ISCTE IUL
Instituto Universitário de Lisboa

Department of Information Science and Technology

Characterizing the Personality of Twitter Users based on their Timeline Information

Anzhela Zhusupova

A dissertation submitted in partial fulfillment of the requirements for the degree of **Master in Computer Engineering**

Supervisor:
PhD Fernando Batista, Assistant Professor,
ISCTE-IUL – Instituto Universitário de Lisboa

October, 2016

Resumo

A personalidade traduz-se num conjunto de características que diferenciam uma pessoa de outras. Pode ser identificada pelas palavras que as pessoas usam numa conversa ou em publicações que fazem nas redes sociais. A maioria dos trabalhos existentes na literatura estão focados na previsão de personalidade analisando textos em Inglês. Neste estudo, foram analisadas publicações dos utilizadores Portugueses na rede social Twitter. Tendo em conta que o limite de 140 caracteres imposto aos tweets pode dificultar a classificação dos sentimentos dos textos produzidos, foi decidido usar diferentes características e métodos tais como locais, tempo de publicação, quantidade de seguidores, quantidade de amigos, etc., para obter uma imagem mais completa da personalidade. Este documento apresenta um método para fazer a previsão da personalidade de utilizadores do Twitter, com base na informação existente e sem qualquer esforço do lado desses utilizadores. A personalidade pode ser calculada através da informação pública disponível.

Abstract

Personality is a set of characteristics that differentiate a person from others. It can be identified by the words that people use in conversations or in publications that they do in social networks. Most existing work focuses on personality prediction analyzing English texts. In this study we analyzed publications of the Portuguese users of the social network Twitter. Taking into account the difficulties in sentiment classification that can be caused by the 140 character limit imposed on tweets, we decided to use different features and methods such as the quantity of followers, friends, locations, publication times, etc. to get a more precise picture of a personality. In this paper, we present methods by which the personality of a user can be predicted without any effort from the Twitter users. The personality can be accurately predicted through the publicly available information on Twitter profiles.

Palavras Chave
Keywords

Palavras chave

Traços de personalidade

Perfil do utilizador

Utilizadores portugueses do Twitter

Análise de Sentimentos

Keywords

Personality traits

User profile

Portuguese Twitter users

Sentiment Analysis

Agradecimentos
Acknowledgements

First and foremost, I would like to express my deepest gratitude and appreciation to my advisor, Fernando Batista, for his great support, inspiration and guidance throughout such the interesting research area the study of which turned into my hobby. I would like to thank him for all that he has done for the success of our work.

I would also like to extend my appreciation to professor Ricardo Ribeiro for his advice and support during my research.

A special thank to my parents Irina and Virgilio for their love, confidence in me, support and inspiration. Also I will always be grateful to my grandmother for everything she had done for me during my life that I will never forget.

And finally, I would like to thank my friends for the encouragement and support during this project.

Thank you everyone else that helped and provided support.

Lisboa, 25 de outubro de 2016
Anzhela Zhusupova

Contents

List of Figures

List of Tables

1 Introduction

> "I applied my thoughts to the puzzling question - one, probably, which will puzzle me for ever - why it is that, while all Greece lies under the same sky and all the Greeks are educated alike, it has befallen us to have characters so variously constituted" (Diggle, 2004)

Several years ago some things that we nowadays consider as trivial, existed only in minds of scientists and in films. But everything has changed with appearance and fast development of computers, of methods for processing of large amounts of data, with appearance of Internet and social networks that made the whole world interconnected, and obviously simplified the life of humanity.

Researchers always were interested in finding more efficient and rapid solutions for different problems. With appearance of social media, approaches for resolving problems of psychological research such as the identification of the type of personality, social behaviour analysis, identification of cognitive styles, are being improved continuously (Tausczik and Pennebaker, 2010a). Social media is the easy and fast mean for people to express their thoughts that "opens doors" into their life. In this work we are interested in making the sentiment and emotional analysis and the personality identification of user profiles in social network Twitter and also it would be interesting to find associations between various characteristics of a user profile with the personality type.

Personality is a set of characteristics that differentiate a person from others. It is a psychological science term that has been the focus of many studies in which have been found relationships between personality and psychological disorders, job performance, satisfaction, romantic success, amongst others (Golbeck et al., 2011a). The process of identification of personality type is not a simple task, taking into account the fact that every person has an individual set of different psychological characteristics (Solera-Ureña et al., 2016). As was noted in the literature (Solera-Ureña et al., 2016), the difficulty of identification of the psychological types varies, i.e. some of them are simpler to recognise than other types. For example, according to some literature sources (Hovy and Hovy, 2015), it is easier recognise traits such as Introvert/Extravert and Thinking/Feeling than other ones.

It is clear in previous work, that the personality type of a human influences on all the sides of his life. For example, on music tastes, selection of movies, books, on the way a person behaves in real life and online, on how one constructs relationships with other people (Quercia et al., 2011) and also on the word usage in daily life. As was pointed out by Mairesse et al. (2007) the sentences make available a big amount of information about the speaker personality together with semantic data.

Other psychological studies also have revealed that the words that we use in daily life reflect our thoughts and emotions. Words are very important features used in psychology to gain better understanding of human beings. Much of our personality can be identified by the words that we use in conversation or in publications that we make in social networks. They also can reveal social relationships, thinking styles, individual differences, the things at what we are focused at in a given moment and what emotions we are experiencing (Tausczik and Pennebaker, 2010b). For example, teenagers are more focused on motion, new technologies, games; people that have problems frequently use pronouns such as “I” or “Me”; positive ads use more frequently future tenses and in negative ones are used past tenses; positive emotion words can show us levels of agreement; lying persons use negative emotion words together with words that express motion such as “go”, “arrive” amongst others; if a person have a close relationship with others uses the pronoun “You” (Tausczik and Pennebaker, 2010b).

For the purpose of simplifying the process of identification of personality, psychologists have discovered correlations between personality traits and lexical categories, twitter user types , n-grams amongst others (Mairesse et al., 2007; Vosoughi et al., 2015; Yarkoni, 2010).

According to Solera-Ureña et al. (2016), there is a great amount of applications that can assess personality traits during the analysis of interactions between people and during interactions of a user and a computer. A progress in artificial intelligence had led to the appearance of applications that provide such resources as virtual agents that conduct their behaviour taking into account some traits of a user for the better practice of communication with him (Solera-Ureña et al., 2016). The social networks provide a great amount of data produced by users that can be studied and leveraged for extracting the necessary information for particular tasks in different areas. For example, last time a social network Twitter, as was noted in the literature, becomes every day more popular and useful especially for studying the behaviour and attitude of people, and for studying the personality traits (Vosoughi et al., 2015).

According to Schwartz et al. (2013), the number of users of Facebook and Twitter equals to 1/7 of the population of Earth. As Golbeck et al. (2011a) pointed out, in 2005 the total number of users on the web approximately was 115 million and already in 2010 on Twitter had registered more than 200 million of users. This social network was launched in 2006, and represents a place where users can read and write millions of short messages that are not longer than 140 characters. The messages are called tweets (Tumasjan et al., 2010). There are different types of information that can be extracted from a profile of a user in social media, for example location, specific interests, political preferences, consumption preferences, style of life, also it is possible to monitor the changes in mood across hours, days and months. This information can be used to predict a personality type of users (Golbeck et al., 2011b; Schwartz et al., 2013). As was noted in the work of Golbeck et al. (2011a), users of Facebook don't try to make profiles that could show only the best sides of their personality, on the contrary, as was shown in previous works Golbeck et al. (2011b), the profiles reflect real characteristics of users, as they are in offline life. In Twitter, unlike Facebook, users must not reveal true information about them, like name and age, so they can make fake accounts and feel free to public all the things they think about. It is expected that the personality traits extracted from users profiles of Twitter are the reflection of their actual personalities (Golbeck et al., 2011a), because everybody uses mobile phones that speed up and make more informal the process of making the tweets, without having to worry about choosing the words for describing their thoughts. This micro - blogging service can be used as a research tool for tracking diseases, for making political predictions, social unrests, to track the mood to predict levels of happiness for the purpose of planning the strategy of stock market and also it is useful for predicting personality traits (Tumasjan et al., 2010; Schwartz et al., 2013). To the opinion of the authors Mairesse et al. (2007), the usefulness of applications for automatic personality detection is also great for helping to solve other more global problems such as analysing of personality of suspected terrorists, criminals; for analysing conversations on dating sites for helping people to find the best match; for make a service of tutoring platforms more adapted to personality of users.

The great part of studies has performed a sentiment classification of tweets for many languages. There were used different standard classification techniques, without taking into account that Twitter consists of very short messages, where are always used specific forms of words and symbols, adapted for more rapid and short thoughts expression.

According to Quercia et al. (2011), personality in social networks is usually studied using a psychological model, named "The Big Five", because it is the most comprehensive model and concentrates the main characteristics of personality: Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism. As was noted in more recent work, Big Five model over the approximately 50 years, has been considered by many researchers as a standard in psychology (Mairesse et al., 2007). For example, it was pointed out that for an extraverted person the social networks are useful and very simple to use (Quercia et al., 2011). To note, in other literature sources (Hovy and Hovy, 2015) was concluded that introverted people find the usage of social networks simpler and useful as they tend to avoid the direct interactions with people.

So this situation proves one more time that the process of identification of a personality type is not an easy one.

Besides tweets, Twitter provides also many other characteristics, which were named in literature as metadata. The metadata, namely a geolocation, a temporal, author information, a quantity of followers, followees and favourite count was used in some works as the additional data for reaching more precise results for sentiment classification and prediction of personality type (Vosoughi et al., 2015). We also became inspired by this approach and tried to apply it to our work.

The remainder of this section describes our goals and motivation and how, in our opinion, this study can contribute to the personality research area.

1.1 Motivation and Goals

We aim at creating a computer application that can automatically extract personality traits of Portuguese Twitter users. They will not need to fill in any long questionnaires, on the contrary of many other existing applications because, for example, it is not practical for business organisations to ask millions of customers to respond questions for the purpose of gaining a deeper understanding of personalities and suggesting appropriate services and products(Gou et al., 2014). Additionally to this, as was supposed in previous work of Gou et al. (2014), nowadays does not exist any effective test with questions that could derive with a good precision traits of personality.

Despite the fact that there are approximately 220 million native speakers and 260 million total speakers of Portuguese, being the sixth most natively spoken language in the world (see www.ethnologue.com/statistics/size), currently, to the best of our knowledge, there are only few studies that analyse the type of personality through the analysis of the publications in Twitter for Portuguese users. According to Tausczik and Pennebaker (2010b), the bulk of the work relies on judges ratings for evaluating text, but even after several experiments, judges did not always agree with each other. Moreover, the work of judges is slow and expensive. Our goal is to create the application that will be able to make a preliminary personality prediction of the users of Twitter, taking into account the following aspects:

- Content produced by each user;
- The periodicity of production of tweets;
- The number of profiles the user follows and the number of followers;
- Gender;

- Age of user;
- Sentiments that user expresses in each tweet: positive, negative or neutral;
- Localisation.

1.2 Contribution

Our contribution in the area of research of automatic personality recognition is as follows:

- A platform that performs a bilingual analysis of twitter user profiles have been developed. After reviewing many existing works, we noted that there is a little number of research linked with analysing textual content produced by Portuguese twitter users. The absolute majority of existing works was elaborated for English and many other languages, but for Portuguese exist a little number of works that can be counted on one hand (Silva and TEAM (2011),Solera-Ureña et al. (2016),Morgado (2012))
- The application that we had elaborated, considerably simplifies the process of preliminary personality prediction and save the time of users, because the analysis of information of twitter user profiles is made automatically. To the best of our knowledge, a majority of applications for identification the type of personality is based on long and tiresome questionnaires.
- The program performs the preliminary twitter user personality type prediction, also it makes the sentiment analysis of every tweet and provides the user temporal activity visualisation. It also shows some additional characteristics, such as twitter user following and followers count, total corpus lexical diversity, number of words per tweet, number of positive and negative emoticons used in all publications, number of swear words and other ones. All these characteristics may help to better understand a personality of a user.
- Twitter publications have restrictions of length, they can not be longer than 140 characters. It makes the process of sentiment analysis and therefore the process of personality identification more difficult and less precise. For the purpose of getting a preliminary result of a personality type and for performing a sentiment analysis of tweets it was assumed that for achieving more precise results would be better to join different tools described in the previous works. We used the dictionary LIWC together with correlations between personality traits and LIWC dictionary categories that were previously achieved by other authors. For sentiment analysis have been used two dictionaries together with the analysis of emoticons. The more detailed description of methods will be performed in the following chapters.

1.3 Personality Detection Approaches

As was pointed out in literature there are some approaches that help to infer the personality of a person, such as non-verbal and text-based approaches (Vinciarelli and Mohammadi, 2014). According to the literature, is possible to detect the type of personality by considering the distance between speakers, by noting intonation, gesticulation together with different postures of a body and eye-gaze direction (Vinciarelli and Mohammadi, 2014). In this work, we will only perform the text-based approach for predicting a type of personality.

Various psychological models exist for detecting types of personality of people, being the "Big Five personality model" and the "Myers-Briggs Type Indicator" (MBTI) two of the most popular models. The following sections describe them in more detail.

1.3.1 MBTI

The "MBTI" model consists of such characteristics as Thinking - Feeling, Sensation - Intuition, Introversion - Extraversion, Judging - Perceiving (see more on http://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/). Each characteristic can describe a person from different angles. Thinking - Feeling shows the way the person prefers to resolve problems: if one relies on the logic or the decision-making depends on circumstances and opinion of others. Sensation - Intuition shows the way a person processes the information: if one absorbs the main concepts or prefer adding some additional meaning. Introversion - Extraversion shows if one is more focused on the inner world or, on the contrary, interested in interaction with the surrounding world. Judging - Perceiving characteristic helps to understand if an individual tends to have a predefined opinion about all things or, on the contrary, he is an open-minded person that accepts all new information (see more on http://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/).

The Myers - Briggs personality type can be expressed as a combination of four letters, each one of them represents one of the characteristics explained above. Generally, there are 16 types of "MBTI" types of personality. This model was developed by Isabel Briggs - Myers in 1940 and still is in the process of ongoing research. As was noted by Hovy and Hovy (2015), MBTI is less expressive than a Big Five model that we will discuss next.

1.3.2 Big Five Personality Traits

As was noted by Alam et al. (2013) personality is a set of particular characteristics that differentiate one person from another one. These differences are reflected in their perception of surrounding world, in their thoughts, actions and also in the words that they use to describe what are they feeling and what are they thinking about in the given moment of time. There are various approaches of analysing words for prediction personality traits, but the most well-researched and the most used in previous work is the “Big Five” model.

