Characterizing the Personality of Twitter Users based on their Timeline Information

Ângela Jusupova, ISCTE-IUL, Portugal, Anzhela_Zhusupova@iscte.pt

Fernando Batista, ISCTE-IUL and INESC-ID, Portugal, fernando.batista@iscte.pt

Ricardo Ribeiro, ISCTE-IUL and INESC-ID, Portugal, ricardo.ribeiro@iscte.pt

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.

Keywords: Personality traits; Twitter user profile; Portuguese Twitter users; Sentiment Analysis

1. INTRODUCTION

The fast development of computers, of methods for processing large amounts of data, of the Internet and social networks helped to simplify some tasks of psychological research, such as the identification of the type of personality, social behavior analysis, identification of cognitive styles, etc. [Tausczik and Pennebaker, 2010]. Especially, it became simpler to understand who is using digital social networks, their reasons and intentions [Hughes et al., 2012].

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, personality and job performance, and personality and satisfaction, among others [Golbeck et al., 2011]. The words that we use in daily life reflect our thoughts and emotions. Words are very important features used in psychology to understand human beings. Our personality can be identified by the words that we use in conversation or in publications that we do in social networks. They also can reveal social relationships, thinking styles, individual differences, the things at what we are focused at a given moment and what emotions we are experiencing [Tausczik and Pennebaker, 2010]. 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

ISSN 2183-489X
DOI http://dx.doi.org/10.18803/capsi.v16.292-299

words together with words that express motion [go, arrive etc.]; if a person have a close relationship with others uses the pronoun "You" [Tausczik and Pennebaker, 2010].

Twitter is one of the most popular social networks, launched in 2006 with more than 20 million unique monthly visitors, where users read and write millions of short messages that are not longer than 140 characters, called tweets [Tumasjan et al. 2010]. In Twitter, unlike Facebook, users do not need to reveal true information about them, so they can make fake accounts and feel free to public everything they think about everything. This microblogging service can be used as a research tool for tracking diseases, for making political predictions, social unrests, and even for predicting personality traits.

Motivation and Goals

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 analyze the type of personality through the analysis of the publications in Twitter for Portuguese users. According to Tausczik and Pennebaker [2010], the bulk of the work relied 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.

We aim at creating an application that can extract personality traits of individuals. Users do not need to fill in any long questionnaires. The application is able to analyze the personality traits 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;
- Localization.

Quercia et al. [2011] noted that the study of the relationship between social networks and personality has commonly relied on The Big Five personality test. We analyze the relationship between the Big Five personality types, namely: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness and different 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), mentioned by Quercia et al. [2011]. Benet-Martinez and John [1998] describe extraversion by traits related to activity and energy, dominance, sociability, expressiveness, and positive emotions. Traits such as altruism, tendermindedness, trust, and modesty characterize agreeableness. Conscientiousness includes trait such as impulse control. Neuroticism combines emotional stability and a large variety of negative effects, such as anxiety, sadness, irritability, and nervous tension. Openness describes the breadth, depth, and complexity of an individual's mental and experiential life. The content of users' profiles that is publicly available can tell us a lot of information, but some users do not share their personal information. To solve this problem, authors [Quercia et al. 2011] access only the information about what they are following, the followers, and listed counts. Our work also allows characterizing the Portuguese community of Twitter in terms of personality. Moreover, it allows contextualizing the users on the different regions of the country.

This paper is structured as follows: Section 2 presents the related work; Section 3 describes the proposed methodology; Section 4 presents a preliminary assessment of the developed system; and, Section 4 closes the document with the conclusions.

2. RELATED WORK

Recently, Twitter became a useful tool for researchers from different disciplines such as psychology, social media, marketing, etc. In this section, we considered only work about personality.

Vosoughi et al. [2015] used contextual information to perform sentiment classification of English tweets, including geo-location, temporal information, and information about the author. To create the sentiment classifier the authors used a Bayesian approach that allowed combining the obtained results with linguistic features. The results showed that this approach obtains better results than a standard linguistic classifier.

Quercia et al. [2011] study the relationship between the Big Five personality traits and Twitter users' types. Five types of users were identified: listeners, popular, highly-read, and two types of influential (Klout: shows retweets and replies on tweets; TIME: ranks public figures). For this study was used information from 335 profiles of users. It was concluded that popular and influential are extroverts and emotionally stable, moreover popular users are imaginative, and influential users are organized. It was shown that is easy to predict Openness 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: following, followers, and listed counts.

