What Does the Bird Say?
Exploring the Link Between Personality and Language Use in Dutch Tweets

Sofie Vandenhoven and Orphée De Clercq

LT^3, Language and Translation Technology Team
Ghent University
Ghent, Belgium
Email: firstname.lastname@ugent.be

***Abstract*—The aim of this paper is to ascertain whether the use of language in Dutch tweets can offer researchers insight into the personality of the user posting those tweets. A database was created, containing the tweets of twenty Belgian, Dutch-speaking Twitter users with an equal representation of both genders. All subjects filled in a personality test based on the Big Five Model of personality and two linguistic analyses were performed on the Dutch tweets. In a first analysis, a more abstract representation of the language was created by means of Part-of-Speech tagging. For the second analysis typical sentiment and personality-charged words were derived from the tweets based on well-known lexicons. Though our database is rather limited, we were able to find some interesting correlations between certain personality traits and language use.**

Keywords–Personality; Big Five; Sentiment analysis.

I. Introduction

Social media are an important aspect of modern-day communication, which is proven by the rising number of monthly active users. This high number of users has logically drawn the attention of researchers, since people share a lot of information about themselves online: how they perceive the world, what they think of current events and how they react on other people are only a few examples. Even more, social media might also offer a deeper insight into their personality, by revealing specific character traits.

Consequently, different sorts of sociolinguistic research on social media have already been conducted: personality, gender and age [1], the use of social media among teens and young adults [2], even the motivation of older adolescents to use social network platforms [3] and also the language used on these social media [4], [5].

The focus of this paper is on personality research. The main objective, however, is not to study the explicit content of messages in order to find out what people talk about online, but to investigate what kind of language is used and whether this language use can reveal something about the personality of the person behind a social media profile. To this purpose, a dataset comprising tweets from twenty respondents -ten males and ten females- was collected. All subjects were asked to fill in a personality test and their tweets were processed using techniques from Natural Language Processing, after which correlations between specific language use and personality were investigated.

The personality model used throughout this paper is known as the Big Five Model [6]. This is one of the most well-researched measures of personality structure of the last decades [4] and it "provides an integrative descriptive model for personality research" [7, p1222]. The personality model contains five traits, marked with the anagram OCEAN or CANOE. Each trait equals a category which is labelled with one substantive. However, the category itself represents a broad range of meaning, captured within this one substantive [7]. For example, the O stands for *Openness*, which includes among others intellect and independence. The different categories are briefly listed in Table 1.

TABLE I. Overview of the Big Five Personality Traits

Trait	*Characteristic*
O for *Openness*	intellectual, polished, independent, open-minded
C for *Conscientiousness*	orderly, responsible, dependable
E for *Extraversion*	talkative, assertive, energetic
A for *Agreeableness*	good-natured, cooperative, trustful
N for *Neuroticism*	not calm, neurotic, easily upset

In the remainder of this paper we will first discuss how the relation between personality and social media has been studied in the past (Section 2). Next, we will explain how Twitter data has been collected and processed from twenty respondents who all filled in an online personality test based on the Big Five (Section 3). In Section 4, we discuss the results, after which this paper is concluded and prospects for future research are offered (Section 5).

II. Related Work

Four main reasons make social media interesting for research. The increasing popularity of social media in the last decade has created an enormous database of personal information [8]. The content in this database, which is widely available through public profiles, is user-generated [9]. The language used on these social media, which fluctuates between spoken and written language but really is neither of them, is a new form of communication [10] and the messages often contain very personal and emotional content [11]. It is highly possible that those four elements caused or at least coincided with a surge in research regarding the Big Five and social media.

However, an often heard criticism is that online profiles might also depict a false and better image of a user, making personality research on social media useless. In the Facebook study presented in [12], no evidence was found to support this presumption. On the contrary, the results show that "people are

not using their social network profiles to promote an idealized virtual identity". This would mean that the personality traits displayed online should correspond to the actual personality of the user.

