Université de Montréal

Personality Extraction Through LinkedIn

par

Frédéric Piedboeuf

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures et postdoctorales
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Informatique

Mai 2019

Université de Montréal

Faculté des études supérieures et postdoctorales

Ce mémoire intitulé

Personality Extraction Through LinkedIn

présenté par

Frédéric Piedboeuf

a été évalué par un jury composé des personnes suivantes :

Michel Boyer
(président-rapporteur)

Philippe Langlais
(directeur de recherche)

Guy Lapalme
(co-directeur)

Nadia El Mabrouk
(membre du jury)

Mémoire accepté le :
Juin 2019

Sommaire

L'extraction de personnalité sur les réseaux sociaux est un domaine qui n'a que récemment commencé à capturer l'attention des chercheurs. La tâche consiste à, en partant d'un corpus de profils d'utilisateurs de réseaux sociaux, être capable de classifier leur personnalité correctement, selon un modèle de personnalité tel que défini en psychologie. Ce mémoire apporte trois innovations au domaine. Premièrement, la collecte d'un corpus d'utilisateurs LinkedIn. Deuxièmement, l'extraction sur deux modèles de personnalités, MBTI et DiSC, l'extraction sur DiSC n'ayant pas encore été faite dans le domaine, et finalement, la possibilité de passer d'un modèle de personnalité à l'autre est explorée, dans l'idée qu'il serait ainsi possible d'obtenir les résultats de multiples modèles de personnalités en partant d'un seul test.
Mots-clés: Extraction de personnalité, MBTI, DiSC, LinkedIn, Réseau sociaux, profilage d'auteur.

Summary

Personality extraction through social networks is a field that only recently started to capture the attention of researchers. The task consists in, starting with a corpus of user profiles on a particular social network, classifying their personalities correctly, according to a specific personality model as described in psychology. In this master thesis, three innovations to the domain are presented. Firstly, the collection of a corpus of LinkedIn users. Secondly, the extraction of the personality according to two personality models, DiSC and MBTI, the extraction with DiSC having never been done before. Lastly, the idea of going from one personality model to the other is explored, thus creating the possibility of having the results on two personality models with only one personality test.
Keywords: Personality extraction, MBTI, DiSC, LinkedIn, social network, author profiling.

Contents

List of tables

List of figures

Acknowledgements

Without a doubt, these years have been for me the most formative and the most transformative of my life - the best and the worst. I met incredible people and had mind blowing discussions, hours of fun, and hours of intense work. I started unsure of what I wanted to do, and fell in love with the field, and the research. And through all of it I have many people responsible for it.

First of all of course, I want to thank my director Philippe Langlais, who helped me at first restrain my wild reaseach ideas to focus and settle down on the good ones, and then who, in an unknown partnership with my mother, who I also thank infinitely, had to go through my writing riddled with the most horrendous mistakes. Then I want to thank my co-director, Guy Lapalme, as well as the Michel Boyer for their advices and comments, and all the people at the RALI who helped me forward my thinking and my project by endless discussion, sometimes non-sensical, sometimes about the deep sense of life, especially Ilan, David, Fabrizio, Louis and Zack.

Introduction

The possibility of understanding human behavior through a set of fixed rules has long been of interest to researchers. To do so, they developed what are called personality models, which try to capture the different aspects of the personality of a person. The first personality model was created in 1917, when the American army tried to identify soldiers that would be more susceptible to nervous breakdowns during bombardment [Gibby and Zickar, 2008]. Since then, psychologists have developed better personality models that take in more dimensions, and computer scientists have started to use artificial intelligence to extract the personality from text, image, video, sound, or other.

The personality of a person is usually described as the set of responses to external stimuli [Kaushal and Patwardhan, 2018], meaning that with a fully realized personality model, it should be possible to correctly predict how someone would react to a given event. Obviously, humans are very complex machines and our personality models are pretty far from capturing that complexity, but given enough time, and with potentially the help of computers, it may be possible to come close. Different models of personality are presented in Section 1.1.

Although it is not possible yet to determine what a person will do in certain conditions knowing their personality, the inverse is possible. Personality extraction is the task of trying to determine the personality of a person based on something they did. That something can be some text they have written, some conversations they had, or even their facial expressions.

One source of information computer scientists use for personality extraction is social networks. The recent advent of social networks, such as Twitter or Facebook, has given ways for computer scientists to have access to vasts amounts of information in record time. In fact, it is estimated that, in 2014, there were 30 petabytes of data on Facebook, and that each day there is a new 4 petabytes of data generated by its users [Wiener and Bronson, 2014]. To get that much information on people before the advent of social networks would have been virtually impossible.

Facebook and Twitter are probably the social networks the most used for the task of personality extraction, since they are the most popular [Maina, 2016], and they provide an easy way for researchers to collect the profile of a user. The term profile describes here all of the information that a user has put on the social networks, or at least as much information as a researcher can collect. Such a profile is usually composed of some text posts that the user has written, and of some meta-data such as the age, the location, or the gender of the user.

The task of personality extraction on social networks is therefore composed of usually 3 distinct sub-tasks:

(1) Obtaining a labelled corpus, which is a set of profiles and the personalities of the users according to a defined personality model,

(2) Pre-processing of the data to make it usable in a classifier,

(3) Classification, or regression, using machine learning techniques.

There are essentially three ways to obtain a labelled corpus for personality extraction. The first consists in recruiting social network users, have them complete a personality test, and collect their profile. This is a highly reliable method because the researcher has total control on all phases of the process. However, it is very time consuming and generally will provide small corpora.

The second way consists in creating a personality test through a web application and have people on a social network take it and also give access to their profile, which can then be collected. Although this will give a good and reliable corpus, the success of it relies on the sharing of the test on the social network, and there are chances that the test won't spread and in turn, that the corpus collected will be of small size.

The last way is to automatically collect the profiles of public users that themselves take a personality test and publish the results on social networks. In this thesis, this method is referred to as automatic corpus collection. The drawbacks of automatic corpus collection are that we cannot verify how reliable the test is and we cannot check if the user is saying the truth. However, following the assumptions that the user is not lying and that the test is reliable enough, this method allows for the collection of a corpus which size is many times what could be accomplished with the first and second methods. For example,the researchers of [Plank and Hovy, 2015] have collected 1500 Twitter profiles in only 1 week, which is a much bigger corpus than what had been collected in past studies. For that method to work, however, a popular personality model has to be used, or the size of the corpus will not be significant.

In this thesis, we work with the social network LinkedIn to perform personality extraction. To do so, we collect a corpus using the third method, a process that is described in Chapter 3. This is, to our knowledge, the first corpus for personality extraction collected on LinkedIn.

The successful extraction of personality on LinkedIn could have many beneficial uses, such as helping guide the users to jobs and companies that will fit their personality better and increase their overall job satisfaction, or use the personality to find ideal team members that would be the most efficient for a job. It could also personalize both the experience of the user and the communication between, for example, recruiters and potential recruits, since the response rate of recruitment through the internet is generally low due to the lack of personalization [Joinson and Reips, 2007][1]. In fact, the goal of personalizing letters issued by recruiters was the starting point of this project.

The initial choice of personality model for the collection of the corpus was the DiSC personality model, which is a personality model that has been developed to explain what motivates people in their work environment. As such, it is ideal for the goal of personalizing letters. It is described in more details in Section 1.1.3

This model has, however, not been studied in depth in psychology and in computer science, even if it is used fairly often by companies[2,3]. Due to the few studies found in the

[1] Results on the effect of personalization on answer rates are equivocal in the literature, but generally positive, allowing us to believe the extraction of personality could be beneficial.

[2] One of the reason it is less used in academia is probably due to its very restricted field of action. In fact, the personality models that are seen more often in studies tend to be personality models that try to explain the behavior of people through all aspects of their life, and not only in a specific context as the DiSC model does with the workplace.

[3] There are no official statistics on which personality models are used by companies or by researchers, and so this assumption is based on a review of the literature surrounding each personality model. While looking at the DiSC model, the only studies found were about how to use the DiSC personality in the workplace, while for the MBTI model, there was a mix of how to use the model in the workplace and of academic studies using the MBTI model.

literature, it is impossible to compare our results to other personality extraction study using the DiSC personality model. For that reason, and because it simplified the collection process, we collect a corpus labelled with both the DiSC personality model and the MBTI one.

The MBTI personality model, described in Section 1.1.2, is a more general personality model that has been more studied in both psychology and computer science. It is also used fairly often in companies. Since we collect a corpus with both personality models, we perform personality extraction on both with the goal of

(1) Using the results on the MBTI personality model to compare to other personality extraction studies,

(2) Comparing the results of the DiSC extraction to the results of the MBTI extraction to see how significant the DiSC results are.

The extraction is done with feature based methods, and the feature engineering process is described in Chapter 3. Some algorithms allowing to learn without feature engineering, such as neural networks, were tried, but did not achieve a good generalization performance, most likely due to the small size of the dataset. For that reason, we focus on feature based methods and extract both semantic features and general non-textual features from the LinkedIn profiles.

In Chapter 4, we perform personality extraction on both the MBTI and the DiSC personality model, as well as analyze and compare the results, and show that we can achieve competitive results on the personality extraction task, on LinkedIn.

Chapter 5 explores the novel idea of trying to classify on one personality model based on the results on another personality model, thus dispensing with the need for multiple personality tests. We denote this task cross-personality extraction, and we show that although we obtain results only slightly better than for the regular personality extraction, adding the other personality model as features gives a significant boost when the classifiers are combined in a voting system, especially for the DiSC classification.

Finally, in Chapter 6, we perform an analysis of the corpus and the personality types, as well as mention several error sources that could have influenced our results. We perform what is, to our knowledge, the first correlation study between the MBTI personality model and the DiSC one, and perform an analysis that show that the third method of collection mentioned earlier and used in this thesis, that is to say to look for users that themselves pass a personality test, introduce a bias, at least on LinkedIn, towards people that fill fairly thoroughly their profile. While this does not affect the other results mentioned in this thesis, as our goal was to show that personality extraction, and cross-personality extraction, is doable on LinkedIn, it still indicates a need for further studies on this potential bias in other social networks.

A paper, also entitled 'Personality Extraction Through LinkedIn', was written on the first part of this thesis dealing with personality extraction and presented at the Canadian AI 2019 conference Piedboeuf et al. [2019]. The paper won the 'Best paper award 2019'.

Chapter 1

Theoritical Background

This chapter reviews some of the theoretical background necessary to better understand the rest of the thesis. In Section 1.1, some personality models are presented, and Section 1.2 briefly describes LinkedIn profiles. Then some tools for textual analysis are reviewed in Sections 1.3, and Section 1.4 explains the algorithms used in the thesis.

1.1. Personality Models

Through the years, multiple personality models have been elaborated to try and break down the behavior of an individual in sub-characteristics that would be common to everyone. These characteristics are called traits, or dimensions, and, in the task of personality extraction, the value that each of these traits takes are the labels we are trying to determine. This section presents 3 different personality models that have been developed through the years and that are important either to understand past research or to understand the work done in this thesis.

1.1.1. Big-5

One of the most important personality model that has been researched in psychology is the Big-5, also called the OCEAN model [Goldberg, 1993]. It is one of the personality models that has been the most used in psychology and probably the one that has been the most verified[1].

[1] By verified, we mean here not only that multiple studies have been done on this model yielding coherent results, but also that it passed some of the reliability tests that exist in psychology to make sure that a measurement is valid. One example of these is the test-retest check, that says basically that if the model is

The Big-5 personality model decomposes personality in 5 distinct traits, hence its name. These traits are:

(1) Openness to experience, or how much is an individual open to live new experience and explore new things in their life.

(2) Conscientiousness, or how much is an individual meticulous and detail-oriented in their life.

(3) Extroversion, or how much is an individual people-oriented and likes to talk to people and meet new people.

(4) Agreeableness, or how much is an individual nice and friendly to other people.

(5) Neuroticism, or how much is an individual prompt to feel negative emotions, such as shame or sadness.

Each trait is given a score from 1 to 5. The higher the score, the more related to the trait the person is. For example, a person who has 5 in Neuroticism will very often experience negative emotions, while a person who has 1 will rarely experience them.

Although the test is often used by psychologists, its continuous and complex nature makes it difficult to be fully adopted by the general population and therefore, automatic corpus collection with this personality model is generally too complex of a thing to be done efficiently. However, it is still found in personality extraction quite often, coming in two forms : either a regression across the different traits to find the exact score obtained on the Big-5, or a classification across the different traits, where each trait has been split in two along, for example, the median, saying that a user has that trait if they are over the median and doesn't have it otherwise.

1.1.2. Myer-Briggs

The Myer-Briggs personality model, more commonly called the MBTI, is a categorical model that evaluates the personality of an individual over 4 dimensions, efficiently separating the people in 16 distinct categories, or types [Furnham, 1996].

The four traits are :

(1) Extroversion(E) vs Introversion(I), or does an individual prefer to focus on the external stimuli or internal ones.

valid, then two measurements taken at a reasonable distance of time should give the same result [Gnambs, 2014].

(2) Intuition(N) vs Sensing(S), or does an individual prefer to analyze the events to find patterns and meaning or simply accept them as happening without an underlying cause.

(3) Feeling(F) vs Thinking(T), or does an individual prefer to follow their instinct when taking decision or do they prefer to think it over and analyze the possible consequences.

(4) Perceiving(P) vs Judging(J), or does an individual prefer the outside world to be spontaneous and flexible or do they prefer for it to be structured and organized.

Together these four traits can be combined to make the 16 different categories, identified usually by the letters of the traits the person has, such as ENFP or ISFJ.

The biggest advantage of this model is without a doubt its popularity with the general public. Between 1962 and 2012, there have been around 50 million people who took a MBTI test[2], and it is estimated that about 2 more million people a year take it [Cunningham, 2012]. This popularity, added to the fact that the personality codes are easy to search for in a profile using regular expressions, probably makes it the best personality model to perform automatic corpus collection.

1.1.3. DiSC

The personality model that is the most interesting in the context of personality extraction with LinkedIn is the DiSC personality model, as it has been developed to better understand work interactions and what motivates people in their professional environment. As such, it seems ideal while working with LinkedIn, which is a social network that focuses on work. The DiSC personality model is also more intuitive to work with in general than the MBTI or the Big-5, making it a good choice for many companies.

Just like the MBTI, the DiSC personality model also separates personality in 4 traits. However while the MBTI traits try to cover the behaviors throughout all aspects of the life of an individual, the DiSC personality model focuses solely on the workplace, meaning that it is a less general personality model than the MBTI or the Big-5. The four traits of the DiSC personality model are Dominance, Influence, Steadiness, and Conscientiousness [Sugerman, 2009]. People who have the Dominance trait tend to be motivated by obtaining

[2] The test can be taken quite easily on the web, for example on `https://www.16personalities.com/free-personality-test`

Tab. 1.1. Examples of Vocabulary Associated with the DiSC Personality Model

Dominance	Influence	Steadiness	Conscientiousness
D	i	S	C
Driver	Appraise	Stable	Cautious
Developer	Promoter	Specialist	Compliance
Achiever	Inspirational	Counselor	Thinker
Creative	Persuader	Practitioner	Objective

results, while those who have the Influence trait tend to be motivated by relationship with others and influencing others. Those who score high in Steadiness tend to be motivated by cooperation with other people, and finally those who have the Conscientiousness trait tend to be motivated by the quality of the work done.

Contrary to the MBTI model, the DiSC personality of an individual is the trait that has the strongest score, with possibility of supporting traits if some other traits also have high scores. For example, someone who scores 66% on Dominance, 35% on Influence, 10% on Steadiness, and 55% on Conscientiousness could say that their personality is Dominant with secondary Conscientiousness. In practice, people often get several traits that score very closely and will express their personality as a mix of these traits, such as 'I am high D and C'.

The DiSC personality also has a very extensive vocabulary associated with it, as shown in Table 1.1. For sake of clarity the vocabulary used in this thesis is restrained to the words Dominance, Influence, Steadiness, and Conscientiousness, except when both the Big-5 and the DiSC personality models are mentioned at the same time, where the word Cautious is used instead of Conscientiousness. There are also several ways to express the same personality type, since it is sometimes given as a number and sometimes as words, as shown in Figure 1.1.

The DiSC personality model is a lesser known model than the MBTI. As such, we found few studies by psychologists, and most of the independent research existing on the DiSC personality model is industry research, where they study what kind of personality is successful or how to use personality to increase cooperation between employees [Reynierse et al., 2000; Sugerman, 2009]. This means that we have only a vague idea of how reliable the DiSC personality model is and of how doable personality extraction is with that model.

Fig. 1.1. Different Ways to Express the D-type Personality

Disc Test: D55% I8% S10% C26%	Disc Assessment: (5|1|3|7)	Disc Profile: [Natural D Adaptive I]	My DiSC assessment reveals I am a persuader
High DI and above the line C on DISC	Disc: “SDIC”	Disc: (D) Driver followed by (C) Conscientious	DiSC: Result-Oriented Pattern

1.2. Structure of a LinkedIn Profile

For a better understanding of the work done in this thesis, this section describes the information found in a LinkedIn profile, once dumped in a json format, and what can be expected to be found in it compared to other social networks such as Facebook or Twitter. The scraping and dumping process are not detailed, as they are not the subject of the thesis, but Section 3.1 gives further details on how the LinkedIn profiles were selected.

LinkedIn is different from other social networks such as Facebook or Twitter in many ways, but three of them could potentially affect the results of personality extraction, as described in [van de Ven et al., 2017]. Firstly, people are more controlled in what they write on LinkedIn than what they write on other social networks due to the possibility of an employer seeing their profiles. Secondly, LinkedIn is a less dynamic social network than most other social networks, and so there are less traces of past behavior than on other social networks. There is a way to share posts and make updates on LinkedIn but, at least in the corpora collected in this thesis, most people do not use it, resulting in very few traces of the interaction habits of the individual. Lastly, due to the same dynamic constraint, there is limited information available on each LinkedIn profile, as opposed to a more dynamic social network such as Twitter, where a theoretically infinite number of posts can be written.

A slightly modified and simplified LinkedIn profile, once dumped in JSON, is given in Listing 1.1. The simplification is only a reduced number of entries for each sections, limiting that number to 1 or 2, while in reality each section may have more entries.

Listing 1.1. LinkedIn Profile Slightly Modified for Anonymity, after Being Saved in JSON

```
{
 "Summary": "I am recognized by my peers as a tech-savvy intuitive
    problem-solver. I thrive in environments where rapid change and
    the need to constantly adapt and learn new information are
    considered the norm. My DiSC personality is Dominant",
  "Personal Branding Claim": "Problem Solver",
  "Connections": "500+",
  "Followers": "0",
  "Skills": ["Leadership 110 endorsements 99+ who are skilled 2
     colleagues", "Strategy 30 endorsements"],
  "Recommendation": "Received (12) Given (3)",
  "Voluntary experiences": {
    "Guest Speaker": {
      "From_date": "Feb 2005",
      "To_date": "Feb 2005",
      "Description": "Guest Speaking at East High School"}},
  "Educations": {
    "Northeastern College of Professional Studies":{
      "Field of study": "Organizational Leadership",
      "From_date": "1987",
      "To_date": "1993",
      "Description": "4 best invested years of my life"}},
  "Experiences": {"Teacher":{
    "From_date": "2015",
    "To_date": "Present",
    "Employer": "Neverends Education",
    "Description": "Teaching software to kids"}},
  "Interests": ["Profyle Tracker 349 followers", "Leoprino Foods
     18802 followers"],
  "Achievements": {
    "Course": {
      "Name": "Team management",
      "Description": "null"},
    "Test": {
      "Name": "MBTI",
      "Result": "ENFP"}}}
```

Each field corresponds to something that users can input on their LinkedIn profile. There are more possible fields than presented, since in the scraping process some information is lost but, in the profiles scraped at least, these are fields not many people use and so the information lost would probably not have made a great difference for the classifiers.

The different fields and what is expected to be found in them are described here:

- Summary: This is a section where users can put free text describing themselves and what they do professionally. This is one of the main sources of free text in the LinkedIn profiles.
- Personal branding claim: Usually a job title, but more generally a very short indication of what the person does professionally.
- Connections: The number of connections of the user. A connection is generally someone met in a professional setting that the user wishes to potentially remain in contact with. For users that have more than 500 connections, LinkedIn simply shows "500+". This corresponds to most of the profiles in our corpus.
- Followers: Followers are people who want to see the public posts of the user, but are not necessarily one of their connections. People who have followers usually use LinkedIn more intensely, writing posts as one would do on Facebook or Twitter. For most of our corpus, the number of followers is 0.
- Skills: On LinkedIn, there is an option to input a number of skills that one would be proficient in, skills such as *Programming*, *Python*, or *System Analysis*. The connections of a user can then endorse them in these skills, and special mentions are given to the connections who have a lot of endorsements in that same skill, as well as to the connections who are colleagues and work at the same place as the user.
- Recommendations: These are short paragraphs that a user can write describing how it was to work with another user. Due to the complexity of retrieving them, the texts are not scraped from LinkedIn, but the number of recommendation received and given by the users are[3].
- Voluntary experiences: Any number of voluntary experiences, with fields for the dates and a short optional text for a description of the experience.
- Experiences: This section represents all the work experiences a user had in their life. Each work experience can be accompanied by some information. The most common,

[3] LinkedIn is a highly dynamic social network, meaning that content is generated only when the user asks for it, and so when scraping LinkedIn many automations are needed to retrieve information. Some fields would need so many specifics automations in order to find that information that it is simply not worth the time.

and those that are dumped in json, are the beginning and ending dates, the employer, and a short text describing the job.

- Interests: The interests section displays the companies, group, influencers, etc., that the user is following. Due to the complexity of the task, only the six first displayed interests are dumped in json, with the number of followers each of the interests has.
- Achievements: The achievements are occurrences like courses, test results, publications, etc. Each of these have specific fields that are dumped in json as well.

1.3. Tools for Textual Analysis

Text cannot be inputted directly in the traditional machine learning algorithms described in Section 1.4. That is because these algorithms work in a vector space and so a way has to be found to transform the text into vectors. In this section, some of the ways to process the text as to render it in vectorial form are presented. There are many more ways that exist, but only those that are used in this research are described. The axes of the vector space are referred to as features.

1.3.1. Normalization

Before going into the vectorization process, 2 normalization techniques that will minimize the loss of information as much as possible are briefly presented. As with the vectorial transformations, there are many more that are possible, but the ones presented here are those that are used in this thesis.

1.3.1.1. *Stop-Words*

In any text there are words that, for the purpose of classifying, will not give any relevant information and so that can often be safely removed from the text. This is because these words are usually so common that everyone uses them in great quantity and so keeping them in the text is not only useless, but will sometimes worsen the results. They are usually referred to as stop-words. The list of stop-words used in this thesis is presented in Appendix A.

1.3.1.2. *Case Normalization*

The other normalization process used in this thesis is case normalization. This is a simple process that consists of putting all words in lower-case (or upper-case), so that variations of the same word count as one word. For example, without case normalization, the words dog and Dog would count as two different words. This process is applied to all text used in this thesis.

1.3.2. Bag of Words

The first, and one of the easiest, way to represent text as a vector is by using a Bag of Words. A Bag of Words vector space is one space where each axis is a word, and the length of the text vector along each axis is the number of times the word appears. For example, given the two sentences "The fox is brown" and "The fox jumped", the vector space would be [the, fox, is, brown, jumped] and so the two sentences would be represented, respectively, as the vectors [1,1,1,1,0] and [1,1,0,0,1].

1.3.3. TF-IDF

The next step, once a Bag of Words representation has been obtained, is to put the words in a vector form called Term-Frequency Inverse-Document-Frequency (TF-IDF). The intuition behind the TF-IDF is that if a word is common in a document, but rare in the corpus, it is probably very relevant to the central ideas formulated in that document.

The TF-IDF is computed in two different parts. The first part is the Term Frequency, which is calculated as the number of time a word appears in a document divided by the total number of words in that document.

The other part is the Inverse Document Frequency, which is calculated as the total number of documents in the corpus divided by the number of documents in which the word appears. The less a word appears in the corpus, the higher the IDF will be, and vice versa.

Finally the TF-IDF is computed by taking the product of both parts.

1.3.4. General Inquirer

The General Inquirer[4] is a tool that allows analysis at a semantic level [Stone et al., 1962]. For each word, there are a total of 184 categories that the word could potentially belong to, and the General Inquirer returns the number of words belonging to each category in the text, both as an absolute and as a relative count. These categories include things such as Anxiety, Family, Sadness, Health, etc, and it is easy to see how the counts could relate to

[4] `http://www.wjh.harvard.edu/~inquirer/`

the personality. For example, one could imagine that a user scoring high in the Dominance trait of the DiSC personality model would use more words semantically related to action and results. The General Inquirer also performs lemmatization in its analysis, which is the action of finding the root of a word when it is written with some variations, for example, swimming which is a variation of the root swim and therefore has the same semantic meaning. One of the main motivation for using it is because it has been shown, in the past, to help for personality extraction and classification tasks [Markovikj et al., 2013; Stone et al., 1962].

1.3.5. Part Of Speech

The Part Of Speech (PoS) features simply refers to the different grammatical use of words in a sentence, such as Noun, Pronoun, or Adverbs. Several studies on personality extraction have taken them as features and the frequency of usage of several PoS have been found related to personality traits [Markovikj et al., 2013; Pennebaker and King, 1999; Mairesse et al., 2007].

1.4. Algorithms

1.4.1. Naive Bayes

The Naive Bayes algorithm is a very simple and very old one. The idea is to calculate, for each class c, the probability that the data point x belongs to the class. Mathematically that can be represented as :

$$p(C_k|x_1, ..., x_m) \tag{1.4.1}$$

where C_k is the class k and $x_1, ..., x_m$ are the different features of the input. The Naive Bayes classifier uses both Baye's theorem and the Naive assumption in order to perform an easier classification. Baye's theorem states that

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)} \tag{1.4.2}$$

where $x = x_1, ..., x_m.$, and the Naive assumption is

$$p(x_j|x_1, ..., x_{j-1}, x_{j+1}, ..., x_m, C_k) = p(x_j|C_k) \tag{1.4.3}$$

which is to say that each feature is conditionally independent from every other feature.