Another study [Argamon et al. 2005] was focused on identifying the level of extraversion and neuroticism using text. The authors observe that words can describe gender, age, feelings, thoughts, and personality of a person. Four parameters were used: a "standard function word list", "conjunctive phrases", "modality indicators", "appraisal adjectives" and "modifiers". A Support Vector Machine was used to separate the two classes. The work shows that it is better to use "appraisal" for predicting neuroticism, and the "function words"-based feature is the most suitable for predicting extraversion.

Roberts et al. [2012] describe the creation of a manually annotated corpus of tweets. The corpus was annotated with seven emotions: anger, disgust, fear, love, joy, sadness, and surprise. The distribution of emotions in this corpus was compared with the distribution of emotions of another annotated corpora. Machine-learning methods were used to automate the emotion detection process. They used a method proposed by Roberts et al. [2012] that was developed to detect emotions in suicide notes. As noted by Pennebaker JW [1999], all people differ in linguistic style. Roberts et al. contributed by adding to that study the observation of styles of expression of emotions. To classify each word into different categories, such as emotionality, attentional focus, social relationship, honesty, deception, they used the Linguistic Inquiry Word Count method [Pennebaker JW, 1999].

3. PROPOSED METHODOLOGY

We selected for analysis a subset of accounts of one thousand of Portuguese users of Twitter. To download the timelines of the users we used the Twitter API. To perform the analysis the system receives the ID of the user as input, and then looks for it in the system's folder. If it is present in the folder, the system performs the analysis of the user's timeline. Otherwise, it refers to the Twitter's API searching for the user, then downloads the user's timeline into the system and makes the analysis of this timeline. For each of these users we analyzed the messages produced over the past two years. Classification takes into account the produced content, in agreement with Big Five personality traits, following the methodologies in the related work.

The 140 character limit imposed on tweets makes sentiment classification a challenge using standard linguistic methods. We decided to use different features and methods. Through sentiment classification and analysis of different features, such as quantity of followers, friends, locations, times, analysis of usage of parts-of-speech etc., we can obtain a picture of a user's personality.

Sentiment analysis

This term can be defined as the automatic extraction of information about sentiments, especially polarity, from unstructured text. It can be used by companies in social networks for the extraction of opinions about their products, for scientific research, etc.

We classify tweets into positive, negative or neutral. For English tweets we used the TextBlob library (Python 2 and 3), a tool for processing English text, namely, sentiment analysis, part-of-speech tagging, words inflexion and lemmatization, spelling correction, parsing, etc. To perform sentiment analysis for Portuguese we used a lexicon-based approach combining two lexicons, the SentiLex-PT and the NRC Emotion Lexicon, and also emoticons found in tweets. SentiLex-PT is a lexicon that contains different attributes such as polarity, polarity target, polarity annotation, etc. In every tweet, we compared the words with the lexicon and calculate the probability of 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. We also studied the dependencies between sentiment polarity and tweets' geo-location and sentiment polarity and time of the day, day of the week, month: it can help to better understand the personal traits of a person. Figures 1, 2, and 3 show different aspects of the production of Tweets by Portuguese users.