There has been research on which personalities are mainly drawn to social media. Hamburger and Ben-Artzi [13] found that users of social media are in general introverted and neurotic. Moreover, they also showed a significant difference between genders: female users of social services are generally introverted and highly neurotic, whereas men are quite the opposite. Gender differences were not considered in the study by Ross [14], where almost 90% of the subjects were female. The most important conclusion drawn from this study is that *Openness* positively correlates with the general use of Facebook. In a more extended study on social media use [15] concluded that it is more easily used by people scoring higher on *Openness* and *Extraversion*, whereas it is less used by people who are emotionally stable. For the network site Twitter, Hughes [16] found out that it is more appealing to users scoring higher on *Openness* and lower on *Conscientiousness* when used for social contacts. People using Twitter for information were found to be more introverted and more conscientious.

However, most of these studies take more than only linguistic features into account, or they study anything but the language used on social media. Golbeck et. al., claim to be the first to test whether all information displayed on a profile could predict one's personality. They conducted two studies, one on Facebook [4] and another on Twitter [5]. Since our paper focuses on Twitter we will only discuss those findings. For this research not only the tweets as such were collected, but also public account data such as followers, mentions and so on. However, a linguistic analysis of the tweets formed the major part of the study. Some intuitively logical correlations between the tweets and the Big Five were discovered using the Linguistic Inquiry and Word Count tool (LIWC) [17]. They found that *Conscientiousness* was negatively correlated with words about *death* (e.g. bury, coffin, kill), meaning that the more conscientious a user is, the less he or she will refer to death. Moreover, the same trait was also negatively correlated with *negative emotions* and *sadness*. Hence both findings suggest that highly conscientious people abandon unhappy subjects. Another finding concerning that personality trait revealed a more frequent usage of the pronoun *you*, indicating that highly conscientious people talk more about others. Also, scoring high in *Agreeableness* indicated a significant use of the pronoun *you* and those users were also less likely to talk about the LIWC categories *achievements* and *money*. When trying to predict personality, the linguistic features contributed most to the task.

In more recent years much research has been performed trying to predict personality based on language, such as [18] and [19]. Though personality prediction is beyond the scope of this paper, we believe that the dataset that was collected for this research will be of use for future research in that direction. Important to note is that most previous research has been conducted on English, whereas we want to know whether Dutch language use without any other profile information, can reveal something about someone's personality. And if this is the case, we want to find out which aspects of language are important to take into consideration.

III. Data Collection and Processing

We convinced twenty Dutch-speaking, Flemish persons to participate in our research. All participants were highly active on Twitter and tweeted mostly in Dutch. Relying on the statistics presented in Fig. 1, originally posted by the Belgian Country lead at Twitter, we made sure that half of our respondents belonged to the first age category (ages 16 to 24) and the other half to the second category (ages 25 to 34). Since there is no consensus on whether gender influences personality [13], [20], both genders were equally represented in our database: ten males and ten females.

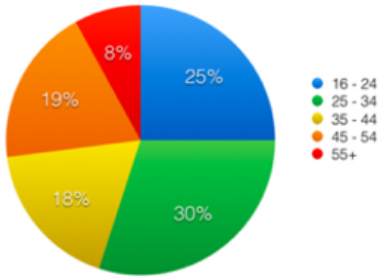

Figure 1. Twitter statistics about the Belgian twitter user profile according to age category in 2015

In order to measure the personality of our respondents, a general Big Five personality test with 46 questions was chosen. The chosen questionnaire uses a Likert-scale from 1 (strongly disagree) to 7 (strongly agree) and had to be filled in online.[1] By agreeing to participate in the research all subjects also agreed to donate their tweets, which were crawled with the Twitter API. After these tweets had been downloaded we made sure that only tweets written in Dutch were retained. In total, our dataset amounted to 8,759 female and 8,780 male tweets.

For this research we first studied whether it is possible to draw a general image of a social media user based on the personality scores that were obtained by our subjects. Next, two linguistic analyses were conducted on the tweets. For both analyses the same two steps were performed. First, a more qualitative analysis was performed by comparing the lowest and highest male and female scorers per personality trait with the outcomes of the linguistic analyses. This more intuitive analysis was followed by measuring Pearson correlations in a second step.