Putting Equations 1.4.2 and 1.4.3 together, it can be seen that

$$p(C_k|x) = \frac{p(C_k)\prod_{j=0}^{m} p(x_j|C_k)}{p(x)} \tag{1.4.4}$$

Given that $p(x)$ is identical for every class, and that only the ratio between the results for each class is needed, it can be removed without affecting the results, resulting in the final equation :

$$p(C_k|x) \propto p(C_k)\prod_{j=0}^{m} p(x_j|C_k) \tag{1.4.5}$$

Once this equation is obtained, there are only two things left to do in order to be able to calculate the probability that a data point x belongs to a class. First, a probability density function is chosen, and then, the probability of the point belonging to each class is evaluated.

1.4.1.1. *Gaussian Distribution*

Features can follow in theory an infinite number of distributions over the space. However, in order to predict the class using the Maximum A Posteriori (MAP) estimation, a probability density function for each class is needed, preferably one that is close to reality. Such a a function gives, for a data point x, the probability that it would occur given its parameters. In this thesis, as the Naive Bayes algorithm is mostly used as a classification baseline, the focus is not put on fine-tuning the classifier and finding the real distribution. Instead, the well known Gaussian distribution, also called the normal distribution, is used. The formula corresponding to this distribution is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{1.4.6}$$

where μ is the mean of the distribution and σ is the standard deviation. Given a training set, the μ and the σ of each feature can be found with the following formulas:

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{j_i} \tag{1.4.7}$$

$$\sigma_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{j_i} - \mu_j)^2} \tag{1.4.8}$$

,

where $E[X]$ is the expected value of X, or the mean.

1.4.1.2. *Maximum A Posteriori*

With the Gaussian distribution as the assumed distribution, the probability of a feature x_j of a new point belonging to the class c can be calculated by replacing the variables in Equation 1.4.6 by the real values. The class the data point is most likely to belong to is then calculated simply by applying the principle of Maximum A Posteriori, which consists of calculating for each class the probability and selecting the maximum value, according to the formula

$$\hat{y} = \arg\max_{k} p(C_k) \prod_{j=0}^{m} p(x_j|C_k) \tag{1.4.9}$$

1.4.2. Support Vector Machine

Support Vector Machine, or SVM, is an algorithm that consists, in its simpler form, in trying to find two parallel hyperplanes in the vector space that will best separate two classes. An hyperplane is described mathematically as

$$\vec{w} \cdot \vec{x} + b = 0 \tag{1.4.10}$$

where $\vec{w}$ is the normal vector of the hyperplane, and b is the bias that allows moving the hyperplane away from the origin of the vector space.

1.4.2.1. *Hard-Margin*

Given $t_i \,\epsilon\, \{-1, 1\}$ the expected label of the data, and given that the classes are linearly separable, that is to say that

$$\exists w, b | \forall i \,\epsilon\, \{1, ..., n\}, (t_i(\vec{w} \cdot \vec{x}_i + b) \geq 1) \tag{1.4.11}$$

then the goal is to find two parallel hyperplanes such that

$$\forall x_i \mid y_i = 1, \;\; \vec{w} \cdot \vec{x_i} + b \geq 1 \tag{1.4.12}$$

and

$$\forall x_i \mid y_i = -1,\ \ \vec{w} \cdot \vec{x_i} + b \leq -1 \tag{1.4.13}$$

That is to say we want to find two hyperplanes such that all points belonging to a class are on one side of them and all points belonging to the other class on the other side, with no points laying between them. The margin is defined as the set of $\vec{x}$ such that

$$-1 \leq \vec{w} \cdot \vec{x_i} + b \leq 1 \tag{1.4.14}$$

and correspond to the space between the two parallel hyperplanes. The distance between these is

$$\frac{2}{||\vec{w}||} \tag{1.4.15}$$

When fitting an SVM, the biggest margin possible is desired. Since it can easily be seen that the bigger the distance between the hyperplanes, the bigger the margin, the problem of maximizing the margin can be formulated as the problem of minimizing $||\vec{w}||$, under the constraint that all points are classified correctly.

1.4.2.2. *Soft-Margin*

In the real world, data is rarely linearly separable, and therefore, the hinge loss function is used. It is defined as

$$max(0, 1 - t_i(\vec{w} \cdot \vec{x}_i + b)) \tag{1.4.16}$$

which gives 0 if the point is correctly classified and a positive value proportional to the distance to the margin otherwise. The problem then becomes of minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(max(0, 1 - t_i(\vec{w} \cdot \vec{x}_i + b))) + \lambda||\vec{w}||^2 \tag{1.4.17}$$

where the λ factor determines a trade-off between the size of the margin and the correct classification.

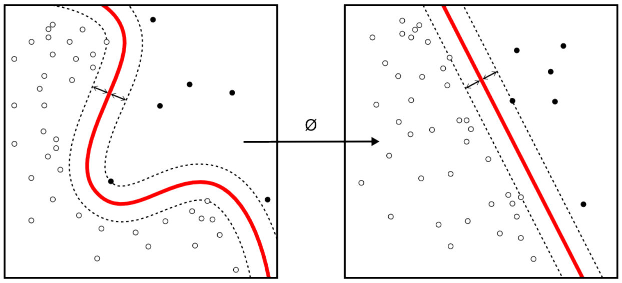


Fig. 1.2. An Illustration of the Kernel Trick. To the left, the hyperplane in the original vectorial space is seen, while to the right we see the hyperplane in the kernel space. Taken from Wikipedia

.

1.4.2.3. *Kernel Trick*

The basic principle of the kernel trick is that, instead of computing the SVM in the original vector space formed by the features, it is computed in a higher, possibly infinite, dimensional space. Doing so allows the SVM to find correlations that are non linear, as shown in Figure 1.2. The goal is then to find a way, for every x, to compute $\tilde{x}$, the coordinates of x in the new chosen space. The kernel trick helps reduce the computational load by supposing that, for any scalar product, the result can be computed in the new vector space without having to compute the two original vectors in that space.

Several kernels are used in practice. In this thesis, the kernel chosen is the Gaussian radial basis function (RBF), which projects the results in an infinite vector space. The formula of the kernel is

$$k(\vec{x}, \vec{x'}) = exp(-\gamma||\vec{x} - \vec{x'}||^2), \gamma > 0 \tag{1.4.18}$$

where γ is a hyper-parameter to be chosen.

1.4.3. Random Forest

Random Forests (RF) is an algorithm that creates an ensemble of Classification or Regression Trees. In this thesis, we use Classification Trees, as personality extraction is seen as a classification problem. The principle of a Classification Tree is first explained, and then the Random Forest algorithm is detailed.

1.4.3.1. *Classification Tree*

A Classification Tree, or more generally a Decision Tree, works by deciding, at each step of the algorithm, which feature would best split the training dataset into its correct classes according to a specific metric. There are two different things to consider when building a tree: the metric to choose, and when to stop. The metric directs the general going of the algorithm, and the stopping point is important because stopping too late causes the algorithm to overfit the data, while stopping too early causes the algorithm to underfit it.

In this research, the metric used is the Gini impurity, which is defined for a node in the tree as the probability of obtaining two different results given that a point that falls into the node is randomly chosen. Mathematically it can be expressed as

$$G(N) = 1 - p_0^2 - p_1^2 \tag{1.4.19}$$

where p_0 and p_1 represent, respectively, the probability of a data point belonging to the class labelled 0 and to the class labelled 1. The Gini impurity is also usable in a context where there are more than two classes, but since we only perform binomial classification, the generalized version is not mentioned any further in this thesis. When a node only contains one class, the Gini impurity is equal to 0, and when there are an equal number of data points belonging to each classes, the Gini impurity is maximal.

The next important thing for a decision tree is to know when to stop splitting. Obviously, if the process keeps going until the Gini impurity is 0, then the error on the training dataset would be null, but the chances are high that the algorithm would also be overfitting and would not be able to generalize correctly. The accepted solution is usually to choose a threshold value, and if the split reduces the impurity by less than that threshold, the tree is not splitted.

1.4.3.2. *A Forest of Trees*

Once we can build a decision tree, we can build what we call a Random Forest (RF). A Random Forest is built by a process called Feature Bagging. This process consists of training several decision trees with a subset of the training dataset and a subset of the features, and the final decision is a majority vote of all the decision trees.

1.4.4. AdaBoosting

Similar to the Random Forest idea of building several classifiers that then take a vote, AdaBoost is an algorithm where the final classifier F_T is a combination of T several weak classifiers f_t, according to

$$F_T(x) = \sum_{t=1}^{T} f_t(x) \tag{1.4.20}$$

where the output of each classifier is either -1 or 1, and so the sign of the sum will be the final prediction. However, unlike RF, when a new classifier is trained, the data points are weighed with a weight proportional to the current error on that data point. This effectively makes the new weak classifier focus more on data points that the algorithm is not yet able to classify.

Chapter 2

Literature Review

This chapter examines the literature concerning personality extraction. The field of personality extraction through social networks is relatively young, having started around 2010, but the field of personality extraction goes back to before the year 2000. The goal of the task is, for a given personality model, to be able to correctly classify or predict the personality of a person given an input. Sometimes this input can be text, as in Section 2.2, a social network profile, as in Section 2.3, or a video, as briefly discussed in Section 2.3.2. In view of the particular corpus labelled with the two personalities, DiSC and MBTI, that are used in this research and described in Section 1.1, Section 2.1 also reviews some studies that draw correlations between personality models. This chapter is concluded by presenting in Section 2.4 some studies that use personality extraction as a sub-task, to show that automatic personality extraction has many uses in computer science and can help with many sub-fields.

2.1. Correlation Studies between Personality Models

Correlation studies allow the evaluation of how closely two models are overlapping. In an industry setting, it is easy to see that if two personality models represent the same thing, then there is no point in utilizing both of them in, for example, team-building exercises. For researchers, it allows to see if some traits are missing from a personality model or if two personality models are representing the same thing. For example, one critic raised about the MBTI model is that it is missing the Neuroticism trait present in the Big-5 personality model, which is confirmed by drawing correlations between the two models, as shown in [Furnham, 1996].

The study of [Furnham, 1996] looks at the correlation between the traits of the two personality models. The authors find that except for the Neuroticism trait of the Big-5 model, all of the traits are closely correlated to one trait of the MBTI personality model. The correlations go as follows: Agreeableness correlates with Thinking-Feeling, Conscientiousness with Judging-Perceiving, Extroversion with Introversion-Extroversion[1], and Openness to experience with the Sensing-Intuitive trait. These correlations have been, to some degree, verified by multiple other studies through the years [Furnham et al., 2003; McCrae and Costa, 1989; MacDonald et al., 1994].

In [Harvey et al., 1995], the researchers calculate the correlations between the questions in the test for the MBTI personality model and the Big-5 personality model, making it possible to obtain from one test the scores on 2 different personality models. By analyzing the results of 1258 participants, they find that if they subdivide the MBTI questions in 13 categories, then they can make each of the categories correlate to one of the sub-traits in the Big-5 test.

Although no study comparing the MBTI and the DiSC personality models were found, the researchers of [Jones and Hartley, 2013] work with the Big-5 and the DiSC models. Using a sample of 89 students who filled both tests, they find multiple correlations. Introversion is correlated with the Influence and Cautious trait of the DiSC personality model, Openness with Influence and Steadiness, and both Agreeableness and Neuroticism with Dominance and Steadiness.

2.2. Personality Extraction through Texts

One of the first study on personality extraction is the one of [Pennebaker and King, 1999], which looks at the relationship between text and personality. Using the Big-5 personality model and a variety of texts, they draw correlations between the categories of the first version of the LIWC, a tool similar to the General Inquirer described in Section 1.3.4, and each of the traits. They do not perform regression, but nevertheless lay the foundation for personality extraction, and show correlations between language and personality.

In [Mairesse et al., 2007], six different learning algorithms are used to try to determine personality based on linguistic cues, more specifically a mix of the LIWC word categories

[1] as one would expect.

and of the MRC psycholinguistic features chosen based on previous studies, the MRC being another tool like the General Inquirer. Using a set of texts and of conversation extracts labelled with the Big-5 personality traits, they find results varying between 52% and 62% of precision depending on the trait. Overall, the algorithm giving the best score is the Support Vector Machine (SMO).

The researchers of [Noecker et al., 2013] use a centroid-based model to classify, according to their MBTI personality, an ensemble of essays written by 145 students. The essays are written in Dutch and several feature sets are tried, including BoW and character n-grams. Character n-grams gives them the best performance, with precision scores ranging between 76.0% and 85.0%.

While the studies mentioned until now look at personality through long texts, a lot of the data on the internet is actually conversational text, which differs in several ways from free-flowing text. The authors of [Su et al., 2016] try to exploit these differences in order to successfully predict personality with the Big-5 personality model. To do so, they use a mix of Recurrent Neural Network and Coupled Hidden Markov Model to respectively extract the short-term temporal evolution and the long-term turn-taking temporal evolution of the information. As features, they use the Chinese version of the LIWC, as the corpus is in Chinese, and report an average precision of 79.7% over the Big-5 traits.

2.3. Personality Extraction through Social Networks

Personality extraction through social networks is, in effect, very close to personality extraction through texts, but there still are some differences. For starter, instead of one well written and coherent text, the social network generally gives a multitude of independent texts, sometimes even limited to a number of characters. Moreover, there is usually what is called in this thesis non-textual data, which denotes the set of data one can get about the user, such as the number of connections or the number of jobs they had in their career. There's also the fact that there's usually an inconsistent number of data between the users, since there's no guarantee that a user will fill all the fields on the social network platform (for example, a user could choose not to input their age or location). Finally, there's the possibility to work with multiple social networks at the same time for given users, capturing in that way more information than on only one social network.

One of the first research that dwells on personality extraction through social networks is the one of [Quercia et al., 2011], which predicts personality of Twitter users by using the number of followers, following, the number of Facebook friends, as well as the number of times the user had been listed. Using a Regression Tree, they achieve a Normalized Root Square Error going from 0.88 to 0.66 depending on the traits.

The team of [Celli et al., 2013] made strides towards establishing the field of personality extraction through social networks as a united field by organizing a shared task where participants were asked to extract the personality of Facebook users, using the Big-5 model. The dataset is composed of 250 profiles containing raw text and several social network measures, such as the number of friends or the transitivity, which indicates how interconnected the social network of the user is. The personality is given both as a score going from 1 to 5 and as a class, splitting each trait along the median and saying that the user is, for example, open to experience if they have a score higher than the median and not otherwise.

The authors of [Markovikj et al., 2013] achieve great results on the classification task by using a SVM with feature selection on 725 features that they extract from the texts and the profiles. The feature selection reduces that number to the best 5 to 16 features, and they obtain results going from 86% to 95% of precision over the different traits.

One of the contributions that is the most important for this thesis is the one of [Plank and Hovy, 2015], where a way to automatically collect profiles with a labelled personality is introduced. The researchers collect 1500 Twitter profiles labelled with a MBTI code and use logistic regression on the profiles, getting results going from 55.4% to 77.4% of precision depending on the trait. As mentioned before, this technique allows corpus of greater sizes to be collected with a good efficiency, and is also the technique used in this thesis to get the corpus, as described in Section 3.1.

The researchers of [Lima and Castro, 2016] use the technique described in [Plank and Hovy, 2015] to collect a corpus of Twitter users that put their personality model results online. However, instead of classifying each trait of the MBTI personality model, they attempt to classify the user into one of the 16 different personality types directly. They try multiple algorithms for the task, achieving about 70.0% of precision and obtaining the best results mostly with a SVM.

Amongst studies that have taken as object other personality models than the MBTI or the Big-5, there is the one of [Ahmad and Siddique, 2017], which downloads an ensemble of tweets associated with keywords representing each of the traits of the DiSC personality model. This is based on the assumption that people who use the vocabulary associated with a personality trait would likely have that personality trait. Performing data analysis, the researchers then extend the vocabulary associated with each of the traits by using the most common words used by each types. Although they do not perform classification on the data, they still show a correlation between the vocabulary used and the personality type, laying the foundation for other personality studies using the DiSC personality model.

Another personality model that is less often seen in automatic personality extraction, but is also very important, is the Dark Triad personality model. This model depicts the traits of Machiavellianism, Narcissism, and Psychopathy. In [Sumner et al., 2012], a corpus of Twitter users is used to try and predict the three traits of the dark Triad. They show that there is a correlation between the language used and the traits. For example, users that swear more and use more words associated with anger tend to score higher in Psychopathy and Machiavellianism. They then attempt to predict personality on the Dark Triad by using, and comparing, several machine learning algorithms, but report results barely above chance.

The only study found that looks at personality models on LinkedIn directly is the one of [van de Ven et al., 2017], which evaluates how much people can determine a user personality when looking at a LinkedIn profile, using the Big-5 personality model. They find that only extroversion can be determined by people. This, however, does not indicate how accurate a classifier would be on the same task, since it has been demonstrated that computers are more efficient at the task than humans [Youyou et al., 2015].

The researchers of [Song et al., 2015] work on classification through multiple social networks, which allows the creation of a more complete user profile, as it has been shown that people tend to change the facet they present depending on the social network used [van Dijck, 2013]. The main problem with this, however, is to palliate the possible lack of information, since there is no guarantee that a user will be active on every social network, or even will have an account on every social network. To do so, the authors create a latent space shared by all social networks, allowing inference on the missing data. They then obtain personality predictions using a linear combination of S predictive models, where S is the number of social networks. Overall, they obtain about 85.6% on the F1-measure, which is higher then what they obtain with individual social networks using the same algorithms.

Personality is a very complex subject and each of the traits is responsible for a certain degree of the behavior of the individual. To account for that, the researchers of [Xue et al., 2017] use Label Distribution Learning algorithms, which have been developed to take into account the different levels of influence the classes can have on a data point. To evaluate the algorithms, they collect a corpus on the Chinese social network Sina Weibo labelled with the Big-5 personality model, and report state of the art results on that social network.

In [Park et al., 2015], an open-vocabulary approach is used to infer automatically the linguistic features that would be of importance for personality extraction. To do so, they work with a corpus of 66,000 Facebook profiles and extract the most frequent words and groups of words with the help of the Latent Dirichlet Allocation (LDA) algorithm, before reducing the dimension of the obtained matrix to keep only the most important features. Finally, they calculate the correlation between these features and the 5 traits of the Big-5 personality model, and by doing so, confirm that there indeed is a correlation between the vocabulary used and the personality.

A comparative study of different algorithms is performed in [Ngatirin et al., 2016]. The data is collected from 100 undergraduate students who filled a Big-5 personality test and gave their Twitter username, and the study is only done using the extroversion trait and 5 features, those being the number of followers, following, tweets, lists, and favorites. After trying several algorithms, they conclude that OneR [Holte, 1993], an algorithm that tries to find a single rule that best classify the data, is the most accurate, with about 84% of precision.

Working with the dataset of [Celli et al., 2013], the team of [Tadesse et al., 2018] evaluates different sets of features by trying to predict the personality of users. They use the LIWC, the Social Network Analysis (SNA), and the Structured Programming for Linguistic Cue Extraction (SPLICE), all of these being feature extractors. They find that globally the SNA work best, but also that depending on the algorithm, other features sometimes are more useful, showing the importance of feature selection tailored to the selected algorithm.

2.3.1. Neural Networks

Although all of the algorithms mentioned above obtain pretty good results on the task, they most often need manual feature extraction, which is time consuming and generally only works in one language for textual features. Neural networks, on the other hand, can be trained to learn by themselves the important features and are not limited by the language.

The texts of [Pennebaker and King, 1999], a set of texts labelled with the Big-5 personality model, are used in a deep convolutional neural network in [Majumder et al., 2017]. The network contains 7 layers and extract unigrams, bigrams, and trigrams, and also takes as input the 84 semantic textual features identified by [Mairesse et al., 2007]. By trying different configurations of the network, the researchers successfully obtain better results on the 5 traits than the previous state of the art results.

In [Kalghatgi et al., 2015], a simple Multilayer Perceptron is used to try and classify personalities with the Big-5 personality model. They work on a corpus of tweets, extracting both meta-data and grammatical attributes to input as features in the neural network. They conclude that for the network to work well, at least 20 tweets by users should be provided.

The researchers of [Yu and Markov, 2017] also work with the data of [Celli et al., 2013], using this time a bidirectional neural network and a convolutional network with an embedding layer, introducing at the same time the non-textual features present in the dataset. The fully connected layer overfits, and better results are achieved using a convolutional layer with average pooling, but overall, they cannot outperform the results of [Markovikj et al., 2013], where a SVM with feature selection is used.

The researchers of [Liu et al., 2016] develop a language independent deep-learning model for personality extraction. To do so, they bypass any classical word embedding layer and instead use a "Bag-of-characters" that feeds into a bi-directionnal RNN, that then feeds into a word-level Bi-RNN, that finally represents a sentence by the concatenation of the last and first hidden states of the two directions of the Bi-RNN. This is topped by a Multilayer Perceptron. The researchers report slightly better results than the state of the art algorithms. Furthermore, they point out that the model has the advantage of being language independent and requires no fine-tuning in order to work.

In [Hernandez and Knight, 2017], the corpus is obtained from the internet forum `https://www.personalitycafe.com/`, a forum where people identify themselves with their

MBTI personality types and can then discuss with other users. The dataset is composed of 8,600 users and of 50 posts per user, and the researchers use a recurrent neural network with pre-trained word embedding. The classification is made by breaking down the personality model into its 4 traits and results ranging from 62.0% to 78.0% are obtained, which is a bit lower than other studies on the MBTI personality model.

2.3.2. Non-Textual Personality Extraction

In this section, studies that take as input not only, or not at all, textual features, are examined. Although these techniques are not used in this thesis, they are still worth mentioning to better understand what can or cannot be done to help predict personality. For starter, there is the study of [Ferwerda and Tkalcic, 2018], which use the Google Vision API[2] to analyse Instagram images of users that filled a personality test. The researchers find several interesting correlations between traits of the Big-5 personality model and the pictures. For example, they find that the presence of clothes is positively correlated with Agreeableness and negatively with Neuroticism, and that the presence of musical instruments is positively correlated with the Openness to experience trait.

Working with videos is more complex than working with images, but it also reveals more information. In [Biel et al., 2012], videos of vloggers are analyzed according to the Big-5 personality model. To do so, emotions are first extracted from the facial expressions and then correlations are drawn between these emotions and each of the traits, such as smiling being correlated with Extroversion and Agreeableness, or Anger showing negative correlation with Extraversion, Openness to experience, and Agreeableness.

Working with text, audio, and visual cues, the authors of [Kampman et al., 2018] extract the personality of Youtube vloggers, once again on the Big-5 personality model. The videos are about 15 seconds in duration, and three different Convolutional Neural Network are used to extract personality from the text, the audio, and the video, before using neural network fusion methods to combine the three. They report an average among all traits of 0.0938 Mean Square Error, outperforming the neural networks that use only one channel as input. They also report that in the corpus, language and speech patterns are the less relevant features, and video frames are the most relevant features, for personality extraction.

[2] https://cloud.google.com/vision/

2.4. Using Personality Extraction

Although personality extraction is very interesting in itself, it also has various uses. For example, knowing the personality of a LinkedIn user, it would then be possible, and even beneficial, to use that information to alter the style used in communications with the user to better appeal to their interests and ease the conversation. In this section, some works that have used the output of personality extraction for another system are mentioned.

The researchers of [Celli and Zaga, 2013] use an unsupervised approach in order to predict the personality of the user with the Big-5 personality model, before using it to predict sentiment polarity. They predict the personality according to each tweet individually, and the predictions of all the tweets of a user give a global picture of the personality of the user, classifying based on features that have been shown to be correlated with each traits of the personality (negatively or positively) in past researches. The system also has 2 measures of performance: the precision (defined just like the standard precision), and the variability, which represents how much all the tweets of a user tend to express the same personality. Given the complexity of personality, it would be very difficult to do unsupervised studies without the help of features extracted in the past. After predicting the personality, they use it as input for predicting sentiment polarity, observing that hashtag position and Conscientiousness work best to predict sentiments, with results of about 60%.

Another work that uses personality as features in sentiment classification is the one of [Lin et al., 2017], which focuses on a corpus of Chinese micro-blogs. They first predict personality with the help of a rule-based method before using ensemble methods for sentiment classification, getting about 90% on the F-value for that task.

The identification of influential communities, which are communities where users can influence other users, is an important topic for a company wanting to do viral marketing. In [Kafeza et al., 2014], communities are detected and then the personality of the users are evaluated on the Big-5 scale. Communities that have more heterogeneous users are most likely to have more influential power from the marketing company point of view, and so the use of personality extraction allows to detect more accurately which are the influential communities.

Personality extraction through the gaming platform Steam is performed in [Yang et al., 2017], with the goal of better recommending games to users. To do so, they take as input the textual reviews written by the users, as well as the the tags they give to games. Then, two personality models are constructed: one for the games, meaning the personality the users playing this game are most likely to have, and one for the players, and the system can match the games and the players that have the same personality.

All the research presented in this section shows that personality extraction is a prolific field that is gaining importance in the world of data we are living in. In order to correctly talk to and reduce the impersonal barriers present due to the distance and lack of information humans use for better understanding who they are talking to, personality extraction could be a useful tool to guide conversations towards a more agreeable discussion.

More technically, the field of personality extraction generally uses either feature based approaches using semantic extractor, such as the LIWC or the General Inquirer, or neural network for automatic feature extraction, which requires however a large amount of data not always readily available, due to the complexity of collecting a corpus for personality extraction.

Chapter 3

Methodology

3.1. Corpus Collection

Through the years, multiple corpora of social network profiles labelled with a personality have been developed and published to be used for personality extraction [Celli et al., 2013; J, 2018]. However, due to the ethical implications of sharing a corpus containing often personal information on people, most researchers collect and work on their corpus independently. As no studies on personality extraction on LinkedIn were found, we first need to collect a corpus on LinkedIn labelled with the DiSC personality before trying to predict personality. The DiSC personality model is preferred as label because it has been used and developed for the purpose of work interaction, which seems very appropriate while working on LinkedIn, a social network centered around work.

In order to collect a corpus of LinkedIn profiles labelled with the DiSC personality model, we follow the method proposed in [Plank and Hovy, 2015]. Using the Google search python API[1], we find all public LinkedIn profiles that contain the words "personality" and "DiSC", which gives a total of 38,000 profiles. However, a look at 100 of the profiles show that only 4 of them have a DiSC personality mentioned in the profile, and that the rest of them are noise. The noise is generated by people using the word 'DiSC' in another context than for the DiSC personality and at the same time using the word 'personality' in their profiles, an occurrence not uncommon if one think of sentences such as "I have been told I have a great personality", which could easily be included in a LinkedIn profile. Figure 3.1 shows some examples of noise using the word 'DiSC'.

