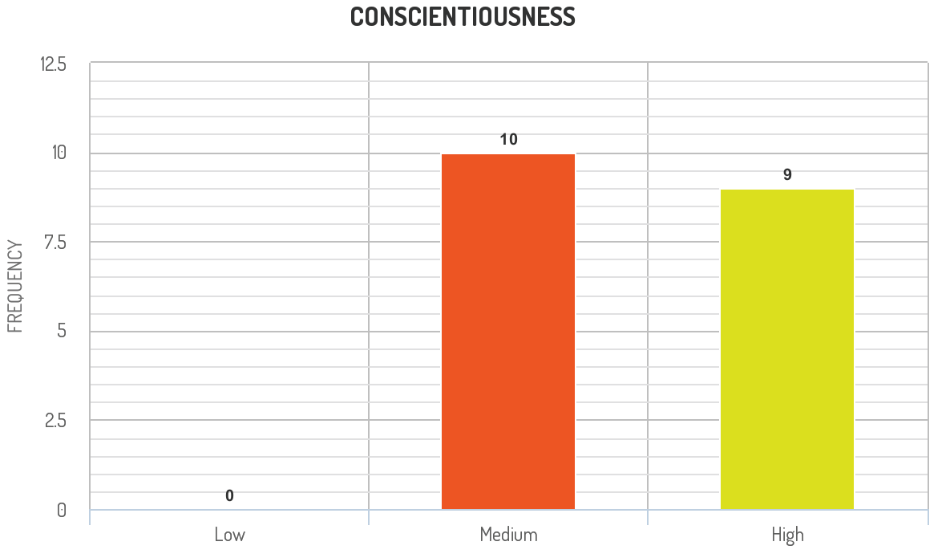


Figure 25.: Representation of male user's conscientiousness trait frequency

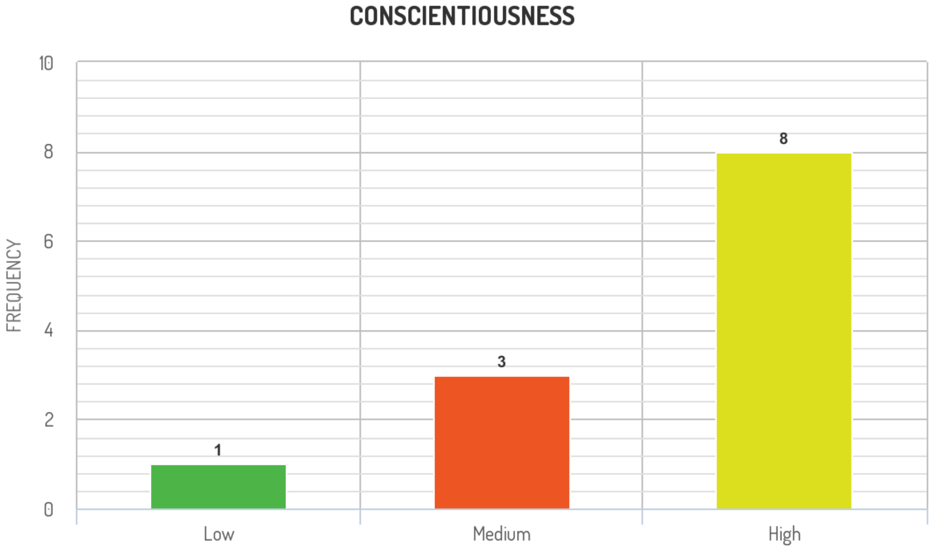


Figure 26.: Representation of female user's conscientiousness trait frequency

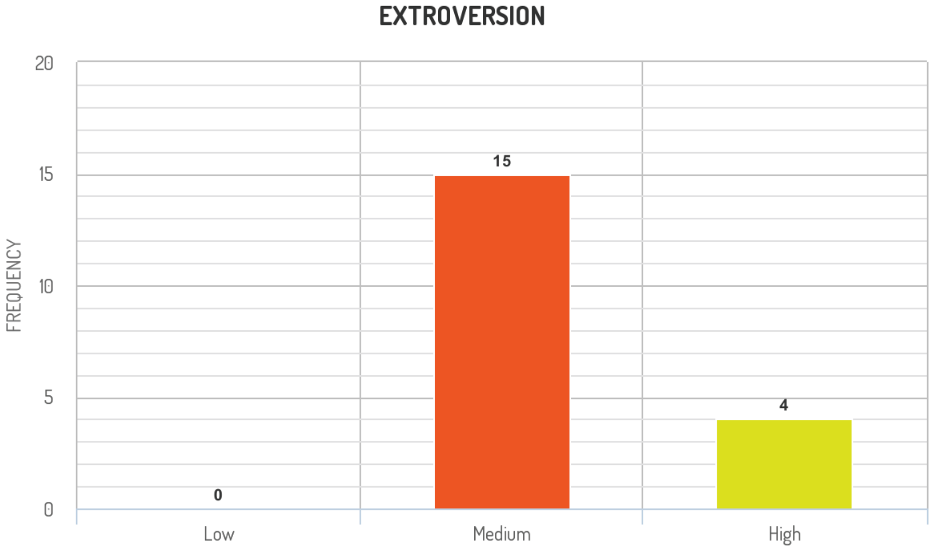