For the first linguistic analysis we rely on the frequencies of the different word forms or Parts-of-Speech (PoS) used in the tweets of our test subjects, in order to derive whether personality can be connected to particular grammatical choices. To this purpose all tweets were tagged with the LeTs Preprocessing Toolkit [21], the PoS module of this tool automatically assigns morphosyntactic labels to each token. Since LeTs is normally used to process standard text material, the output of the tool was adapted in order to deal with Twitter-specific tokens such as hashtags, mentions, emoji's,...

The second linguistic analysis performed on the Twitter data focuses more on the occurrence of words that are known to be charged with a certain sentiment or personality on the basis of lexicons. As sentiment lexicons, we made use of the

[1]The test can be consulted at www.outofservice.com/bigfive

only two existing sentiment lexicons for Dutch, namely the Duoman lexicon [22] and the Pattern lexicon [23]. The Duoman lexicon comprises nouns, adjectives, verbs and adverbs that have been manually labelled by two human annotators as either positive, negative or neutral. The Pattern lexicon is a list of adjectives that were manually assigned a polarity value between -1 (negative) and +1 (positive). In order to perform the analysis all tokenized tweets were processed and all positive and negative lexicon matches retained. As personality lexicon we used the Linguistic Inquiry and Word Count or LIWC [17], which was also used in previous research [4], [5]. An analysis with the LIWC results in a categorisation of all words used into lexical dimensions, accompanied by their relative percentages. Examples of those dimensions are *negemo* for negative emotions, *future* for future tenses and *cogmech* for cognitive processes. This analysis could reveal that people scoring particularly high or low on a personality trait might be recognized by the use of some lexical dimensions.

IV. Results

A. General Social Media Image

The results of the online personality test filled in by our respondents, should be interpreted as follows: scoring above 50% is considered as scoring high on a particular trait and scoring lower than 50% as low. The general averages assigned to each personality trait of our twenty subjects and the average male and female scores are presented below.

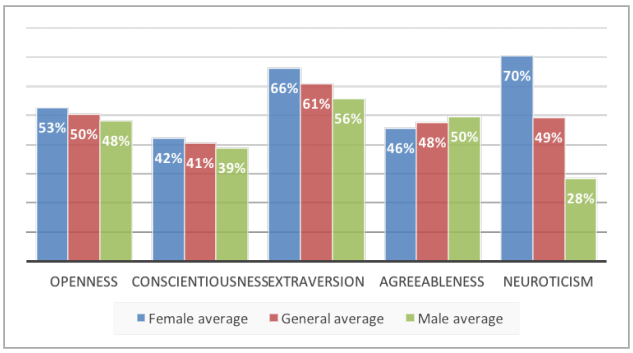

Figure 2. Bar charts representing the general, female and male averages of the Big Five scores from our twenty subjects.

As was said in previous research by [13], [14] and [15], people scoring high on *Extraversion*, *Openness* and *Neuroticism* are the individuals more easily drawn to social media in general. It has to be said that the general averages of our database are not quite convincing to either confirm or deny these results. In general, the twenty subjects do score high on *Extraversion*: 61% on average. They score neither high nor low on *Openness* with an average of 50%. The same is true for *Neuroticism*: on average, the subjects score 49%.

When zooming in on the gender differences, we see that both genders score almost the same for all traits except for *Extraversion* and *Neuroticism*. In both instances the female subjects score higher and for the trait *Neuroticism* the average of 70% is 2.5 times higher than the male average. These findings are in line with some previous research [20] though it should be kept in mind that the database used for this research is very limited and, as a consequence, no generalizations can be made.

B. Part-of-Speech Analysis

Though our idea was to analyse the frequencies of the PoS-tags, some preliminary analyses convinced us to narrow down our research to the category of pronouns, which have also proven indicative of personality in previous research [24], [5].