[1] `https://developers.google.com/api-client-library/python/apis/customsearch/v1`

Fig. 3.1. Example of Noise in LinkedIn Profiles while Looking for the DiSC Personality

DiSC Personality Inventory Facilitator	Owner at Sam's Disc Golf	Works with DISC, MBTI, others	The Liberty Lumbar Disc clinical trials
Radio DJ personality and Sound Disc jock	Lessons are saved on a convenient compact disc	Perform disc surgery	Disc Jockey

The noise can be reduced by surmising that people who put their MBTI code in their LinkedIn profile are more likely to put their DiSC personality than those who do not, and so that the presence of the word "DiSC" in a LinkedIn profile is more relevant if the user also has a MBTI personality mentioned in the profile. Since the MBTI personalities are most often expressed as 4-letters codes very distinct from any other common words, as shown in Section 1.1.2, we are sure that results returned by the Google API contain a MBTI personality. Moreover, this also gives the possibility to perform a correlation study between the DiSC and the MBTI personality and to try and perform classification from one personality model to another.

The final form of the requests submitted through the Google API are

"site:https://www.LinkedIn.com/in/ DISC personality code_mbti",

where code_mbti is replaced by each of the MBTI codes, giving in practice 16 different requests. The requests can be read as such: in all urls that start with https://www.LinkedIn.com/in/, which is the set of the urls denoting user pages, look for profiles that contain the words Disc, personality, and one of the MBTI codes. The inclusion of the word "personality" is for reasons of both reducing further the noise, and trying to constraint the corpus to English-speaking (or at least writing) users, as our textual feature extractors are in English.

The next step is to scrape the pages of the LinkedIn users that are returned by the google API. To do so, a LinkedIn scraper found on github[2] is modified as to include more information and as to be able to handle a large, continuous, scraping process.

This gives a total of 1253 profiles labelled with the MBTI personality, which form the MBTI corpus. To be sure to correctly eliminate all the noise present in the DiSC corpus, or the set of profiles labelled with a DiSC personality, all the profiles containing the words "HR", "certified", "jockey", "administering", "workshop", "golf", "spine", and "coaching" close to the word "DiSC" are removed, leaving 841 profiles. There are 10 profiles returned by the API that do not include a MBTI personality, but include a DiSC personality. This is most likely due to the fact that the MBTI personality may have been written in one of the sections that is not scraped.

Both the DiSC corpus and the MBTI corpus are of rather small size, which can be explained by two factors. Firstly, the fact that LinkedIn is used as the social network accounts greatly for the size. Not only are there less people on LinkedIn than on other social networks, such as Facebook [Maina, 2016], but due to the more professional nature of LinkedIn, people are more controlled and more professional in what they write [van Dijck, 2013], and so it is not unreasonable to assume they would be less likely to include their personality results.

Secondly, there is also the fact that the profiles sought include both the DiSC and the MBTI personality. While this is done out of the necessity to trim noise out of the DiSC corpus, it certainly cuts out a large number of profiles as opposed to getting a corpus labelled simply with the MBTI or with the DiSC personality model, given that this could have been done efficiently[3].

[2] `https://github.com/joeyism/LinkedIn_scraper`

[3] A quick google search with the requests 'site:https//www.LinkedIn.com/in/ code_mbti', where once again code_mbti represent in turn the 16 different types, returns 74,690 results. There may be some noise or profiles in other languages, but even after eliminating that noise that would leave a corpus which size is many times what was collected for this thesis. The scraping process however would be more problematic, since scraping only the 1,252 profiles took about 1 week as LinkedIn blocks people who suspiciously look at too many profiles in a short amounts of time.

3.1.1. Labelling of the Corpora

The labelling of the MBTI corpus, or the action of extracting from the profiles the labels on which the algorithms will later classify, is fairly straightforward, as the MBTI codes can be directly captured with regular expressions. The only real choice is to decide on which side of the scale to put a 1 and on which to put a 0, since the dimensions go from one trait to another trait. The convention that is followed is that a 1 is used if the user has the trait of Introversion, Intuition, Thinking, or Judging, and a 0 otherwise. In the following sections, "IE", "NS", "TF", and "JP" are written to denote the Introversion-Extroversion, Intuition-Sensing, Thinking-Feeling, and Judging-Perceiving dimensions.

The labelling of the DiSC corpus presents more difficulties. First, the written personality has to be interpreted. In the personality model, as presented in Section 1.1.3, an individual has one main trait and one or more supporting traits. However, users on LinkedIn often express their personalities as if they have several main traits, such as "I am high C and D", or "My main traits are S, C, and I". The guideline that is followed is that the labelling should reflect what the users say, and so the label is DI if a user says "I am equal part D and I". In the case of supporting traits, such as in the sentence "My DiSC personality is D with supporting C", the idea is once again that if the user mention Conscientiousness as a trait, it must be strong enough to be significant and so in this case the profile is labelled with "DC".

In Section 1.1.3, many ways of expressing a DiSC personality are presented. Add to that the fact that several of the words used are common words, and labelling becomes a very difficult problem. Some attempts using regular expressions are made, but at the end capture noise instead of the personality and so the corpus is labelled manually by looking, for each profile, at a window of ± 150 characters around the word DiSC, and noting the personality written.

In Table 3.1, the percentages of the different traits for the MBTI personality and for the DiSC personality are presented. The classes are imbalanced, but at least for the MBTI personality, it is concordant with the statistics found in [Plank and Hovy, 2015] and given by the Myers-Briggs Foundation[4]. Both are also shown in Table 3.1.

[4] https://www.myersbriggs.org/my-mbti-personality-type/my-mbti-results/how-frequent-is-my-type.htm?bhcp=1

Tab. 3.1. Percentages of the Different Traits for the MBTI and DiSC Personality

MBTI	LinkedIn	Plank 2015	MBTI Foundation	DiSC	LinkedIn	Disc Profile website
I / E	35.7	36.0	50.7	D	60.5	24.8
N / S	72.0	73.0	73.3	I	57.2	25.1
T / F	55.6	58.0	59.8	S	27.3	25.7
P / J	34.0	41.0	45.9	C	36.9	24.4

According to the DiSC Profile website[5], the main traits are expected to be roughly equally distributed, with 25% of users in each personality type. However, a radically different distribution is found in the DiSC corpus collected from LinkedIn. This could be explained by two factors. Firstly, the fact that no discrimination is made between main and supporting traits changes the statistics. For instance, it might be possible that the Dominant trait appears more often as a supporting trait than the Steadiness trait or the Conscientious one. Secondly, it is also possible that Dominant and Influent users are simply more likely to put their personality test results on LinkedIn than Steady and Conscientious users, or that they use LinkedIn more. Further statistics on the corpus are presented in Section 6.1.

3.2. Feature Engineering

This section presents the features extracted from the LinkedIn profiles that will be used in the classification. The features are split in two parts: the textual features, which is to say all features that are derived from the texts, and the non-textual features, which represent all the other features[6].

3.2.1. Non-Textual Features

From each LinkedIn profile, 38 non-textual features are extracted:

(1) Number of connections,

(2) Presence of personal branding (binary),

[5] `https://www.discprofile.com/what-is-disc/faq/`

[6] The only features described here are the ones that seemed to have helped with the classification. For example, embeddings such as Fast-Text have also been tried, but did not give a good precision and therefore, are not presented here.

(3) Presence of a summary (binary),

(4) Number of entries in the work section,

(5) Average duration of jobs,

(6) Number of distinct companies in the work section,

(7) Number of descriptions normalized for the work section (number of descriptions divided by the number of jobs),

(8) Duration of the shortest job,

(9) Duration of the longest job,

(10) Number of entries in the education section,

(11) Number of distinct establishments in the education section,

(12) Number of descriptions normalized for the education section (number of descriptions divided by the number of entries),

(13) Highest level of education achieved (0 for under-university, 1 for undergraduate, 2 for master, 3 for doctorate and above),

(14) Number of entries in the voluntary experience section,

(15) Average duration of the voluntary experience,

(16) Number of descriptions normalized for the voluntary experience section (number of description divided by the number of entries),

(17) Number of skills,

(18) Total number of endorsements,

(19) Average number of endorsements,

(20) Number of endorsements by skilled people,

(21) Number of endorsements by colleagues,

(22) Number of recommendations given,

(23) Number of recommendations received,

(24) Number of accomplishments,

(25) Number of titles,

(26) Number of honor titles,

(27) Number of languages,

(28) Number of publicity,

(29) Number of tests taken,

(30) Number of courses,

(31) Number of patents,

(32) Number of sections, or number of different types of accomplishment,

(33) Number of interests,

(34) Average number of followers of the interests,

(35) Minimum number of followers of an interest,

(36) Maximum number of followers of an interest,

(37) Whether the user likes the company he is currently working at,

(38) Whether the user likes an influencer.

Some of the features appear only in a couple of profiles, such as "Number of publicity". Although intuitively it seems these won't be of use due to the limited number of users who fill them, they are nevertheless included in the non-textual features, and the decision on the usefulness of the feature is left for later.

3.2.2. Textual Features

For each LinkedIn profile, a text is collected, which is a combination of 1- the summary written by the user, and 2 - the descriptions of work experiences, educations, and voluntary experiences if those are present. As such, the text is not continuous, but since only frequency analysis is performed, it does not affect the classification.

Several textual features that have been shown to work well on this task in past studies are tried. Some work well only with some algorithms, while others do not work well at all with our corpora and algorithms. As mentioned in Section 1.2, LinkedIn text is fairly different from the texts from other social networks, since its users tend to be more reserved in what they write, and so features that work well on other social networks cannot necessarily be expected to work well here.

Before extracting the textual features however, the textual indicators that would give away the personality need to be removed. Due to the method of collection of the corpora, most of the texts contain the personality written in it, which has to be removed as the goal is to obtain a classifier that is able to generalize to profiles where users do not write their personality results.

To do so, the texts pass through several sets of regular expressions designed to identify and then remove any indicators of the personality. Although this affect the syntax of the sentences, it does not matter much for the chosen features, which do not take into account the structure of the sentences but only the individual words.

Once again, removing the MBTI indicators in the texts is very easy due to how they are written. The regular expressions are shown in Listing 3.1.

Listing 3.1. Regular Expressions for the MBTI

```
mbti=["(intj|INTJ)", "(intp|INTP)", "(entj|ENTJ)", "(entp|ENTP)", "(
    infj|INFJ)", "(infp|INFP)", "(enfj|ENFJ)", "(enfp|ENFP)", "(istj|
    ISTJ)", "(isfj|ISFJ)", "(estj|ESTJ)", "(esfj|ESFJ)", "(istp|ISTP)"
    , "(isfp|ISFP)", "(estp|ESTP)", "(esfp|ESFP)"]
```

For the DiSC personality indicators however, things are a little bit more complicated, since there are many ways to express each personality. As described earlier, one of the problems with using regular expressions for the DiSC labelling is that they capture things that are not the actual personality, but either results on other personality models that have a shared vocabulary with the DiSC personality model, or just regular words used in the texts. For example, "Achiever" and "Developer" are both in the vocabularies of the StrengthFinder personality test[7] and the DiSC personality model, which also includes common words such as "Result", "Creative", and "Agent". In the case of removing important indicators from the text however, it is acceptable to use a set of regular expressions that cuts out more than just the DiSC personality. It is of course a source of information that is lost, but allowing the classifier to generalize efficiently is without a doubt more important than to keep all of the information. The regular expressions for the DiSC personality are shown in Listings 3.2 to 3.5. One hundred randoms profiles are manually checked to ensure that all indicators of the personality are indeed removed from the texts.

Listing 3.2. Regular Expressions for the D Trait of the DiSC Personality

[7] https://www.gallupstrengthscenter.com/home/en-us/cliftonstrengths-themes-domains

```
D=["High[- ]D", "[sS]econdary D", "\"D\"", "[Dd]ominan[a-zA-Z]*",  "[
   ( +][Dd][ )+]", "[/-] [Dd] ", " [Dd][CIiS][CIiS ]", " [CIiS][D]",
   " [CIiS][CIiS][Dd] ", "[Dd][CIiS]?/", " [Dd][,;]", "\([CIiS]?D\)",
    "\(D[CIiS]?\)", "\[D\]", "[Cc]reative", "[Rr]esult(s?)( | - )[oO]
   rient", "[Ii]nspirational",  "[Ii]nvestigator", "[pP]ersuader", "[
   dD]-orient", "[dD]eveloper", "[Aa]chiever", "D[0-9]",  "&D",  "D&"
   ]
```

Listing 3.3. Regular Expressions for the I Trait of the DiSC Personality

```
I=["High[- ][Ii]", "[Ss]trong [iI]",  "[sS]econdary [iI]" , "\"[Ii]
   \"", "[Ii]nfluen[a-zA-Z]*", "[nN]atural I", ": [Ii]", ":  [iI][
   DSC]", "[Aa]daptive I", "[/-] [Ii] ", "[Ii] [/-]", " [Ii][CDS][CDS
    ]", " [Ii][CDS][CDS]", " [CDS][CDS][Ii] ", " [CDS][Ii] ",  "[Ii][
   CDS]?/", " I ,", "\([CDS]?[Ii]\)", "\([Ii][CDS]?\)", "\[[iI]\]", "
   [Aa]gent", "[Aa]ppraiser", "[Cc]ounselor", "[Ii]nspirational", "[
   pP]ersuader", "[pP]ractitioner", "[pP]romoter", "[Rr]esult(s?)( |
   - )[oO]rient", "[Ii]-orient", "(&|and) [Ii][DSC ]", "&[Ii]", "[Ii]
   &"]
```

Listing 3.4. Regular Expressions for the S Trait of the DiSC Personality

```
S=["High[- ]S", "\"S\"",  "[sS]econdary S", "[Ss]tead[a-zA-Z]*", " S
   ", "[/-] S ", " S[CIiD][CIiD ]", " [CIiD]S", " [CDIi][CIiD]S ", "S
   [CIiSD]?/", " S,","\([CIiD]?S\)", "\(S[CIiD]?\)","\[S\]", "[Aa]
   chiever", "[Aa]gent", "[Cc]ounselor", "[Ii]nvestigator", "[pP]
   erfectionist", "[pP]ractitioner", "[sS]pecialist", "[sS]-oriente",
    "S[0-9]", "&S",  "S&"]
```

Listing 3.5. Regular Expressions for the C Trait of the DiSC Personality

```
C=["High[- ]C", "\"C\"",  "[sS]econdary C", "[Cc]onsc[a-zA-Z]*", "[Cc
   ]autious", "[Cc]ompliance", " C ", " C ", " C[DIiS][DIiS ]", " [
   DIiS]C", " [SDIi][SIiD]C ", "C[DIiS]?/", " C,",
"\([DIiS]?C\)", "\(C[DIiS]?\)", "\[C\]", "[Aa]ppraiser", "[Cc]reative
   ", "[Ii]nvestigator", "[Oo]bjective", "[Tt]hinker", "[pP]
   erfectionist", "[pP]ractitioner", "[Cc]-orient", "C[0-9]", "&C",
   "C&"]
```

Once the indicators from the texts are removed, three kinds of features are extracted from it. Features are obtained from the General Inquirer, the PoS, as well as the TF-IDF on the 5000 most frequent words. The choice of 5000 words as a meta-parameter is justified in Chapter 4.

3.3. Combination of Features

Each algorithm is run with 5 distinct subsets of features. The features are separated for two reasons. The first one is that the collection method of the corpora, that is to say to take users who indicate their personality type online, seems to have biased the corpora towards users who fill their LinkedIn profiles fairly thoroughly, as shown in Section 6.2. Although this is not as critical for the validity of the research as one would think, since it is very difficult to collect a corpus involving humans that is not biased[8], and that the goal is simply to show that personality extraction on LinkedIn is doable, it still means that when deploying the system in the real world, profiles that are less complete might be problematic. However, regardless of the completeness of the profiles, the non-textual features are always applicable, with some modifications to the algorithm to take into account the lesser number of data, giving us an idea of the results obtained for someone that did not write any text in their profile. Of course that would mean that, when using the system in the real world, the point at which the text is not enough to obtain a good performance would need to determined to then switch to the non-textual classifier.

There's also, in the separation of the non-textual features from the textual features, a possibility to eventually perform personality extraction with any language. Performing

[8] In most psychological studies, there is a bias introduced by the method of collection. For example, even if students had been asked to take personality tests, it would still have introduced a bias in our corpus in the sense that it would have been a corpus of highly educated students. In a similar way, posting an announcement in a newspaper and promising money in exchange for taking a personality test would bias the corpus towards people who need money quickly, which may correspond to a particular class of the population.

personality extraction to the same level on another language would require some adaptations of the features used. An adaptation of the General Inquirer and of the PoS would have to be found for the other language, and there is no guarantee that the TF-IDF features would perform as well in another language than it does in English, nor that the other language can easily be parsed into a TF-IDF vector. The one invariant aspect, however, are the non-textual features, which are easy to extract no matter the language. There may be some cultural variations that affect which features are important[9], but ultimately it is expected that, by retraining the classifier, very similar results could be achieved.

The separation of the TF-IDF features from the textual features is made for a more practical reason, being the issue of running time. Using feature selection with the SVM classifier, as presented in Section 4.2.1 takes a long computing time with only the 418 features of the textual feature set, and so it would be quite impractical to run it on the TF-IDF feature set, which contains 5000 features. For that reason, it is separated from the textual features.

The different subsets of features are:

(1) The non-textual features,

(2) The textual features, which regroup the PoS and the General Inquirer counts,

(3) The TF-IDF features,

(4) All hand-crafted features together (the features of group 1 coupled with the features of group 2),

(5) All features together (the features of group 1, 2, and 3 together).

[9] It has been shown that there are variations in the expression of the same personality trait across cultures and languages. For example, American extroverts make fewer pauses when talking, while German extroverts make more pauses than German introverts [Kaushal and Patwardhan, 2018].

Chapter 4

Prediction of Personality on LinkedIn

This chapter describes our experiments trying to predict the personality of users based on their LinkedIn profile. The train-test sets are obtained with a 20-fold cross validation, and 8 classifiers are trained per algorithm; 1 per trait, 4 per personality model, which act as binary classifiers to decide if the user has that trait or not. The algorithms tested are Naive Bayes, SVM, and Random Forest. Despite their reported successes in the literature, preliminary experiments with several architectures of neural networks gave results that could not compete with the other algorithms presented in this chapter[1].

4.1. Naive Bayes

As mentioned in Section 1.4.1, the Naive Bayes algorithm is mostly used to provide a classification baseline and therefore not much time is spent trying to optimize it. The Naive Bayes is run with all the sets of features described in Section 3.3, and results are reported for each personality traits in Table 4.1, as well as the Majority rule baseline. AdaBoost is also applied to the Naive Bayes to see if it would help the performance, as shown in Table 4.2.

Worthy of note is the very unequal performance of the AdaBoost algorithm. Depending on the trait and on the subset of features, using AdaBoost either achieves a little increase in precision or a large decrease, as observable by the fact that the Naive Bayes with AdaBoost can rarely surpass the Majority Rule baseline, while the one without AdaBoost always does.

[1] The architectures tried include several configurations of fully-connected feed forward neural networks, and several configurations of LSTM neural networks, both at the character and at the token level with pre-trained embeddings. The maximum precision reached by any of these neural networks on the Dominance trait of the DiSC personality model is 61.3%, and the majority baseline on that trait is 60.5%.

Tab. 4.1. Precision (%) over the Different Traits for the MBTI and the DiSC Personality Models for the NB Algorithms

Traits	Majority Rule	Non-Textual	Textual	TF-IDF	Hand-Crafted	All
D	60.5	46.9	47.5	**69.6**	46.5	46.5
I	57.2	58.6	52.2	**68.3**	57.6	57.6
S	72.7	37.9	42.5	**76.9**	41.1	41.1
C	63.1	63.4	47.7	**71.9**	52.3	52.3
IE	64.3	62.6	53.6	**73.3**	59.2	59.2
NS	72.0	47.1	59.5	**77.3**	69.3	69.3
TF	55.6	48.5	50.2	**72.6**	48.3	48.3
PJ	66.0	65.3	42.0	**72.1**	41.6	41.6

Tab. 4.2. Precision (%) over the Different Traits for the MBTI and the DiSC Personality Models for the NB Algorithm with AdaBoost

Traits	Majority Rule	Non-Textual	Textual	TF-IDF	Hand-Crafted	All
D	**60.5**	48.7	52.3	47.9	48.0	45.9
I	**57.2**	53.0	47.5	54.1	49.3	52.2
S	**72.7**	51.8	57.5	70.8	55.1	55.5
C	63.1	51.5	53.9	**66.3**	50.3	49.3
IE	64.3	51.1	57.1	**72.1**	52.0	51.5
NS	**72.0**	52.2	56.0	55.7	51.6	48.2
TF	**55.6**	47.7	52.3	48.9	50.5	48.9
PJ	66.0	52.5	48.8	**70.4**	53.5	54.8

There are still, however, places where the AdaBoosted Naive Bayes performs better, for example, on the textual features, for most of the traits. This lack of consistent performance of the Naive Bayes with the AdaBoost algorithm is concordant with the findings of [Ting and Zheng, 2003], where a detailed analysis of the behavior of the Naive Bayes with AdaBoost is presented.

Taking the Dominance trait as an example, we can see that for most feature sets, the Naive Bayes performs worst than the majority baseline, which is in this case 60.5%. The TF-IDF feature set however, without AdaBoost, gains a bit of precision and reaches 69.6%. That is true for all the traits, and shows that the Naive Bayes baseline is a stronger baseline than the Majority Rule one, as better results are obtained across all traits.

4.2. Support Vector Machine

In previous work on personality extraction, the SVM have often been found to be the best classifier [Markovikj et al., 2013; Lima and Castro, 2016]. In this thesis, a feature selection algorithm is used on a SVM with a Gaussian kernel to increase its precision, since it has been shown SVMs can perform badly if they receive too many irrelevant features [Weston et al., 2001].

4.2.1. Feature Selection

The feature selection algorithm works in three steps: Initialisation, Reranking, and Removing and Adding.

The initialisation is the only step that is run only once. The algorithm first picks k features and, out of these k features, one feature f_0. The precision of the SVM is calculated based solely on that feature, and then based on every pair of features (f_0, f_i), where $f_i \epsilon (1, k)$. The features are finally ranked based on their Precision Gain, defined as

$$PG(f_i) = p(f_0, f_i) - p(f_0) \tag{4.2.1}$$

meaning that a feature will come first if adding it to the mix of features seems to improve the performance.

Once the k features are ranked in that manner, the algorithm enters the phase of reranking. Reranking consists of running k classifiers, adding one after the other the features previously ranked, and then of reranking them according to the precision obtained on the classifiers, without moving the first m features, which obviously will have worst ranks due to the very limited number of features. It is logical that the SVM will perform worst when trained with only one or two features, even if they are good ones, and so for these first features, the precision reached is not very representative of their performance. Reranking is performed five times. Several variations on this step are tried, such as randomly swapping one of the first m features for another feature, or taking the precision gain instead of the precision for reranking, but do not bring any improvement and so are not kept. Multiples values of m are also tried, and at the end it is fixed to 3.

The final step is the removal and adding of features, in which the worst c features are removed and another c features chosen at random is added to the considered features, before jumping to the step of reranking, until all features have been tested.

All along, the best performance as well as the best combination of features is recorded, and at the end, the feature selection algorithm returns them. To account for the variability in this process, it is run a total of 15 times, and the best result of all is the final result. For the non-textual feature set, k and c are fixed at 15 and 5, and for the other feature sets, at 30 and 10. The best features are listed for each trait in Appendix B.

This feature selection algorithm is used instead of more traditional feature ranking, such as performing chi-squared tests, because it allows to, first, perform feature selection while taking the features as a group and assessing the interactions between them and, second, perform feature selection in the kernel space used by the SVM. Experiments also showed a better performance while using this feature selection algorithm versus ranking the feature individually with a statistical test. A more complete analysis of feature selection and an overview of the traditional techniques is presented in [Guyon and Elisseeff, 2003].

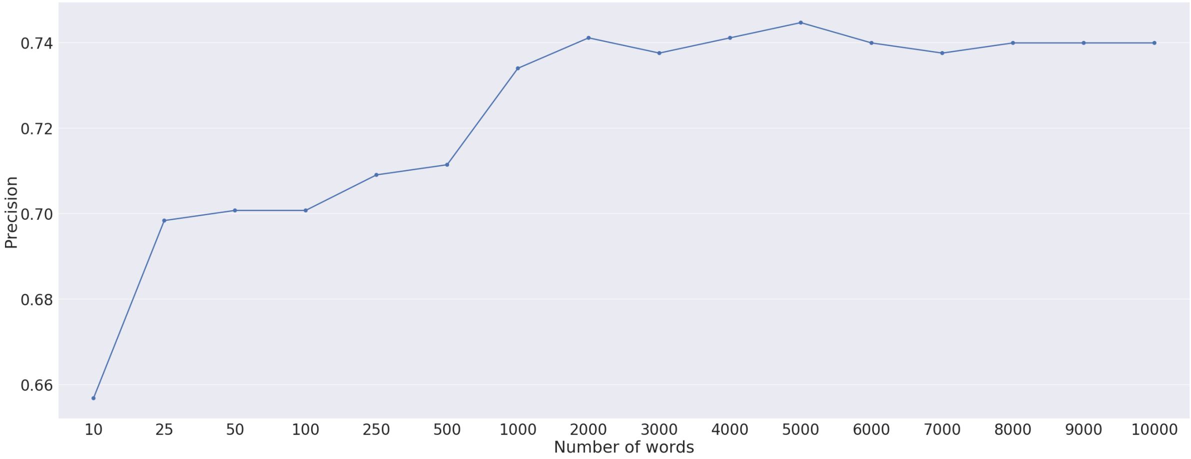
4.2.2. Optimization

Once the feature selection algorithm has returned the best features, a grid search over λ and γ is performed to find the best possible configuration. The considered values of λ are $[2^{-5}, 2^{-3}, 2^{-1}, 1, 2^{1}, 2^{3}, 2^{5}, 2^{7}, 2^{9}, 2^{11}, 2^{13}, 2^{15}]$, and the considered values of γ are $[2^{-15}, 2^{-13}, 2^{-11}, 2^{-9}, 2^{-7}, 2^{-5}, 2^{-3}, 2^{-1}, auto, 2^{1}, 2^{3}, 2^{5}]$. The *auto* value is the default value in the python package sklearn[2], and corresponds to $\frac{1}{n_{features}}$. The optimization gives a boost of on average 1.11% of precision.