- Benet-Martinez and John (1998) stated that people high in *extraversion* can be described with the following characteristics: activity and energy, dominance, sociability, express-iveness, and positive emotions. Extraverted people tend to be outgoing, find friends in a simple way, like to talk and to be the centre of attention and also usually participate in social activities. Introverted ones, on the contrary, seem to be closed and tend to avoid social contacts.
- Characteristics such as altruism, tender-mindedness, trust, and modesty characterise the trait *agreeableness*. People of this type of personality similarly to extraverted ones tend to emit positive emotions, avoiding expressing negativity. The thing that better then others characterise agreeable people is the love to help others and to adapt to their needs. Disagreeable ones, according to literature, are not cooperative, are focused on needs of themselves, and do not depend on social expectations (Farnadi et al., 2014).
- The people that are high in *Conscientiousness* are fond of work, are organised, honest, reliable, like to make plans and are focused on the achievement of the goals. People that characterised by the personality trait opposite to *Conscientiousness* tend to be more creative, don’t like to make plans and follow the rules (Farnadi et al., 2014).
- *Neuroticism* combines a large variety of negative effects such as anxiety, sadness, irritability, and nervous tension. This type of people tends to be depressive, have the unpredictable mood and also use words that reflex negative thoughts and emotions. But, emotionally stable people, on the contrary, have a calm character, they are more self-confident and tend to emit positive emotions (Farnadi et al., 2014).
- People high in *Openness*, obviously, are opened to new experience, have a good imagination, are very creative and curious and also have a good sense of aesthetics (Farnadi et al., 2014).

1.4 Sentiment Analysis and Emotion Detection

Sentiment analysis and emotion detection can be defined as the automatic extraction of information about sentiments and emotions from unstructured text. Since the Twitter publications can not be longer than 140 characters, users tend to use more informal language similar to a language used in oral daily conversations, and every publication carry an emotional weight and sentiment polarity. There are several types of basic emotions which were mentioned in a work of Farnadi et al. (2014), such as: fear, joy, trust, anger, anticipation, disgust, surprise and sadness. Concerning sentiment polarity, it is common to distinguish between 2 types: positive and negative, but in our work we also considered the neutral one. Analysis of a content voluntarily produced by users of social networks, blogs provides an unprecedented source of data that can be leveraged by some companies and organisations that want to know their strong and weak sides for making their strategy more nimble and holistic for improving the profit. One of the possible ways to do it, is tracking sentiments and emotions expressed by twitter users in the feedback about their products or services. Knowing emotional states of people also can be useful for a large number of applications, including the identification of a suicidal mood, the prediction of results of political elections or for analysing emotions during sportive events. We made a try in tracking the sentiments and emotions expression in every tweet produced by users over two years. The detailed process and results of this analysis will be described in the following chapters.

1.5 Lexical Complexity Analysis

The level of lexical complexity varies from trait to trait. It consists of various characteristics, but in this study we considered such metrics as lexical diversity, a number of words per tweet and the average length of each word used by the user. According to Russell (2013), lexical diversity indicates the richness of the language of a user, i.e. the diversity of a user's vocabulary.

There are different opinions relatively the lexical diversity. Vaezi and Kafshgar (2012) stated that extraverts tend to write more complexly structured publications, while the speech of the people high in Introversion is poor in terms of lexical structure and variety of used words.Conclusion of **?** is contradictory to Vaezi and Kafshgar (2012). In this study authors stated that language of extraverts is characterised with poor lexicon and low lexical diversity, while introverts have a rich lexicon with a high level of lexical diversity.

There have been made more interesting findings about the lexical diversity in a previous work. For example, in the work Russell (2013) authors state that this term can be used for understanding of the competence of a person during the discussion of some problems. If one repeatedly avoids providing some details and instead of this tend to generalise information, the interlocutor may think that a person is not well-informed about a discussed subject and have not enough competence to resolve a problem.

Bradac et al. (1979) pointed out that this characteristic may indicate different levels of stress experienced by a person in a moment of conversation. There are several researches that proved that "cognitive stress on the part of a source is inversely related to lexical diversity" (Bradac et al., 1979). There were performed some experiments in previous work, during which several persons were interviewed in different conditions. In one case, interviewers showed aggression during the interview, that caused a low level of diversity of speech of respondents, while in a control group of respondents it was higher (Bradac et al., 1979). Authors also proposed a hypothesis, that there is a probability of the lexical diversity of respondents being low in very high or very low stressed situations, but a moderately stressed situation may only cause a high lexical diversity. In another experiment, was noted, that the higher level of anxiety causes a repetition of words. In another work was found that if a person has a high writing apprehension uses a fewer number of words, while not apprehensive one, tend to produce longer messages using more uncommon words. Tausczik and Pennebaker (2010b) also noted, that high-status speakers are expected to use a language with a high level of lexical diversity for example, students of a college, unlike the people with a lower level of education (Bradac et al., 1979).

1.6 Structure of the Document

The remainder of this document is structured as follows. In Chapter 2 we review the existing work about approaches for personality identification and sentiment and emotion analysis. Chapter 3 describes the structure, the functionality parts, and methodologies employed in our platform and also discusses the results and some suggestions about other possible methodologies that could be also used here. Finally, Chapter 4 shows the results achieved so far by the platform and in Chapter 5 we made some conclusions about all the performed work.

Related Work 2

Last years it was shown an increasing interest in personality study and sentiment analysis. There were developed many psychological models for personality prediction and different approaches for sentiment analysis. This chapter overviews the literature that serves as supports for this work. We observe the main approaches for creating a user profile, for performing sentiment analysis and for predicting personality traits. We also report some examples of web applications for personality detection that work by analysing Twitter user profiles.

2.1 Big Five Personality Type Prediction

There are several studies in which was investigated a relation between personality and linguistic cues. Already in the early 1950s, some psychological researchers concluded that the word usage in daily life can reveal us many details about the mental and physical state of people (Argamon et al., 2005). More specifically, the usage of function words is a very important descriptor of a person's psychological state. Function words were defined by Argamon et al. (2005) as the words, that are met very frequently in sentences and have "primarily grammatical function in the language", so due to these facts, people in their speech are not able to control the usage of function words. For example, people with depression tend to use first person singular pronouns, while people in some situations of a great stress such as terror or disaster mostly use first person plural pronouns.

A review by Tausczik and Pennebaker (2010b) explain the way the daily usage of words can characterise a person in terms of thinking styles, attentional focus, emotionality, social relationships amongst others. So the analysis of these linguistical cues allows to predict an approximate state of mental health of people. Results can not be very precise because of some linguistic differences between cultures, genders and ages (Tausczik and Pennebaker, 2010b).

Walter Weintraub, a physician, became very enthralled in studying a relationship between everyday usage of words such as articles and pronounces (Tausczik and Pennebaker, 2010b). To perform this study, he manually annotated and counted words used in various interviews. The conclusion was that not only the content words, that consist of nouns, adjectives, adverbs, regular verbs and others, can tell us about a psychological state, but function words (style words), that consist of conjunctions, articles, pronouns, prepositions, auxiliary verbs etc., should also be considered. Because usage of function words can show us on whom or on what has focused a person, show the level of communication between them, who is dominant in the conversation while content words transmit a meaning of what they are saying. As have been noted in this study, if someone uses the first-person pronoun "I" it can be an indicator of a person being self-focused and hence, probably depressed. There have been noted more interesting things in the study (Tausczik and Pennebaker, 2010b). For example, higher and lower-status individuals can be distinguished by counting in their speech the frequency of usage of words related to themselves and to other people. The pronoun usage can reveal a quality of a relationship between romantic partners for example, frequent usage of a word "we" tells about a higher quality of a relationship, when the usage of pronoun "you" on the contrary predicts a lower quality of the relationship. Low-statused individuals use self-references more frequently than their higher-status colleagues, language of which mostly consists of statements referenced to the other people (Tausczik and Pennebaker, 2010b).

Tausczik and Pennebaker (2010b) noted in their work that in positive ads future and present tenses are used more frequently, while in negative ads addressed to opponents are mostly used past tenses. Assents and positive emotions are mentioned as indicators that can help to identify a level of agreement. There was also concluded that liars use motion and negative words, words that are not much descriptive and they also avoid self-reference and third-person pronouns, while honest people in their speech tend to use negations and exclusive words such as "but", "without", "never","none" amongst others.

A level of a statement complexity reveals the depth of thinking and can be defined by the usage of certain words. Words of cognitive category ("remember", "cause", "ought", "know"), causal category ("because", "due to", "since"), words longer than six characters and prepositions demonstrate a level of complexity of the individual's language. Also, authors pointed out that the usage of the tentative category of words reveals how much a person feels uncertain or insecure in the moment.

In the study of Pennebaker JW (1999) is clear that categories of words used in daily conversations are linked with big-five personality traits. If an individual tends to use many words and most of them are longer than six letters, and at the same time tends to avoid categories such as articles, causal and social words, negations it can be predicted one of the big-five personality traits, named extraversion. There have been noted that the words of extraverted people express more positive emotions than negative ones, while people in depression mostly use first-person singular pronouns and express negative emotions.

As pointed out by Argamon et al. (2005), extraversion is the trait that is more difficult to predict than neuroticism. According to this authors, extraverted people are more focused on words about completeness/incompleteness, certainty/uncertainty, while neurotic people are more concerned with themselves, that is notable by a high number of first-person pronouns in their usual speech.

Our work is most closely related to a recent work elaborated by Mahmud et al. (2014). Authors created an "intelligent information collection system" that generated some questions to obtain the desired information, for example about events visited by the user or about a quality of some goods. The purpose was to choose "the right users at the right time" that are more likely to give the needed information. For performing this, first, the system analyses a twitter stream to choose the tweets that contain the information of interest, then processes tweets of a timeline of authors of the chosen tweets to calculate a type of user's personality. For computing the relevant features for defining a type of personality in this research authors have used Linguistic Inquiry and Word Count LIWC-2001 dictionary. To note, retweets during this process were excluded. Next, were calculated Big Five traits and their facets. After performing that, the system shows a list of recommended users to make questions to. Authors supposed that only users with particular personality traits such as extraversion and friendliness are more likely to respond.

The content of user profiles that is publicly available can provide a lot of information for personality prediction, but some users do not share their personal information. So, to solve this problem, some authors in more recent work (Quercia et al., 2011) have analysed only the information about a number of people that they are *following*, *a number of followers*, and *listed counts*. There have been stated that personality type of people influences the way they interact with others in the real world and also in the virtual world, so the values of metadata variables mentioned above may vary depending of particular personality trait. The authors have studied the relationship between the Big Five personality traits and Twitter user types. There have been distinguished some types of twitter users, such as *the listeners* (people that follow many users), *the popular* (those who are followed by many users), *the highly-read* (users that are saved in reading lists of others) and two types of influential users (Klout: shows retweets and replies on tweets; TIME: ranks public figures) (Quercia et al., 2011). The information from 335 profiles of users have been extracted for performing this study. Authors concluded that popular and influential types of users are extraverts and emotionally stable, moreover, popular users are imaginative, and influential users are organised. It was shown that Openness is easy to predict while Extraversion is more difficult. Moreover, it was attempted to predict the user type without using tweets, relying only on parameters that are publicly available: number of following, followers, and listed counts. There have been achieved correlation values between Big five personality traits and Twitter user types that can be used in future personality studies.

There is also work of Argamon et al. (2005) focused on identifying the level of extraversion and neuroticism analysing text. The authors pointed out that analysis of words usage is able to describe gender, age, feelings, thoughts, and type of a personality of a person. In this study 4 parameters were chosen for the analysis: a "standard function word list", "conjunctive phrases", "modality indicators", "appraisal adjectives" and "modifiers". The work shows that it is better to use "appraisal" for predicting neuroticism, and the "function words" is the best feature for predicting extraversion.

The study of Mairesse and Walker (2006) also have shown that the language of a person can describe a personality. The authors aimed to develop models for automatic personality prediction. This study as many previous ones have been based on the Big Five inventory, because to the opinion of the authors of this work, this model is the best in terms of showing individual differences. There were extracted features using categories of the LIWC utility, "MRC Psycholinguistic database" (Mairesse and Walker, 2006) and also there were tagged sentences from "EAR corpus" with "speech act categories" (Mairesse and Walker, 2006). Authors have reported that prosodic features allow to model extraversion, agreeableness, and openness to experience with the a considerably good precision. LIWC categories are helpful in prediction of extraversion and emotional stability, while MRC perform good in modelling extraversion and conscientiousness. It is clear in this study that extraversion is the easiest trait to predict.

In a research of Hovy and Hovy (2015) have been suggested one more approach that differentiates from the previous ones with a methodology of defining personality types of users. The approach used in this work is named Myers-Briggs Type Indicators (MBTI). The purpose was to find out correlations between "personality traits and demographic and linguistic features" (Hovy and Hovy, 2015). The data used in this research have been collected from 1200 Twitter user profiles. Each of them have been previously annotated by its owner with an MBTI personality type. The authors have made an open-vocabulary data-driven personality research and created a corpus that is constituted of 1,2M tweets with gender and personality annotation. It was reported that the most easily predictive features are gender, and such traits as Introvert – Extrovert and Thinking – Feeling. For performing the identification of user's personality trait, as was pointed out in this research, 100 user tweets are sufficient and having more data can advance the accuracy of results. The metadata was also taken into account for this research. The authors concluded that a number of followers that equals to 100- 500 users, may indicate extraverted personality type. There was also made an interesting conclusion, that the number of tweets that varies between 1000 and 5000 predicts the introverted psychological type and if a user has less than 500 tweets, it means that the user is extravert. Also, it was noted that if the user is in 5-50 lists then it can be an indicator of a user being introverted, while extraverts can be detected in a case of being in less than 5 lists. These findings lead to one more conclusion that for introverted people it is easier and more comfortable to express themselves in the virtual world than during an interaction with people in the real life (Hovy and Hovy, 2015). Using the results of this research can be helpful for health care and some applications, taking into account that there is a little number of annotated data for personality detection and also it is very expensive (Hovy and Hovy, 2015).

Qiu et al. (2012) aimed to measure the Big Five personality traits exploring the association between them and some linguistic features extracted from the tweets. There were employed judges for making an assessment of personality type for owners of twitter profiles. The linguistic analysis for predicting the personality was made using the mentioned above LIWC2007 program. As a result, it was concluded that the Extraversion is strongly linked to the usage of words associated with social processes and the usage of positive emotion words, and at the the same time is negatively correlated with the usage of articles. To the opinion of authors, extraverted people avoid using complex lexical structures. Agreeable people avoid using negations, neurotic ones tend to be focused on themselves, openness is negatively correlated with the swear, affect, and non-fluency words, but is strongly correlated with usage of prepositions.

The approach described in a study of Schwartz et al. (2013) allows to extract personality traits, knowing age, a location and psychological characteristics obtained by analysing publications in social media. The method used in the study was called by Schwartz et al. (2013) the "open-vocabulary" analysis, because the lexicon is based on the words used in user publications, and not on predefined categories of words, as in previous work. Authors state that it helps to make more deep insight into personality tendencies of language usage. In this research have been analysed 15,4 millions of Facebook messages of 75,000 authors (Schwartz et al., 2013). Least squares regression have been used to link categories of words, extracted from publications, with personality and other user characteristics. As was mentioned in this work, the explanatory variables, in this case, were categories of LIWC, while personality traits served as dependent variables. The frequency of usage of a word of each category was calculated by dividing a number of occurrences of a word from a category by the total number of words used by a participant. The coefficient of explanatory variable served as a weight in a linear function that links explanatory and dependent variables. The results of this work proved that the open-vocabulary approach provides more detailed information than other models of research where categories of words are predefined. Also were provided correlation values between age, gender, and personality (Schwartz et al., 2013).