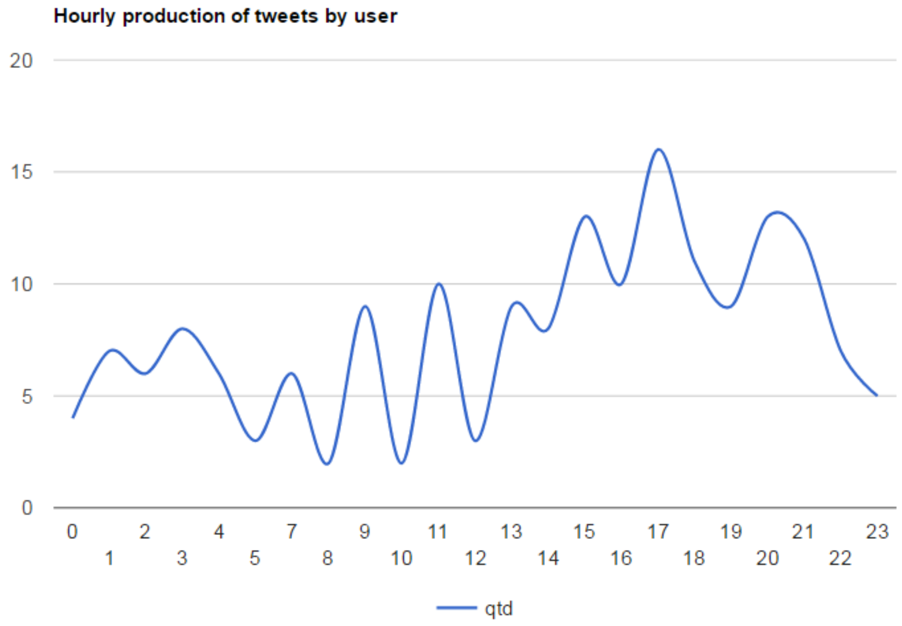

Figure 2 — Hourly production of tweets by user

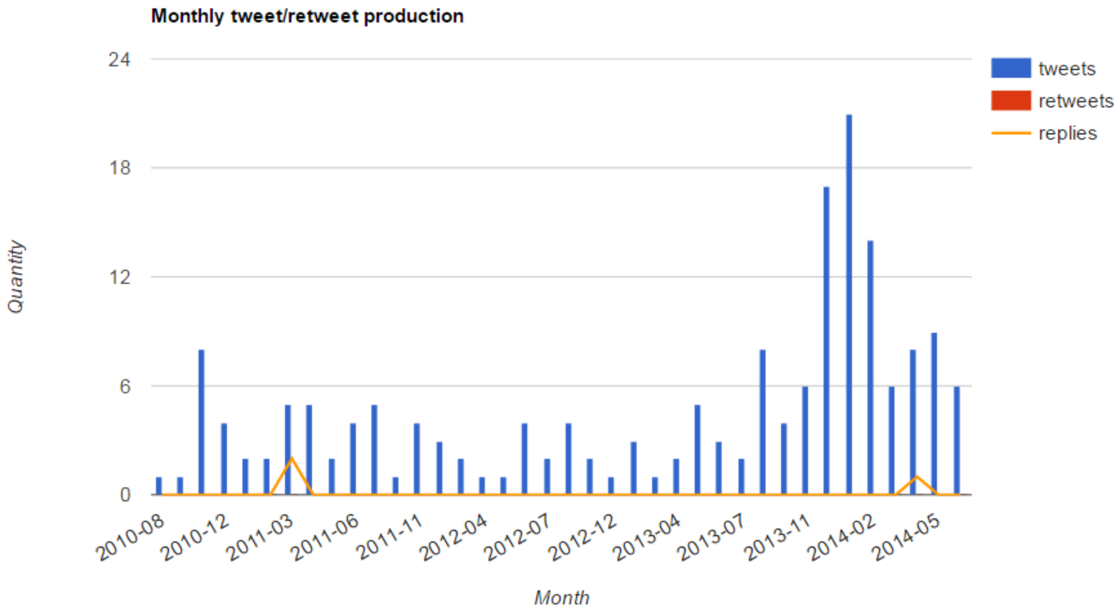

Figure 3 — Tweets per month

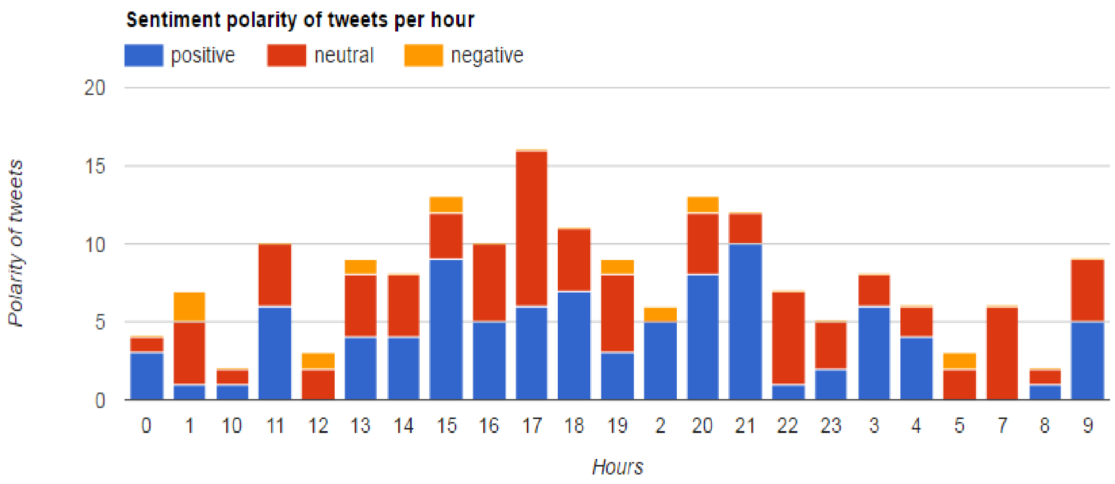

Figure 4 — Sentiment of tweets per hour

Emoticons

Many approaches for Sentiment Analysis use machine-learning techniques. However machine-learning can only be effective in case of an adequate matching between the training and test data: most of the classifiers are topic-dependent, domain-dependent, and even temporally dependent [Read, 2005]. This is frequently overcome by using emoticons as features. We also hypothesize that emoticons can contribute to improve the results of sentiment classifiers.