The feature selection algorithm is used with the non-textual, textual, and hand-crafted feature sets. Due to the time taken by the feature selection algorithm, it is not used with the TF-IDF set of features. For the set of features denoted *All*, which represents the hand-crafted

[2] `https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html`

Fig. 4.1. Precision Reached by the SVM Depending on the Number of Words Included in the TF-IDF Vector for the Dominance Trait of the DiSC Personality Model



and the TF-IDF appended together, two configurations are tried, being all the hand-crafted features appended with the TF-IDF, as well as simply the 30 features returned by the feature selection algorithm appended to the TF-IDF features, and the best results are returned.

Concerning the number of words used in the TF-IDF, Figure 4.1 shows that the results seem fairly stable after 2000 words. The number of words is fixed at 5000 nevertheless, both to have a common ground with the Random Forest, presented in the next section, and because adding more words does not decrease precision, but simply augment the time needed for the calculation, but not enough to be an inconvenience.

Results for each trait and each set of features are shown in Table 4.3. Overall, the SVM performs well, when compared to the two baselines. In fact, it always surpasses the Naive Bayes baseline, with either the TF-IDF features or the Hand-Crafted ones. Using once again the Dominance trait as an example, the Naive Bayes baseline gives a result of 69.6%, and the SVM reaches 74.5% of precision, with the TF-IDF features. The SVM was used mostly because it was found to be useful in other studies on automatic personality extraction and that the use of the Gaussian kernel allows to find non-linear correlations between the features and the traits.

The fact that the TF-IDF features, without feature selection, perform often better than the other sets of features raises the question of how necessary is the feature selection step. To prove its usefulness, two demonstrations of its necessity are now presented. First, Table 4.4 shows the results on the non-textual, the textual, and the hand-crafted feature sets,

Tab. 4.3. Precision (%) Over the Different Traits for the MBTI and the DiSC Personality Models for the SVM. * Means Feature Selection is Applied to That Feature Set

	Naive Bayes	Non-Textual*	Textual*	TF-IDF	Hand-Crafted*	All
D	69.6	72.3	66.5	**74.5**	73.7	71.5
I	68.3	73.2	66.7	**74.3**	73.8	67.6
S	76.9	80.5	77.7	**80.6**	80.5	77.9
C	71.9	75.5	74.1	**77.3**	76.7	72.0
IE	73.3	77.5	71.8	76.7	**77.9**	71.6
NS	77.3	82.0	77.1	82.1	**83.7**	77.3
TF	72.6	69.4	68.3	**77.4**	71.7	68.8
JP	72.1	77.4	73.6	77.8	**78.4**	73.6

with and without feature selection with an optimized SVM. Without it, the SVM performs barely over the majority baseline but, with it, it considerably gains in precision. Secondly, Figure 4.2 shows the result on an optimized SVM when adding one after the other the best features returned by the feature selection algorithm for the non-textual feature sets with the Dominance trait of the DiSC personality model. Adding the 7th and the 13th feature decreases precision significantly and the precision is never able to go back to what it was before, showing that adding unnecessary features does indeed hurt the overall performance[3].

4.3. Random Forest

Random Forests have a lot of variability in their results depending on the split of features and data made during training. That is even more true when working with a small dataset. Since the goal is to have an accurate estimate of the generalization performance, a search over the number of trees, or estimators, is performed to find the point after which there is not much variance in the results. As shown in Figure 4.4, after 4000 trees the variance of the results seems to stabilize, and so to be safe all further experiments are run with 5000 trees.

[3] The 7th feature in this case correspond to the average duration of voluntary experiences, and the 13th to the average number of followers for the interest section.

Tab. 4.4. Precision (%) over the Different Traits for the MBTI and the DiSC Personality Models for the SVM With and Without Feature Selection. NT stands for Non-Textual, T for Textual, HC for Hand-Crafted, and w/ and wo/ Stands for With Feature Selection and Without Feature Selection

	NT wo/	NT w/	T wo/	T w/	HC wo/	HC w/
D	61.2	**72.3**	60.9	**66.5**	60.9	**73.7**
I	58.2	**73.2**	57.7	**66.7**	57.7	**73.8**
S	73.2	**80.5**	72.9	**77.7**	72.9	**80.5**
C	63.8	**75.5**	63.3	**74.1**	63.3	**76.7**
IE	64.3	**71.8**	64.5	**71.8**	64.5	**77.9**
NS	72.0	**77.1**	72.1	**77.1**	72.1	**83.7**
TF	55.8	**68.3**	56.2	**68.3**	56.2	**71.7**
PJ	66.0	**77.4**	66.2	**73.6**	66.2	**78.4**

Fig. 4.2. Precision vs the Number of Features on an Optimized SVM for the Dominance Trait of the DiSC Personality Model, With the Features Returned by the Feature Selection Algorithm For the Non-Textual Feature Set

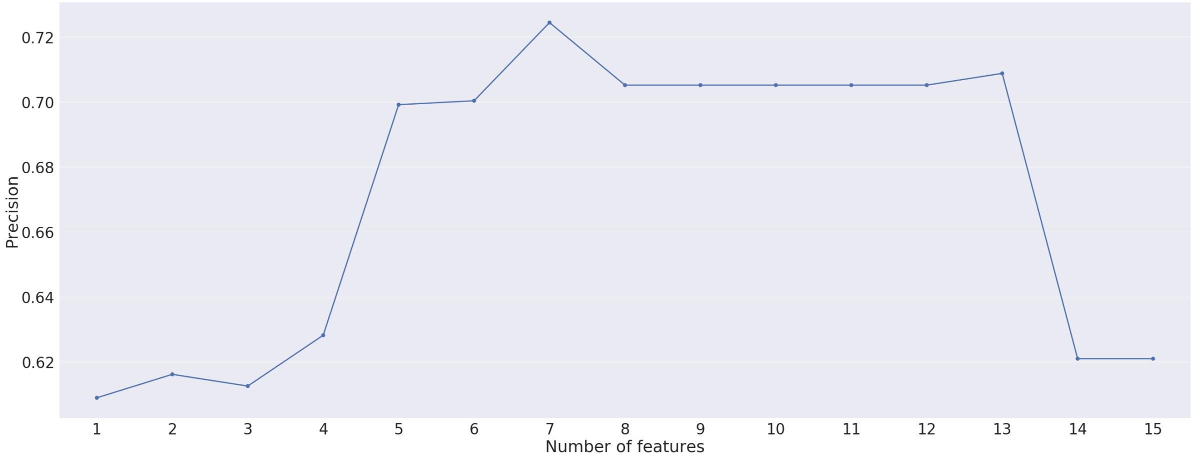


Figure 4.4 shows the precision of the Random Forest depending on the number of words included, for 4 different kinds of treatments, those being: a natural Random Forest, a Random Forest where stop-words are removed before the TF-IDF transformation, a Random Forest with AdaBoost, and a Random Forest with AdaBoost and stop-words. The graphic

Fig. 4.3. Precision vs the Number of Estimators Used in the Random Forests with the Dominance Trait of the DiSC Personality Model and the TF-IDF Feature Set. Each Point Corresponds to an Experiment

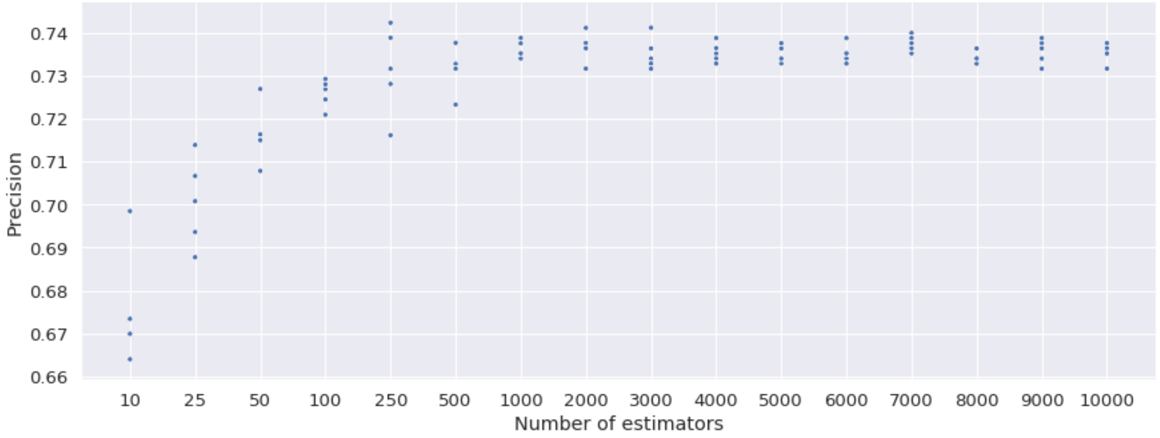
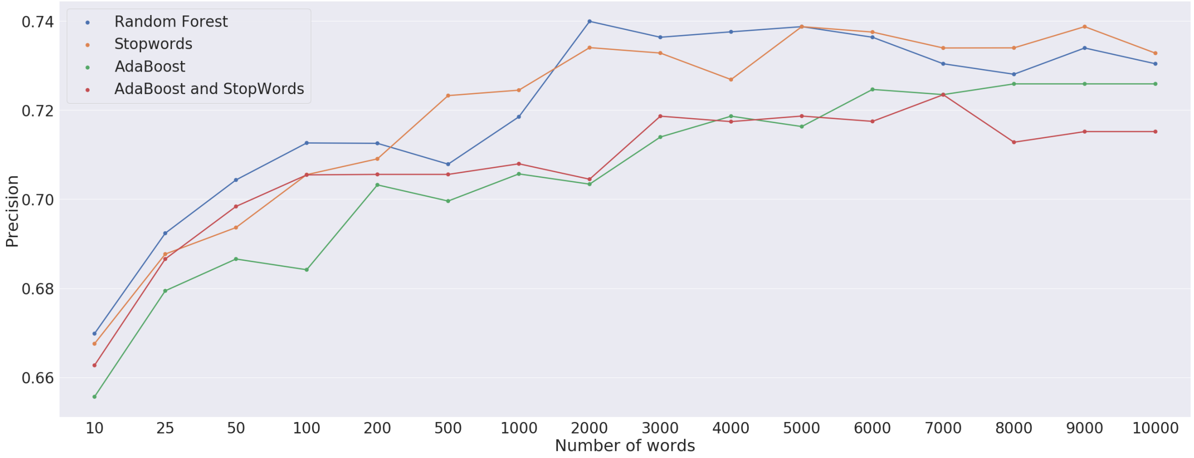


Fig. 4.4. Precision vs the Number of Words for Different Kind of Treatments with the Random Forest with the Dominance Trait of the DiSC Personality Model



offers two pieces of information. First, the use of the 5000 most frequent words for the TF-IDF seems to offer an excellent precision, and second, the normal Random Forest with no treatment seems to work the best. Stop-words don't seem to make a difference, probably due to the capacity of Random Forest to identify them as useless features, and AdaBoost seems to worsen the performance, and takes much longer to run.

Tab. 4.5. Precision (%) Over the Different Traits for the MBTI and the DiSC Personality Models for the RF

	NB	SVM	Non-Textual	Textual	TF-IDF	Hand-Crafted	All
D	69.9	**74.5**	72.5	72.7	73.6	72.5	72.1
I	68.3	74.3	75.1	74.5	**75.2**	75.1	**75.2**
S	76.9	**80.6**	79.3	79.3	79.9	80.5	79.3
C	71.9	**77**.3	74.1	74.1	76.0	74.1	74.3
IE	73.3	77.9	75.2	82.6	78.3	**84.4**	76.7
NS	77.3	**83.7**	79.1	81.5	80.5	76.6	79.8
TF	72.6	77.4	68.9	78.3	77.7	**83.5**	76.5
JP	72.1	78.4	72.4	76.5	77.4	**80.7**	76.4

As with the SVM, the Random Forest is run with all possible sets of features. The results are shown in Table 4.5, where the Naive Bayes baseline and the best results obtained with the SVM are also shown.

The Random Forest performs much better than expected. It was originally tried simply because it is known for having a good capacity to perform feature selection, which is essential in this kind of complicated classification, where there are many features and it is difficult to know which ones would be useful. Furthermore, it is able to control well the over-fitting problem, due to its many predictors. However, as few works that used Random Forests were found, we did not expect that it to perform so well. The original theory was that it would perform worse than the SVM, as it only looks for correlation in a linear space and not in a kernel space. In practice, however, it gets results close to the SVM, surpassing it for half of the traits. An analysis and comparison of the algorithms is presented in the next section.

4.4. Analysis

Table 4.6 reports the best results by feature type, regardless of the algorithm. This allows for an easy comparison of which sets of features perform the best. For the DISC personality extraction, the TF-IDF features constantly get the best results, and for the MBTI, the hand-crafted ones work best.

Tab. 4.6. Precision (%) over the Different Traits for the MBTI and the DiSC Personality Models Depending on the Feature Subset

	Non-Textual	Textual	TF-IDF	Hand-Crafted	All
D	72.5	72.7	**74.5**	73.7	72.1
I	75.1	74.5	**75.2**	75.1	**75.2**
S	80.5	79.3	**80.6**	80.5	79.3
C	75.5	74.1	**77.3**	76.7	74.3
IE	77.5	82.6	78.3	**84.4**	76.7
NS	82.0	81.5	82.1	**83.7**	79.8
TF	69.4	78.3	77.7	**83.5**	76.5
JP	77.4	76.5	77.8	**80.7**	76.4

For the non-textual features, the SVM with feature selection seems to perform better than the Random Forest, while it seems to perform worst for all the other feature sets. The only difference between the feature selection for the non-textual features and the feature selection for the other feature sets is the final number of features, which is 15 instead of 30, and the number of features eliminated each round, which is 5 instead of 10. One could imagine that running the SVM on the other feature sets with those same hyper-parameters would then boost the SVM performance higher than what the Random Forest can achieve, but experiments show that this is not the case and that the hyper-parameters of 30 and 10 perform better for the textual and hand-crafted sets of features. This then suggests that there may be an optimal number of final features and of features eliminated which would boost performances, and that a thorough grid-search could possibly increase the performances of the SVM higher than the Random Forest. Unfortunately, running time issues prevent the finding of these optimal hyper-parameters.

Table 4.7 reports, for each of the classifiers, the best performances for each traits, regardless of the subset of features used. The Random Forest is the most efficient for the MBTI classification, but for the DiSC classification, the SVM seems to perform better. A reason for this could be that there are more correlations in a non-linear space than in the linear space, and so the Random Forest is not able to outperform the SVM.

Tab. 4.7. Precision(%) over the Different Traits for the MBTI and the DiSC Personality Models Depending on the Algorithm

	Majority Rule	Naive Bayes	Random Forest	SVM
D	60.5	69.6	73.6	**74.5**
I	57.2	68.3	**75.2**	74.3
S	72.7	76.9	80.5	**80.6**
C	63.1	71.9	76.0	**77.3**
IE	64.3	73.3	**84.4**	77.9
NS	72.0	77.3	81.5	**83.7**
TF	55.6	72.6	**83.5**	77.4
PJ	66.0	72.1	**80.7**	78.4

Independently of the algorithms, the results in themselves seem pretty good. One thing to notice is that results are a little bit worse for the DiSC personality model than for the MBTI one, when comparing the traits that have approximately the same majority baseline. For example, comparing the Steadiness trait and the Intuition-Sensing dimension, which have respectively 72.7% and 72.0% as majority baselines, the Steadiness trait obtains a maximum precision of 80.6%, and the Intuition-Sensing of 83.7%. This tendency is observable across all traits. As mentioned in Section 1.1.3, there have been few studies on the DiSC personality models, and no independent studies were found on the validity of the model, and so one of the explanation for the worst results is that it is less representative of the personality of an individual and therefore, more difficult to capture. Section 6.3 mentions some more sources of errors that could explain why we have a difference in the performance of the algorithms between the two personality models.

As there are no other studies on LinkedIn, it is not possible to compare the results obtained to state-of-the-art performances. The next best thing is to get studies on other social networks that have for object the MBTI personality model to compare to these. Table 4.8 reports the best results obtained in this section, the results of [Noecker et al., 2013] (Study 1), the results of [Hernandez and Knight, 2017] (Study 2), and the results of [Plank and Hovy, 2015] (Study 3). The three studies are described in more details in Chapter 2, but Study 1 is performed on a corpus of text written by students, Study 2 on a corpus of a forum

Tab. 4.8. Comparison between the Results on LinkedIn and Results of Other Studies for the MBTI Personality Model

Traits	LinkedIn	Study 1 (Texts)	Study 2 (Forum)	Study 3 (Twitter)
IE	**84.4** (+18.1)	76.4 (+21.2)	67.0	72.5 (+8.5)
NS	**83.7** (+11.7)	76.5 (+22.7)	62.0	79.5 (+2.5)
TF	**83.5** (+27.9)	80.5 (+8.1)	77.8	61.2 (+3.2)
PJ	80.7 (+14.7)	**84.5** (+3.8)	63.7	58.2 (-0.8)

based around the MBTI personalities, and Study 3 on Twitter. Since the classes are very imbalanced, the improvements over the majority baseline are also reported in parentheses, when these are available, to be able to say how much of the difference between the results of the studies is because of the algorithmic differences or because the baseline was higher. Even then, however, the comparison is to be taken lightly since the studies are performed on different social networks, which could also explain differences of results.

The results obtained on LinkedIn are better on all the traits and across all studies, except for the Perception-Judging dimension, for which Study 1 gets a better precision. The natural conclusion is still that not only is it possible to extract correctly the personality from LinkedIn profiles, but the results of the extraction can be competitive with other studies on the MBTI and with other social networks.

Chapter 5

Cross-Personality Extraction

Chapter 4 explores whether it is possible to extract a personality model from a LinkedIn profile according to two personality models, DiSC and MBTI. In this chapter, the global idea is pushed a bit further, to see whether it is possible to predict the personality using another personality model, action that we denote here cross-personality extraction. Cross-personality extraction could have multiple applications, such as obtaining the results on several personality models at the cost of only one personality test, or being able to obtain a bigger corpus with high precision using a bootstrapping technique.

The algorithms used in this chapter are the same as in the preceding one, except that each feature set is appended with the other personality model (DiSC when performing MBTI classification, and vice-versa). The classification is done both ways, from DiSC to MBTI and from MBTI to DiSC. The Naive Bayes is not used, as Chapter 4 shows it does not perform very well and the baseline will instead be the scores of the SVM and Random Forest obtained in Chapter 4, with the goal of estimating if adding the other personality model gives a boost of precision when compared to regular personality extraction.

The corpus used is composed of 831 profiles that are labelled with both the MBTI and DiSC personalities, and 20-fold cross validations is once again applied to get the training and testing dataset. Section 5.1 presents some correlations between the DiSC personality model and the MBTI personality model that hint that using the other personality as feature could be beneficial, Section 5.2 shows the results achieved on individual classifiers, and Section 5.3 demonstrates that a system of experts can boost the precision significantly at the cost of recall.

In order to make it easier for the reader, the expression "personality extraction" refers solely to extracting the personality from LinkedIn profiles, as presented in Chapter 4, while the expression "cross-personality extraction" refers to trying to predict the personality on one model starting with the personality on another model, using as well additional information.

5.1. Correlations

In order to know if cross-personality extraction is feasible and has an advantage over personality extraction, the first step is to check whether there are correlations between the two personality models. Although even an absence of linear correlation still wouldn't mean that the other personality model is not a good feature, as it is possible, for example, that there are non-linear correlations, finding some linear correlations will allow to try cross-personality extraction with more confidence, and as explained in Section 2.1, correlation studies like this are useful for many reasons. Table 5.1 shows the results of the chi-squared test between the dimensions of the two personality models. Since each dimension is a binary variable, 16 independent tests are run. There are a lot of correlations between the two models, with strong confidence levels. The only dimensions that are not correlated are Intuition and Judging, with Dominance and Steadiness. In order to know the strength and the direction of the correlations, 16 Phi tests are run in a similar fashion, for which results are presented in Table 5.2.

The Introversion dimension of the MBTI presents the strongest correlations, especially with the Influence, Steadiness, and Conscientiousness dimensions of the DiSC model. The 16 contingency tables used for calculating the correlations are shown in Appendix C .

The MBTI and DiSC personalities can also be seen as, respectively, the 16 and 12 types composing them. For the DiSC personality, the 12 types are all combinations of the traits, excluding the people who said they had 3 high scores, and the case where someone would say they have no high scores, or 4 high scores, which happened only once. The result of the chi-squared test between the 16 types of the MBTI and the 12 types of the DiSC personality models allow to affirm that there is correlation between the two at $p < 0.0001$. However, since the expected counts are small, this p-value is to be taken lightly.

The correlations hint that it is possible to use the other personality model as features to increase the precision of our classifier. However, it is doubtful that good precision scores can

Tab. 5.1. Correlation Between the Traits of the MBTI and DiSC Personalities. Y Means There is a Correlation, N Means That There is no Correlation, p is the Confidence Level

	D	i	S	C
IE	Y (p<0.0001)	Y (p<0.0001)	Y (p<0.0001)	Y (p<0.0001)
NS	N	Y(p<0.0001)	N	Y (p<0.001)
TF	Y (p<0.0001)	Y(p<0.0001)	Y (p<0.0001)	Y(p<0.0001)
JP	N	Y(p<0.0001)	N	Y(p<0.0001)

Tab. 5.2. Phi Coefficients Between the Traits of the MBTI and DiSC Personalities

	D	i	S	C
IE	-0.19	-0.44	0.31	0.42
NS	0.02	0.18	-0.07	-0.12
TF	0.26	-0.20	-0.22	0.16
JP	0.03	-0.19	0.04	0.19

be obtained using only the other personality model, since the correlations are not strong, but medium or weak.

When running the SVM and Random Forests, all subsets of features are the same as in Chapter 4, with the exception that they are appended with the other personality model results (DiSC for the MBTI classification and vice-versa). Two ways of doing so are tried: by appending the personality as the four individual traits (D, I , S, and C, or IE, NS, TF, and PJ), or as the types (a one-hot encoding for each types saying if the person is or not the type, where the type is the combination of the 4 traits, such as ENFJ or DC). The four individual traits give better results, so those are used in the feature sets. In the following tables, the feature sets denoted "MBTI" and "DiSC" mean the classification is attempted using as feature only the other personality model.

5.2. Cross-Personality Extraction Using Individual Classifiers

5.2.1. Random Forest

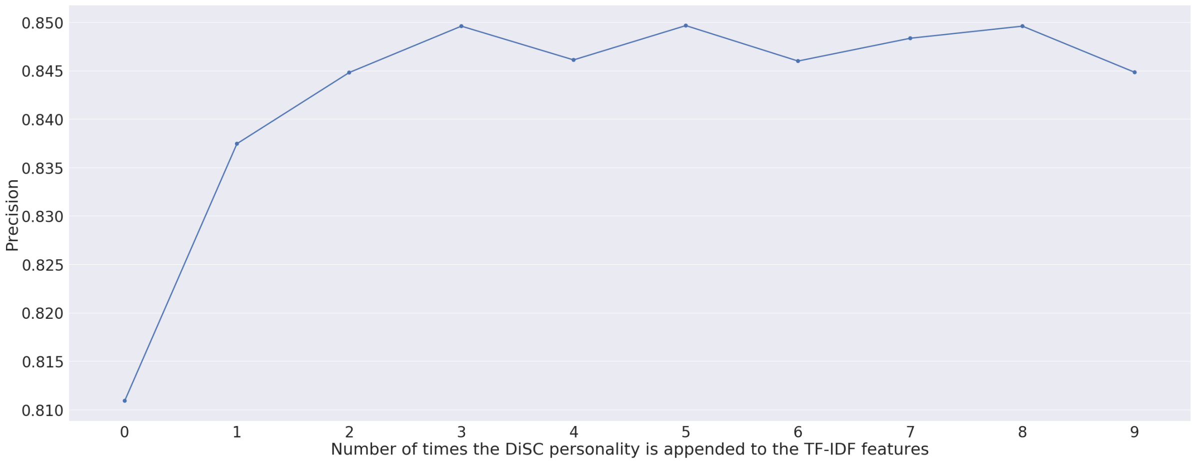
Initial attempts with the TF-IDF appended with the other personality model yielded a greater increase in precision for the DiSC classification than for the MBTI one, when comparing to the personality extraction results for the same traits. By analyzing the data, one can see that one thing that could create such a difference is how the personality is expressed in the text. For the MBTI, it is expressed as a clear 4 letter code, but for the DiSC personality model, as presented in Section 1.1.3, the expression of the personality textually is more diverse and more difficult, for both a human and an algorithm, to capture. That means that in the TF-IDF features for the DiSC classification, most likely the 16 codes representing the MBTI personality are present and represent good features, but due to the multiple ways the DiSC personality is expressed in the text, the TF-IDF may not be able to capture the DiSC personality types.

Since the other personality is already appended to the TF-IDF features, the logical conclusion is that as they are correlated features, having them more than once improves the odds the RF will select them when choosing a subset of features, a process described in Section 1.4.3, and therefore, it will increase the global precision reached.

That intuition proves correct, as seen in Figure 5.1, which represents the precision reached on the Random Forest versus the number of times the DiSC personality model is appended to the TF-IDF feature set for the classification on the Introversion trait of the MBTI personality model. The technique is also used for the DiSC classification, appending more than once the MBTI personality model, with a similar increase in precision. For the SVM, however, it does not improve classification and therefore, the other personality model is appended only once.

The results of the experiments with the Random Forest are presented in Table 5.3. One thing to observe is that, when compared with the best results from Chapter 4, independently of the algorithms, there is a slight increase of the precision reached for the DiSC personality model, and a slight decrease for the MBTI personality model. For example, when performing the personality extraction task, the SVM reaches 74.5%, and the Random Forest 73.6% on the Dominance trait of the DiSC personality model, but while performing cross-personality

Fig. 5.1. Precision vs the Number of Times the DiSC Personality Is Included as Features for a Random Forest Trained on the TF-IDF Features with the Introversion Trait of the MBTI Personality Model



Tab. 5.3. Cross-Personality Results for Random Forest Depending on the Input for the DiSC and MBTI Personality Models. PE Stands for Personality Extraction

	PE	MBTI	NT	TF-IDF	Text	Hand-Crafted	All
D	74.5	67.2	74.6	**75.6**	73.0	72.8	72.6
I	75.2	73.1	78.8	**81.6**	76.9	76.9	77.4
S	80.6	78.6	80.4	**81.9**	79.8	79.5	79.5
C	77.3	73.9	80.2	**82.3**	75.7	75.6	75.1

	PE	DiSC	NT	TF-IDF	Text	Hand-Crafted	All
IE	**84.4**	76.2	83.3	84.2	79.3	79.3	79.3
NS	**83.7**	68.7	78.1	80.0	78.4	78.4	78.7
TF	**83.5**	67.9	74.6	80.3	77.1	76.6	76.5
JP	**80.7**	65.2	75.2	80.1	77.7	77.7	77.5

extraction, the Random Forest reaches 75.6%. This is not as significant of a gain as originally expected, and so in Section 5.3, a voting system is created with the goal of increasing further the precision reached. Before that however, the cross-personality extraction task is performed with the SVM, to see if it yields a greater increase of the precision than the Random Forest.