Figure 27.: Representation of male user's extroversion trait frequency

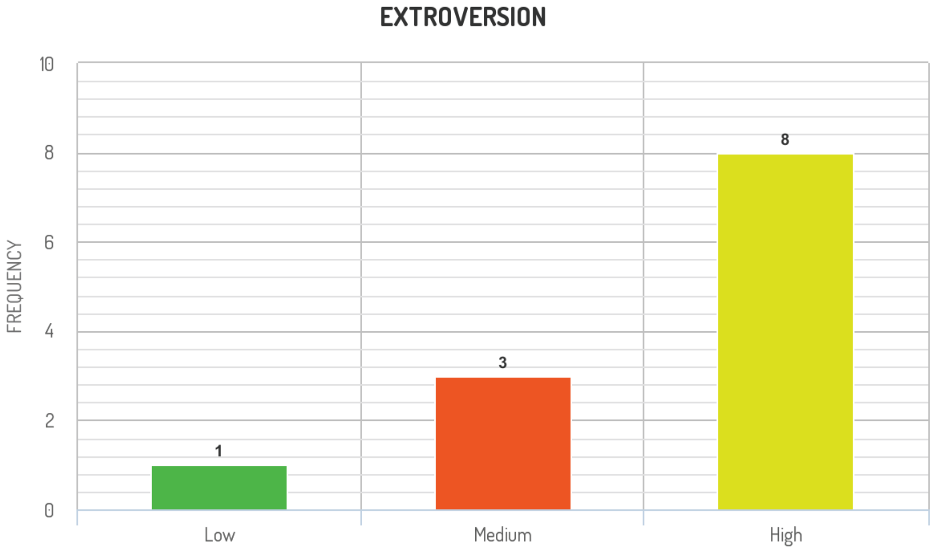


Figure 28.: Representation of female user's extroversion trait frequency

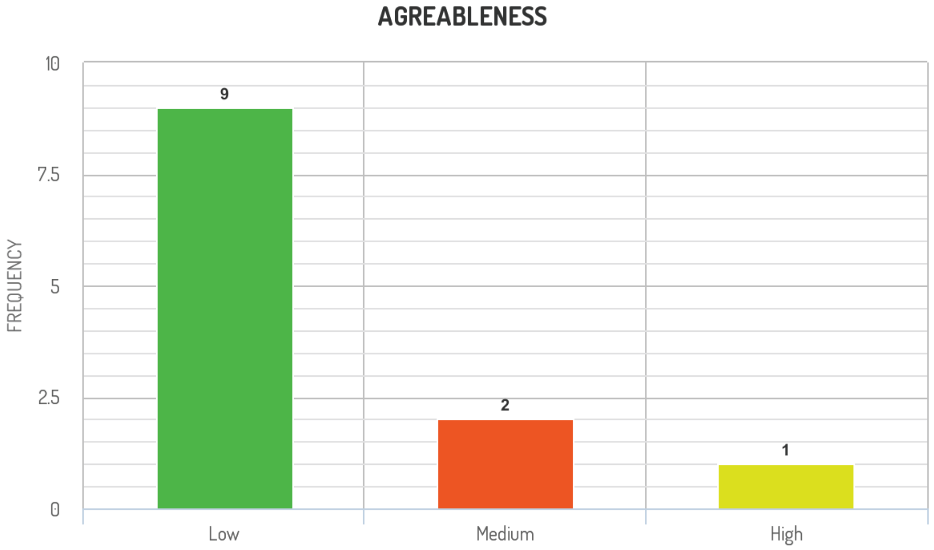


Figure 29.: Representation of male user's agreableness trait frequency

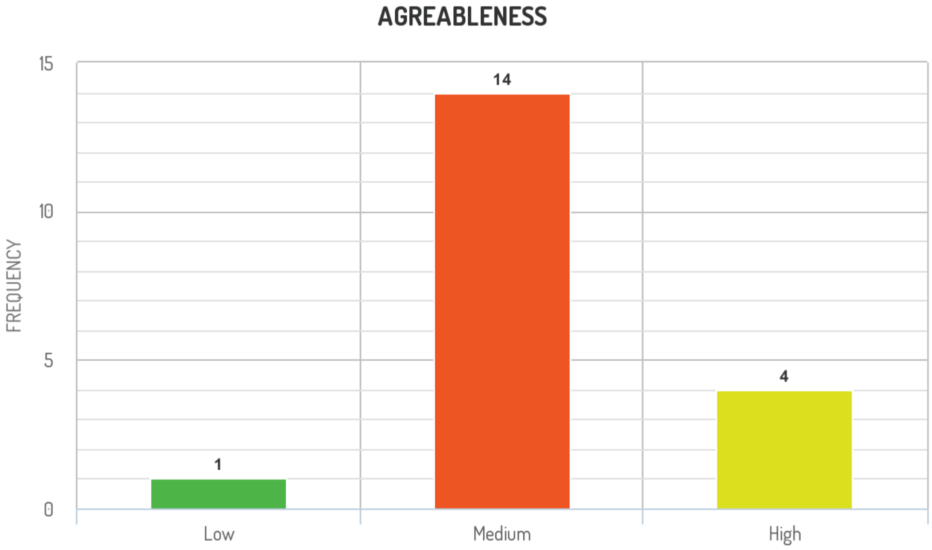