In the study of Yarkoni (2010) were achieved correlation values that link the preferences in words usage with a type of personality of a person. There already existed some similar studies about personality, but unlike previous works, here had been analysed 694 blogs that permitted to make more deep and detailed analysis of a personality. The previous studies were based on writing topics on particular themes, that were chosen by participants. But, according to the authors, it is not clear how much a personality type can influence on a selection of a type of a topic, so the results obtained using this method can not be considered precise enough. So, to obtain more believable results were performed other researches based on more “naturalistic” materials, such as recordings of speech of people. The results of that studies showed a relation between personality and the words used in different situations. But the defect of methods used in previous works is that all of them were based on analysis of speech samples obtained in short period of time, so considering this fact it is not possible to say with certainty if the results of this studies remain stable during a longer time (Yarkoni, 2010). Authors pointed out, that the majority of previous studies only considered the general traits of personality, as the Big Five traits. So, one of the main differences of this study from other ones is that here were studied associations between language cues with not only Big Five traits but also with their low-level facets. In this work was used a questionnaire for obtaining some information like age, gender, personality, and information about participants. To note, according to Yarkoni (2010) the participants were chosen in the following manner: only those bloggers that left the e-mail publicly available were contacted, and only those who had responded were included in this experiment. So, the fact that some types of personality are more likely to be contacted by e-mail and more likely to respond than other ones makes this method of selection and hence, the results of research not so much precise as was desired (Yarkoni, 2010). After the blogs had been selected, there was made a category-based analysis, during which were analysed 66 categories of LIWC - dictionary and revealed strong correlations between Big Five personality traits and frequency of usage of words from different LIWC- categories (Yarkoni, 2010). In this study were also achieved results related to correlations between 30 low-level facets of Big Five and 66 LIWC categories. Many of that results one more time proved the correctness of correlations that were obtained in previous work between the Big Five traits and the word usage. In this study also had been conducted a word-based analysis to reveal a relationship between a personality and word-usage preferences. The authors have concluded that a personality is an important factor that influences either on behaviour of an individual in the virtual world as on behaviour of one in the real world.

In the work of Schwartz et al. (2013) were achieved some interesting discoveries about personality preferences. For example, if there are mentions in user’s publication about some activities and sports interests, this information may be an indicator of emotional stability. Another one notion that helps to understand better a person is about a hobby linked with the Japanese part of the culture like anime, pokemons, mangas, specific type of emoticons, widely used by anime lovers. If a person is interested in such type of things it can be supposed that the type of personality is Introversion. According to this work, males use possessive pronoun “my” with “girlfriend” and “wife” more frequently than females use it with words “boyfriend” and “husband” (Schwartz et al., 2013).

2.2 Sentiment Analysis

Various studies about sentiment classification of english texts are reported on the literature, but only a few can be found for Portuguese text analysis. For that reason, the following review focus on studies related to both languages: Portuguese and English.

There is an interesting work about sentiment analysis of Portuguese text where authors Morgado (2012) describe the process of classification of on-line news sentences. The sentences have been classified into 3 categories such as positive, negative and neutral. The approach described in this work is similar to the one applied in our work. Morgado (2012) also counted numbers of positive, negative and neutral words in the sentence and then calculated a probability of a sentence belonging to one of the 3 categories. But the difference of this approach from ours is that during the process of calculation of this probabilities in this study polarities of the neighbour sentences were also taken into account because, as was stated in this study, the context influences on precision of sentiment classification. Considering the specificity of Twitter publications there was no necessity to do the same thing in our analysis because the absolute majority of tweet examples that we had to analyse does not include more than one short sentence.

Vosoughi et al. (2015) in their study hypothesised that exists a dependency of sentiment polarity from the contextual characteristics such as a geo-location, temporal information, and information about the Twitter user. Two datasets of tweets have been annotated with sentiment polarity for each category of the metadata such as an hour of day, a day of week and other items of metadata. One dataset have been annotated by humans and another one have been annotated by the trained sentiment analyser. Then this annotated datasets have been joined for the training of a sentiment classifier. The Bayesian approach have been used for combining the results obtained by sentiment classifier with metadata characteristics. To note, for annotation of tweets with sentiment polarities authors have used emoticon analysis. In total, there were taken into account six basic and the most used emoticons such as “:)”, “:(’, “:-)”, “:-(“, “:)”, “: (“ (Vosoughi et al., 2015). There was noted that happiness level is usually different in particular temporal periods such as hour of day or day of week, or even month. For instance, on weekends people tend to be happier than during the working days and they usually are depressed when approximates the end of holidays. Also, geolocation can make some influence on mood and psychological states of people. There was made a map of average sentiment distribution over all states in the USA. The results showed that this approach obtains better results than a standard linguistic classifier (Vosoughi et al., 2015).

More recently Roberts et al. (2012) created a manually annotated corpus of tweets. This corpus was annotated with seven emotions: anger, disgust, fear, love, joy, sadness, surprise and love. The authors have chosen only the topics that to their opinion might contain all this emotions. During the process of downloading of tweets hashtags were being used as criteria. Next on the stage of preprocessing the hashtags, punctuation, URLs and similar tweets have been removed. In the process of annotation that consisted of several stages participated several annotators and also was used an annotation tool created by the authors.

2.3 User Analytics Web Sites

Personality information can be useful for various applications. We will consider some applications for personality analysis such as "Analyse words", "Watson Personality Insights for Twitter" and a tool for the sentiment analysis named "Twitómetro".

"Analyse words" (URL://analyzewords.com/) is a web-application that is based on the program LIWC (Linguistic Inquiry and Word Count) (Tausczik and Pennebaker, 2010b), that makes a count of the words that belong to some particular psychological categories. This program was created at the University of Texas at Austin and the Auckland Medical School in New Zealand. It is mostly based on counting the words of such categories of words as prepositions, pronouns, articles and some other groups of words to which was not paid much attention in the previous works based on word counting. The input of application must be a twitter username. As output, the program shows in a web-page with some characteristics of a user that are associated with his emotional, social and thinking styles. For example, the program demonstrates the degree of upbeatness, of depression, anger, arrogance and other psychological states.

"Watson Personality Insights for Twitter" (URLs://personality-insights-livedemo.mybluemix.net/) also makes the analysis of Twitter user timeline when receives an ID of a user. This tool can also make the analysis of some written text produced by a person in whose personality traits information we are interested in. For obtaining correct results the text should contain more than 100 words. The output is a web-page with a brief description of personality portrait of a person and also are shown results about "Big Five" personality characteristics that we have already discussed in previous chapters. The second part of results are characteristics of "Needs" category such as curiosity, harmony, self-expression, stability, closeness that are the visualisation of reasons why a customer will buy a product. And finally, the third part of the shown results is the "Values" category that visualise the motivating factors such as achievement, openness to change, self - transcendence, self - enhancement, and hedonism. To note, the language of tweets must be in Arabic, English, Spanish or Japanese language (Gou et al., 2014).

Recently have been created a tool for sentiment analysis of opinions of Portuguese people relatively politician leaders for the purpose to predict the results of elections (Silva and TEAM, 2011). The analysis of opinions is made every 24 hours for the purpose to update the "index of sentiments" of people for every politician. In this work sentiment classification was based on a manually annotated corpus of tweets that contains 881 tweet. The authors reported about 71% of right sentiment analysis results. This visualisations of daily sentiment statistics are available on the web-page of "Twitómetro" (Silva and TEAM, 2011).

2.4 Main Conclusions

We have analysed various works related to personality detection and sentiment analysis of a text. Exist various psychological models of research of personality, but the most studied and expressive one is the "Big Five" model. The first personality research and sentiment analyses of texts were performed using such examples of texts as essays, questionnaires, interviews amongst others. Nowadays, with the emergence of social networks and blogs, the problem of lack of data had disappeared, but the problem of a personality traits detection had not become easier, because of specific features, such as shortened internet language, use of smiles, hashtags amongst others. We also observed 2 existing web platforms for the twitter user's profile analysis that do not perform analysis of Portuguese texts and one web-application for Portuguese sentiment analysis.

We have observed some works that make a disclosure about the importance of words used by people in daily life for psychology (Gou et al., 2014). It is clear in previous work that analysis of word usage can discover for us many details about the mental and physical state of people. It was also noted that analysis of function words also have an important role in the process of predicting personality because it is impossible to control their usage in statements (Tausczik and Pennebaker, 2010b). Making a linguistic analysis of linguistic footprints it is possible to better understand thinking styles, attentional focus, emotionality, social relationships and even to predict the level of honesty of a person. It can be helpful for prediction a state of mental health of one, or even for prediction some physical illnesses (Tausczik and Pennebaker, 2010b).

We have observed various studies about Big Five personality identification. In some studies, it was pointed out that some traits is simpler to predict than others. For example, it was concluded that extraversion is more difficult to predict than neuroticism.

The work of Mahmud et al. (2014) have been described a method of calculating personality traits using Linguistic Inquiry and Word Count LIWC-2001 dictionary, that we also applied to our system.

In the Quercia et al. (2011) study authors established the relationship between the Big Five personality traits and Twitter user types that can be useful for predicting a user personality type without having to analyse tweets, in case of a user profile being closed to the public. Another work of Argamon et al. (2005) showed how to predict extraversion and neuroticism analysing a text. There have been made a conclusion that analysis of words can help to identify either feelings, thoughts, and personality of a person as a gender and an age.

For making a comparison between the models of personality identification we observed the work of Hovy and Hovy (2015) that described a method named Myers-Briggs Type Indicators (MBTI). In this study the author have made a conclusion that a number of followers that equals to 100 - 500 users, can be a characteristic of extraverts while introverted find social networks more comfortable for expressing themselves, so the number of followers varies from 1000 to 5000. In the work of Qiu et al. (2012) for predicting a Big Five personality type have been used the LIWC2007 program. There was concluded that the extraverted people are more involved in social processes are likely to use positive emotion words and use a little number of articles and also tend to use complex lexical structures. People high in agreeableness avoid using negations, neurotic ones are always self-focused. Also were made more interesting findings such as openness is negatively correlated with the swear, affect, and non-fluency words, but is strongly related to prepositions. In a study of Schwartz et al. (2013) we have been observed an approach for extracting personality traits, knowing age, a location by analysing publications in social media sources In this study authors tried to link the LIWC categories of words with personality and other user characteristics and provided correlation values between age, gender, and personality.

In the work of Yarkoni (2010) have been provided correlational values that link linguistic cues of a person with a type of personality. One of the main differences of this study from other ones is that here besides the Big Five traits, were also considered the low-level facets of Big Five personality traits for exploring the association between them and linguistic cues. In the work of Schwartz et al. (2013) were achieved some interesting discoveries about personal preferences and personality types. As example can be considered people interested in the anime, pokemons, mangas. According to this study the most probable type of personality of such people is Introversion.

We also have observed some previous works about sentiment analysis. Roberts et al. (2012) have created a corpus of tweets annotated with 5 basic emotions and 2 sentiments that we have used in our study for performing the analysis of emotions and sentiments of tweets. In another more recent study authors Vosoughi et al. (2015) considered the metadata, namely geolocation, temporal information, and information about the Twitter user for discovering a dependency of emotional polarity from this contextual characteristics. During the sentiment classification authors had made the analysis of emoticons. There was noted that happiness level is usually different in particular temporal periods. For instance, on weekends people tend to be happier than during the working days and they usually are depressed when approximates the end of holidays. Also, geolocation can make some influence on mood and psychological states of people.

So far we discovered that there are some applications for personality analysis such as “Analyse words” and “Watson Personality Insights for Twitter” and a web - platform for sentiment analysis of Portuguese news named “Twitómetro”. “Analyse words” (URL://analyzewords.com/) is an application based on the program LIWC (Linguistic Inquiry and Word Count) that receives a username and shows a web-page with some psychological characteristics of a user. “Watson Personality Insights for Twitter” also outputs a web-page with a brief description of personality portrait and results about “Big Five” personality characteristics; about consumer needs characteristics, and values characteristics. But this program is not adapted for the processing of Portuguese texts. “Twitómetro” is aimed to show a daily changes of sentiments of Portuguese people in relation of politician leaders. The visualisation of results is available on the web-page of this application and can help to predict the results of elections.

3 Twitter User Profile Analyser

This chapter describes the tools and methodologies of implementation of our system. The system consists of 3 main modules, each of them has a contributive role in the process of forming a general picture of a personality trait of the user. In the following sections, we will describe the process of creating the personality analysis, the sentiment analyser module and also the temporal activity analyser module.

3.1 Personality Analyser Platform

As the "Big five" model of personality characteristics has been researched in many works such as, for example, Benet-Martinez and John (1998); Hovy and Hovy (2015); Hughes et al. (2012); Schwartz et al. (2013); Yarkoni (2010) and many others,. We have inspired and based our work on these models and also on our earlier experiments, previously reported in Jusupova et al. (2016). We selected for analysis a subset of accounts of one thousand of Portuguese users of Twitter. For collecting the Twitter user timelines we had used the Application Programming Interface (API), which allows us to have an access to the Twitter user data having possibility to read and write tweets, look for users, tweets, hashtags and to make different other requests.

To perform the analysis, the system receives the ID of the user as input and then looks for it in the system folder. If the user ID is present in the folder, the system performs the analysis of the user's timeline. Otherwise, it refers to the Twitter API searching for the twitter user timeline, then downloads it into the system and makes the analysis of this timeline. We now have available the content of user timelines produced over the past two years, which can be used for further studies and experiments. for the purpose of identification of the Big Five personality traits, following the methodologies described in some related works. Figure 3.1 shows the overall architecture of our personality analyser platform that we propose in this study.

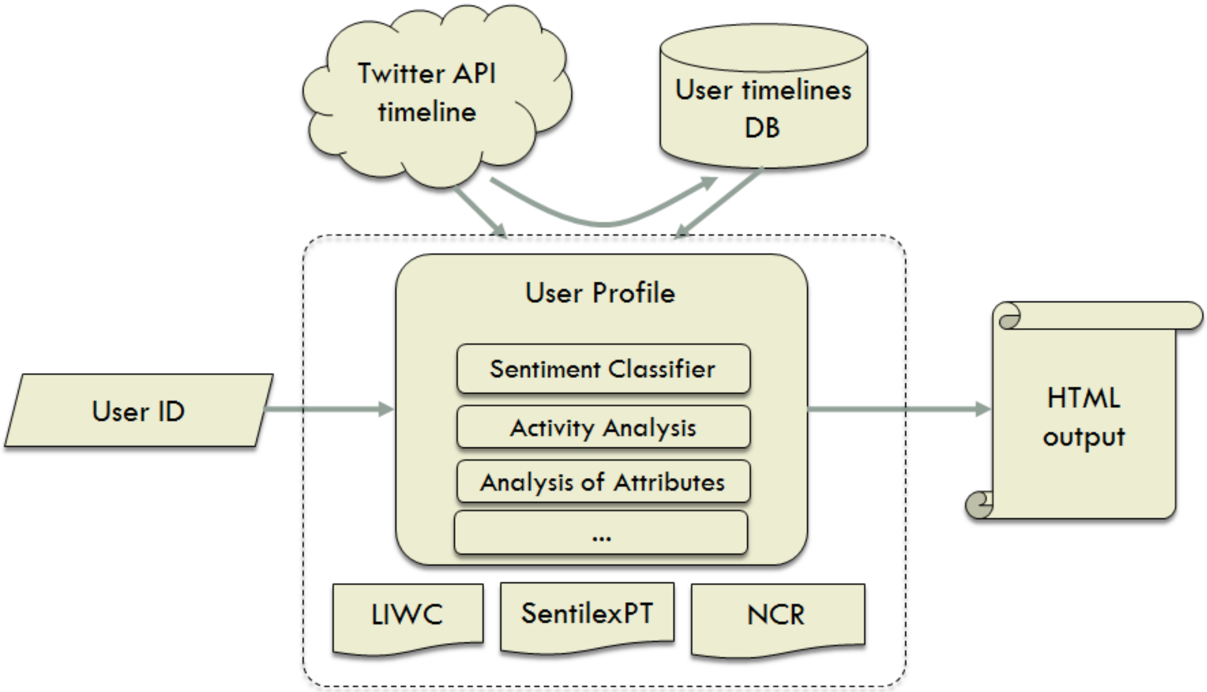


Figure 3.1: Scheme of the Platform.