Tab. 5.4. Cross-Personality Results for Support Vector Machine Depending on the Input for the DiSC and MBTI Personality Model. PE Stands for Personality Extraction

	PE	MBTI	NT	TF-IDF	Text	Hand-Crafted	All
D	74.5	68.5	74.6	**78.8**	71.6	72.7	68.1
I	75.2	73.6	76.3	**81.7**	74.8	73.7	71.6
S	80.6	78.7	81.9	**83.5**	78.8	79.9	77.3
C	77.3	74.5	76.9	**82.8**	75.6	76.1	69.9

	PE	DiSC	NT	TF-IDF	Text	Hand-Crafted	All
IE	84.4	77.3	79.7	**85.8**	77.3	78.3	73.6
NS	**83.7**	69.8	82.6	80.9	76.8	79.3	76.2
TF	**83.5**	69.8	74.1	81.2	75.8	73.9	68.1
JP	**80.7**	68.0	78.9	79.4	76.9	77.9	75.5

5.2.2. SVM

The SVM is also run in a similar fashion as in Chapter 4, rerunning the feature selection algorithm and the grid-search optimization for all traits and all feature sets, appended with the other personality model. The results for the SVM are presented in Table 5.4.

There is a greater increase of the precision with the SVM than with the Random Forest. Taking once again the Dominance trait, the precision obtained is now 78.8%, while the Random Forest obtains 75.6%, and the best result obtained on the personality extraction task is 74.6%, showing thus that there indeed is an advantage to adding the other personality model as features.

Table 5.5 shows the performance of the algorithms with and without the use of personality model as features, and in parentheses, the increase or decrease of precision when adding the other personality model as features. These results are the best regardless of the feature set. One thing to point out is that the prediction with only the other personality model alone is not sufficient to surpass the results of the simple personality extraction, as was surmised in Section 5.1 by observing the correlations between the personality models.

Regrouping the results by personality models and by algorithms gives the following average increases: 4.98% for the DiSC classification with the SVM and 4.75% with the RF,

Tab. 5.5. Comparison of the Results With and Without the Addition of Another Personality Model

	SVM without MBTI	SVM with MBTI	RF without MBTI	RF with MBTI
D	74.5	**78.8** (+4.3)	72.7	75.6 (+2.9)
I	74.5	**81.7** (+7.2)	75.1	81.6 (+6.5)
S	80.6	**83.5** (+2.9)	80.5	81.9 (+1.4)
C	77.3	**82.8** (+5.5)	74.1	82.3 (+8.2)

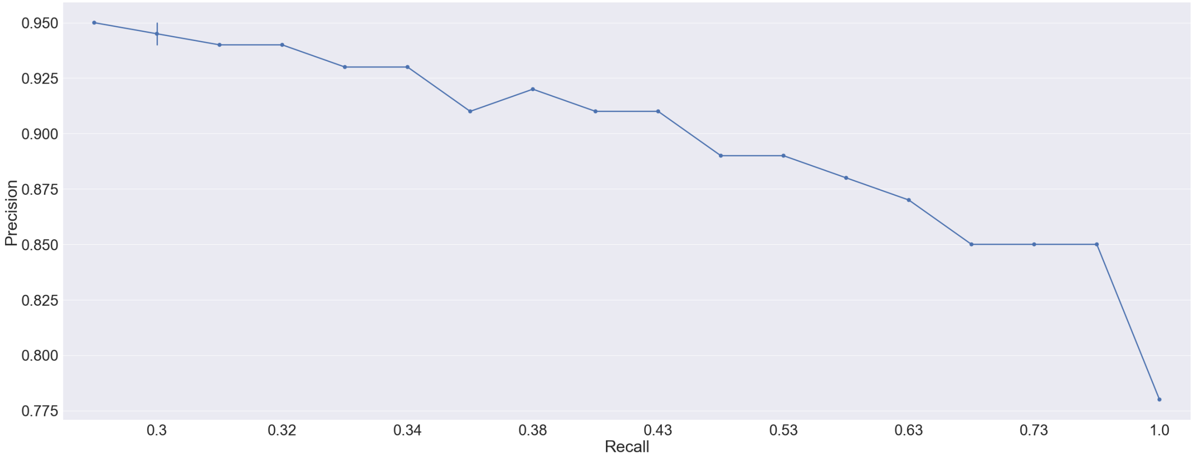
	SVM without DiSC	SVM with DiSC	RF without DiSC	RF with DiSC
IE	77.9	**85.8** (+7.9)	84.4	84.2 (-0.2)
NS	**83.7**	82.6 (-1.1)	81.5	80.0 (-1.5)
TF	77.4	81.2 (+3.8)	**83.5**	80.3 (-3.2)
PJ	78.4	79.4 (+1.0)	**80.7**	80.1 (-0.6)

and 2.90% for the MBTI classification with the SVM, and -1.38% with the RF. Except for the RF for the MBTI classification, there is a slight improvement for both personality models. About 2% more is gained on the DiSC classification compared to the MBTI classification, but that is mostly due to Intuition and Judging dimensions, where either a decrease in performance or a very small improvement is obtained, which makes sense since these two dimensions are also the ones presenting the weakest correlations with the DiSC personality model.

5.3. Voting System

Since the individual classifiers give a smaller boost in performance than what was expected, the next step is to see if it is possible to increase the precision further by combining them into a voting system. The voting system only returns a result if all the systems agree on it, and discard the data otherwise, effectively trading recall for precision. The subset $\tilde{Y}_{test}$ being the answers the voting system agrees on, the precision is defined as the number of correct answers in $\tilde{Y}_{test}$ divided by the number of elements in $\tilde{Y}_{test}$, and the recall as the number of elements in $\tilde{Y}_{test}$ divided by the number of elements in Y_{test} the original testing set.

Fig. 5.2. Precision-Recall Curve for the Voting System for the Dominance Trait of the DiSC Personality



The voting system is designed by adding the classifiers shown in Tables 5.3 and 5.4 in decreasing order of the precision reached, and then adding Random Forests with a random subset of 50 features to the mix. If any of the systems added do not increase, but decrease precision, it is removed and another one is added instead. Figure 5.2 shows the precision-recall curves obtained from our voting system for the Dominance trait of the DiSC personality model[1]. Starting from a Recall of 100%, each point corresponds to the adding of a new classifier to the system. The precision-recall curves for the Influence, Steadiness, and Conscientiousness traits, as well as for the MBTI classification, are shown in Appendix D.

The maximum precision reached, and the recalls associated to it, are shown for each of the traits in Table 5.6, as well as the precision and recalls for the same voting system, but run on feature sets that do not contain the other personality model, meaning the results of a voting system for the personality extraction task, instead of the cross-personality extraction task. Except for the Intuition trait, all voting systems achieve a greater precision when the feature sets are augmented with the other personality model, and the Intuition only achieves 0.2% less precision, and has a recall 7.1% higher.

The maximum precision is only useful in the case where one is ready to sacrifice recall indiscriminately. If that is not the case, an analysis of the precision-recall curves is necessary.

[1] Although a classifier is only added if the precision augments, the decrease in recall at some points in the graphic is explained by classifiers that increase precision AND recall, most likely by some randomness in the algorithms making them perform better than the previous iteration.

Tab. 5.6. Maximum Precision and Recall Reached for Each Trait. CP Stands for Cross-Personality Extraction, and PE Stands for Personality Extraction

	Precision CP	Recall CP	Precision PE	Recall PE
D	89.7	36.4	83.9	48.3
i	95.2	29.1	93.8	25.7
S	88.2	70.6	84.1	77.3
C	92.8	34.1	87.3	51.8
IE	95.7	43.0	93.8	53.5
NS	82.6	89.6	82.8	76.5
TF	98.5	20.0	95.7	17.8
JP	81.8	72.7	78.6	89.5

Fig. 5.3. Precision-Recall Curves for the Voting System for the Dominance Trait of the DiSC Personality, With and Without the Inclusion of the MBTI Personality as Features

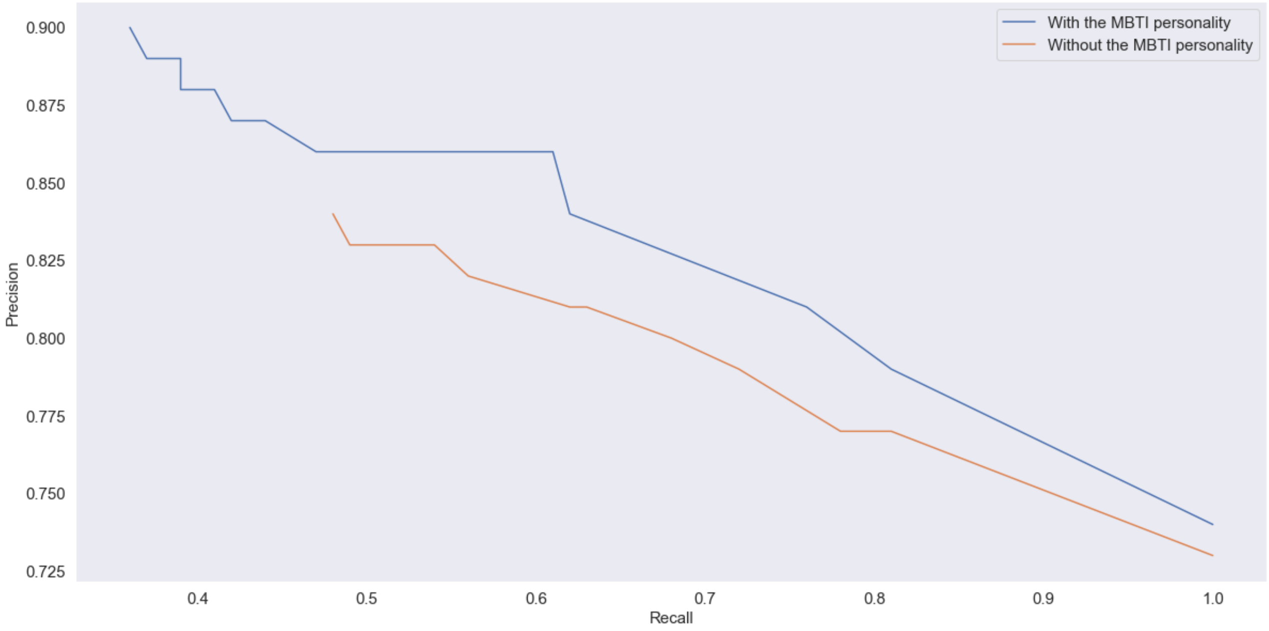


Figure 5.2 shows the precision-recall curve for the voting system for the Dominance trait of the DiSC personality model, when run with and without the inclusion of the MBTI personality model as features. We can see that adding the MBTI personality model as feature helps and that for the same recall, the cross-personality voting system always achieves about

5% more in precision, and that for the same precision, it gains up to 10% on recall, which is a significant boost.

In Appendix D, the precision-recall figures containing both the cross-personality and the personality extraction voting system are shown for all the other traits. With the exception of the Introversion and the Intuition traits, which at some point surpass the cross-personality voting system in both precision and recall, all voting systems prove more efficient when augmented with the other personality model, with various degrees of efficiencies.

Chapter 6

Analysis

6.1. Analysis of the Corpus

Sections 3.1 and 5.1 present some statistics on the distribution of the personality types in the corpus that help understand the task a little better. In this section, a more detailed analysis is presented, since our LinkedIn corpus has the fairly unique characteristic of being labelled with both the MBTI and the DiSC personality models.

The MBTI personality model has been criticized by some researchers [Boyle, 1995; Pittenger], mostly due to the categorical nature of the test, meaning it is less precise than some alternatives, such as the Big-5, and that there is less chance that it will succeed on the test-retest, a criterion in psychology that specifies that a test is valid if the results are the same when taken twice at a reasonable interval of time. The validity of the MBTI as a personality model is outside the scope of this thesis, but understanding the limitations of personality tests, and more specifically of the MBTI and of the DiSC personality models, can help give some insight to sources of errors during classification. In this section, some of these limitations are explained, and more are presented in Section 6.3.

Table 6.1 shows the results of chi-squared tests to see if the 4 dimensions of the MBTI in the corpus are independent. For a personality model such as the MBTI, the independence of dimensions means that every dimension does capture a different facet of the personality. The only dependent dimension is the Judging-Perceiving, which is correlated with both the Intuition-Sensing and the Thinking-Feeling. The Phi coefficients associated are respectively -0.23 and 0.15. For the correlation between NS and JP, that is consistent with the findings of [McCrae and Costa, 1989]. No report of the correlation between TF and JP have

Tab. 6.1. Results of the Chi-Squared Test between the Traits of the MBTI Personality. Y Means there Is a Correlation, N Means there Is no Correlation, p is the Confidence Level, and - Means that the Result Is Found in Another Case, as the Table Is Symmetric

	IE	NS	TF	JP
IE	Y	-	-	-
NS	N	Y	-	-
TF	N	N	Y	-
JP	N	Y ($p<0.0001$)	Y ($p<0.0001$)	Y

been found in the literature, but the correlation is weak enough that it may be a characteristic of the dataset.

The DiSC personality is different in its interpretation and its axes. When running the chi-squared tests and calculating the Phi coefficients, what is expected is not independence of dimensions but rather negative correlations between all dimensions, since the model pushes people to express their personality as only one or two traits of personality. That means that someone who is Dominant probably has less chances to be Influent, Steady, and Conscientious, and someone who presents two main traits has very little chance of having the other two. As a matter of fact, every dimensions are correlated except for the Influence-Dominant pair, as shown in Table 6.2. Just like for the MBTI model, the Phi coefficients are also calculated, as shown in Table 6.3. There is a surprising positive correlation between Steadiness and Conscientiousness, and the strongest negative correlations are between Dominance and Steadiness, and Conscientiousness and Influence. Although not always the case, the DiSC personality model is sometimes presented with these two pairs being opposed personality traits, which could explain the stronger correlations.

As mentioned in Section 1.1.3, there are few studies validating the DiSC personality model. There are reports by companies providing DiSC training showing that it is a valid personality model [noa, 2013], and industry research talking about the use of DiSC in the professional world [Reynierse et al., 2000; Sugerman, 2009], but no independent study working on the DiSC personality model from the psychology point of view was found. As such we had, when starting the project, no real idea of how well personality extraction would work with the DiSC personality. The fact that it can be extracted with a fairly good precision

Tab. 6.2. Results of the Chi-Squared Tests between the Traits of the DiSC Personality. Y Means there Is a Correlation, N Means there Is no Correlation, p is the Confidence Level, and - Means the Result Is Found in Another Case, as the Table is Symmetric

	D	i	S	C
D	Y	-	-	-
I	N	Y	-	-
S	Y ($p<0.0001$)	Y ($p<0.0001$)	Y	-
C	Y ($p<0.0001$)	Y ($p<0.0001$)	Y ($p<0.05$)	Y

Tab. 6.3. Phi Coefficients Between the Traits of the DiSC Personality

	D	i	S	C
D	1.00	X	X	X
I	-0.04	1.00	X	X
S	-0.42	-0.23	1.00	X
C	-0.17	-0.62	0.11	1.00

means that there are differences between someone who has a personality trait and someone who does not have it, but this study should not be taken as a proof of validity of the DiSC personality model.

In Table 6.4, the percentages of each MBTI personality type in the MMBTI corpus are shown, as well as the percentages found in [Macnab, 2008] (Study 1), and on the MBTI foundation website[1]. The two things to notice here are that the percentages found in the MBTI corpus do not agree at all with the other studies, and that the studies don't agree that much between themselves depending on the trait.

Table 6.5 reports the findings of [Schaubhut et al., 2012], which asks the question "Do you actively use LinkedIn" and measures which percentage of each type answered positively. In certain cases, the statistics found explains the over-representation, or under-representation, of certain types in the corpus. For example, there are more INTJ, ENTJ, and ENFJ than found in the general population, but these also correspond to the three types that use the

[1] https://www.myersbriggs.org/my-mbti-personality-type/my-mbti-results/how-frequent-is-my-type.htm

Tab. 6.4. Percentage of Each Type of the MBTI

	Our Corpus	Study 1	MBTI Foundation		Our corpus	Study 1	MBTI Foundation
INTJ	90 - 10.8%	4.4%	2.1%	ENTJ	111 - 13.4%	5.8%	1.8%
INTP	27 - 3.25%	5.7%	3.3%	ENTP	75 - 9.03%	7.5%	3.2%
INFJ	49 - 5.90%	2.8%	1.5%	ENFJ	97 - 11.6%	4.1%	2.5%
INFP	39 - 4.7%	5.7%	4.4%	ENFP	93 - 11.2%	9.6%	8.1%
ISTJ	44 - 5.29%	14.8%	11.6%	ESTJ	90 - 10.8%	11.4%	8.7%
ISTP	9 - 1.08%	5.0%	5.4%	ESTP	9 - 1.08%	2.2%	4.3%
ISFJ	21 - 2.53%	6.2%	13.8%	ESFJ	55 - 6.62%	6.4%	12.3%
ISFP	5 - 0.60%	3.6%	8.8%	ESFP	18 - 2.17%	4.9%	8.5%

Tab. 6.5. Percentage of People who Said Yes to the Question "Do You Actively Use LinkedIn". Taken From [Schaubhut et al., 2012]

INTJ	42.0%	ENTJ	53.9%	ISTJ	27.0%	ESTJ	40.0%
INTP	39.8%	ENTP	46.7%	ISTP	24.2%	ESTP	39.8%
INFJ	16.4%	ENFJ	48.2%	ISFJ	22.0%	ESFJ	28.1%
INFP	29.6%	ENFP	40.6%	ISFP	13.0%	ESFP	17.3%

most LinkedIn. In a similar fashion, there are less ISTP, ISFP, and ESFP than found in the general population, and those correspond to the types who don't use LinkedIn much. However, there still are some deviations that we can't explain, such as the INFJ, which is over-represented in our dataset but is not, according to [Schaubhut et al., 2012], a type that uses LinkedIn a lot. There is possibly, in addition to the use of LinkedIn, the factor of which types would be more likely to post their personality results online that influences the percentages.

Although no statistics on the DiSC personality model have been found, except for those presented in Section 3.1, we still present the distribution of the types of the DiSC in Table 6.6. Contrary to Table 3.1, people who are in the D category will have only the D type, and will have low scores on I, S, and C.

Tab. 6.6. Percentage of Each Type of Personality in the LinekdIn Corpus According to the DiSC Personality Model

D	54 - 6.50%	IS	65 - 7.82%
I	87 - 10.47%	IC	31 - 3.73%
S	21 - 2.53%	SC	66 - 7.94%
C	43 - 5.17%	DIS	11 - 1.32%
DI	261 - 31.42%	DIC	9 - 1.08%
DS	27 - 3.25%	DSC	24 - 2.89%
DC	119 - 14.32%	ISC	12 - 1.44%

As expected, there are few people that say they scored high on 3 different traits. The biggest category is by far the DI one, where 31.42% of the LinkedIn users find themselves. This isn't surprising given the high number of people that said they were high D or high I, as shown in Table 3.1. Although it is not in the table, there is also one person that claims they got high scores on all 4 traits.

6.2. Analysis of the Collection Method

As explained in Section 3.1, the corpora are collected by scraping the pages of users that put voluntarily their personality results in their LinkedIn profiles. This is a fairly common way of collecting corpora when performing personality extraction on social networks, used for example in [Plank and Hovy, 2015; Lima and Castro, 2016; Hernandez and Knight, 2017]. Since we have access to a corpus of 1000 random LinkedIn profiles where people did not a priori put their personalities, we use the two corpora to show that there are differences between the two of them, with people putting their personality results on their LinkedIn profiles also putting more information globally than their counterparts. This is not a source of errors and it doesn't affect the study, since the goal is to show that personality extraction on LinkedIn is feasible, but it does imply that the personality extraction system may not generalize well if deployed in the real world. For example, as shown in Table 6.7, people who put their personality model results on their profiles also input on average 2.62 more work experiences than those who don't, which means that if the number of work experiences is

Tab. 6.7. Statistic over a Corpus with MBTI and one without MBTI

	Corpus with MBTI	Corpus without MBTI
Number of Connections	416	367
Percent of user with a summary	98.1	58.8
Average number of words per filled summary	147	119
Percent of work experience with a description	81.2	68.8
Average number of work experiences per user	6.45	3.82
Average number of educations per user	2.52	1.34
Average number of skills	64.4	11.7

used as a feature, the separation in the vector space would be at the wrong place when using the system in the real world.

6.3. Sources of Possible Errors

Although the results are pretty good and comparable to results on other social networks, there are still multiple sources of possible errors in the study that should be mentioned. Obviously, the personality of an individual is something very difficult to define and capture, and the personality tests and models that are used today are not perfect, and so from the start we expected to have difficulties accomplishing personality extraction correctly. There are, however, concrete sources of errors that are worth mentioning, that relate mostly to the method of the corpus collection or to the personality models.

The first minor source of possible errors is the changing of personality over time. While personality was, at first, believed to be something that was constant through the life of an individual, psychologists found in fact that the personality of an individual does change, and that the longer the interval between 2 personality tests, the more variations there will be [Watson, 2004]. As such, since there is no guarantee that the tests, the texts, and the profiles of the LinkedIn users were taken, written, and filled at the same time, they may have been written by someone whose personality changed and then gave confusing features to the algorithm. This is probably a very minor source of error, but it still is worthy of mention.

A bit more likely, but still statistically small, would be the borderline cases, meaning the people who are barely enough on one side to be qualified as having that dimension of the personality. That holds especially for the MBTI personality model where one is required to be one or the other, and so someone who is situated in the center of a dimension will still be given a letter for that dimension, but being borderline they may, by taking the test on another day and in another mood, receive the other letter. For example, someone who is neither extroverted nor introverted (the so-called centroverts [Zack, 2010]), would, on the MBTI scale, receive a code IXXX one day and EXXX another day, introducing noise in the corpus.

While the DiSC model doesn't have that problem due to the relativity of the traits (meaning that someone will usually only put as their personality type the strongest or the two strongest traits), it has other problems brought on by that same characteristic. These problems are most likely more common, since they hold true for all users and not only in special cases like the borderline cases. The thing with the relativity of the DiSC personality model, or with users only putting their strongest or two strongest traits instead of their continuous scores, is that it makes it a very difficult classification process when taken as four different classifiers as done in this thesis. In Section 4.4, it is mentioned that the lower results on the DiSC personality model could be explained by the fact that the it is less well defined than the MBTI, and that still holds true, but the way users input their DiSC personality is probably a greater source of the difference between the results on the two models. For example, if a user gets 90%, 20%, 50%, and 40% on respectively each of the traits of the DiSC personality, they most likely will put that their DiSC personality type is Dominant, due to the fact that the score on that dimension is so much higher than on the other three. However, if someone instead gets 50%, 20%, 50%, and 30%, then they probably will put that their personality is Dominant and Steadiness, despise the score for Dominance being very far from the score of the first user, and the score for Steadiness being the same. The problem here is that the classifiers are trained on individual traits, while in fact the scoring on one trait greatly depends on the other 3 traits, since a user will decide if they have or not that trait gauging this with the difference with the other traits.

Another possible source of errors is the fact that we do not know how reliable the tests are. Personality tests range from being administered by trained psychologists to tests created on the internet by people without the qualifications to do so. This is especially true for the MBTI personality model, who has been heavily popularized and so there are more ways to find out your personality online with the MBTI model than with the DiSC personality model.

The fourth and final source of errors concerning the method of collection of the corpus comes from the self-perception aspect of LinkedIn. As mentioned in Section 1.2, people tend to be more controlled and professional in what they write on LinkedIn, due to the fact that their profile is going to be seen by potential employers. That being so, and with the DiSC personality model instinctively putting D and I types as the "Go-getter" proactive types that are more often perceived as the "good employee" [Minuk, 2018], some of the reasons we see more Dominant and Influent people on the corpus could be that:

(1) People who are D and I tend to put more their personality model results on LinkedIn,
(2) People lie and change slightly their personality to appear like a more valuable potential employee,
(3) People answer differently than their true nature when taking the test, thinking that D and I may be more valuable than S and C, and therefore, get personality results that are not necessarily representative of their true personality.

For all the reasons mentioned in this section, there is a possibility that the personality label given in the profiles do not correspond to the real personality of the user. The only way to reduce these errors would have been to administer the personality tests in person, but time and money constraints prevented that. It is to note however that the fourth source of error is purely speculative and that it is possible that Dominant and Influent people are simply more likely to use LinkedIn than Steady and Conscientious people, in a similar way to the fact that not all MBTI types are equally likely to use LinkedIn, as shown in Section 6.1.

6.4. Ethical Considerations

The goal of this thesis is to make communication between people on LinkedIn more efficient and more pleasant by tailoring messages to appeal to the personality of the person. Nevertheless, personality extraction and personality assessment, especially on LinkedIn, have ethical ramifications, some of them we would like to address here.

There is the possibility of misuse of this research for trimming incorrectly candidates while hiring. As pointed out in [Coe, 1992], the best way to use personality assessment in the trimming of candidates is to look for potential clashes between the manager's personality and the applicant's personality, and be sure to create a work environment where these clashes will be resolved quickly and efficiently. In fact, in studies trying to find correlations between job performance and personality with the Big-5 personality model, only Neuroticism have been found to be negatively correlated to job performance [Tett et al., 1991]. As such, this study should not be used to encourage discrimination based on personal bias when hiring candidates, but rather in order to facilitate communication and workflow.