Figure 30.: Representation of female user's agreableness trait frequency

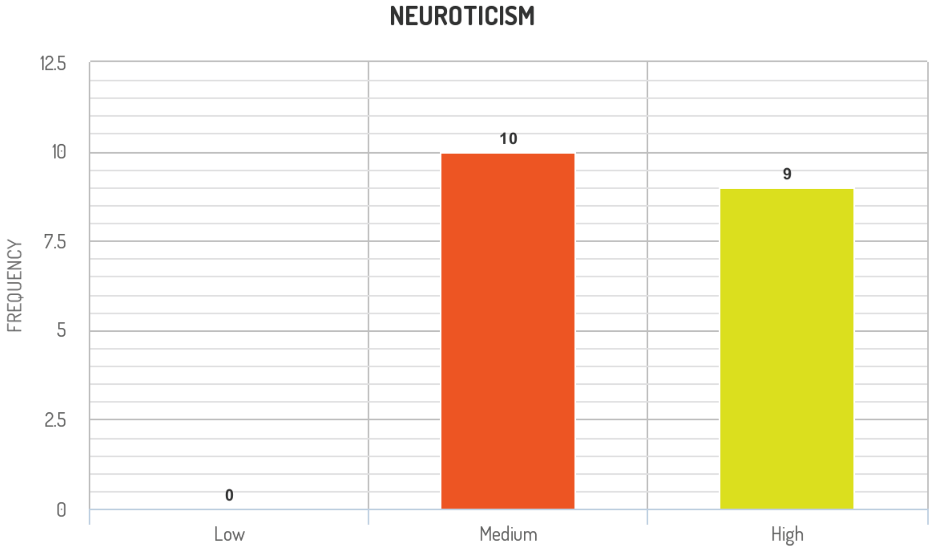


Figure 31.: Representation of male user's neuroticism trait frequency

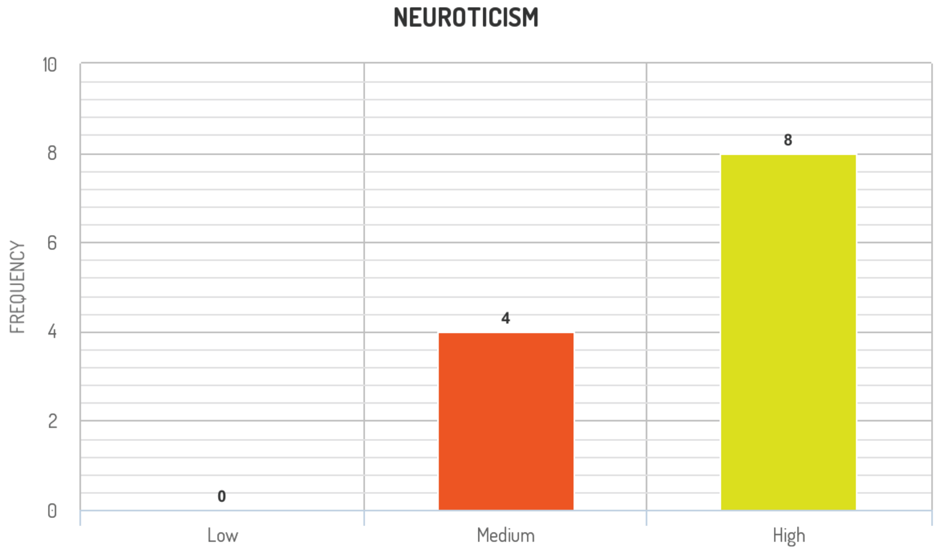


Figure 32.: Representation of female user's neuroticism trait frequency