3.1.1 Technologies Involved

For constructing our web-platform we used a micro-framework named Flask. It provides a core with all basic services and also has extensions that a user can add if it is necessary for performing some specific tasks (Grinberg, 2014). It is based on Werkzeug and Jinja2 libraries. Werkzeug provides Web Server Gateway Interface subsystems, routing and debugging. Template support is provided by Jinja2 (Grinberg, 2014). The reason why we had chosen exactly this framework is that we have an opportunity to pick up the necessary extensions, to change them and even to create ones for our needs, on the contrary of other frameworks. Before installing the Flask it is necessary to install the Python on a computer (Grinberg, 2014).

For downloading the tweets of users, we used the REST API. According to documentation (https://dev.twitter.com/rest/public) it allows to read and write data in the Twitter. For authorised access to Twitter API is used OAuth. The data returned by the REST API has the JSON (JavaScript Object Notation) format. The REST API permits to search the most popular tweets by indices, the tweets made in some particular places, it also can return timeline of a user. As a timeline can be very big, it is possible to apply restrictions for downloading only the desired quantity of tweets.

3.2 Preprocessing Stage

Preprocess is the important and the first stage in the process of sentiment and personality analysis if it is based on linguistic features. It is clear in previous work that stop-words (Saif et al., 2014), punctuation have a negative impact on sentiment analysis performance. In our study to perform the preprocessing stage, we followed several steps described below.

The first step was tokenisation that is a subdividing of strings into lists of substrings. There are 2 types of tokenisation: sentence and word tokenisation. Sentence tokenisation is used for dividing the text into a list of sentences. For performing a sentence tokenisation for one of 17 languages can be used pre-trained tokenisers that are included in "nltk. tokenize.punkt" module.

tokeniser = nltk.data.load('tokenizers/punkt/Portuguese.pickle')

After performing the sentence tokenisation, the following step to perform should be word tokenisation. Word tokenisation is the process of subdividing a sentence into words. For this purpose can be used tokeniser named TreebankWordTokenizer that is also from "nltk.tokenize" module.

Taking into account the fact that the majority of existing tokenisers are topic dependent, i.e. were created basing on newspapers, books and other sources of texts, we were interested to find a tool that could correspond to 2 important requirements:

- It can be applied to social network language processing, i.e is able to deal with hashtags, mentions, URLs and other specific internet features in an appropriate way,
- It could be applied for processing tweets in Portuguese and English, as well.

We find that a regex-based tokeniser elaborated by Krieger and Ahn (2010) better corresponds to these requirements than tokenisers mentioned previously, so we used it for our task.

After finishing the process of tokenisation we have removed some features such as stop-words, punctuation, URL links (tokens started with "http") , hashtags (tokens started with "#") and user mentions (tokens started with "@") to deal only with words we care about. After studying previous work we concluded that TextBlob and Natural Language ToolKit (NLTK) are effective tools for text preprocessing that are simple to use but the TextBlob library is only adapted for English text processing, and NLTK can be adapted for processing text in 17 languages amongst of which is Portuguese.

For the purpose of finding the Portuguese and english stop-words, and punctuation can be used NLTK library lists. Some examples of code are shown below.

stop-words = nltk.corpora.stop-words.words('portuguese'), for Portuguese stop-words = nltk.corpora.stop-words.words('english'), for english.

Before beginning the preprocessing of tweets, we manually analysed some Twitter profiles and concluded that many twitter users tend to make publications in more than one language. So it is clear that our program should use different approaches for Portuguese and english twitter user profile analysis. We concluded that it is crucially important before executing the sentiment and personality identification to make a language identification of tweets. We have included a module that uses NLTK's sets of stop-words from different languages and, basically, counts how many times are found stop words from every set. This module returns the language with the higher score. Score in this case is a number of found stop-words that belong to a particular language. After performing all these preprocessing steps it is already possible to begin the process of extracting the necessary features for sentiment analysis and personality identification. The detailed description of this process is provided below.

3.3 Extracted Features

Twitter users express their personality in various manners. There are different types of information that should be analysed for effective personality prediction. It is important to take into account not only what users share but also the ways how the information is shared. For example, it can be useful to analyse such statistics as a number of followers and followees, hashtags, a tendency of sharing links, a number of tweets, retweets amongst others (Maruf et al., 2014). The design of our study, in order to characterise the user, is based on a calculation of a set of features, ranging from features based on sentiment analysis to personality detection. The remainder of this section presents the details of the feature extraction process.

3.3.1 Hashtags

Hashtag represents a keyword that starts with a symbol "#". It serves for marking and grouping messages about the same issue for easier search. It provides messages with some short significant information, i.e context, about the content of messages and as was stated in the study of Maruf et al. (2014) this type of metadata makes easier a personal expression. Although a meaning of a hashtag may carry some emotions, sentiments, interests, the things a user is focused on, ads of brands or other significant information (Maruf et al., 2014), it has not attracted much attention of personality researchers in previous studies. For example, as was reported in a work of Maruf et al. (2014), the hashtags related to anger, negative emotions, affective processes such as "#afraid", "#violence", "#war", "#irritated" are positively correlated to Neuroticism. Extraverted users are more likely to use hashtags that reflect positive emotions, hashtags that are related to social processes such as “#love”, “#friends”, “#walking”, “#gonna” amongst others(Maruf et al., 2014). In the present study, we have extracted the most used hashtags in the users content to make a generalised personality description.

3.3.2 Lexical Complexity

As was mentioned in the previous chapter, lexical complexity consists of various characteristics. We have calculated 3 main values such as a lexical diversity, an average number of words per tweet and an average length of words (number of characters per word). The value of lexical diversity shows the percentage of unique information provided during a period of time. It is calculated by dividing the number of unique words, i.e. different lexical words, by the total number of the words from all user tweets (Vaezi and Kafshgar, 2012).

3.3.3 Sentiment Analysis

Tweet sentiment classification is the popular area of research, because analysing the polarity of tweets helps to make insight into the character, behaviour, attitudes and make a general picture of personality. Sentiment classification, according to literature, can be performed in two ways:

- using machine learning,
- perform lexicon-based approach.

Machine learning techniques are used for the purpose of making automatic prediction based on previously extracted patterns. Exist various algorithms of machine learning such as Naive Bayes, random forests, decision trees, support vector machines amongst others that are already included in some libraries of Python.

In this work was decided to use 2 methods for performing the sentiment analysis: the lexicon-based approach for Portuguese tweets and for analysis of english tweets we used the TextBlob tools.

We classify each tweet into the positive, negative or neutral category. The detailed description of the process of the analysis of tweets is described below.

In this study, to perform sentiment analysis for Portuguese we have chosen a lexicon-based approach in which we have used two lexicons the SentiLex-PT, the NRC Emotion Lexicon and emoticons found in tweets.

SentiLex-PT is a lexicon that contains different attributes such as polarity, polarity target, polarity annotation amongst others. We have used a score of each word from a lexicon found in tweets for calculating the probability of each tweet being positive, negative or neutral. NRC Emotion Lexicon is a set of English words translated with Google Translator into over twenty languages that contains emotion characteristics for every word, such as anger, fear, anticipation, trust, surprise, sadness, joy, and disgust, and two sentiments, positive and negative. To note, a word can belong to several categories at the same time.

We have compared each word in all tweets with words from both lexicons. Below we describe the full process of identifying a sentiment polarity of tweets.

If the program finds a word from a tweet in the dictionary of Sentilex-flex-PT02 it extracts the polarity of this word and adds it to a tweet score variable. The value of a polarity can be equal to -1, to 0, or to 1.The same process is performed using the dictionary of NRC Emotion Lexicon, which contains the unique value for all words that is equal to 1 in case of a word belonging to a positive or negative category. The score is added to another tweet score variable. This process is repeated for each word in a tweet. In case of the program finding only one word from a tweet in lexicons then the polarity of this word is assigned to this tweet. If in a tweet was not found any word from a lexicon, that means that the program will classify a tweet polarity as neutral. To note, the misspellings for example, "LOOOL" or "Feliiiiz", in this work were not considered. Then we look for a negation that is located before a punctuation sign. If a program finds a negation, the values of the both score variables change the sign to a contrary one. For example, if the value of a variable "score" is negative, hence in case of presence of one negation before a punctuation sign it is changed to positive. Finally, we sum both of scores and verify the resulting score if it is positive, negative or equals to zero. If the result is positive(or more than zero), then we increment positive sentiment tweets counter. If it is negative, then the negative sentiment tweets counter is incremented. If equals to zero, it means that the program can not find any word in both dictionaries. In this case, the program starts looking for emoticons that are specific features constituted of various sets of characters that provide easily identifiable sentiments. If the sentiment analyser module finds an emoticon from positive emoticons category or from negative one, then the counters of positive or negative emotion are incremented. If the program finds neither words from dictionaries, mentioned above, nor emoticons then neutral sentiment tweets counter is incremented.

Sentiment classification of tweets is one of the most important parts of our application. Considering the fact that the NLTK sentiment classification tools are the most used in previous work, we became curious about experimenting with the functionality of the TextBlob library (Python 2 and 3) that is aimed for sentiment classification, a tool for performing different types of processing of English texts, namely, sentiment analysis, part-of-speech tagging, words inflection and lemmatisation, spelling correction, parsing, amongst others.

Many approaches for the sentiment analysis can only be effective in a case of adequate matching between the training and test data: most of the classifiers are topic-dependent, domain-dependent, and even temporally dependent (Read, 2005). According to literature, this problem can be overcome by using emoticons as additional features. We also hypothesise that emoticons can contribute improving the results of sentiment classifiers. We created a program module that is able to extract some of the most popular emoticons and calculate a frequency with what it is used in all user's publications. Emoticons are subdivided into two following categories: positive and negative.

3.3.4 Emotion detection

The term emotion was defined by Scherer (2000) as "relatively brief episode of response to the evaluation of an external or internal event as being of major significance". According to Farnadi et al. (2014), at least one type of emotion can be extracted from a tweet. As was stated in the study of Farnadi et al. (2014) the emotion and sentiment expression cues of every personality trait are different. According to this study, users high in Neuroticism are less emotional than opened and extraverted ones. It was also demonstrated that these statistics are similar for users high in Conscientiousness and Agreeableness. There was also noted that the tweets of extraverted, conscientious and agreeable users contain anticipation words, but such emotions as disgust, sadness, and other negative sentiments are more likely to be expressed by neurotic users. People opened to experience frequently speak about their fears and anger (Farnadi et al., 2014).

Inspired by the previously mentioned study of Farnadi et al. (2014), we performed the emotion extraction from tweet publications, using a high-quality NRC word-emotion lexicon that contains a list of words translated from English into many languages. This lexicon consists of eight categories of emotions such as joy, surprise, sadness, anger, fear, disgust, anticipation, trust and two types of sentiments: positive and negative. A purpose of this analysis is providing an additional information about the frequency of emotions calculated for all tweets of a particular user. Despite the fact that the annotations of emotions and sentiments in this lexicon were made for English language and then translated to Portuguese and other languages, the author of a lexicon Mohammad and Turney (2013) states that "a majority of affective norms are stable across languages", so we also expect that the quality of the results for all languages remain relatively similar to english ones.

In our study for twitter user profiles with tweets made in English and in Portuguese we used english and Portuguese versions of NRC-lexicon respectively. The diagram represented

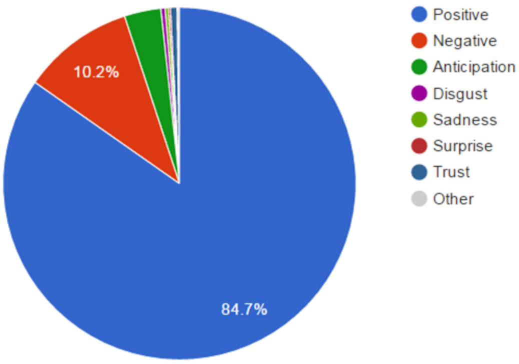


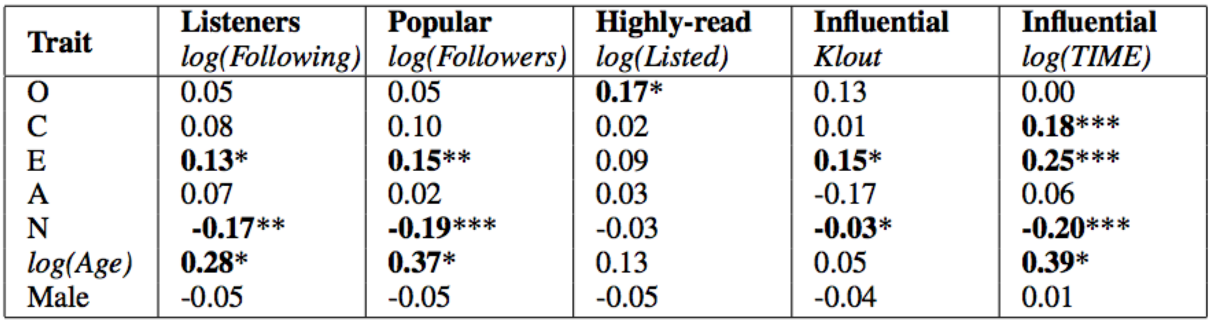
Figure 3.2: Emotion expression statistics for all tweets (used only NRC Lexicon).

below demonstrates the frequency of emotion and sentiment expression in tweets by a user over 2 years. Analysing this information we can achieve a more deep vision of a user personality.

3.3.5 Most used Parts of Speech

As we were pursuing the purpose of better understanding of the personality of a user, we tried to study this problem from various sides. Oberlander and Gill (2004) supposed that lexical information may be more valuable for analysing linguistic style than content. Hence, POS tag information may be more important for detecting some of the personality traits, than information about the usage of words. The authors pointed out that in oral language high extraverts tend to use more adverbs, pronouns, verbs, while the usage of nouns, adjectives and prepositions is very low. High neurotics in their speech prefer to use pronouns and conjunctions, low neurotics use more adjectives and nouns. So we decided to implement this method as the complementary one for obtaining an additional characteristic related for a personality. We explain a method of part-of-speech tagging that we used in this work below.