The sharing of the dataset collected was initially a goal when starting the project, but due to the method of collection, which takes in non-volunteer users, it became too complex to share it ethically. As pointed out in [Zimmer, 2010], working with social networks can be ethically dangerous, and we do not wish to trample the privacy of the users. Although all of them are public users who accepted implicitly that their profiles could be seen by anyone, they still did not agree to be part of the study and as such, it is a delicate to share the dataset. To do so, a full anonymization of the corpora would have to be performed, and due to the uniqueness of the life experiences and the uniqueness of the text written in the profiles, that would in itself be a project too long and laborious to be reasonably considered in the time it takes to do a master thesis. As for the simple accessing and scraping of the profiles, we consider it both ethical and legal as a 2017 ruling by a California federal court has declared it legal to scrape public parts of LinkedIn [Lee, 2017], since all public users agreed to have their profiles seen by anybody.

Overall, personality extraction, and, more generally, author profiling, can be used by companies to target vulnerable users, influencing them for whatever reason, or as a deanonymization technique, revealing authors that may not wish to reveal themselves. It can also, however, be used to help people, facilitating communication or guiding users to a job that would make them happy. The point here is the technology presented in this master thesis has many uses, some would be considered bad, and others good. As such, nothing more can be done than warn against unethical use of personality extraction, and finish this chapter with the famous quote of David Wong:

> «*New technology is not good or evil in and of itself. It's all about how people choose to use it.* »

Chapter 7

Conclusion

Personality extraction is a relatively young field, and personality extraction through social networks even more so. The task can be defined as a classification problem, where the input is a profile coming from one of the many social networks existing today, and the output is the prediction over one of the personality models existing in psychology.

In this master thesis, the social network is LinkedIn, a professional social network, and the personality models are both the MBTI and DiSC, which are used quite often in the professional world. Three innovations are presented which, to our knowledge, have never been done before:

(1) The extraction of a LinkedIn corpus, as presented in Chapter 3,

(2) The successful personality extraction on LinkedIn, as presented in Chapter 4,

(3) The personality extraction on the DiSC personality model, also presented in Chapter 4.

In Chapter 5, the unique LinkedIn corpus labelled with two personality models is used to:

(1) study the correlation between the DiSC personality model and the MBTI personality model,

(2) Test whether cross-personality extraction is feasible and more efficient than regular personality extraction,

where cross-personality extraction denotes the action of predicting the results on a personality model based on the results on another personality model, as well as some additional information. Finally, a bias in the way that corpora are usually collected in personality extractions through social networks studies is showed, at least on LinkedIn, which raises the question of whether or not this bias also exists when the method is used on other social networks.

Following this master thesis, there are many ways the field could evolve, some of them being listed here:

(1) cross-personality studies could be done between other personality models,

(2) studies of the bias of the collection method could be performed on other social networks,

(3) in order to have a common benchmark for researchers, the anonymization of a LinkedIn dataset could be done, allowing to share it ethically,

(4) a generalization performance study with a corpus collected following the method of [Plank and Hovy, 2015], and a manually collected corpus could be performed in order to see if the bias introduced hurts the performance, and techniques to reduce error when using the algorithms in a real-world setting could be developed,

(5) the cross-personality extraction technique could be used to bootstrap a larger dataset and use more powerful machine learning algorithm on it, such as neural networks.

Bibliography

About Everything DiSC: Theory and Research. Technical report.

Research Report for Adaptive Testing Assessment. Technical report, 2013. URL `https://www.discprofile.com/DiscProfile/media/PDFs-Other/Research%20Reports%20and%20White%20Papers/EverythingDiSCResearchReport.pdf`.

Nadeem Ahmad and Jawaid Siddique. Personality Assessment using Twitter Tweets. *Procedia Computer Science*, 112:1964–1973, 2017. ISSN 18770509. doi: 10.1016/j.procs.2017.08.067. URL `http://linkinghub.elsevier.com/retrieve/pii/S1877050917314114`.

Joan-Isaac Biel, Lucía Teijeiro-Mosquera, and Daniel Gatica-Perez. FaceTube: predicting personality from facial expressions of emotion in online conversational video. In *Proceedings of the 14th ACM international conference on Multimodal interaction - ICMI '12*, page 53, Santa Monica, California, USA, 2012. ACM Press. ISBN 978-1-4503-1467-1. doi: 10.1145/2388676.2388689. URL `http://dl.acm.org/citation.cfm?doid=2388676.2388689`.

Gregory J. Boyle. Myers-Briggs Type Indicator (MBTI): Some Psychometric Limitations. *Australian Psychologist*, 30(1):71–74, 1995. ISSN 1742-9544. doi: 10.1111/j.1742-9544.1995.tb01750.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1742-9544.1995.tb01750.x`.

Fabio Celli and Cristina Zaga. Be Conscientious, Express your Sentiment! page 9, 2013.

Fabio Celli, Fabio Pianesi, David Stillwell, and Michal Kosinski. Workshop on Computational Personality Recognition: Shared task. page 4, 2013.

Charles K. Coe. The MBTI: Potential Uses and Misuses in Personnel Administration. *Public Personnel Management*, 21(4):511–522, December 1992. ISSN 0091-0260. doi: 10.1177/009102609202100407. URL `https://doi.org/10.1177/009102609202100407`.

Lillian Cunningham. Myers-Briggs: Does it pay to know your type? *Washington Post*, 2012. URL `https://www.washingtonpost.com/national/on-leadership/myers-briggs-does-it-pay-to-know-your-type/2012/12/14/eaed51ae-3fcc-11e2-bca3-aadc9b7e29c5_story.html`.

Bruce Ferwerda and Marko Tkalcic. You Are What You Post: What the Content of Instagram Pictures Tells About Users' Personality. page 5, 2018.

Adrian Furnham. The big five versus the big four: the relationship between the Myers-Briggs Type Indicator (MBTI) and NEO-PI five factor model of personality. *Personality and Individual Differences*, 21(2):303–307, August 1996. ISSN 01918869. doi: 10.1016/0191-8869(96)00033-5. URL `http://linkinghub.elsevier.com/retrieve/pii/0191886996000335`.

Adrian Furnham, Joanna Moutafi, and John Crump. THE RELATIONSHIP BETWEEN THE REVISED NEO-PERSONALITY INVENTORY AND THE MYERS-BRIGGS TYPE INDICATOR. 2003. URL https://jfdeschamps.files.wordpress.com/2012/09/correl-ocean-mbti-furnham-2003-6p.pdf.

Robert E. Gibby and Michael J. Zickar. A history of the early days of personality testing in American industry: An obsession with adjustment. *History of Psychology*, 11(3):164, 2008. ISSN 1939-0610. doi: 10.1037/a0013041. URL http://psycnet.apa.org/fulltext/2008-10736-002.pdf.

Timo Gnambs. A meta-analysis of dependability coefficients (test–retest reliabilities) for measures of the Big Five. *Journal of Research in Personality*, 52:20–28, October 2014. ISSN 0092-6566. doi: 10.1016/j.jrp.2014.06.003. URL http://www.sciencedirect.com/science/article/pii/S0092656614000543.

Lewis R Goldberg. The Structure of Phenotypic Personality Traits. *American Psychologist*, page 9, 1993.

Isabelle Guyon and Andre Elisseeff. An Introduction to Variable and Feature Selection. page 26, 2003.

Robert J. Harvey, William D. Murry, and Steven E. Markham. *Myers-Briggs Type Indicator*. 1995.

Rayne Hernandez and Ian Scott Knight. Predicting Myers-Briggs Type Indicator with Text Classification, 2017. URL https://web.stanford.edu/class/cs224n/reports/6839354.pdf.

Robert C. Holte. Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. *Machine Learning*, 11(1):63–90, April 1993. ISSN 1573-0565. doi: 10.1023/A:1022631118932. URL https://doi.org/10.1023/A:1022631118932.

Mitchell J. (MBTI) Myers-Briggs Personality Type Dataset, 2018. URL https://kaggle.com/datasnaek/mbti-type.

Adam Joinson and Ulf-Dietrich Reips. Personalized salutation, power of senderand response rates to Web-based surveys. *Computers in Human Behavior*, 23:1372–1383, 2007. doi: 10.1016/j.chb.2004.12.011. URL https://reader.elsevier.com/reader/sd/pii/S0747563204002304?token=2EA80035050CD424E8C75519FFAEB9FBD8A331BF861E980CC2C1407182FFB8E0E9D7164E2C4C40C4D7056FB602842319.

Cathleen S. Jones and Nell T. Hartley. Comparing Correlations Between Four-Quadrant And Five-Factor Personality Assessments. *American Journal of Business Education (AJBE)*, 6(4):459, July 2013. ISSN 1942-2512, 1942-2504. doi: 10.19030/ajbe.v6i4.7945. URL https://clutejournals.com/index.php/AJBE/article/view/7945.

E. Kafeza, A. Kanavos, C. Makris, and P. Vikatos. T-PICE: Twitter Personality Based Influential Communities Extraction System. In *2014 IEEE International Congress on Big Data*, pages 212–219, June 2014. doi: 10.1109/BigData.Congress.2014.38.

Mayuri Pundlik Kalghatgi, Manjula Ramannavar, and Dr Nandini S Sidnal. A Neural Network Approach to Personality Prediction based on the Big-Five Model. *International Journal of Innovative Research in Advanced Engineering*, 2(8):8, 2015.

Onno Kampman, Elham J. Barezi, Dario Bertero, and Pascale Fung. Investigating Audio, Video, and Text Fusion Methods for End-to-End Automatic Personality Prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 606–611, Melbourne, Australia, July 2018. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P18-2096.

Vishal Kaushal and Manasi Patwardhan. Emerging Trends in Personality Identification Using Online Social Networks—A Literature Survey. *ACM Transactions on Knowledge Discovery from Data*, 12(2):1–30, January 2018. ISSN 15564681. doi: 10.1145/3070645. URL http://dl.acm.org/citation.cfm?doid=3178544.3070645.

Timothy B. Lee. Court rejects LinkedIn claim that unauthorized scraping is hacking, August 2017. URL https://arstechnica.com/tech-policy/2017/08/court-rejects-linkedin-claim-that-unauthorized-scraping-is-hacking/.

A. C. E. S. Lima and L. N. de Castro. Predicting Temperament from Twitter Data. In *2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 599–604, July 2016. doi: 10.1109/IIAI-AAI.2016.239.

Junjie Lin, Wenji Mao, and Daniel Zeng, D. Personality-based refinement for sentiment classification in microblog, 2017.

Fei Liu, Julien Perez, and Scott Nowson. A Language-independent and Compositional Model for Personality Trait Recognition from Short Texts. *arXiv:1610.04345 [cs, stat]*, October 2016. URL http://arxiv.org/abs/1610.04345. arXiv: 1610.04345.

Douglas A. MacDonald, Peter E. Anderson, Catherine I. Tsagarakis, and Cornelius J. Holland. Examination of the Relationship between the Myers-Briggs Type Indicator and the Neo Personality Inventory. *Psychological Reports*, 74(1):339–344, February 1994. ISSN 0033-2941. doi: 10.2466/pr0.1994.74.1.339. URL https://doi.org/10.2466/pr0.1994.74.1.339.

Dr Donald Macnab. Myers-Briggs Type Indicator (MBTI) in Canada. page 19, 2008.

Antony Maina. 20 Popular Social Media Sites Right Now, May 2016. URL https://smallbiztrends.com/2016/05/popular-social-media-sites.html.

Francois Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. page 44, 2007.

N. Majumder, S. Poria, A. Gelbukh, and E. Cambria. Deep Learning-Based Document Modeling for Personality Detection from Text. *IEEE Intelligent Systems*, 32(2):74–79, March 2017. ISSN 1541-1672. doi: 10.1109/MIS.2017.23.

Dejan Markovikj, Sonja Gievska, Michal Kosinski, and David Stillwell. Mining Facebook Data for Predictive Personality Modeling. page 4, 2013.

Robert R. McCrae and Paul T. Costa. Reinterpreting the Myers-Briggs Type Indicator From the Perspective of the Five-Factor Model of Personality. *Journal of Personality*, 57(1):17–40, March 1989. ISSN 0022-3506, 1467-6494. doi: 10.1111/j.1467-6494.1989.tb00759.x. URL http://doi.wiley.com/10.1111/j.1467-6494.1989.tb00759.x.

Amanda Minuk. How to be the best damn employee, 2018. URL https://www.bmeaningful.com/blog/2018/05/how-to-be-the-best-employee/.

N. R. Ngatirin, Z. Zainol, and T. L. C. Yoong. A comparative study of different classifiers for automatic personality prediction. In *2016 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, pages 435–440, November 2016. doi: 10.1109/ICCSCE.2016.7893613.

J. Noecker, M. Ryan, and P. Juola. Psychological profiling through textual analysis. *Literary and Linguistic Computing*, 28(3):382–387, September 2013. ISSN 0268-1145, 1477-4615. doi: 10.1093/llc/fqs070. URL https://academic.oup.com/dsh/article-lookup/doi/10.1093/llc/fqs070.

Gregory Park, H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Michal Kosinski, David J. Stillwell, Lyle H. Ungar, and Martin E. P. Seligman. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, 108(6):934–952, 2015. ISSN 1939-1315, 0022-3514. doi: 10.1037/pspp0000020. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/pspp0000020.

JW Pennebaker and LA King. Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, 1999. URL https://www.ffri.hr/~ibrdar/komunikacija/seminari/Pennebaker,%20King,%201999%20-%20Linguistic%20stiles.pdf.

Frédéric Piedboeuf, Philippe Langlais, and Ludovic Bourg. Personality Extraction Through LinkedIn. In Marie-Jean Meurs and Frank Rudzicz, editors, *Advances in Artificial Intelligence*, Lecture Notes in Computer Science, pages 55–67. Springer International Publishing, 2019. ISBN 978-3-030-18305-9.

David J Pittenger. Measuring the MBTI... And Coming Up Short. page 7.

Barbara Plank and Dirk Hovy. Personality Traits on Twitter—or—How to Get 1,500 Personality Tests in a Week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–98, Lisboa, Portugal, September 2015. Association for Computational Linguistics. URL http://aclweb.org/anthology/W15-2913.

D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft. Our Twitter Profiles, Our Selves: Predicting Personality with Twitter. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 180–185, October 2011. doi: 10.1109/PASSAT/SocialCom.2011.26.

James H. Reynierse, Dennis Ackerman, Alexis A. Fink, and John B. Harker. The Effects of Personality and Management Role on Perceived Values in Business Settings. *International Journal of Value-Based Management*, 13(1):1–13, January 2000. ISSN 1572-8528. doi: 10.1023/A:1007707800997. URL https://doi.org/10.1023/A:1007707800997.

Nancy Schaubhut, Amanda Weber, and Rich Thompson. Myers-Briggs® Type and Social Media Report. Technical report, 2012. URL https://ap.themyersbriggs.com/content/Research%20and%20White%20Papers/MBTI/MBTI_Social_Media_Report.pdf.

Xuemeng Song, Liqiang Nie, Luming Zhang, Mohammad Akbari, and Tat-Seng Chua. Multiple Social Network Learning and Its Application in Volunteerism Tendency Prediction. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 213–222, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3621-5. doi: 10.1145/2766462.2767726. URL http://doi.acm.org/10.1145/2766462.2767726.

Philip J. Stone, Robert F. Bales, J. Zvi Namenwirth, and Daniel M. Ogilvie. The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, 7(4):484–498, 1962. ISSN 1099-1743. doi: 10.1002/bs.3830070412. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bs.3830070412.

M. H. Su, C. H. Wu, and Y. T. Zheng. Exploiting Turn-Taking Temporal Evolution for Personality Trait Perception in Dyadic Conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):733–744, April 2016. ISSN 2329-9290. doi: 10.1109/TASLP.2016.2531286.

Jeffrey Sugerman. Using the DiSC® model to improve communication effectiveness. *Industrial and Commercial Training*, 41(3):151–154, April 2009. ISSN 0019-7858. doi: 10.1108/00197850910950952.

C. Sumner, A. Byers, R. Boochever, and G. J. Park. Predicting Dark Triad Personality Traits from Twitter Usage and a Linguistic Analysis of Tweets. In *2012 11th International Conference on Machine Learning and Applications*, volume 2, pages 386–393, December 2012. doi: 10.1109/ICMLA.2012.218.

M. M. Tadesse, H. Lin, B. Xu, and L. Yang. Personality Predictions Based on User Behaviour on the Facebook Social Media Platform. *IEEE Access*, pages 1–1, 2018. ISSN 2169-3536. doi: 10.1109/ACCESS.2018.2876502.

Robert P. Tett, Douglas N. Jackson, and Mitchell Rothstein. Personality Measures as Predictors of Job Performance: A Meta-Analytic Review. *Personnel Psychology*, 44(4):703–742, 1991. ISSN 1744-6570. doi: 10.1111/j.1744-6570.1991.tb00696.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1744-6570.1991.tb00696.x.

Kai Ming Ting and Zijian Zheng. A Study of AdaBoost with Naive Bayesian Classifiers: Weakness and Improvement. *Computational Intelligence*, 19(2):186–200, May 2003. ISSN 0824-7935, 1467-8640. doi: 10.1111/1467-8640.00219. URL http://doi.wiley.com/10.1111/1467-8640.00219.

Niels van de Ven, Aniek Bogaert, Alec Serlie, Mark J. Brandt, and Jaap J.A. Denissen. Personality perception based on LinkedIn profiles. *Journal of Managerial Psychology*, 32(6):418–429, August 2017. ISSN 0268-3946. doi: 10.1108/JMP-07-2016-0220. URL https://www.emeraldinsight.com/doi/full/10.1108/JMP-07-2016-0220.

José van Dijck. 'You have one identity': performing the self on Facebook and LinkedIn. *Media, Culture & Society*, 35(2):199–215, March 2013. ISSN 0163-4437, 1460-3675. doi: 10.1177/0163443712468605. URL http://journals.sagepub.com/doi/10.1177/0163443712468605.

David Watson. Stability versus change, dependability versus error: Issues in the assessment of personality over time. *Journal of Research in Personality*, 38(4):319–350, August 2004. ISSN 00926566. doi: 10.1016/j.jrp.2004.03.001. URL https://linkinghub.elsevier.com/retrieve/pii/S0092656604000261.

Jason Weston, Sayan Mukherjee, Olivier Chapelle, Massimiliano Pontil, Tomaso Poggio, and Vladimir Vapnik. Feature Selection for SVMs. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 668–674. MIT Press, 2001. URL http://papers.nips.cc/paper/1850-feature-selection-for-svms.pdf.

Janet Wiener and Nathan Bronson. Facebook's Top Open Data Problems, 2014. URL https://research.fb.com/facebook-s-top-open-data-problems.

D. Xue, Z. Hong, S. Guo, L. Gao, L. Wu, J. Zheng, and N. Zhao. Personality Recognition on Social Media With Label Distribution Learning. *IEEE Access*, 5:13478–13488, 2017. doi: 10.1109/ACCESS.2017.2719018.

Hsin-Chang Yang, Cathy S. Lin, Zi-Rui Huang, and Tsung-Hsing Tsai. Text Mining on Player Personality for Game Recommendation. pages 1–6. ACM Press, 2017. ISBN 978-1-4503-4881-2. doi: 10.1145/3092090.3092132. URL http://dl.acm.org/citation.cfm?doid=3092090.3092132.

Wu Youyou, Michal Kosinski, and David Stillwell. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences of the United States of America*, 112(4):1036–1040, January 2015. ISSN 0027-8424. doi: 10.1073/pnas.1418680112. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4313801/.

J. Yu and K. Markov. Deep learning based personality recognition from Facebook status updates. In *2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST)*, pages 383–387, November 2017. doi: 10.1109/ICAwST.2017.8256484.

Devora Zack. *Networking for People Who Hate Networking: A Field Guide for Introverts, the Overwhelmed, and the Underconnected*. ReadHowYouWant.com, December 2010. ISBN 978-1-4587-2547-9. Google-Books-ID: Mx0mDtfdf3wC.

Michael Zimmer. "But the data is already public": on the ethics of research in Facebook. *Ethics and Information Technology*, 12(4):313–325, December 2010. ISSN 1572-8439. doi: 10.1007/s10676-010-9227-5. URL https://doi.org/10.1007/s10676-010-9227-5.

Appendix A

List of stop-words

’you’, "didn’t", ’her’, ’again’, ’and’, ’no’, "she’s", ’hadn’, "you’ve", "weren’t", "you’re", ’hasn’, ’more’, ’down’, ’just’, ’shouldn’, ’was’, ’these’, ’there’, ’m’, ’its’, "you’d", ’the’, ’against’, ’or’, ’above’, ’been’, ’have’, ’aren’, ’did’, ’how’, ’weren’, ’ours’, ’yourselves’, ’mightn’, ’be’, ’don’, ’nor’, "that’ll", ’couldn’, ’few’, ’very’, ’into’, ’do’, ’is’, ’re’, ’then’, ’being’, ’through’, ’why’, ’can’, ’at’, ’our’, ’own’, ’will’, ’as’, ’herself’, ’ll’, ’should’, "wouldn’t", ’below’, "needn’t", ’what’, ’wasn’, ’didn’, ’needn’, ’ourselves’, ’they’, ’has’, ’himself’, ’not’, ’yourself’, ’theirs’, ’during’, ’same’, ’by’, "you’ll", ’from’, ’had’, ’an’, ’haven’, ’hers’, ’we’, ’o’, ’over’, ’if’, "shan’t", ’in’, ’s’, "hasn’t", ’this’, ’does’, "aren’t", ’that’, ’all’, ’ma’, "won’t", ’t’, ’to’, ’their’, ’until’, ’won’, ’while’, ’whom’, ’for’, ’are’, ’itself’, "don’t", ’isn’, ’than’, ’it’, ’only’, "wasn’t", "haven’t", ’myself’, ’between’, ’yours’, ’where’, "mightn’t", ’because’, ’of’, ’about’, "mustn’t", ’doing’, ’me’, "doesn’t", ’but’, ’wouldn’, ’she’, "it’s", ’am’, ’a’, ’once’, "hadn’t", ’ve’, ’them’, ’y’, ’with’, ’d’, ’doesn’, ’after’, ’my’, ’i’, ’before’, ’each’, ’those’, ’further’, "couldn’t", ’mustn’, "shouldn’t", ’him’, ’both’, ’on’, ’under’, ’his’, "should’ve", ’now’, ’who’, ’having’, ’too’, ’themselves’, ’shan’, ’off’, ’most’, ’up’, ’out’, "isn’t", ’your’, ’such’, ’he’, ’so’, ’ain’, ’here’, ’when’, ’any’, ’were’, ’some’, ’other’, ’which’

Appendix B

Features returned by the feature selection algorithm

Note: In the following sections, the suffix _x represent the absolute counts returned by the General Inquirer, while the suffix _y represent the relative counts.

B.1. Personality Extraction

- Dominance
  - Non-Textual: Number of titles, Number of test, number of sections in the achievement section, Number of description normalized for voluntary experience, number of description normalized, number of accomplishments, presence of personal branding, number of honor title, Number of colleagues endorsements, number of connections, number of skills, average duration of education, total number of endorsements.
  - Textual: PRP$, Vary_x, Ngtv_x, Self_x, Means_y, NegAff_y, Feel_y, PowGain_x, If_x, Negativ_y, Try_x, Stay_y, Our_y, COLL_y, Negate_x, TIME_x, Intrj_y, PowPt_x, FREQ_y, PowCon_y, No_x, MD, WltTot_y, IAV_y, MALE_x, Relig_y, Vehicule_x, Quan_y, Fall_x, NatrPro_y
  - Hand-Crafted: :, IndAjd_x, Quan_x, nb of endorsements by skilled people, Travel_x, HU_y, Need_y, nb of description normalized for voluntary experienced, COLOR_y, Polit@_y, Causal_x, Travel_y, WltPt_y, Fail_y, number of pub, RcRelig_y, TrnLoss_x, DIST_y, FREQ_y, WltTran_y, Vice_x, like influencer, You_y, AffLoss_x, Fetch_y, Aquatic_y, DIST_x, RcTot_x, PowDoct_x, ArenaLW_x
- Influence

  - Non-Textual: Number of description normalized for schools, average duration of education, number of language, number of Test, number of description normalized for voluntary experience, total number of endorsements, whether the user like the company, number of sections, number of endorsements by skilled people, number of jobs, number of description normalized, whether the user like an influencer, average number of recommendation received, average number of followers, number of honor title.
  - Textual: Bldfpt_x, Exch_x, Negate_x, PowAren_y, WDT, PowAren_x, NatrPro_y, NNPS, JJS, Route_y, EnlEnds_x, TrnGain_y, Tool_y, Anomie_x, leftovers_y, Persist_x, Name_y, WlbPsyc_y, RspLoss_x, RCLoss_x, WlbLoss_y, EnlLoss_x, Race_x, Time@_x, RspLoss_y, PowPt_x, TIME_x, EVAL_y, Sky_y
  - Hand-Crafted: Rel_y, FormLw_y, Weak_y, total number of endorsements, Fail_x, Submit_y, AffLoss_y, FREQ_y, average duration of education, number of titles, RspTot_x, You_y, Anomie_x, Enl0th_x, PowEnds_y, Rsp0th_y, ABS_y, Travel_y, Arousal_x, Ovrst_x, PowAuth_y, Econ@_x, JJ, Exch_x, Goal_y, Devreas_x, PowPt_y, Positiv_y, shortest job, AffTot_y

- Steadiness
  - Non-Textual: Number of Languages, average number of endorsment, number of titles, number of recommendation given, number of recommendation received, number of connections, number of skills, number of endorsements by skilled people, number of endorsements by colleagues, number of publications, number of description normalized, average duration of education, wheter the user like the company, wheter the user like an influencer, average number of followers
  - Textual: Social_x, Strong_y, Space_x, WltPT_x, Intrj_x, SklTot_y, AffLoss_y, Nation_x, PLACE_x, PowCon_y, Time@_x, SklAsth_y, Persist_x, RcLoss_y, COLOR_y, Doctrin_y, EnlLoss_x, Think_y, RCGain_y, BodyPt_x, WlbPsyc_y, Know_x, Ought_x, Yes_y, RspTot_y, Quality_y, Female_y, RP, Vice_y, WlbLoss_x
  - Hand-Crafted: Exert_x, Kin@_x, PowAren_x, total number of endorsements, ORD_y, Pleasur_y, Land_y, EnlEnds_y, WlbGain_y, Persist_y, number

of colleagues endo, Rise_x, CARD_y, Complet_y, Region_y, AffTot_y, Negativ_y, Polit@_y, Quality_y, ComForm_y, MD, Virtue_x, TrnGain_x, Skl0th_y, Exch_x, Fetch_x, Econ@_y, Pstv_x, DIST_x, Rel_x