In general, we found that our male and female subjects talk more about themselves and the groups they belong to, in other words, they use more first person pronouns, both singular and plural. This can easily be explained by Twitter being a microblogging website: it is very logical to talk more about one's own opinions and comments. In a next step, we checked whether there are any correlations that might indicate a relation between personality traits and the use of certain pronouns. We could not find a correlation between a high use of the pronoun *jij* (you) by people scoring high in *Agreeableness* and *Openess*, as found in [5]. The highest correlations we found were with the trait *Neuroticism*: the use of the possessive pronoun *hun* (their) is negatively correlated (-0.54), this correlation, however, is not statistically significant (p-value of 0.09). Other positive correlations between pronouns and *Neuroticism* were found with the pronouns *zij, haar and hij* (she, her and he), i.e., 0.38 and with first person possessives, 0.36. For the other personality traits no specific findings can be reported.

C. Lexicon-Based Analyses

Three different lexicon look-ups were performed: we relied on two Dutch sentiment lexicons and one well-known lexicon for personality research.

Both the Pattern [23] and Duoman [22] sentiment lexicons allowed us to have a closer look at the number of positive and negative words used by our subjects. We found that almost all respondents use more positive than negative words, with the exception of one male subject. However, no links between this finding and the personality traits of our subjects could be discovered.

When processing the data with the LIWC lexicon [17], the outcome is a table listing all LIWC categories that were found in the data, accompanied by a relative percentage. We first performed a more qualitative analysis for which a general overview of the retrieved percentages was created. The highest and lowest percentages per gender were highlighted and compared to each other.

Most of the qualitative results found in our database do make intuitive sense, such as introverted people talking more about *death* (e.g. bury, coffin, kill), *sadness* (crying, grief) and more about *negative emotions* in general (hurt, ugly, nasty). People scoring low on *Neuroticism*, and therefore calmer people, talk more about *friends* (buddy, friend, neighbour), time (end, until, season) and *certainties* (always, never), whereas they also talk more about themselves. In our database, we also discovered that highly conscientious people, talk more about their physical appearance in general: they talk about *eating, food and dieting*, about *physical states* and *grooming*. Since these findings are rather intuitive, we referred to calculating Pearson correlations in a next phase.

In Table 2 we present only the correlations of 0.5 or more that were discovered between a certain personality trait and an LIWC dimension. In our database, we found eight LIWC such categories, correlated mostly with the trait *Openness*: social

processes (*social*, e.g. mate, talk, they, child), humans (*humans*, e.g. baby, adult, boy), sensory and perpetual processes (*senses*, e.g. see, touch, hear), hearing (*hear*, e.g. listening, hearing), present tenses (*present*) and communication (*comm*). For the trait *Conscientiousness*, talking about physical states (*physical*) was found to correlate positively and we also saw that people scoring higher on *Neuroticism* tend to talk more about inhibitions (*inhib*, e.g. block, constrain, stop). Nevertheless, only two of these higher correlations were actually statistically relevant, namely the positive correlation between *Openness* and the mentioning of social processes (*social*) such as mate, talk, they, child; and the positive correlation of that same personality trait with the description of sensory and perceptual processes (*senses*) such as see, touch or listen. This is surprising because our database is only built on the data of twenty people. It is thus definitely worthwhile to conduct a more elaborate study and see whether the highly correlated items will also return in an experiment with a larger database.

TABLE II. CORRELATIONS BETWEEN PERSONALITY AND LIWC DIMENSIONS

Trait	LIWC dimension	Correlation	p-value
Openness	social	**0.6193**	**0.0497**
	humans	0.5254	0.1112
	senses	**0.6242**	**0.0473**
	hear	0.5042	0.1296
	present	0.5227	0.1134
	communication	0.5284	0.1087
Conscientiousness	physical	0.5088	0.1254
Neuroticism	inhib	0.5237	0.1226

Compared to previous research [5], our findings do not support previous results: people scoring high on *Conscientiousness* in our database did not necessarily have a high negative correlation with words about *death* (death; -0.0341), a high positive correlation with *negative emotions* (negemo; 0.1080) or words about *sadness* (sad; 0.0438).