For POS tagging a text in English and Portuguese there is a good library that provides necessary tools, named NLTK. But to obtain POS tags for Portuguese text it was necessary to train POS tagging model. NLTK library contains NgramTagger that allows using various training data. We have been used two available tagged corpora, namely “Floresta” and “Mac_morpho”. For the purpose of obtaining more precise results, we have used a combination of DefaultTagger, UnigramTagger, BigramTagger and TrigramTagger. The UnigramTagger predicts the most frequent tag for every token, the BigramTagger takes into account a given word and a previous word for getting tag for a given word. And, finally, the TrigramTagger considers the two previous words for getting a tag for the test word. As was pointed out in literature, the trigram tagger has a less coverage and hence the precision of results is not high, unlike the unigram tagger. For obtaining more accurate results, we followed the approach represented in previous work, that



Trait	Listeners *log(Following)*	Popular *log(Followers)*	Highly-read *log(Listed)*	Influential *Klout*	Influential *log(TIME)*
O	0.05	0.05	**0.17***	0.13	0.00
C	0.08	0.10	0.02	0.01	**0.18*****
E	**0.13***	**0.15****	0.09	**0.15***	**0.25*****
A	0.07	0.02	0.03	-0.17	0.06
N	**-0.17****	**-0.19*****	-0.03	**-0.03***	**-0.20*****
log(Age)	**0.28***	**0.37***	0.13	0.05	**0.39***
Male	-0.05	-0.05	-0.05	-0.04	0.01

Figure 3.3: Correlation coefficients between Big Five personality traits and five quantities that characterise listeners, popular users, highly read users, and (klout & TIME) influentials. Statistically significant correlations are in bold and their p-values are expressed with *s: p< 0.001 (***), p< 0.01 (**) and p< 0.05(*). This table was extracted from (Quercia et al., 2011).

combines all these taggers, namely the unigram, the bigram and the trigram taggers. In other words, as was explained in the work of Perkins (2010), a new tagger looks for a most frequent tag for a context and apply it to that context. If it is impossible to find and assign any tag it applies a back-off tagger, that applies "NN" tag by default to any context.

To make a POS tagging for english tweets we also have used the library NLTK, that contains already trained tagger. For obtaining the desired result was used an embedded function "nltk.pos_tag (text)".

3.3.6 Metadata: Following and Friends

As it has already been mentioned, there have been discovered some important findings in the previous work related to twitter user types and personality traits such as for example, the popular users may be characterised by the trait Openness, influential ones are high on Conscientiousness amongst others (Quercia et al., 2011). In our study we have chosen such user information as following counts and a number of the user friends as an additional personality characteristic. In case of a user have a closed profile page, these parameters turn out to be indispensable for predicting the personality traits. According to Quercia et al. (2011) the personality prediction can be performed knowing the publicly available counts of followers, following, listed counts and the correlation values between them and Big five personality traits that were provided by Quercia et al. (2011).

3.3.7 Other Twitter's Metadata for Personality Detection

According to Vosoughi et al. (2015), metadata such as temporal information, geolocation and information about the author can help to get a picture of personality traits and a style of life of a person. For example, as was reported by Farnadi et al. (2014), people tend to be more emotional and negative during the days of work and study, while during weekends the intensity of emotion expression decreases and publications become more positive. Concerning months of the year and sentiment expression, there was also found a relationship between this 2 variables. For example, users tend to be more positive on December and during summer vacations the publications almost do not carry any emotions. To note, these observations were made by Farnadi et al. (2014) for english Facebook users. We made the similar analysis for Portuguese Twitter users for the purpose to show this information additionally to the other characteristics.

3.4 Temporal Analysis and Visualisation

After the analysis of individual tweets and extracting overall information, we have created graphical representations for the user activity based on the amount of tweets produced in different periods of time. We have also created visualisations of the sentiment distribution over time, thus revealing the sentiment changes hidden in global statistics. The remainder of this section presents further details about this.

3.4.1 Temporal Activity

We extracted temporal features from time stamps of every tweet and then calculated the frequency of "tweeting" in particular periods of time such as months, days and hours. We assume that the visualisation of activity of a user can be also contributive for creating the personality image.

In the Figure 3.4 it is notable that the user has a tendency being more active in the time interval beginning from 15 p.m to 21 p.m, while beginning from 21 p.m. it begins to decrease. We made such analysis for more users and made a try to establish an association between personality type and activity variations.

In the Figure 3.5 is shown the statistics about user activity by day of week. It can be made an assumption that on such days of week as Tuesday, Wednesday and Thursday the user is more occupied with work or study, while on Friday, Saturday, Sunday and Monday a user is more relaxed and have more time for making publications in networks.

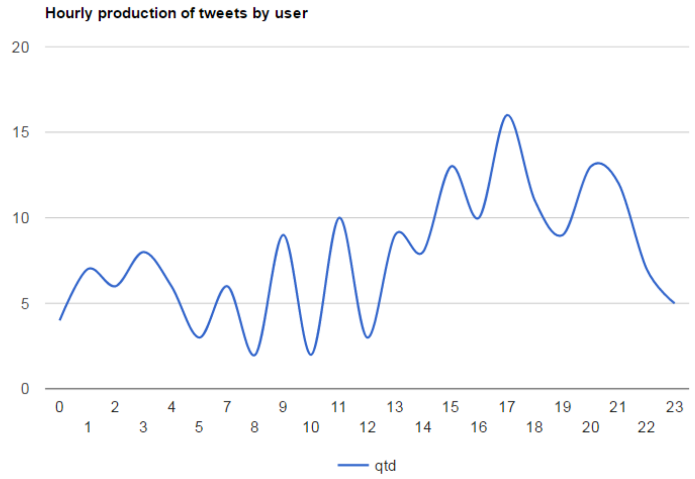


Figure 3.4: Twitting activity per hour.

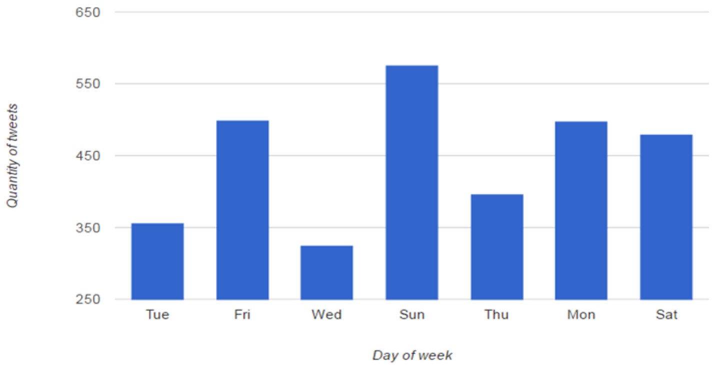


Figure 3.5: Twitting activity per day.

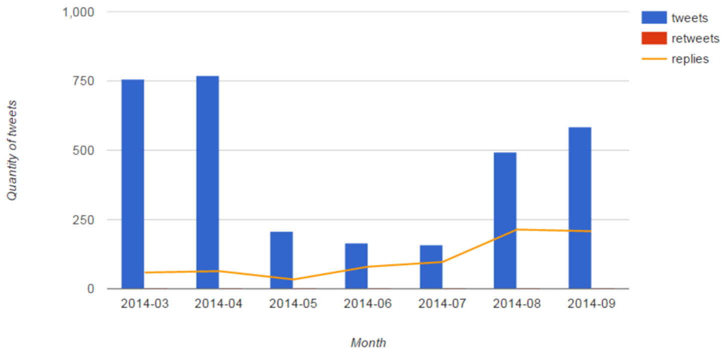


Figure 3.6: Twitting activity per month.

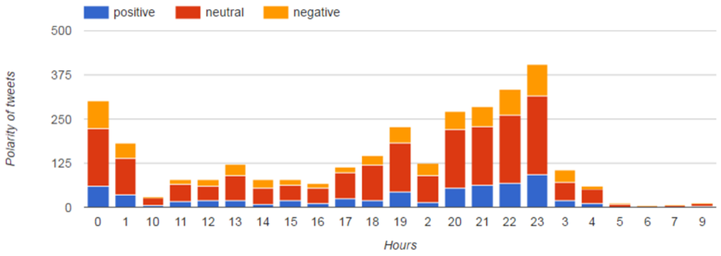


Figure 3.7: Sentiment polarity of tweets per hour.

In the Figure 3.6 the Twitter user activity statistics shows that the maximum number of tweets over two years was made in March, April, while the minimum activity was registered during June and July that proves one more time the assumption made by Farnadi et al. (2014) about the minimal activity in Summer. Also, in the Figure are shown the statistics about replies number of which increases with every month.

3.4.2 Analysis of Sentiments over the Time

As was hypothesised by Vosoughi et al. (2015), different time of twitting, different locations can influence on a mood of the author, hence the polarity of tweets may differ. For the purpose of understanding of the relationship between sentiments expressed in a particular period of time and personality traits, we made the sentiment analysis of tweets produced by users in different time periods. The visualisation of some statistics is shown in the pictures 3.7, 3.9, 3.8 represented below.

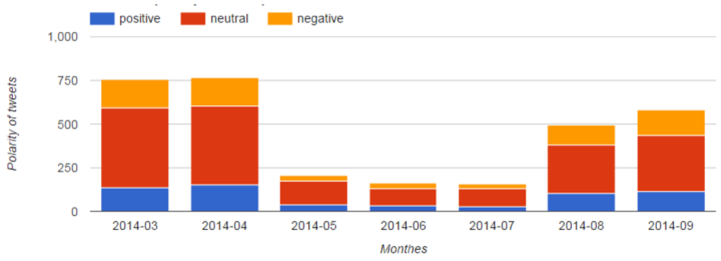


Figure 3.8: Sentiment polarity of tweets per month.

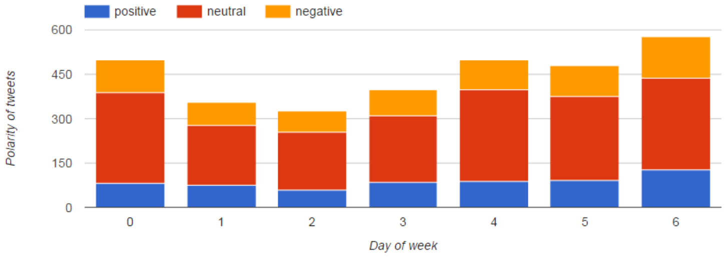


Figure 3.9: Sentiment polarity of tweets per week.

3.5 Personality Prediction

We have used the lexicon-based method of identification of a personality type that allows to predict a type of a user personality by analysing the words that one uses in tweets. We have used a “Brazilian Portuguese LIWC 2007 Dictionary” (http://143.107.183.175:21380/portlex/index.php/en/liwc, link accessed on 26.01.2016), created for the Portuguese language. This version of the dictionary contains 64 categories. We used almost all of them except categories that aren’t important for our work for example, “assent”, “nonfluencies”, “fillers” to analyse and provide a complete image of a user. Similarly to the previous studies that were mentioned above we calculated scores for every psychological category and at the next step used correlation coefficients obtained from the study of Yarkoni (2010) to combine each of the the Big Five personality traits with the obtained scores of categories. And finally, as the result was chosen a trait with the highest value. There are some previous studies in which that have been found associations between the word usage and the personality traits.

The method that we have used in is closest to some methods elaborated by authors Tausczik and Pennebaker (2010b), Yarkoni (2010) and Mahmud et al. (2014) more recently. Tausczik and Pennebaker (2010b) created a program Linguistic Inquiry and Word Count (LIWC) that consists of a dictionary of english words subdivided into different psychological categories, and a program that performs analysis of a text. The program compares every word of the text with words of the dictionary and when finds, increments scores of categories that include that word.

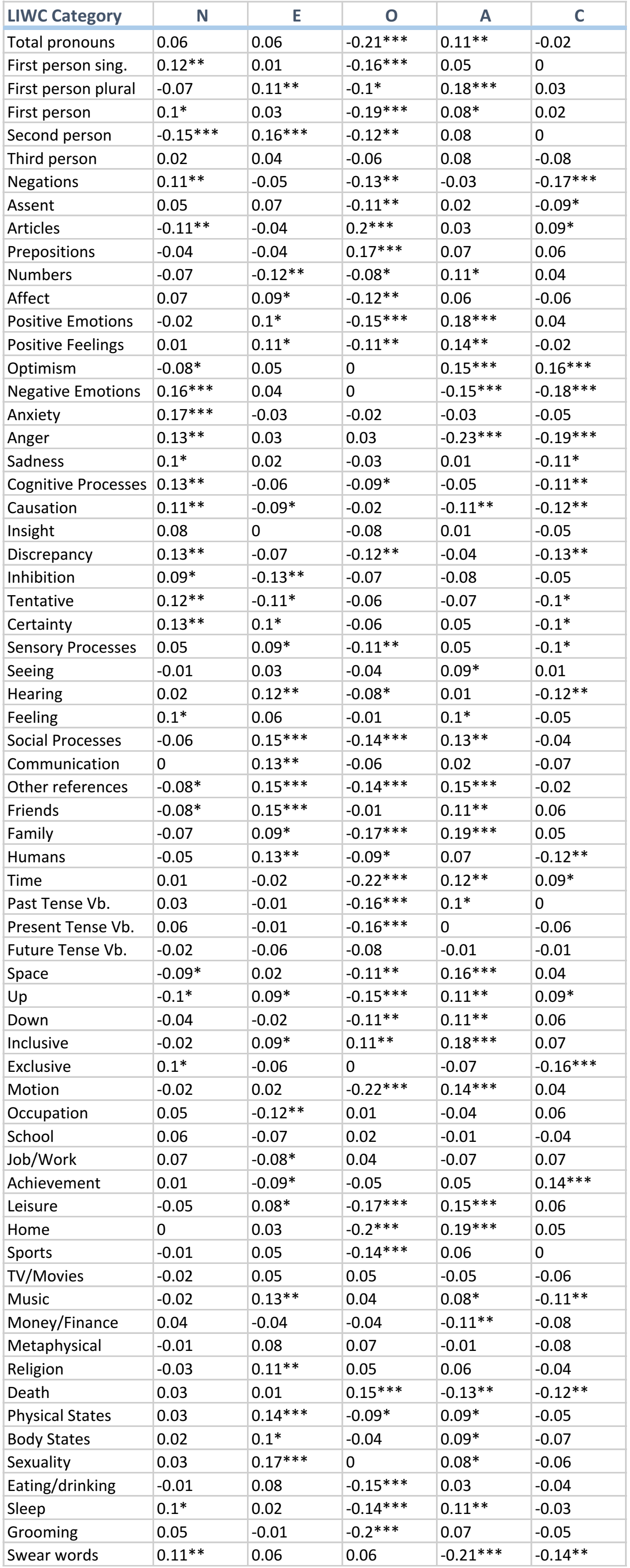
It can be one or more than one categories in which can be found the same word. At the last step is calculated a percentage of different categories of words found in the text. It was calculated by dividing the scores of all words of a particular category by a quantity of words used by the user in all publications. The LIWC is a program that is used frequently for defining a relationship between word use and psychological variables. The LIWC dictionary contains more than 80 categories (Tausczik and Pennebaker, 2010b). In the work of Yarkoni (2010) found correlations between LIWC categories and Big 5 personality traits and their lower-level facets that are shown in the Figure 3.10.

It was noted that neuroticism strongly correlates with the usage of the words associated with the following categories: fear, anxiety, sadness, anger and other words that express negative emotions. Extraversion, unlike the neuroticism, characterised by the usage of the words, included in such categories as positive emotions, friends, social processes, sexuality and others. Concerning such trait as Agreeableness, it is characterised by such categories of words as positive emotions, friends, family and frequent usage of first person plural pronounces that means a kind of inclination toward a"social communality". Openness is related to a high frequency of usage of articles and prepositions. Mahmud et al. (2014) in their study also relied on this correlation table and used the values as weights for the linear combination of Big 5 traits and their facets with the categories of LIWC.