- Conscientious
  - Non-Textual: Number of recommendation received, number of skills, number of publicity, number of titles, number of description normalized, total number of endorsements, number of project, number of description normalized for the voluntary experience section, presence of personal branding, length of the shortest job, highest level of education, length of the longuest job, number of endorsements by skilled people, average number of endorsements, max followers.
  - Textual: Finish_x, Fetch_x, VBZ, $, Exch_x, ”, Ritual_x, RB, Affpt_x, PowAuPt_x, WltTran_y, DIM_x, Quan_y,), Negativ_y, Aquatic_x, Means_Lw_x, PowAren_y, Negate_x, Yes_y, WltTran_x, TrnLoss_y, Abs@_y, WMOT_y, UH, PowPt_y, Know_y, Relig_y, NatObj_x, ComForm_x
  - Hand-Crafted: Average number of endorsements, Enloth_y, Work_y, RspLoss_x, WltPt_y, total number of endorsements, $, Rel_y, Connections, Ovrst_y, AffGain_y, Ngtv_y, number of interests, You_x, RspGain_x, EMOT_x, EnlLoss_x, Eval@_x, li,e company, RcLoss_x, Fetch_y, Our_x, RcEnds_y, PowPt_y, WP, Milit_y, NUMB_x, If_y, Ngtv_x, Negativ_x
- Introversion
  - Non-Textual: Number of recommendation received, Number of endorsements by colleagues, Average number of endorsements, Number of Project, Number of recommendation given, Total number of endorsements, Number of educations, Number of description normalized for school, Number of titles, Number of Interests, Number of languages, Number of endorsements by skilled people, Number of courses, Number of tests, Minimum number of followers for an interest
  - Textual: Name_x, ArenaLw_x, Food_x, Pain_x, Doctrin_x, VBZ, PowEnds_x, Vice_y, Power_y, Econ@_y, VBP, (, Yes_y, COLL_y, Aquatic_x, SV_y, Know_y, AffPt_x, RcEthic_y, Perceiv_y, NatrPro_x,

Try_x, BldgPt_y, Know_x, Negate_y, Fall_y, Social_x, ABS_x, Pleasur_x, Solve_x

  - Hand-Crafted: WDT, COM_y, JJS, total number of endorsements, Means_y, RcGain_y, Means_y, RspGain_x, Nation_x, WltTot_y, , COLL_y, Anomie_x, Decreas_x, UH, WlbGain_x, Kin@_y, TO, nb of description normalized (volExp), Decreas_y, Quality_x, PowPt_y, Exch_x, PLACE_x, Kin@_x, Need_x, Polit@_x, Social_x, Active_y, SureLw_x, Ovrst_y

- Intuition
  - Non-Textual: Average duration of voluntary experiences, Average duration of educations, Number of description normalized for voluntary experience, Total number of endorsements, number of endorsements by colleagues, number of connections, Number of publicity, Number of recommendation given, Number of recommendation received, Number of project, Duration of shortest job, Wheter the user like the company they're currently working on, Number of titles, Number of description normalized for educations, Average number of endorsements.
  - Textual; ", (, JJS, WltTot_y, Try_y, VBP, Goal_x, EndsLW, ComnObj_x, ComnObj_y, EnlEnds_x, PtLw_y, Think_x, Intrj_y, PowEnds_y, PowCoop_x, Land_y, PowDoct_y, TranLw_y, WlbPhys_x, MALE_y, WlbPsyc_y, Abs@_x, Rel_x, Exch_y, VBZ, NegAff_x, RspOth_x, Legal_x, POS_x
  - Hand-Crafted: Perceiv_x, Kin@_x, RcRelig_x, number of colleagues endo, number of recomm given, number of skilled endorsement, PowLoss_y, Nonadlt_y, IPadj_y, AffGain_y, NUMB_y, EnlLoss_y, Exch_y, UH, CD, Goal_y, Weak_y, number of accomplishements, NatrPro_y, NegAff_x, AffLoss_x, Self_x, WlbLoss_x, MeansLw_y, RcTot_y, RcEnds_x, EnlTot_y, SocRel_x, Pstv_y, COM_x
- Thinking
  - Non-Textual: Number of educations, number of description normalized for schools, Number of different establishments in the education section, Total number of endorsements, Average number of endorsement, Average duration of jobs, Number of jobs, Number of voluntary experiences, Average duration of education, Whether the user likes the company he's currently working at, Highest level

    of education reached, Number of sections in the achievement section, Number of publicity, Number of description normalized for the work section, Number of description normalized for the voluntary experience section.

  - Textual: ComnObj_x, Power_y, Affil_y, Vice_x, PowEnds_x, Pleasur_x, Complet_x, Yes_x, Yes_x, Solve_y, AffLoss_y, VBP, Undrst_y, PowPt_x, RspOth_y, SklAsth_y, PowEnds_y, Academ_y, Tim@_y, C, Abs@_y, DAV_y, HU_y, RcEnds_y, Travel_y, Increas_x, MD, PRP$, Strong_y, , TIME_x
  - Hand-Crafted: ”, Complet_y, BodyPt_y, number of recomm received, Abs@_y, Power_y, total number of endorsments, EnlEnds_x, Milit_y, Say_x, Object_y, DIM_x, PowAuth_x, Quality_y, like company, Fall_x, DIM_y, Work_y, Pleasur_y, number of skills, PRP$, RspLoss_x, COM_x, CC, EndsLw_x, Sky_y, BldgPt_y, COLL_y, PowAren_y, ComForm_y

- Judging
  - Non-Textual: Number of publicity, Whether the user like the company he’s currently working at, If there is a summary written, Number of skills, Total number of endorsements, Number of endorsements by colleagues, Highest level of education reached, Average duration of education, Number of description normalized for the work section, Number of recommandation received, Number of educations, Number of project, duration of the longuest job, Minimum number of followers in the interests section, Total number of accomplishements.
  - Textual: Tool_x, NUMB_y, Sky_x, Quan_x, RB, Aquatic_x, PowGain_y, Need_y, Doctrin_y, Hostile_y, Self_y, Social_y, Anomie_x, PowEnds_x, DIM_y, SocRel_y, AffTot_x, ArenaLw_y, WlbPsyc_y, SklPt_y, PowAren_x, PowCon_x, SYM, RspLoss_y, AffLoss_x, Aquatic_y, EnlLoss_y, TrnLoss_y, No_y, Finish_y
  - Hand-Crafted: WltTot_y, Ought_x, Work_y, Object_y, Role_y, Sky_y, PowDoct_x, total number of endorsements, Hostile_y, You_y, Say_x, Fall_y, IndAdj_y, Hostile_x, NotLw_x, DIST_x, nb of description normalized (volExp), RspLoss_x, Exprsv_x, Aquatic_y, average duration of vol exps, POS_y, Object_x, Region_x, Think_y, Nation_x, Econ@_x, number of honor title, Finish_x, number of skills

B.2. Cross-Personality Extraction

- Dominance
  - Non-Textual: Number of educations, Number of recommendations received, Number of recommendation given, Average number of educations, average duration of volutary experiences, Number of Skills, Number of Patents, Whether the user like an influencer, Number of publicity, Judging dimension, Number of skilled endorsements, number of sections in the achievement section, number of Connections, Duration of Longuest job, Number of colleagues endorsements.
  - Textual: Means_y, If_x, WlbPsyc_x, Econ@_y, Feel_x, Pleasur_x, COLOR_x, WRB, Female_x, SklOth_y, PowGain_x, PowOth_y, NatrPro_y, Academ_y, EnlEnds_y, TrnLoss_x, RspLoss_x, Region_x, Finish_y, Arousal_x, Time@_x, MeansLw_x, Self_y, WltTran_y, PosAff_y, RcEnds_x, NNPS, Doctrin_x, RBR, Work_x.
  - Hand-Crafted: :, PowAuth_y, Exert_y, Know_y, Doctrin_y, eval@_x, ANI_x, Exprsv_y, Legal_x, NUMB_y, ArenaLw_y, Sky_y, JJR, Foodx, RspLoss_x, Land_y, MeansLw_y, RcEthic_x, Name_x, NatrPro_y, Average duration of education, Number of recommendations received, Submit_y, SV_y, PowOth_y, Weak_y, Increas_y, Number of skilled endorsements received, Exert_x, IAV_x.
- Influence
  - Non-Textual: Intuition dimension, Judging dimension, number of patents, Introversion dimension, number of sections in the achievement section, Thinking dimension, number of titles, number of publicity, number of educations, average duration of voluntary experiences, number of companies in the work section, whether the user like the company he's currently working at, number of skilled endorsements, number of interests, minimum number of followers of an interest.
  - Textual: Doctrin_y, Name_y, Ritual_y, Introversion, PowCoop_x, CC, Affil_x, AffOth_x, PRP$, Complet_x, DAV_x, Legal_x, PowEnds_y, ComnObj_x, Space_y, FREQ_x, DIM_x, SklPt_y, RcEthic_y, Exch_y, Anomie_x, Vehicule_y, Finish_y, SklTot_x, If_y, WlbPt_x, MALE_y, ANI_y, Fail_y, Power_x.

  - Hand-Crafted: Work_y, SV_y, ANI_y, Introversion dimension, COLOR_y, Solve_y, Perceiv_y, RspLoss_y, PtLw_y, Total number of endorsements received, Sky_y, Exert_x, EnlLoss_x, Exprsv_y, Fall_x, EndsLw_y, PowGain_y, Kin@_y, Presence of personal branding, WlbLoss_x, Polit@_x, FREQ_x, Try_x, Decreas_y, Increas_y, VB, ABS_y, Space_x, HU_y.

- Steadiness
  - Non-Textual: Introversion dimension, Number of voluntary experiences, Presence of personal branding, Judging dimension, number of patents, Thinking dimension, Number of projects, number of recommendations received, number of skilled endorsements received, number of colleagues endorsements received, number of sections in the achievement section, Intuition dimension, number of companies, whether the user like the company he's currently working at. : Solve_x, Polit@_x, Ritual_x, RspTot_y, Positiv_y, COM_y, BldgPt_y, TranLw_x, Social_x, Eval@_y, Exch_y, PRP, Land_y, Intrj_y, FREQ_y, Complet_x, AffPt_y, Need_x, Ovrst_x, COLOR_y, WlbTot_y, VB, Intuition, WltTot_x, Positiv_x, PosAdd_x, WltOth_x, TimeSpc_y, VBZ, Undrst_x.
  - Hand-Crafted: Number of educations, SklPt_x, JJR, Total number of endorsements received, PowGain_y, WlbTot_y, POLIT_x, Weak_x, EnlOth_x, CD, AffLoss_x, Fall_x, IPadj_x, You_x, Stay_y, TmLoss_x, Ngtv_x, TrnGain_y, , Affil_y, Doctrin_y, WlbPsyc_x, Decreas_x, DAV_x, Compare_y, Aquatic_y, ArenaLw_y, POS_y, Polit@_y, TimeSpc_x.
- Conscientiousness
  - Non-Textual: Average number of endorsements, Number of colleagues endorsements, Number of voluntary experiences, Introversion dimension, Number of languages mastered, Number of description normalized in the Voluntary experiences section, whether the user like the company he<s currently working at, Number of honor titles, Presence of personal branding, Highest level of education reached, Number of educations, Number of interests, number of description normalized in the work section, number of description normalized in the education section, maximum number of follower for an interest.

  - Textual: Undrst_y, Academ_y, ComnObj_y, Know_y, Introversion, Econ@_y, EMOT_y, IPadj_y, Decreas_x, Yes_y, TimeSpc_x, Vehicule_x, Decreas_y, SklTot_x, Exert_y, Think_y, Means_y, Need_y, CD, $, Increas_x, Ngtv_y, Self_y, EnlOth_y, WlbPsyc_y, Tool_x, PowAuPt_x, Tool_x, PowAuPt_x, FREQ_x, EndsLw_y, SYM.
  - Hand-Crafted: Number of Patents, Polit@_y, Rel_y, Total number of endorseemnts, Number of titles, EndsLw_y, NegAff_y, Means_y, ", COM_y, Fall_y, TimeSpc_y, Whether the user like an influencer, NotLw_y, POS_y, Race_y, Goal_x, EVAL_x, WlbPhys_y, CD, PowCoop_y, SklOth_y, Kin@_y, Pleasur_y, Exert_y, , , MALE_x, COLL_x, Solve_x, DT.

- Introversion
  - Non-Textual: Number of skills, Average duration of educations, Average number of endorsements, Number of skilled endorsements received, Number of connections, Number of recommendation received, Number of educations, Number of description normalized, Number of sections in the achievement section, Number of recommendations given, Number of interests, Number of description normalized in the education section, whether the user like an influencer, Duration of longest job, Average duration of voluntary experiences.
  - Textual: Feel_x, Region_x, Quality_x, Influence, Milit_y, WltPt_y, Space_y, Pleasur_x, RBS, FREQ_y, RcRelig_x, TrnLoss_y, TranLw_y, EnlPt_y, Arousal_y, SYM, Goal_y, Weak_y, COLL_y, JJS, Positiv_y, RcTot_y, PowEnds_x, DIST_y, PtLw_x, EnlTot_y, NUMB_y, WltTot_x, Quan_y, AffGain_y.
  - Hand-Crafted: Negativ_y, Causal_y, Route_y, PtLw_x, Self_y, Solve_y, Number of skills, ORD_y, RspOth_y, , Number of patents, AffPt_y, PowLoss_x, RspGain_x, ANI_y, BodyPt_y, RcLoss_y, SklTot_x, SklTot_y, PowLoss_y, Intrj_y, PowAuth_x, Total number of endorsements, Think_x, Virtue_y, NotLw_x, Doctrin_x, Arousal_x, Means_x, Positiv_x.

- Intuition
  - Non-Textual: Presence of personal branding, Presence of summary, Number of skilled endorsements, Number of colleagues endorsements, Number of projects,

Number of descriptions normalized, Number of recommendations received, Number of interests, Average duration of educations, Number of skills, Number of description normalized in the education section, Number of voluntary experiences, Maximum number of followers for an interest, average number of followers of the interests, Minimum number of followers for an interest.

  - Textual: PLACE_x, Object_x, :, Pain_x, Stay_x, SocRel_y, Aquatic_y, ANI_x, Begin_y, Region_y, Anomie_x, Weak_y, Need_x, WMOT_y, MALE_y, ORD_x, Need_y, AffGain_y, RBS, Stay_y, ArenaLw_x, Exprsv_y, HU_y, SureLw_x, PowCoop_x, Affil_y, WlbTot_x, Our_x, COM_x, Academ_y.
  - Hand-Crafted: Average number of endorsements, Number of skills, Social_x, Space_y, WltOth_y, If_x, Negativ_x, , Submit_y, Legal_y, Arousal_x, SklTot_x, Self_y, You_x, Fall_y, Exprsv_y, Number of languages, EnlEnds_y, Decreas_y, Work_y, Tool_x, Fail_y, Active_y, ABS_x, AlbLoss_y, RcGain_y, ANI_x, ORD_y, Duration of Longuest jobs, RcLoss_y.

- Thinking
  - Non-Textual: Number of educations, Whether the user like an influencer, Highest level of education reached, Total number of endorsements received, Number of Connections, Average durations of work experiences, Number of description normalized in the work section, Number of publicity, Number of voluntary experiences, Number of sections in the achievement section, Number of work experiences, whether the user like the company he's currently working at, Number of projects, Average duration of voluntary experiences, Number of colleagues endorsements.
  - Textual: PosAff_x, Fall_y, Fail_y, Dominance, Econ@_y, Ovrst_y, EVAL_x, If_x, NatObj_x, Rel_y, Decreas_x, Pstv_y, Route_x, EMOT_x, WltPt_y, PowCoop_y, Milit_x, POS_x, Compare_x, FW, Stay_x, Ritual_y, Rise_x, PLACE_x, WDT, VBG, COLOR_x, PowAuPt_y, TIME_y, Perceiv_y.
  - Hand-Crafted: SV_y, BodyPt_x, Goal_y, Total number of endorsements, Active_y, WltTot_y, Complet_x, WltPt_x, Number of recommendation received, VBP, Perceiv_y, Compare_y, Travel_x, PowAren_x, PowGain_x, Academ_y,

SklAsth_x, NotLw_y, NatObj_y, Objct_x, Number of titles, Race_y, Region_x, WDT, HU_x, Steadiness Dimension, Number of connections, , , NatObj_x, Need_y.

- Judging
  - Non-Textual: Number of skills, Number of establishments in the education section, Average number of endorsements received, Number of publicity, Number of language mastered, Number of recommendations received, Highest level of education reached, Number of colleagues endorsements received, Number of patents, Average durations of jobs, Whether the user like an influencer, Number of voluntary experiences, Presence of summary, Number of work experiences, Number of courses.
  - Textual: PowTot_y, ECON_y, Fetch_x, Solve_x, ComnObj_y, PRP, Positiv_y, RspOth_y, NatrPro_y, SocRel_y, BldgPt_y, IndAdj_y, COLOR_y, Food_y, Weak_y, AffOth_y, POLIT_y, Undrst_x,), Name_y, ”, NatObj_x, PRP$, wordcount_y, Finish_x, PtLw_x, NNS, Academ_x, Arousal_x, Exch_x.
  - Hand-Crafted: Time@_y, Dominant Dimension, AffTot_x, Total number of endorsements, ABS_y, Passive_y, ECON_y, Female_y, FREQ_y, PowCon_x, Whether the user like an influencer, Route_y, Self_y, Number of tests, FormLw_y, SklOth_x, Number of description normalized in the education sections, Milit_x, EnlPt_y, WltTot_x, Ngtv_x, ABS_x, Ought_y, COM_y, Finish_y, RcEthic_y, Relig_y, EnlTot_y, Ngtv_y, Fall_x.

Appendix C

Contingency tables

Tab. C.1. Contingency tables between the DiSC and the MBTi personality models

	$\neg D$	D	$\neg I$	I	$\neg S$	S	$\neg C$	C
Extroversion	177	370	147	400	452	95	426	121
Introversion	148	136	207	77	152	132	100	184
Sensing	102	149	140	111	170	81	136	115
Intuition	223	357	214	366	434	146	390	190
Feeling	199	177	120	256	233	143	270	106
Thinking	126	329	234	221	371	84	256	199
Perceiving	113	162	80	195	207	68	209	66
Judging	212	344	274	282	397	159	317	239

Tab. C.2. Contingency table between the traits of the MBTI personality model

	Extroversion	Introversion	Sensing	Intuition
Extroversion	547	0	172	375
Introversion	0	284	79	205
Sensing	172	79	251	0
Intuition	375	205	0	580
Feeling	262	114	99	277
Thinking	285	170	152	303
Perceiving	195	80	41	234
Judging	352	204	210	346
	Feeling	Thinking	Perceiving	Judging
Extroversion	262	285	195	352
Introversion	114	170	80	204
Sensing	99	152	41	210
Intuition	277	303	234	346
Feeling	376	0	155	221
Thinking	0	455	120	335
Perceiving	155	120	275	0
Judging	221	335	0	556

Tab. C.3. Contingency table between the traits of the DiSC personality model

	$\neg D$	D	$\neg I$	I	$\neg S$	S	$\neg C$	C
$\neg D$	325	0						
D	0	556						
$\neg I$	130	224	354	0				
I	195	282	0	477				
$\neg S$	161	443	216	388	604	0		
S	164	63	138	89	0	227		
$\neg C$	173	353	102	204	402	124	526	0
C	152	153	252	53	202	103	0	305

Tab. C.4. Contingency table between each types of the DISC personality model and each type of the MBTI personality model

	ESFP	ESFJ	ESTP	ESTJ	ENFP	ENFJ	ENTP	ENTJ
C	1	1	0	2	1	1	0	2
S	0	2	0	1	0	2	0	0
SC	0	4	0	3	0	3	0	2
I	3	6	0	7	33	15	5	13
IC	0	6	0	1	2	0	6	8
IS	2	12	0	5	6	8	4	2
ISC	0	1	0	0	0	6	0	1
D	2	1	2	14	2	7	4	13
DC	0	7	0	24	4	4	5	13
DS	0	2	0	1	3	2	3	2
DSC	0	0	2	3	0	1	0	1
DI	10	12	5	24	39	44	46	52
DIC	0	0	0	3	0	0	1	1
DIS	0	1	0	2	3	3	0	1
DISC	0	0	0	0	0	0	1	0

	ISFP	ISFJ	ISTP	ISTJ	INFP	INFJ	INTP	INTJ
C	0	3	0	10	1	1	7	13
S	2	2	0	1	6	1	0	4
SC	1	10	0	9	10	15	4	5
I	1	0	0	0	1	1	0	2
IC	1	0	0	0	0	3	2	2
IS	0	4	0	1	11	6	1	3
ISC	0	0	0	0	2	2	0	0
D	0	0	1	3	0	1	0	4
DC	0	0	4	12	1	7	7	31
DS	0	1	1	1	1	1	3	6
DSC	0	1	0	6	2	3	0	5
DI	0	0	3	1	4	6	2	13
DIC	0	0	0	0	0	1	1	2
DIS	0	0	0	0	0	1	0	0
DISC	0	0	0	0	0	0	0	0

Appendix D

Precision-Recall curve

Fig. D.1. Precision-Recall curve for the voting system for the Influence trait of the DiSC personality

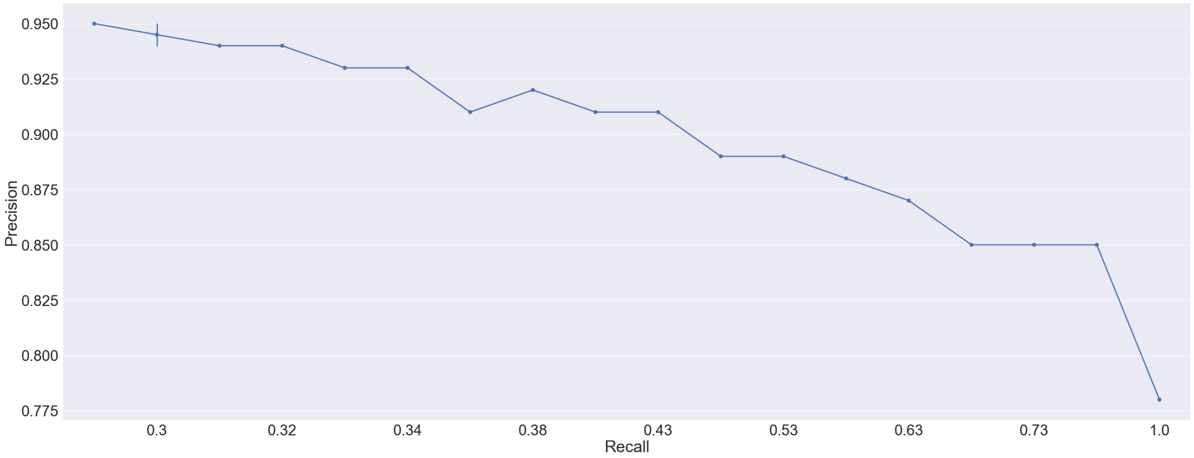


Fig. D.2. Precision-Recall curve for the voting system for the Influence trait of the DiSC personality, with and without the inclusion of the other personality model

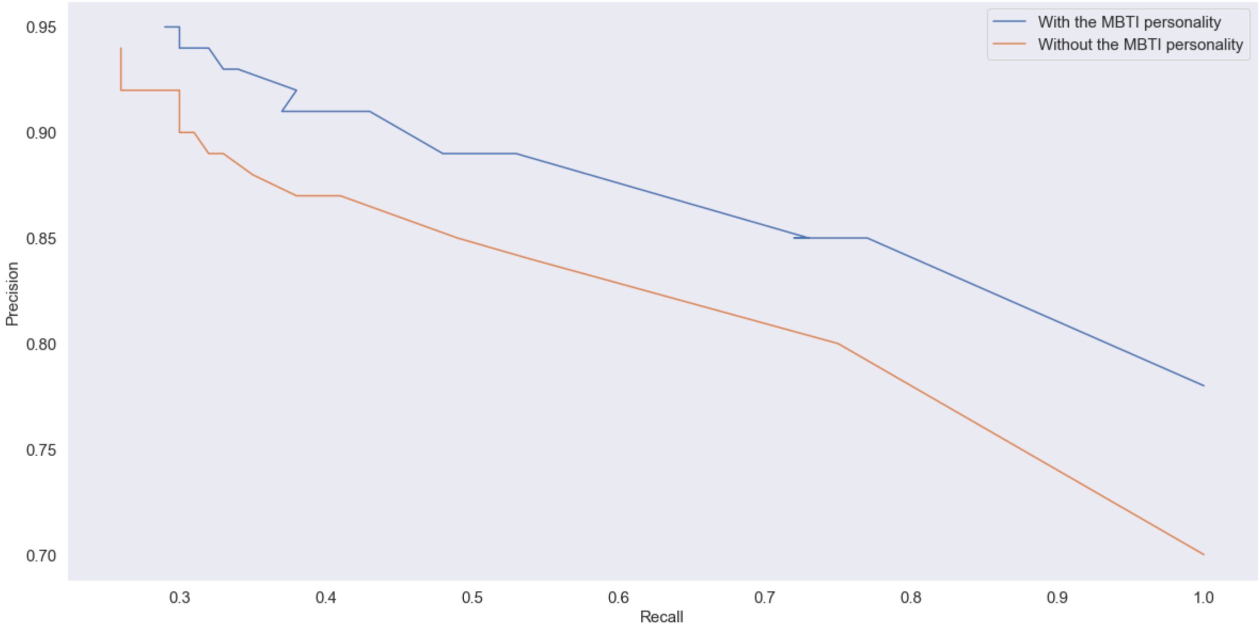


Fig. D.3. Precision-Recall curve for the voting system for the Steadiness trait of the DiSC personality