V. CONCLUSION AND FUTURE WORK

The goal of this research was to investigate whether the language used by a specific person on Twitter can reveal something about this person's personality. And if this is the case, we wanted to find out which aspects of language are important to take into consideration. In order to answer this question we first briefly discussed the Big Five and how it has been used to measure the relation between personality and social media in the past. Next, we explained how twenty respondents, ten male and ten female persons, were persuaded to participate in our research. These subjects filled in a personality test and gave their consent to have their Dutch tweets downloaded and analysed. On these tweets two linguistic analyses were then performed: a more abstract analysis by means of Part-of-Speech tagging and a lexicon-based analysis based on two Dutch sentiment lexicons and one personality lexicon. A close analysis of all available data led to some interesting results.

Firstly, since the results of the Big Five personality test of all twenty subjects were available, our findings were compared with previous research on the link between personality and social media. We tried to answer the question whether it is possible to draw a general image of a social media or Twitter user. When it comes to the Big Five and social media in general, which was researched by [13], [14] and [15], one trait corresponds completely, namely scoring high on *Extraversion*. For the traits *Openness* and *Neuroticism*, however, our database might have been too small: the numbers fluctuate around 50%, which makes it impossible to say whether scoring high on both traits is something frequent on social media. What is remarkable is that both the social media and Twitter user are said to score high in *Openness*, which is not supported by our database: our subjects score on average 50% on this trait.

Secondly, based on the Part-of-Speech analysis of the tweets, we found that the use of pronouns in general did not seem to reveal any correlation with a particular trait; therefore, a deeper research was conducted on the use of personal and possessive pronouns. This more thorough analysis did not reveal any particular link with personality. Both male and female users do talk more about themselves and groups they belong to, in other words, they use more first person pronouns, both singular and plural.

Thirdly, based on the lexicon analyses no clear results were conveyed with two Dutch sentiment lexicons. None of the personality traits had a specifically high or low use of positive and negative words. Moreover, all but one respondent used more positive than negative words. Since that one respondent did not score particularly high or low on a trait, we can only guess about the origins of this difference. The analysis with the Dutch LIWC lexicon, however, did provide us with some interesting findings on how often certain dimensions of words are used with a particular personality trait. These were achieved after first performing an intuitive qualitative research, after which Pearson correlations were measured. In our database, we found eight LIWC dimensions to be highly correlated, six with the trait *Openness* and one each with the traits *Conscientiousness* and *Neuroticism*.

A great challenge lied in working with such a limited database. However, much to our surprise, we did discover two statistically significant correlations. The trait *Openness* is positively correlated with social terms, such as family and friends and also with sensory and perceptual processes such as see, touch or listen. This finding is a great stimulus to continue this research on a larger database: the high correlations could even be more outspoken if only they were researched on more data. In future research, it is thus definitely recommended to collect more data: this will help in defining more concretely the general image of a social media user and, of course, in discovering which language items are typical for specific Big Five personality traits. Our database definitely forms a valuable gold standard to conduct research on personality prediction in the near future.

ACKNOWLEDGMENT

The authors would like to thank all twenty Twitter users who agreed to participate in this research by filling in a personality test and donating their tweets.

REFERENCES

[1] A. Schwartz, J. Eichstaedt, M. Kern, L. Dziurzynski, S. Ramones, and M. Agrawal, "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach," PLoS ONE, vol. 8, 2013, pp. 1–16.

[2] A. Lenhart, K. Purcell, A. Smith, and K. Zickuhr, "Social Media & Mobile Internet Use among Teens and Young Adults. Millennials." 2010.

[3] V. Barker, "Older Adolescents' Motivations for Social Network Site Use: The Influence of Gender, Group Identity, and Collective Self-Esteem," Cyberpsychology & Behaviour, vol. 12, 2009, pp. 209–213.

[4] J. Golbeck, C. Robles, and K. Turner, "Predicting personality with social media," in CHI '11 Extended Abstracts on Human Factors in Computing Systems, ser. CHI EA '11. New York, NY, USA: ACM, 2011, pp. 253–262.