3.6 Conclusions

So far, after studying many literature sources, we had discovered some approaches for making processing of Portuguese users tweets that led us to understand a way of extracting characteristics that help us to obtain a psychological image of a person, without making one to fill questionnaires or another type of work for obtaining results. More particularly, we applied an approach, called in literature "closed-vocabulary analyses", using the LIWC - dictionary, for making a preliminary identification of a user profile personality type. The sentiment analysis was performed using the "Sentilex-PT" lexicon together with the "NRC-PT" lexicon, created previously by other authors. Despite these tools were used by other authors for various tasks, it is good to remember that lexicons and dictionaries are often limited to a specific domain, so it is not always possible to catch the context of tweets to make a more precise sentiment and personality analysis.

The first main difference of our platform from already existing ones is that the unique thing that a user has to do for getting any result about a personality type is entering a user ID. The second difference concerns the analysis that is being performed for Portuguese twitter users' timelines because, as mentioned earlier in this work, the absolute majority of sentiment and



LIWC Category	N	E	O	A	C
Total pronouns	0.06	0.06	-0.21***	0.11**	-0.02
First person sing.	0.12**	0.01	-0.16***	0.05	0
First person plural	-0.07	0.11**	-0.1*	0.18***	0.03
First person	0.1*	0.03	-0.19***	0.08*	0.02
Second person	-0.15***	0.16***	-0.12**	0.08	0
Third person	0.02	0.04	-0.06	0.08	-0.08
Negations	0.11**	-0.05	-0.13**	-0.03	-0.17***
Assent	0.05	0.07	-0.11**	0.02	-0.09*
Articles	-0.11**	-0.04	0.2***	0.03	0.09*
Prepositions	-0.04	-0.04	0.17***	0.07	0.06
Numbers	-0.07	-0.12**	-0.08*	0.11*	0.04
Affect	0.07	0.09*	-0.12**	0.06	-0.06
Positive Emotions	-0.02	0.1*	-0.15***	0.18***	0.04
Positive Feelings	0.01	0.11*	-0.11**	0.14**	-0.02
Optimism	-0.08*	0.05	0	0.15***	0.16***
Negative Emotions	0.16***	0.04	0	-0.15***	-0.18***
Anxiety	0.17***	-0.03	-0.02	-0.03	-0.05
Anger	0.13**	0.03	0.03	-0.23***	-0.19***
Sadness	0.1*	0.02	-0.03	0.01	-0.11*
Cognitive Processes	0.13**	-0.06	-0.09*	-0.05	-0.11**
Causation	0.11**	-0.09*	-0.02	-0.11**	-0.12**
Insight	0.08	0	-0.08	0.01	-0.05
Discrepancy	0.13**	-0.07	-0.12**	-0.04	-0.13**
Inhibition	0.09*	-0.13**	-0.07	-0.08	-0.05
Tentative	0.12**	-0.11*	-0.06	-0.07	-0.1*
Certainty	0.13**	0.1*	-0.06	0.05	-0.1*
Sensory Processes	0.05	0.09*	-0.11**	0.05	-0.1*
Seeing	-0.01	0.03	-0.04	0.09*	0.01
Hearing	0.02	0.12**	-0.08*	0.01	-0.12**
Feeling	0.1*	0.06	-0.01	0.1*	-0.05
Social Processes	-0.06	0.15***	-0.14***	0.13**	-0.04
Communication	0	0.13**	-0.06	0.02	-0.07
Other references	-0.08*	0.15***	-0.14***	0.15***	-0.02
Friends	-0.08*	0.15***	-0.01	0.11**	0.06
Family	-0.07	0.09*	-0.17***	0.19***	0.05
Humans	-0.05	0.13**	-0.09*	0.07	-0.12**
Time	0.01	-0.02	-0.22***	0.12**	0.09*
Past Tense Vb.	0.03	-0.01	-0.16***	0.1*	0
Present Tense Vb.	0.06	-0.01	-0.16***	0	-0.06
Future Tense Vb.	-0.02	-0.06	-0.08	-0.01	-0.01
Space	-0.09*	0.02	-0.11**	0.16***	0.04
Up	-0.1*	0.09*	-0.15***	0.11**	0.09*
Down	-0.04	-0.02	-0.11**	0.11**	0.06
Inclusive	-0.02	0.09*	0.11**	0.18***	0.07
Exclusive	0.1*	-0.06	0	-0.07	-0.16***
Motion	-0.02	0.02	-0.22***	0.14***	0.04
Occupation	0.05	-0.12**	0.01	-0.04	0.06
School	0.06	-0.07	0.02	-0.01	-0.04
Job/Work	0.07	-0.08*	0.04	-0.07	0.07
Achievement	0.01	-0.09*	-0.05	0.05	0.14***
Leisure	-0.05	0.08*	-0.17***	0.15***	0.06
Home	0	0.03	-0.2***	0.19***	0.05
Sports	-0.01	0.05	-0.14***	0.06	0
TV/Movies	-0.02	0.05	0.05	-0.05	-0.06
Music	-0.02	0.13**	0.04	0.08*	-0.11**
Money/Finance	0.04	-0.04	-0.04	-0.11**	-0.08
Metaphysical	-0.01	0.08	0.07	-0.01	-0.08
Religion	-0.03	0.11**	0.05	0.06	-0.04
Death	0.03	0.01	0.15***	-0.13**	-0.12**
Physical States	0.03	0.14***	-0.09*	0.09*	-0.05
Body States	0.02	0.1*	-0.04	0.09*	-0.07
Sexuality	0.03	0.17***	0	0.08*	-0.06
Eating/drinking	-0.01	0.08	-0.15***	0.03	-0.04
Sleep	0.1*	0.02	-0.14***	0.11**	-0.03
Grooming	0.05	-0.01	-0.2***	0.07	-0.05
Swear words	0.11**	0.06	0.06	-0.21***	-0.14**

Figure 3.10: Correlations between Big Five personality traits and LIWC categories. Extracted from (Yarkoni, 2010).

personality analysis is possible to do for english Twitter user profiles and others, but not for the Portuguese ones.

This research can help to find more easily solutions of some problems. For example, knowing what type of personality a person belongs to, can help to understand better the needs of that person and also, probably, to predict health related problems with the purpose of preventing them in the future. The work performed here was based on the text analysis, for the purpose to obtain results about personality type. That means that the user profiles of which are not publicly available can not get the results that this platform is willing to show to the users with opened profiles. So we see that it could be a good idea to improve the coverage of analysis, making it for all types of user profiles using the information about numbers of followers, following and listed counts, as well. The obtained results about opened user profiles were tested in an informal way by different people, so we could not assess its performance in a formal way. In the future we are planning to perform an extensive analysis of the results, using more sophisticated approaches. We recently have discovered a publicly available corpus, created by Verhoeven et al. (2016), that contains data about some users and their results about MBTI personality type. It was developed for research in author profiling. The results are available for profiles in different languages amongst of which is Portuguese. But it is not clear yet for us how MBTI type could be associated with the Big Five personality traits. Knowing that, it would be possible to make some tests of user profiles, results of which already exist in the corpus, by comparing with the results obtained by our platform.

Analysing all characteristics including the user activity and sentiment polarity during different periods of time, a number of unique words, the average length of words and other statistics for the purpose to amend inaccuracies of results about a personality type can drastically improve the results.

4 Manual Evaluation

This chapter describes the results of a manual evaluation of user profile statistics extracted from web-pages generated by our personality analysing platform. We provide a table with results for all user profiles that have been chosen randomly, and then make the generalised analysis for every personality trait. As names of users and user IDs were left as private information, so we mention all users by different numbers. The next step of our evaluation consists of analysis of diagrams that visualise temporal statistics, that have been described below, and language usage statistics.

4.1 Evaluation of Tweets Sentiment Analysis Results

We have performed a manual evaluation of results of sentiment analysis applied to tweets. To do this, we have chosen 10 tweets from each user and manually validated the sentiments assigned to each tweet. We provide below the sentiment analysis results for some tweets extracted from randomly chosen user profile.

☹ negative tweet ⇒ parem de provocar uns e outros

○ neutral tweet ⇒ seguir em frente

☺ positive tweet ⇒ RT @instagranzin: loiras: todo mundo sabe que elas são perfeitas

○ neutral tweet ⇒ @pinkiieb awwwwww obrigadaaaa s2

☺ positive tweet ⇒ RT @pinkiieb: @RitaTeixeira0 está mesmo fixe o teu avatar

☺ positive tweet ⇒ ai que queridos que vocês estão !

○ neutral tweet ⇒ @valha_meDeus ahahhahahah está bem xd

☹ negative tweet ⇒ @valha_meDeus isto hoje ta muito atrasado xd

☺ positive tweet ⇒ vou jantar

☹ ☹ negative tweet ⇒ tou sem imaginação, só dou rt

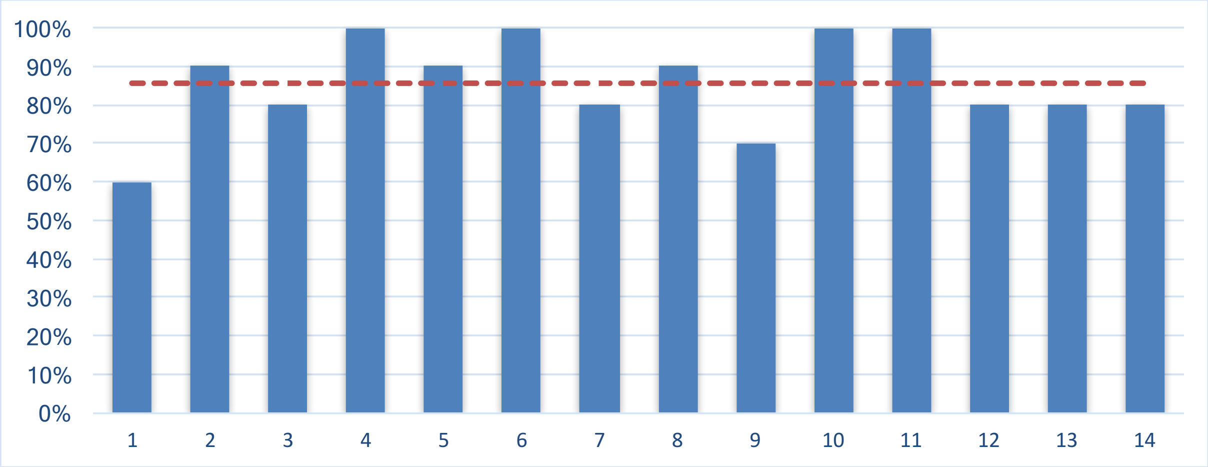
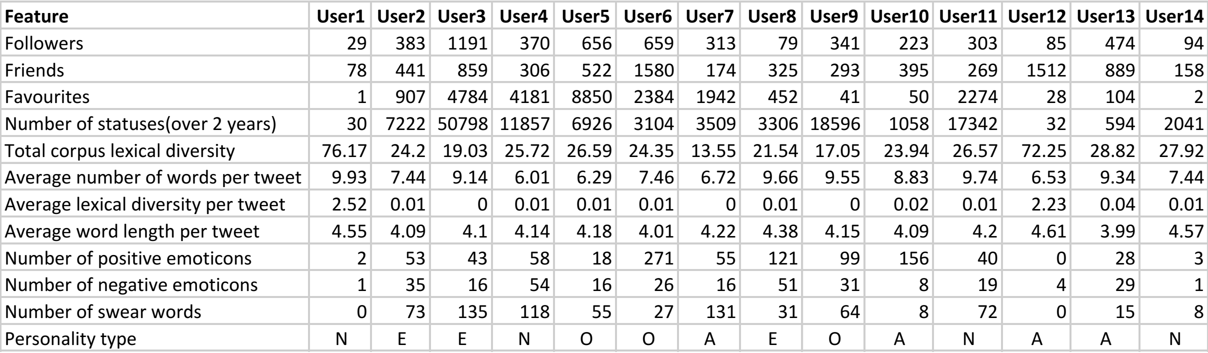


Table 4.1: Manual assessment of the Sentiment Classification for 10 tweets of each user.



Feature	User1	User2	User3	User4	User5	User6	User7	User8	User9	User10	User11	User12	User13	User14
Followers	29	383	1191	370	656	659	313	79	341	223	303	85	474	94
Friends	78	441	859	306	522	1580	174	325	293	395	269	1512	889	158
Favourites	1	907	4784	4181	8850	2384	1942	452	41	50	2274	28	104	2
Number of statuses(over 2 years)	30	7222	50798	11857	6926	3104	3509	3306	18596	1058	17342	32	594	2041
Total corpus lexical diversity	76.17	24.2	19.03	25.72	26.59	24.35	13.55	21.54	17.05	23.94	26.57	72.25	28.82	27.92
Average number of words per tweet	9.93	7.44	9.14	6.01	6.29	7.46	6.72	9.66	9.55	8.83	9.74	6.53	9.34	7.44
Average lexical diversity per tweet	2.52	0.01	0	0.01	0.01	0.01	0	0.01	0	0.02	0.01	2.23	0.04	0.01
Average word length per tweet	4.55	4.09	4.1	4.14	4.18	4.01	4.22	4.38	4.15	4.09	4.2	4.61	3.99	4.57
Number of positive emoticons	2	53	43	58	18	271	55	121	99	156	40	0	28	3
Number of negative emoticons	1	35	16	54	16	26	16	51	31	8	19	4	29	1
Number of swear words	0	73	135	118	55	27	131	31	64	8	72	0	15	8
Personality type	N	E	E	N	O	O	A	E	O	A	N	A	A	N

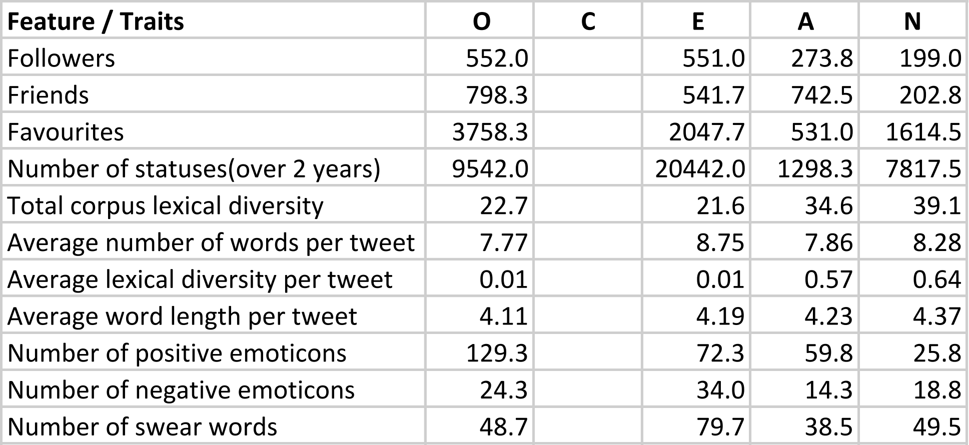
Table 4.2: Manual evaluation results per user.

Table 4.1 shows the percentage of the correct results for each user, together with the average, which corresponds to about 86% accuracy.