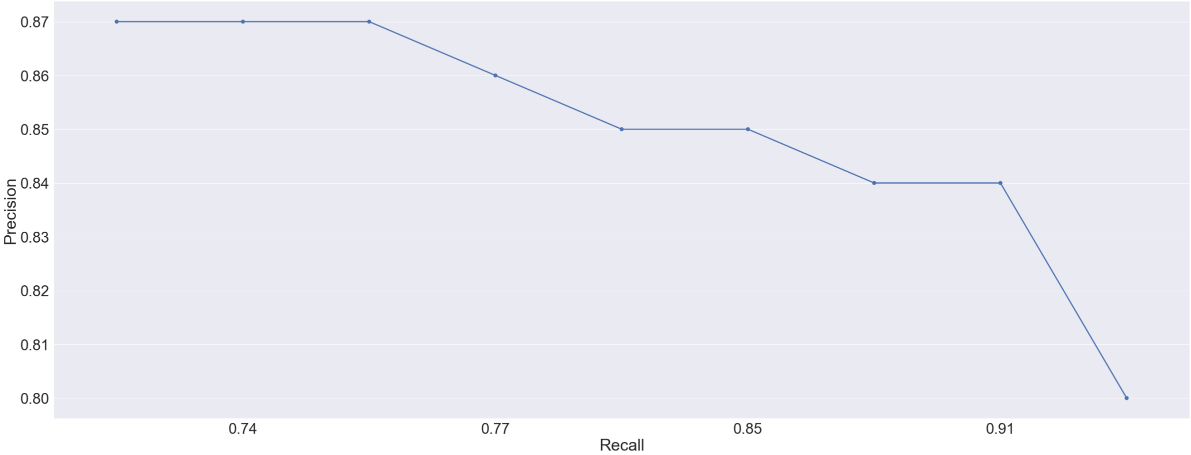


Fig. D.4. Precision-Recall curve for the voting system for the Steadiness trait of the DiSC personality, with and without the inclusion of the other personality model

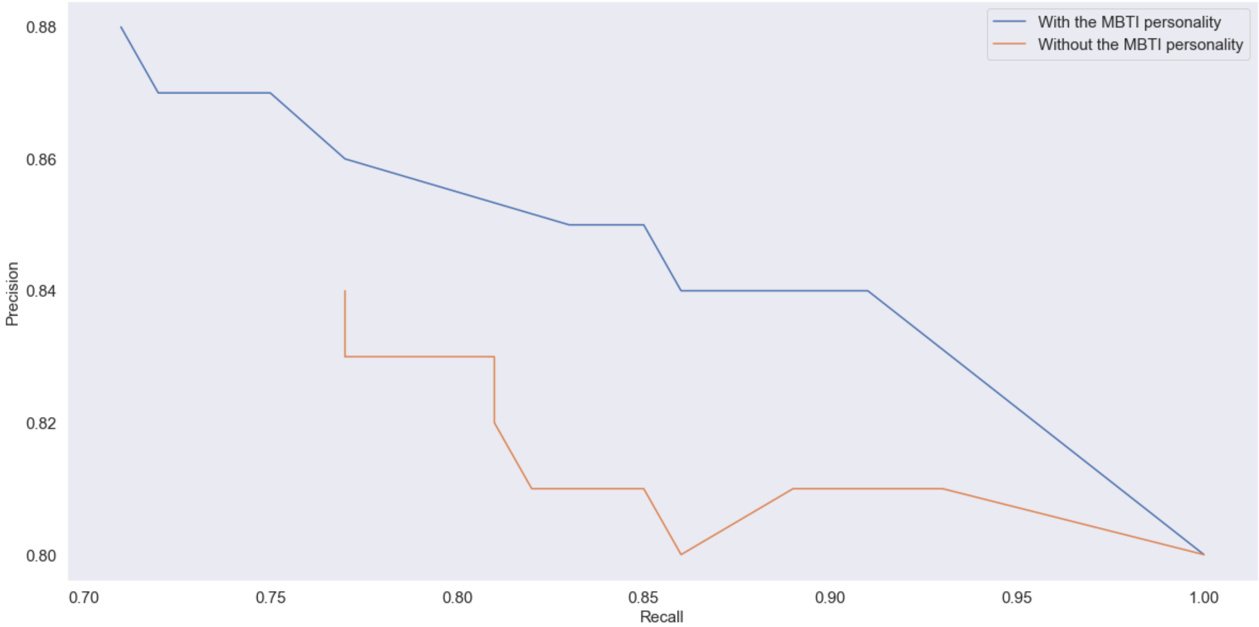


Fig. D.5. Precision-Recall curve for the voting system for the Conscientiousness trait of the DiSC personality

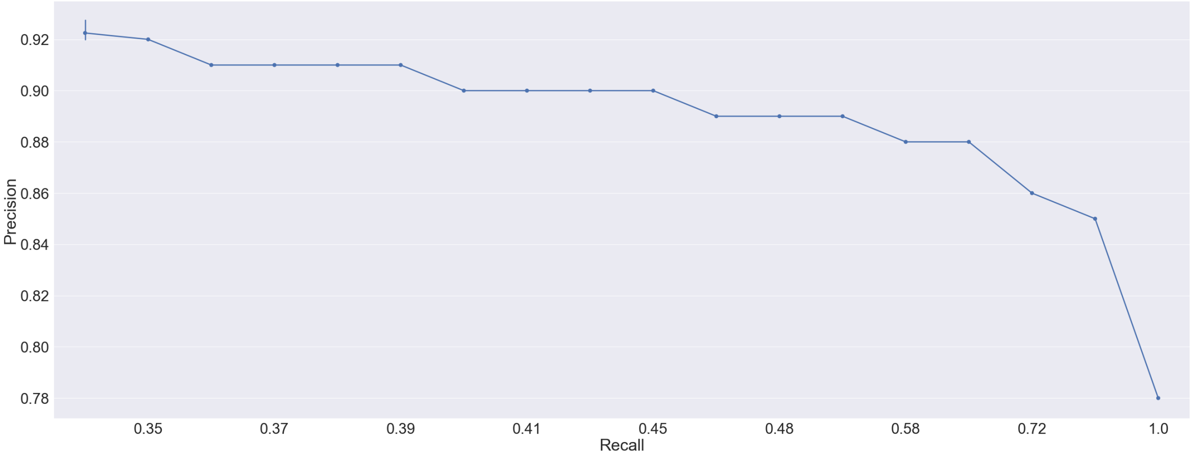


Fig. D.6. Precision-Recall curve for the voting system for the Conscientiousness trait of the DiSC personality, with and without the inclusion of the other personality model

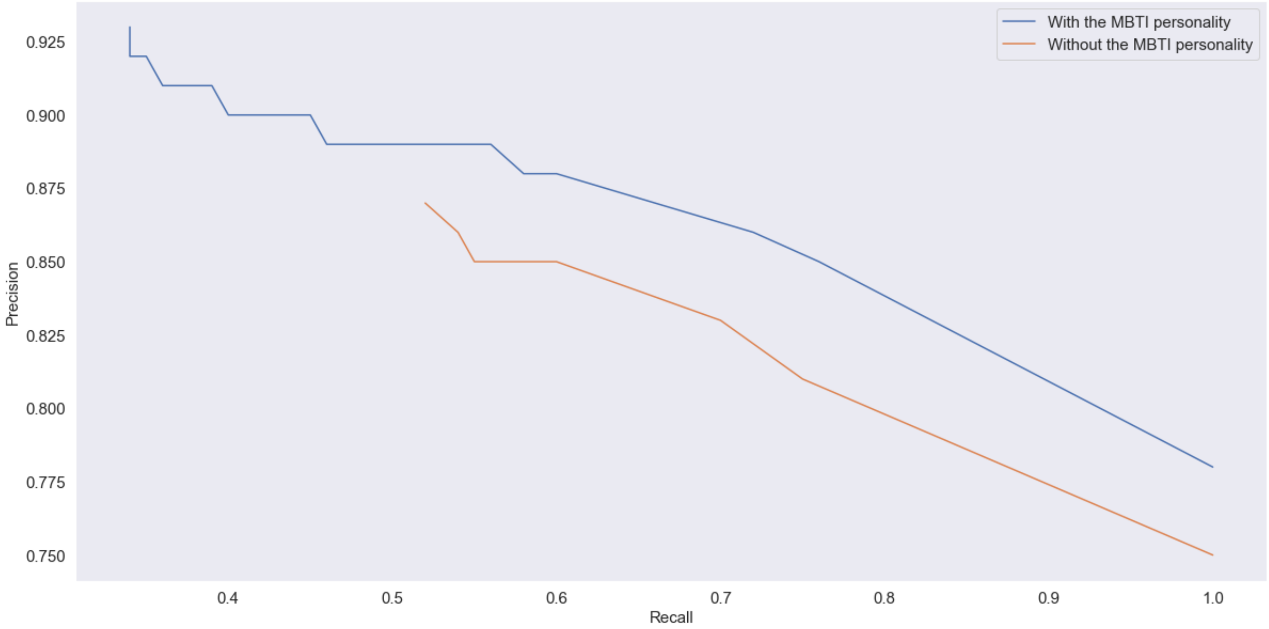


Fig. D.7. Precision-Recall curve for the voting system for the Introversion trait of the MBTI personality

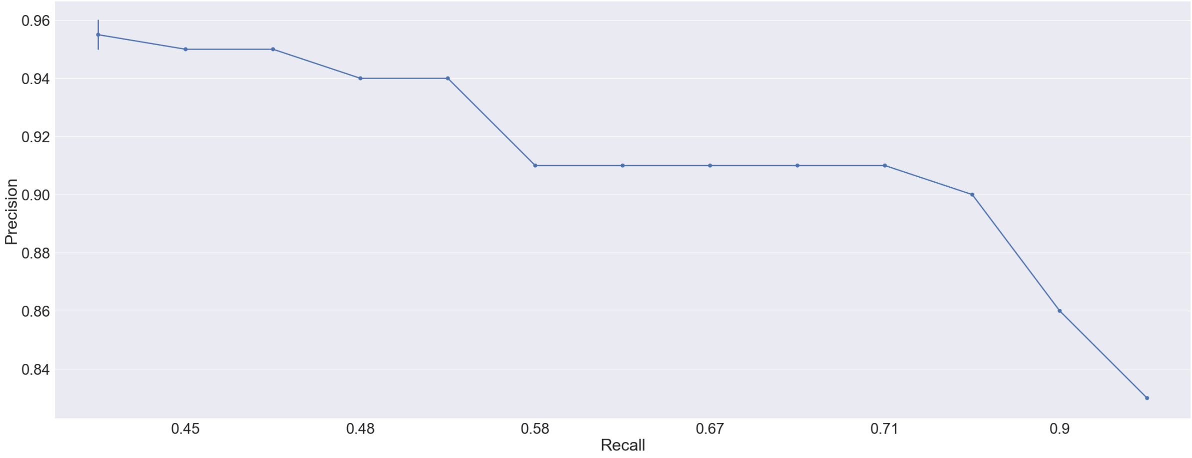


Fig. D.8. Precision-Recall curve for the voting system for the Introversion trait of the MBTI personality, with and without the inclusion of the other personality model

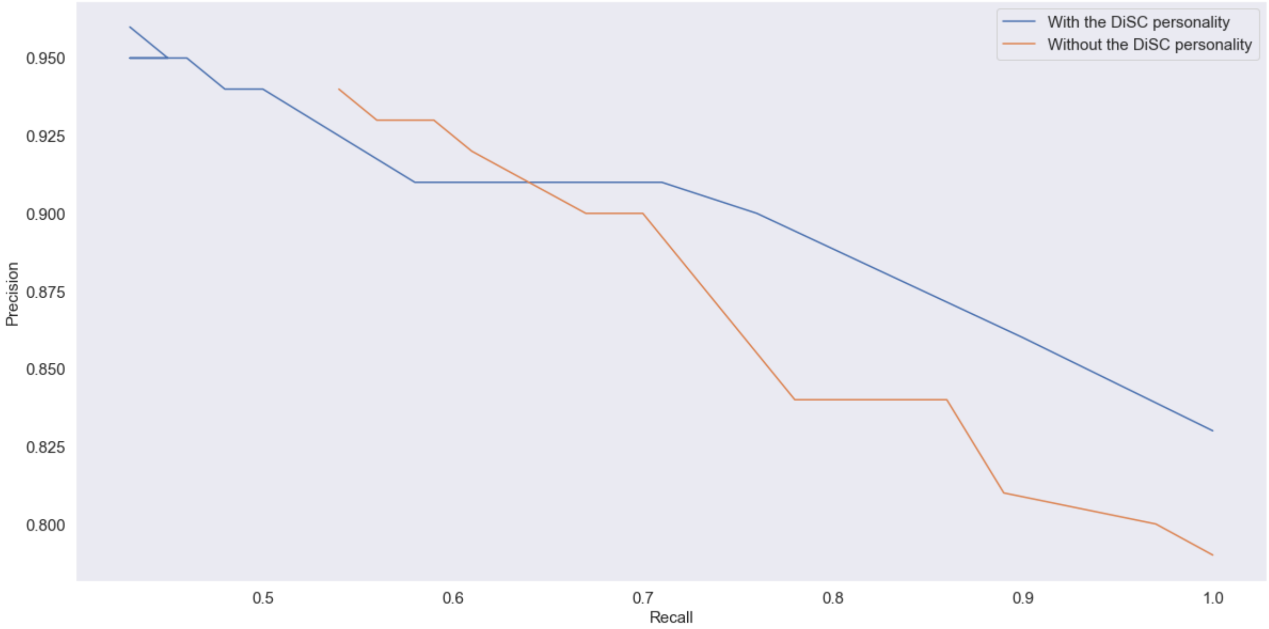


Fig. D.9. Precision-Recall curve for the voting system for the Intuition trait of the MBTI personality

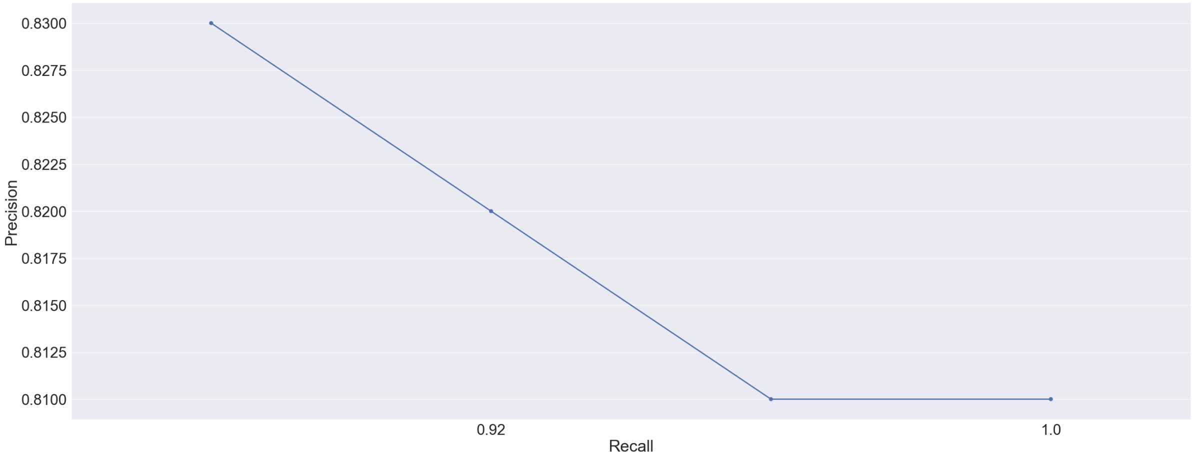


Fig. D.10. Precision-Recall curve for the voting system for the Intuition trait of the MBTI personality, with and without the inclusion of the other personality model

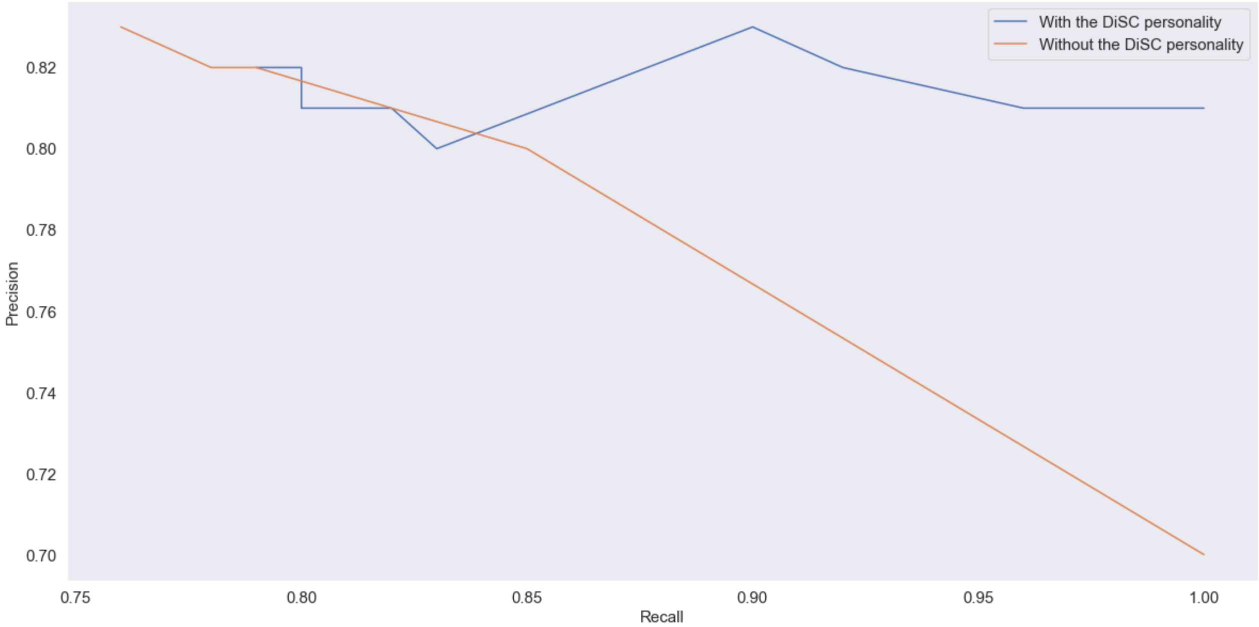


Fig. D.11. Precision-Recall curve for the voting system for the Thinking trait of the MBTI personality

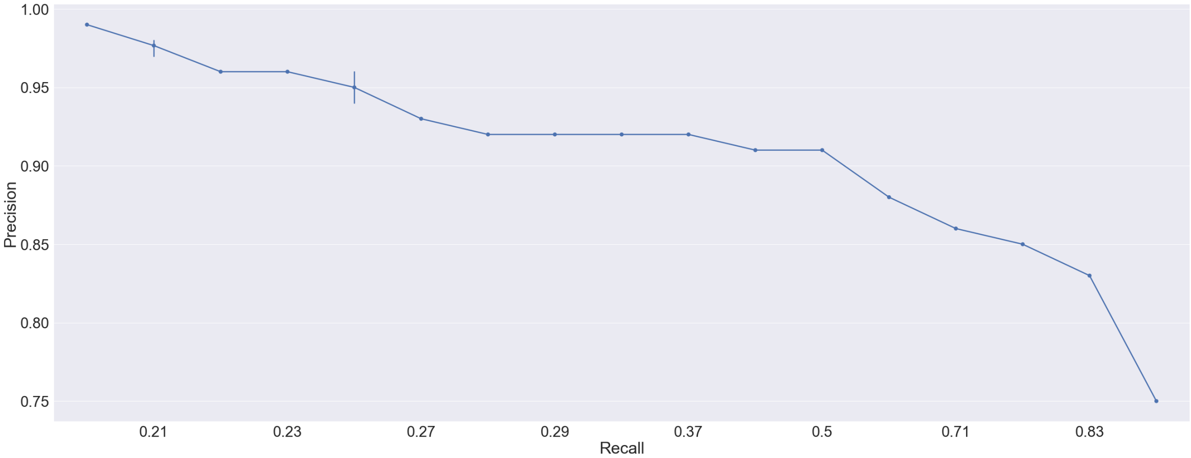


Fig. D.12. Precision-Recall curve for the voting system for the Thinking trait of the MBTI personality, with and without the inclusion of the other personality model

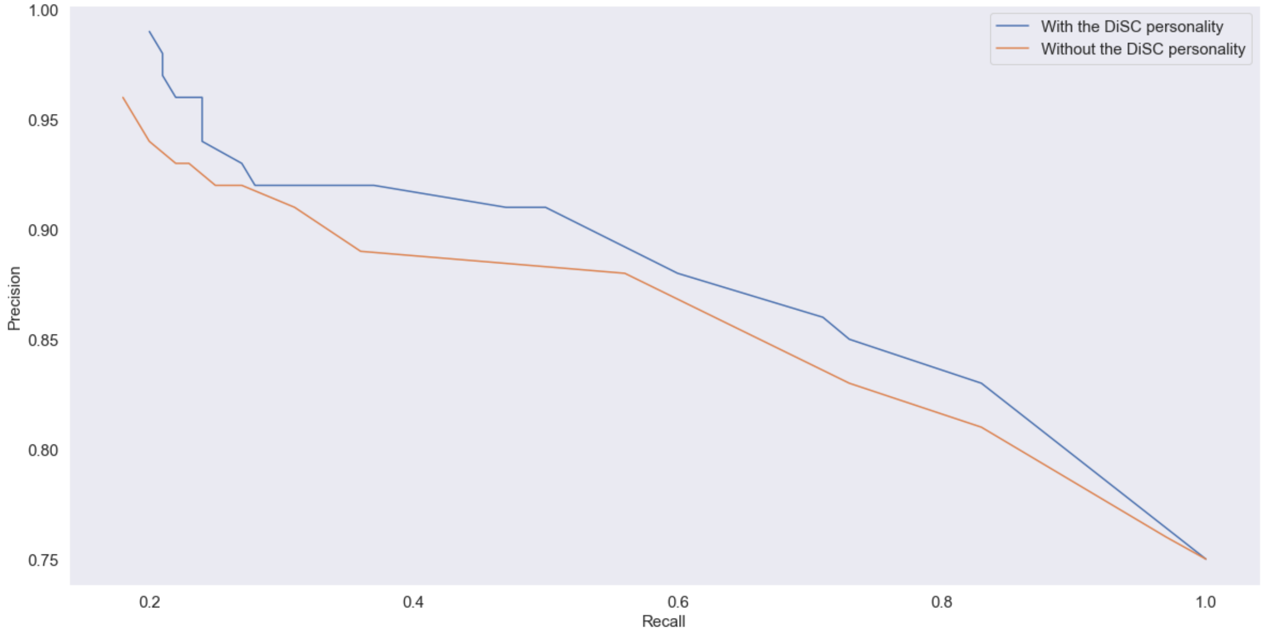


Fig. D.13. Precision-Recall curve for the voting system for the Judging trait of the MBTI personality

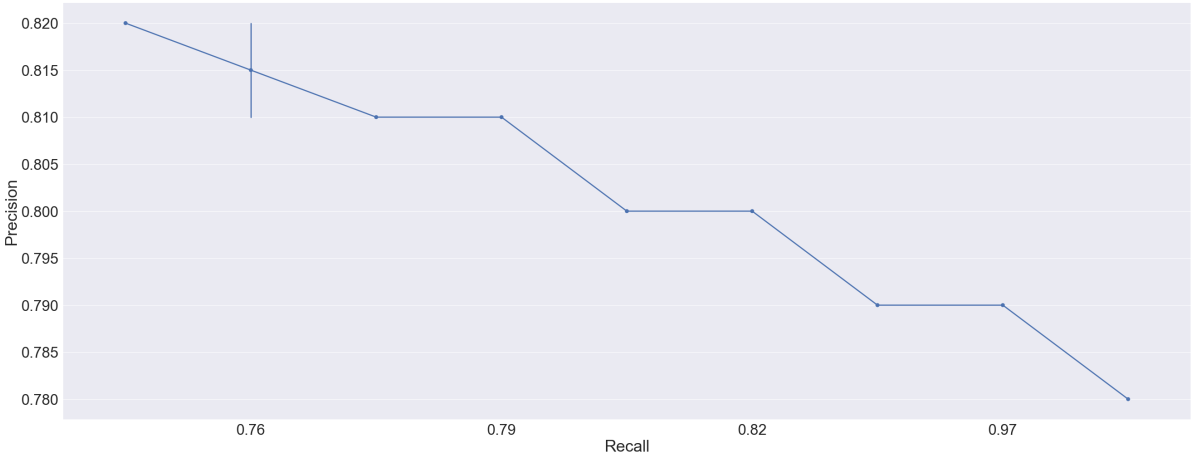
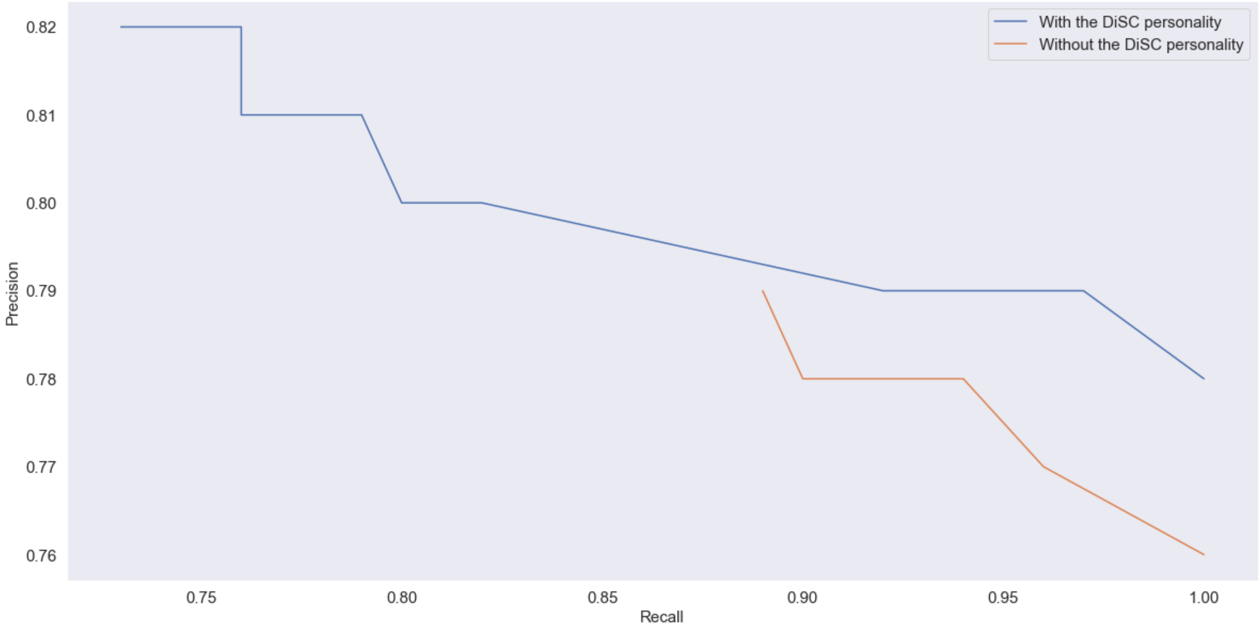


Fig. D.14. Precision-Recall curve for the voting system for the Judging trait of the MBTI personality, with and without the inclusion of the other personality model


With the DiSC personality
Without the DiSC personality
Precision
Recall