Stopwords

Using a generic list of stopwords can have a negative impact on sentiment analysis performance [Saif et al. 2014]. We decided to use the stopwords during extraction of features, but to remove them during the sentiment analysis process.

Part-of-speech tagging

To perform part-of-speech tagging for tweets in English, we decided to use a part-of-speech (POS) tagger of the Natural Language Toolkit NLTK, which includes many tools for text processing, such as a tokenizer, a chunker, a POS tagger, a stemmer, etc., and also corpora in different languages. As the available POS tagger is only good for English, we trained our own tagger for Portuguese. NLTK's data package includes two POS-tagged corpora in Portuguese: the Floresta Sinta(c)tica corpus and the MAC-MORPHO Brazilian Portuguese POS-tagged news text. For better POS tagging accuracy, we used both.

4. EVALUATION AND DISCUSSION

For each personality trait we considered specific features extracted during the analysis of timelines of 10 users. According to Furnham et.al [2010], extroverts use more adjectives, prepositions and nouns, while introverts use more verbs, adverbs and pronouns. Neurotic users' posts contain more negative emotions and their posts tend to express negative sentiments. Extroverts, agreeable users express positive emotions in their posts and have many friends and followers. Open users tend to express all emotions. For now, we obtained promising results in sentiment analysis of tweets and defined the type of personality in terms of extroversion and introversion. We used temporal features to analyse users' activity and emotion distribution during different scales of time.

PORTUGUESE USER	SENTIMENT ANALYSIS OF TWEETS (% OF ACCURACY)	EXTROVERT/ INTROVERT /AMBIVERT
1	80	right
2	90	right
3	75	right
4	70	right
5	80	right
6	65	right
7	95	wrong
8	80	right
9	95	right
10	75	right

Table 1 — Analysis of the timelines of 10 users

5. Conclusion and future work

We introduced the idea of characterizing the Portuguese community of Twitter users in terms of personality using the public information of Twitter's profiles. Drawing on the web application, we are going to make a manual validation of this classification, based on what each user thinks about him.

We believe that there are practical applications of this work in different areas like user interface design, recommender systems, business and marketing — if companies take into account the personality traits of people, they will be able to offer more personalized services and focus on the most appropriate products for each of the consumer groups —, and human resources recruitment.

Our future work is to expand the system to make analysis of timelines of Portuguese users not only of the Twitter but also for other digital social networks.

6. Acknowledgments

Work supported by national funds through FCT – Fundação para a Ciência e a Tecnologia under reference UID/CEC/50021/2013 and by ISCTE-IUL.

7. BIBLIOGRAPHY

Argamon, S., Dhawle, S., Koppel, M., and Pennebaker, JW (2005). Lexical predictors of personality type. In Joint Annual Meeting of the Interface and the Classication Society of North America, 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.

Golbeck, J., Robles, C., and Turner, K. (2011). Predicting Personality with Social Media. In CHI '11 Extended Abstracts on Human Factors in Computing Systems, 253-262, ACM.

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.

Pennebaker, JW and LA, King (1999). Linguistic styles: language use as an individual difference. J Pers Soc Psychol.

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 and 2011 IEEE Third International Conference on Social Computing, 180-185.

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In ACL Student Research Workshop, 43-48, ACL.

Roberts, K., Roach, M., and Johnson, J. (2012). EmpaTweet: Annotating and Detecting Emotions on Twitter. In LREC'12, 3806-3813, ELRA.

Saif, H., Fernandez, M., He, Y., and Alani, H. (2014). On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter. In LREC'14, ELRA.

Tausczik, Y. R. and Pennebaker, JW (2010). 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., Universität, T., and München (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In Fourth International AAAI Conference on Weblogs and Social Media.

Vosoughi, S., Zhou, H., and Roy, D. (2015). Enhanced twitter sentiment classification using contextual information. In 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 16-24, ACL.

Wang, X., Wei, F., Liu, X., Zhou, M., and Zhang, M. (2011). Topic Sentiment Analysis in Twitter: A Graph-based Hashtag Sentiment Classification Approach. In CIKM '11, 1031-1040, ACM.

Furnham, A. and Petrova, E. (2010) Body Language in Business: Decoding the Signals. Palgrave Macmillan.