4.2 Individual Analysis of Results for each Trait

There were selected 14 Twitter user profiles and then subjected to the personality analysis. We manually verified some results produced by our Personality Analysis Platform that are represented in the Table 4.2 that shows individual statistics for each user. The program had produced a web-page with different characteristics for every selected user. We have extracted some of the results and performed an analysis of them within each of traits of personality for the purpose of estimating relations between personality type and other characteristics such as Twitter user type, user activity and linguistic characteristics provided by our platform.

According to the obtained results, 3 from 14 users are characterised by the trait Openness, 3 users are characterised by Extraversion, another 4 users are characterised by Agreeableness,



Feature / Traits	O	C	E	A	N
Followers	552.0		551.0	273.8	199.0
Friends	798.3		541.7	742.5	202.8
Favourites	3758.3		2047.7	531.0	1614.5
Number of statuses(over 2 years)	9542.0		20442.0	1298.3	7817.5
Total corpus lexical diversity	22.7		21.6	34.6	39.1
Average number of words per tweet	7.77		8.75	7.86	8.28
Average lexical diversity per tweet	0.01		0.01	0.57	0.64
Average word length per tweet	4.11		4.19	4.23	4.37
Number of positive emoticons	129.3		72.3	59.8	25.8
Number of negative emoticons	24.3		34.0	14.3	18.8
Number of swear words	48.7		79.7	38.5	49.5

Table 4.3: Manual evaluation results per user.

and 4 by Neuroticism. Users with the trait Conscientiousness were not identified. Below we provide an analysis of statistics for each personality trait that we have mentioned above.

We started our analysis of results from establishing a relationship between a number of followers and some of the personality types. Considering the results obtained by the program that are showed in the Figure 4.3, it is obvious that users characterised by the trait Neuroticism have the less number of Followers(average value equals to 199), while people opened to experience and extraverted have almost the same average number of followers which is the maximum and equal 551,552 respectively. People with the trait Agreeableness have the average number of followers that is 199. It is interesting to note, Hovy and Hovy (2015) in their work concluded that the number of followers of extraverted users varies from 100 to 500. Our results support this finding.

Taking into account the table results 4.3, it is clear that the minimum average number of friends (202) have the users characterised by the trait Neuroticism. People, that are opened to experience, have the maximum average number of friends (798). The number of friends for another traits like Extraversion and Agreeableness varies from 542 to 743.

According to the table results 4.4, the maximum average number of publications that have been made over 2 years belongs to the extraverted Twitter users (20442). Surprisingly, users characterised by the trait Agreeableness, showed the minimum result (average number equals to 1299). The average number of statuses for users characterised by the traits Openness and Neuroticism varies from 7818 to 9542. Considering these results, we suppose that extraverted people find networks easy to use and very interesting, unlike introverts. This conclusion contradicts to the finding of Hovy and Hovy (2015), where was stated that introverted users make 1000–5000 tweets and extraverted make less than 500. To note, the personality prediction in that study was performed by using MBTI model, so it is possible to see the difference between the results achieved by our platform, based on Big Five personality model and results that were produced by another application elaborated by Hovy and Hovy (2015) based on MBTI model.

As it is shown in the Figure 4.2, average values of total corpus lexical diversity for people characterised by the traits Openness and Extraversion are very similar and equal to 22.66 and 21.59 accordingly. But the values of this characteristic for people with the traits Agreeableness, and Neuroticism are the maximum values, 34.64 and 39.1 respectively. The similar conclusions were made relatively to the average values of lexical diversity calculated for every tweet. Agreeableness, and Neuroticism have the maximum values and Openness and Extraversion, on the contrary, have the minimum average values. It was noted by Mairesse et al. (2007) that introverted people in conversation use a language with the high diversity, unlike the people characterised by the trait Extraversion. Our results support the finding related to extraverted users, as they use the language with less diversity.

Considering the results about the average number of words per tweet, shown on Table 4.2, we concluded that the maximum average result belongs to the people characterised by Extraversion (8.75), on the second place are users with the trait Neuroticism (8.28) and finally, people characterised by Openness and Agreeableness have almost the similar average number of words per tweet, 7.77 and 7.86 respectively. In the work of Mairesse et al. (2007) was stated that extraverted people use less words than introverted., but the results of our program show, that they, on the contrary, use the maximum number of words per tweet.

Our expectations that are related to average word length per tweet were not realised. After analysing the results shown in the Figure 4.5 we have made a conclusion that all the traits have similar average values that vary from 4.11 to 4.4.

We also achieved some interesting results related to the usage of positive and negative emoticons in all publications made over 2 years shown in the Figure 4.1. People opened to experience use positive emoticons more frequently than others. The average number of positive emoticons, found by the program, equals to 129. The minimum result was shown for users characterised by Neuroticism (26). Extraverted and agreeable users have the similar tendency in the usage of positive emoticons, the average numbers vary from 72 to 80.

It was also interesting for us to know about the tendency of usage of negative emoticons. There were achieved the results that surprisingly differ from our expectations. Agreeable users, according to the Figure 4.1, in average use the maximum number of negative emoticons (79), and the minimum average number was showed for the trait Neuroticism (19). On the second place are extraverted users that used in average 34 negative emoticons and opened to experience ones used 24 emoticons.

The relation between the frequency of usage of swear words and personality type, to the

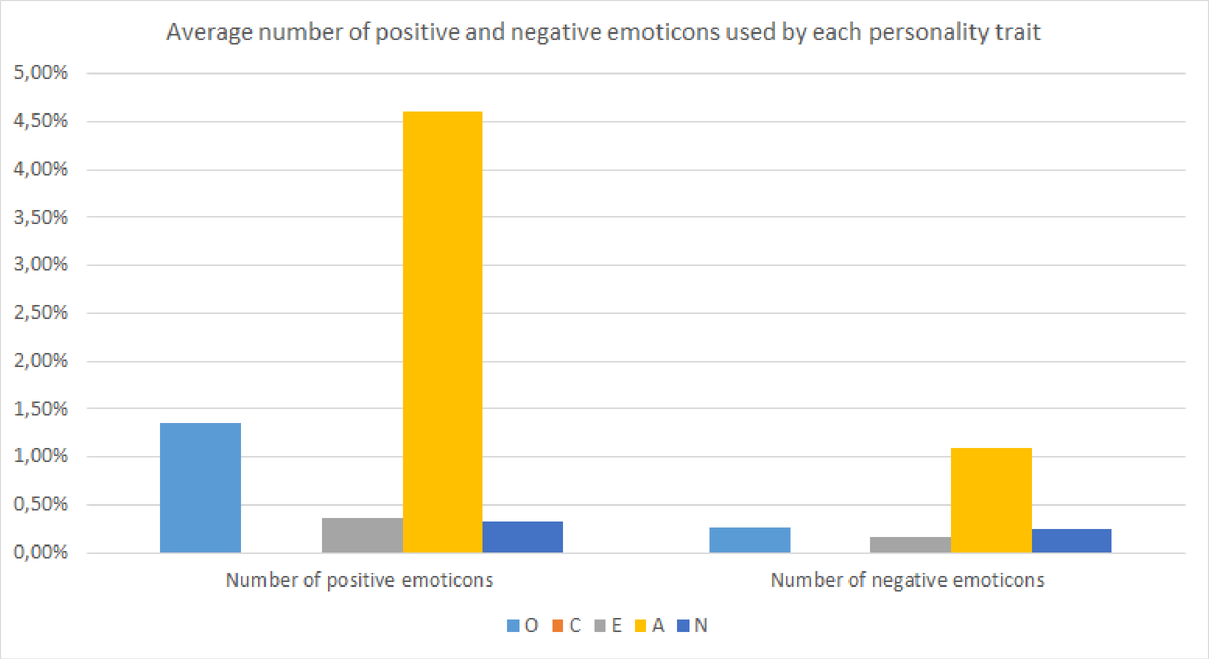


Figure 4.1: Usage of emoticons.

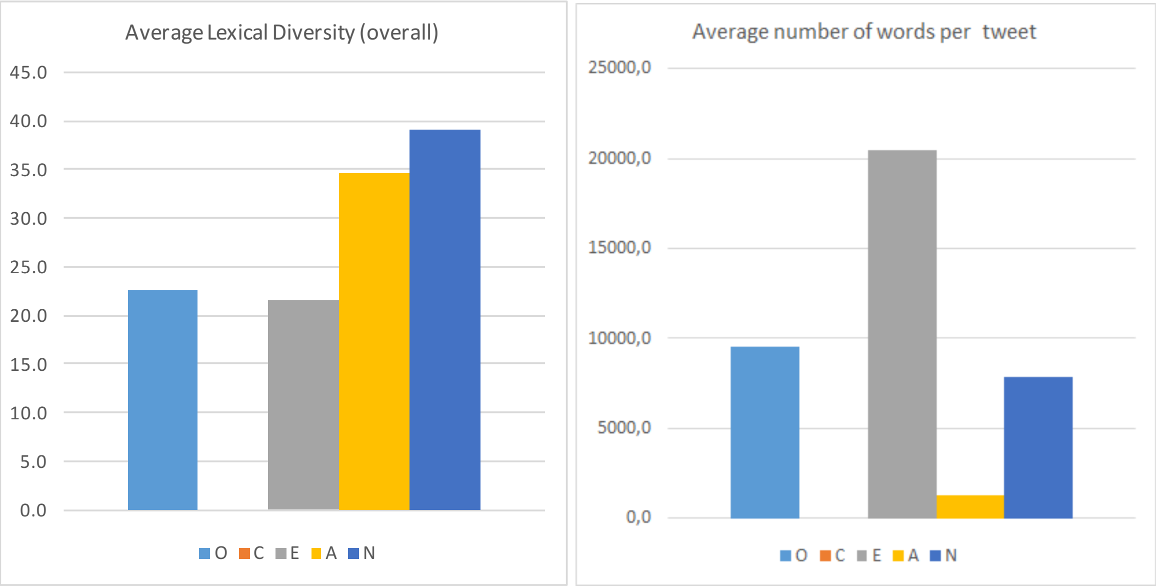


Figure 4.2: Average value of lexical diversity and average number of words per tweet.

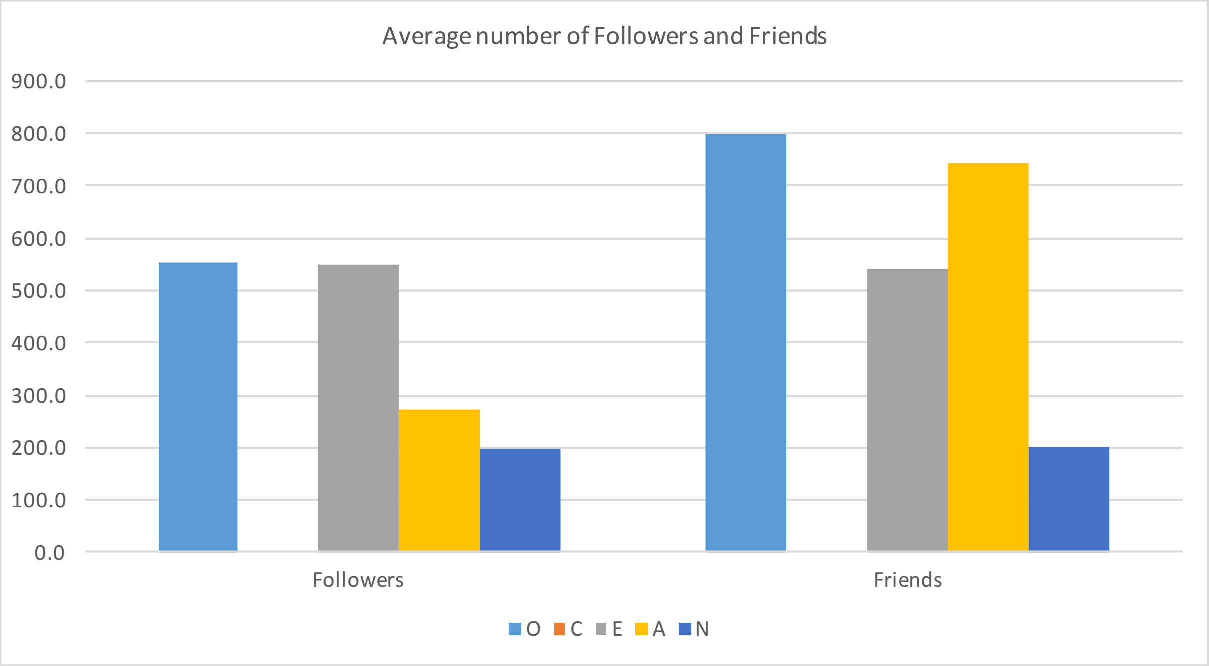


Figure 4.3: Average number of followers and friends.

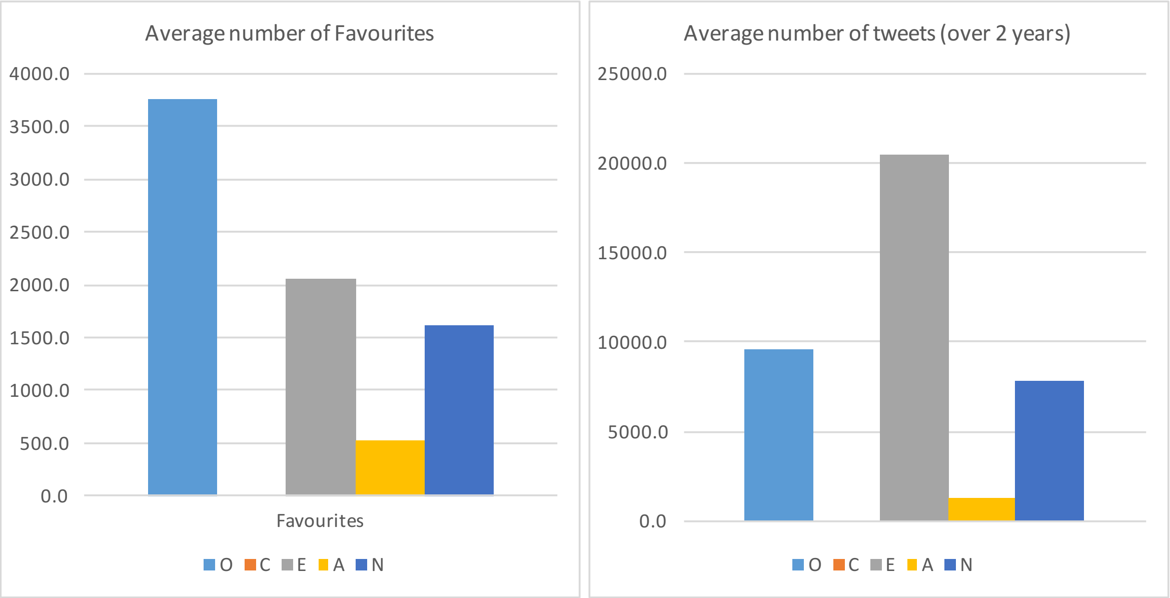


Figure 4.4: Average number of favourited publications and average number of publications made over 2 years.