[5] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting personality with Twitter," in Privacy, Security, 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust, and IEEE International Conference on Social Computing (SocialCom), 2011, pp. 149–156.

[6] L. Goldberg, "An alternative "description of personality": the big-five factor structure," Journal of Personality and Social Psychology, vol. 59, 1990, pp. 1216–1229.

[7] O. John and S. Srivastava, The Big-Five Trait Taxonomy: History, Measurement and Theoretical Perspectives. Guilford, 1999.

[8] A. M. Kaplan and M. Haenlein, "Users of the world, unite! the challenges and opportunities of social media," Business Horizons, vol. 53, no. 1, 2010, pp. 59 – 68.

[9] M.-F. Moens, J. Li, and T.-S. Chua, Eds., Mining user generated content. Chapman and Hall/CRC, 2014.

[10] W. G. Mangold and D. J. Faulds, "Social media: The new hybrid element of the promotion mix," Business Horizons, vol. 52, no. 4, 2009, pp. 357 – 365.

[11] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval., vol. 2, no. 1-2, 2008, pp. 1–135.

[12] M. Back, J. Stopfer, S. Vazire, S. Gaddis, S. Schmuckle, and B. Egloff, "Facebook Profiles Reflect Actul Personality, Not Self-Idealization," Psychological Science, vol. 21, 2010, pp. 372–374.

[13] Y. Hamburger and E. Ben-Artzi, "The relationship between extraversion and neuroticism and the different uses of the Internet," Computers in Human Behavior, vol. 16, 2000, pp. 441–449.

[14] C. Ross, E. Orr, M. Sisic, A. J.M., M. Simmering, and R. Orr, "Personality and motivations associated with Facebook use," Computers in Human Behavior, vol. 25, 2009, pp. 578–586.

[15] T. Correa, A. Hinsley, and H. Gil de Zúniga, "Who interacts on the Web?: The intersection of users' personality and social media use," Computers in Human Behavior, vol. 26, 2010, pp. 247–253.

[16] D. Hughes, M. Rowe, M. Batey, and A. Lee, "A tale of two sites, Twitter vs. Facebook and the personality predictors of social media usage," Computers in Human Behavior, vol. 28, 2012, pp. 561–569.

[17] J. W. Pennebaker, F. M.E., and B. R.J., "Linguistic Inquiry and Word Count (LIWC): LIWC2001," 2001.

[18] G. Park, H. Schwartz, J. Eichstaedt, M. Kern, M. Kosinski, D. Stillwell, L. Ungar, and M. Seligman, "Automatic personality assessment through social media language," J. Pers. Soc. Psychol., vol. 108, no. 6, 2015, pp. 9–34.

[19] G. Farnadi, G. Sitaraman, S. Sushmita, F. Celli, M. Kosinski, D. Stillwell, S. Davalos, M.-F. Moens, and M. De Cock, "Computational personality recognition in social media," User Modeling and User-Adapted Interaction, vol. 26, no. 2, 2016, pp. 109–142.

[20] M. Vianello, K. Schnabel, N. Sriram, and B. Nosek, "Gender differences in implicit and explicit personality traits," Personality and Individual differences, vol. 26, 2013, pp. 994–999.

[21] M. Van de Kauter, G. Coorman, E. Lefever, B. Desmet, L. Macken, and V. Hoste, "LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit." Computational Linguistics in the Netherlands Journal, vol. 3, 2013, pp. 103–120.

[22] V. Jijkoun and K. Hofmann, "Generating a non-English subjectivity lexicon: Relations that matter," in Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2009), 2009, pp. 398–405.

[23] T. De Smedt and W. Daelemans, "Vreselijk mooi! Terribly beautiful: a subjectivity lexicon for Dutch adjectives," in Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012), 2012, pp. 3568–3572.

[24] J. Pennebaker, M. Mehl, and K. Niederhoffer, " Psychological aspects of natural language use: our words, our selves," Annual review of psychology, vol. 54, 2003, pp. 547–577.