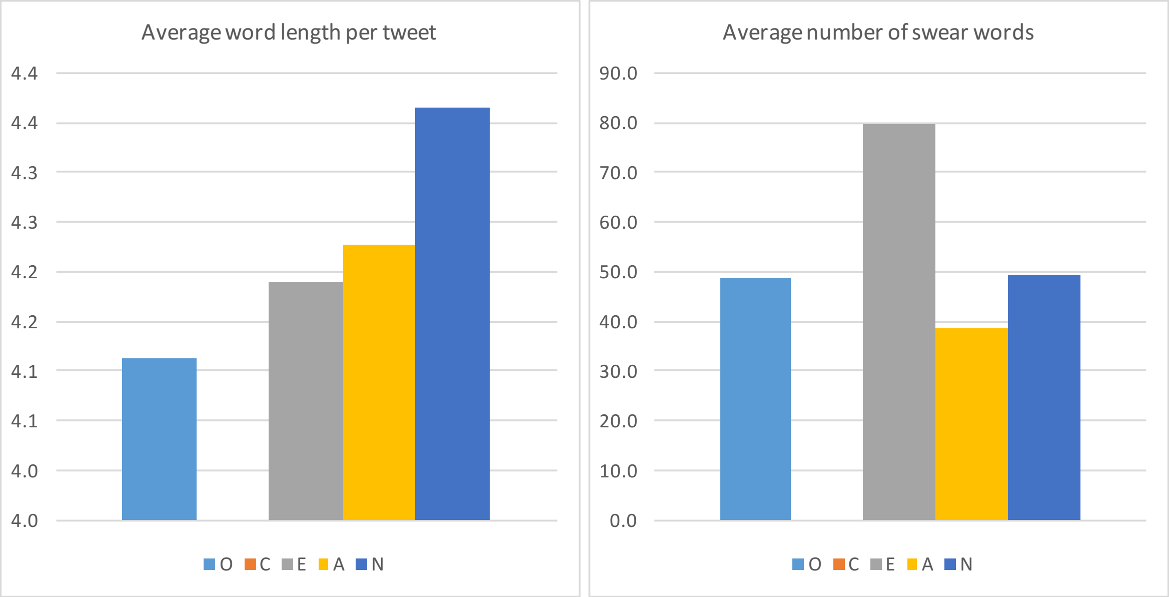


Figure 4.5: Average number of swear words and average word length.

best of our knowledge, was considered in a little work. So, we became interested in knowing is there any dependency between this 2 variables. As was shown in the Figure 4.5 extraverted people tend to use the swear words with the most frequency, the average value of which equals to 80. Agreeable people tend to avoid this type of words in their publications (40). People that are characterised by the traits Openness and Neuroticism have the similar average results on this characteristics, 49 and 50 accordingly. This results partially support the findings of Fast and Funder (2008) where was noted that people who tend to use swear words are characterised as extraverted, while the people characterised by the traits Openness, Agreeableness, and Conscientiousness tend to use less or avoid the usage of this type of words at all.

We made more conclusions after the analysis of characteristics related to temporal activity and the LIWC psychological variables.

Starting from users, opened to experience, it may be said that the maximum number of publications is made in the period between 21 p.m. and 23 p.m, the minimum activity is during early mornings, between 3 a.m. and 5 a.m. To note, there were not noted any tendencies in the usage of specific words from LIWC psychological categories, unlike users characterised by the trait Neuroticism, Agreeableness, and Extraversion.

Those Twitter users high on Neuroticism have no any cue in terms of activity in particular periods of time. Approximate results show that the minimum number of tweets is made between 1am to 11 am and the activity begins to increase from 16 p.m until midnight. They were more likely to use in publications many prepositions, conjunctions, swear words, words related to body, cognitive and biological processes, to insight and relativity words, also tend to use words from such categories as discrepancy, tentative, inhibition, inclusive and exclusive words, words related to ingestion, motion, time, achievement, money, certainty, quantifiers. Some users also have such linguistic cues as usage of pronouns, articles, words that are related to humans, affective processes and causation. It is important to note that this group of users has many similar linguistic cues as agreeable ones.

Agreeable people are less active at the period from midnight till 16 p.m but the maximum activity was noted at night from 21 p.m until 23 p.m. As the users characterised with the trait Neuroticism this group of users tends to use present tense in their tweets, they also use prepositions, conjunctions, numbers, discrepancy, humans. They also can be characterised as tentative, are more likely to use words associated with causation, affective processes. But there are also cues that differentiate these 2 traits. On the contrary of the previously mentioned trait, they are more likely to use pronouns, articles, inclusive, exclusive words, words related to social, biological, perceptual and cognitive processes, words related to space, relativity, motion, time, achievement, words that describe what the user is feeling, state of the health, words associated with the ingestion, leisure, family, and home and also sexual words.

Finally, extraverts were less active in the morning between 6 a.m. and 9 a.m. and more active at night beginning at 20 p.m until 1 a.m. They use many articles, pronouns, prepositions, swear, affective and social words. We also pointed out that this type of users tends to talk about humans, body, ingestion, health and other biological processes. Extraverts frequently make tweets about space and motion using inclusive and exclusive group of words. This type of Twitter users can be described as tentative people, that have a tendency in the usage word of certainty and inhibition.

Taking into account the fact that different people have different free time distribution because of working or studying schedule, so our analysis is not able to identify any preferences about daily activity by every trait and also any tendency of sentiment expression during particular intervals of time.

Conclusions and Future Work 5

In the modern world, as was said by Winston Churchill, "Information rules the world", i.e. those who have more information than competitors in the right time are the most successful. Last decades internet resources such as blogs, social networks, micro-blogs are gaining increased attention of people from different areas beginning from politicians to marketers because of the growing volume of various data that flows throw them. It is caused by the fact that, nowadays, the majority of people tend to express themselves via social networks, such as Twitter, Facebook. They create a great volume of different content, posting links, photos, publications and personal information that reflect emotions, opinions and, sentiments. In previous works, it was supposed that a virtual profile of a user reflects the real personality of one. A type of a user personality can be detected by applying an analysis of a photo of profile or of the textual content. It is important to remember that if the photo of a profile can be false, the language used in comments and posts and, hence the image of personality, is more difficult to fake. As was mentioned in various work, data scientists for the purpose of obtaining the necessary data had to base their personality and sentiment analyses on long and tiresome questionnaires, interviews and essays. With the appearance of such rich data sources as micro-blogs and social networks, the problem of lack of data had disappeared, but the problem of a personality trait's detection hadn't become easier, because of specific issues associated with textual processing of unstructured content that may include shortened internet language, emoticons, hashtags, links and other features that should be preprocessed to obtain the necessary data for personality prediction.

Taking into account the fact that the absolute majority of existing works was elaborated for English and many other languages, but for Portuguese exist a little number of works, it was decided to create a system that could define the one's personality trait by analysing the Twitter profile's publicly available information in Portuguese language.

Trying to achieve the goals we have met some restrictions:

1. Despite the fast rise of interest in the analysis of social networks, there is a little number of lexicons based on the social network's publications. The majority of lexicons are based

on writing samples such as questionnaires, essays, interviews, articles etc. As lexicon used in a social network such as Twitter, is more informal, hence the results produced by our system that uses this lexicon can be not very precise.

2. The second restriction is that almost all the tools for the text analysis, such as the python library NLTK, lexicons, are created for English. It was absolutely necessary for us to create or look for the similar tools for Portuguese.

3. The restriction imposed on the size of tweets make users try to express their thoughts and feelings using the minimum of characters. It has led to the appearance of a specific "internet language", that is not a simple thing to process with existing standard text processing tools.

4. Finally, lexicons, like LIWC, used in many previous researches and also in our work ignore context, irony, sarcasm, and idioms. So, this factor also must be considered during the evaluation of the results.

We have created a Twitter user profile analyser, that shows preliminary result about Twitter user profile personality type with additional characteristics such as the sentiment analysis of tweets, visualisation of a user's temporal activity, of a periodicity of publications, of a distribution of sentiments over time and many other features.

The main differences of our platform from already existing ones is that a user can obtain characteristics of his personality without making one to fill questionnaires or another type of work except entering a user ID, and this analysis can be performed for Portuguese twitter user timeline. To note, we have also performed some analysis for english tweets that will be extended in the future work.

There can be done more work for advancing the achieved results, particularly, we are planning to combine analysis of tweets with analysis, based on correlation of personality traits with numbers of followers, friends, followees. Analysing these characteristics together with all statistics including the user activity and sentiment polarity during different periods of time, a number of unique words, the average length of words and other statistics for the purpose to amend inaccuracies of results about a personality type could drastically improve the results. As Twitter is "invaded" by hashtags, it would be interesting to explore how it can help in revealing a mood, focus and probably even a personality type of a Twitter user. Also, to our opinion, the future work may include the analysis of tweets of all Portuguese users in order to get the characteristics of the Portuguese community of Twitter in terms of personality, contextualising the users on the different regions of the country. For performing assessment of results of personality analysis can be used a publicly available corpus that contains data about some users and their results about MBTI personality type, that was elaborated by Verhoeven et al. (2016). We are planning to explore matchings between the traits of these 2 models of personality. Having this information it could be possible to compare the results about user personality type extracted from this corpus with the results obtained by our platform.

Our platform may be applicable in different areas, because the information it provides can simplify human resources recruitment, the prediction of needs, preferences for products, services, brands, movies, books amongst others and also to predict behaviour in different situations such as for example, during political elections or social unrests. It also can contribute in the understanding of mental health states of users for prediction risks of illnesses.

Bibliography

Alam, F., Stepanov, E. A., and Riccardi, G. (2013). Personality Traits Recognition on Social Network - Facebook. *Seventh international AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 5–8.

Argamon, S., Dhawle, S., Koppel, M., and Pennebaker, J. W. (2005). Lexical predictors of personality type. *Proceedings of joint annual meeting of the interface and The Classification Society of North America*, pages 1–16.

Benet-Martinez, V. and John, O. E. (1998). Los cinco grandes across cultures and ethnic groups: Multitrait?multimethod analyses of the big five in spanish and english. *Journal of Personality and Social Psychology*, pages 729–750.

Bradac, J. J., BOWERS, J. W., and COURTRIGHT, J. A. (1979). Three language variables in communication research: Intensity, immediacy, and diversity. *Human Communication Research*, 5(3):257–269.

Diggle, J. (2004). *Theophrastus: Characters*. Cambridge Classical Texts and Commentaries. Cambridge University Press.

Farnadi, G., Sitaraman, G., Rohani, M., Kosinski, M., Stillwell, D., Moens, M.-F., Davalos, S., and De Cock, M. (2014). How are you doing? emotions and personality in facebook. In *Proceedings of the EMPIRE Workshop of the 22nd International Conference on User Modeling, Adaptation and Personalization (UMAP 2014)*.

Fast, L. A. and Funder, D. C. (2008). Personality as manifest in word use: correlations with self-report, acquaintance report, and behavior. *Journal of personality and social psychology*, 94(2):334–346.

Golbeck, J., Robles, C., Edmondson, M., and Turner, K. (2011a). Predicting personality from twitter. In *SocialCom/PASSAT*.

Golbeck, J., Robles, C., and Turner, K. (2011b). Predicting Personality with Social Media. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, pages 253–262, New York, NY, USA. ACM.

Gou, L., Zhou, M. X., and Yang, H. (2014). Knowme and shareme: understanding automatically discovered personality traits from social media and user sharing preferences. In *CHI*, pages 955–964. ACM.

Grinberg, M. (2014). *Flask Web Development: Developing Web Applications with Python*. O'Reilly Media, Inc., 1st edition.

Hovy, DirkPlank, B. and Hovy, D. (2015). Personality traits on Twitter or how to get 1500 personality tests in a week. *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–98.

Hughes, D. J., Rowe, M., Batey, M., and Lee, A. (2012). A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage. *Computers in Human Behavior*, 28(2):561–569.

Jusupova, A., Batista, F., and Ribeiro, R. (2016). Characterizing the personality of twitter users based on their timeline information. Ambos (Digital e Impresso).

Krieger, M. and Ahn, D. (2010). Tweetmotif: exploratory search and topic summarization for twitter. In *In Proc. of AAAI Conference on Weblogs and Social*.

Mahmud, J., Zhou, M. X., Megiddo, N., Nichols, J., and Drews, C. (2014). Optimizing the selection of strangers to answer questions in social media. *CoRR*, abs/1404.2013.

Mairesse, F. and Walker, M. A. (2006). Words mark the nerds: Computational models of personality recognition through language. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pages 543–548.

Mairesse, F., Walker, M. A., Mehl, M. R., and Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Int. Res.*, 30(1):457–500.

Maruf, H. A., Mahmud, J., and Ali, M. E. (2014). Can hashtags bear the testimony of personality? predicting personality from hashtag use.

Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. 29(3):436–465.

Morgado, I. C. (2012). Classification of sentiment polarity of portuguese on-line news. In *Proceedings of the 7th Doctoral Symposium in Informatics Engineering*, pages 139–150.

Oberlander, J. and Gill, A. (2004). Individual differences and implicit language: Personality, parts-of-speech and pervasiveness. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, pages 1035–1040.

Pennebaker JW, K. L. (1999). Linguistic styles: language use as an individual difference. *J Pers Soc Psychol.*

Perkins, J. (2010). *Python Text Processing with NLTK 2.0 Cookbook*. Packt Publishing.

Qiu, L., Lin, H., Ramsay, J., and Yang, F. (2012). You are what you tweet: Personality expression and perception on Twitter. *Journal of Research in Personality*, 46(6):710–718.

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

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, pages 43–48, Stroudsburg, PA, USA. Association for Computational Linguistics.

Roberts, K., Roach, M., and Johnson, J. (2012). EmpaTweet: Annotating and Detecting Emotions on Twitter. *Lrec*, pages 3806–3813.

Russell, M. A. (2013). *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. O'Reilly Media, second edition edition.

Saif, H., Fernandez, M., He, Y., and Alani, H. (2014). On stopwords, filtering and data sparsity for sentiment analysis of twitter. In Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Scherer, K. R. (2000). Psychological models of emotion. *The neuropsychology of emotion*, 137(3):137–162.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., and Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, 8(9):e73791.

Silva, M. J. and TEAM, R. (2011). Notas sobre a realização e qualidade do twitómetro. Technical report, University of Lisbon, Faculty of Sciences,LASIGE.

Solera-Ureña, R., Moniz, H., Batista, F., Fernández-Astudillo, R., Campos, J., Paiva, A., and Trancoso, I. (2016). Acoustic-prosodic automatic personality trait assessment for adults and children. In *Proc. of IberSpeech 2016 (to appear)*, Lisbon, Portugal.

Tausczik, Y. R. and Pennebaker, J. W. (2010a). The psychological meaning of words: Liwc and computerized text analysis methods.

Tausczik, Y. R. and Pennebaker, J. W. (2010b). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., Welpe, I. M., Universit‰t, T., and M¸nchen (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*.

Vaezi, S. and Kafshgar, N. B. (2012). Learner characteristics and syntactic and lexical complexity of written products. *International Journal of Linguistics*, 4(3):671–687.

Verhoeven, B., Daelemans, W., and Plank, B. (2016). Twisty: a multilingual twitter stylometry corpus for gender and personality profiling. In *Proceedings of the 10th Annual Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. ELRA, ELRA.

Vinciarelli, A. and Mohammadi, G. (2014). A survey of personality computing. *IEEE Transaction on Affective Computing*, 5(3):273–291.

Vosoughi, S., Zhou, H., and roy, d. (2015). Enhanced twitter sentiment classification using contextual information. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 16–24, Lisboa, Portugal. Association for Computational Linguistics.

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality*, 44(3):363–373.