A Survey of Automatic Personality Detection from Texts

Sanja Štajner and **Seren Yenikent***
Symanto Research
Nuremberg, Germany
{sanja.stajner, seren.yenikent}@symanto.com

Abstract

Personality profiling has long been used in psychology to predict life outcomes. Recently, automatic detection of personality traits from written messages has gained significant attention in computational linguistics and natural language processing communities, due to its applicability in various fields. In this survey, we show the trajectory of research towards automatic personality detection from purely psychology approaches, through psycholinguistics, to the recent purely natural language processing approaches on large datasets automatically extracted from social media. We point out what has been gained and what lost during that trajectory, and show what can be realistic expectations in the field.

1 Introduction

Personality is a collection of different constructs such as thoughts, feelings, and values which underlie individual differences and predict human behavior (Roberts and Mroczek, 2008). Due to its complex and multifaceted structure, automatic detection of personality requires a holistic understanding of the construct, which is not an easy task even considering today's technological advancements.

Throughout the personality research history, attempts for personality modelling ranged from traditional psychology methods (e.g. questionnaires), via psycholinguistic approaches (e.g. counting specific word types in texts), to the recent purely natural language processing (NLP) approaches that attempt at detecting personality traits from large amounts of social media data. Especially with recent technological advancements and big data, the research area has extended with the premise that digital footprints could capture not only indirect and natural characteristics but also psychological insights on deeper levels. However, the research has shown a number of problems raising from such approaches which we systematically discuss in this survey.

In Section 2, we introduce the two widely-used personality models, and discuss their similarities and dissimilarities in details. In Section 3, we present use cases of personality detection in the fields of computational linguistics (CL), natural language processing (NLP), and artificial intelligence (AI). Section 4 shows the trajectory of automatic personality detection, from the early pure psychology approaches to the latest NLP approaches on large social media datasets, pointing out strengths and weaknesses of each of the approaches. Section 5 further discusses computational problems raising from using Twitter data for the Myers-Briggs Type Indicator (MBTI) personality modelling, while Section 6 discusses ethical concerns regarding automatic personality detection in general. In Section 7, we revisit the main conclusions of the presented survey.

2 Personality Models

The most widely and frequently used personality models are The Big 5 Model (Costa and McCrae, 1992) and the MBTI model (Briggs-Myers and Myers, 1995). Both of them are long-established psychological models that have attracted attention from CL and NLP fields in hope to offer their wider usage in industry and everyday life.

* Both authors have contributed equally to this work.

Proceedings of the 28th International Conference on Computational Linguistics, pages 6284–6295
Barcelona, Spain (Online), December 8-13, 2020

2.1 The Big 5 Model

The Big 5 Model (also known as the OCEAN model), has originated from lexical approaches discovered and defined by several independent groups of researchers studying and factor-analyzing hundreds of measures of personality traits in order to find the underlying factors of personality (Cattell, 1946; Tupes and Christal, 1961; Goldberg, 1982; Costa and McCrae, 1992). Lexical methods focused on the factor analysis of adjective lists that were rated by participants. The five factor personality model has eventually been accepted since most of these studies resulted in pointing out to five distinctive traits.

The Big 5 Model identifies five broad personality dimensions on a 100-point scale (e.g. 53% Extraversion, 72% Agreeableness):

- **O**penness – liberal and open to new experiences vs. conservative and traditional
- **C**onscientiousness – organized and detail-oriented vs. disorganized and careless
- **E**xtraversion – sociable and outgoing vs. reserved and quite
- **A**greeableness – considerate and cooperative vs. competitive and critical
- **N**euroticism (Emotional Stability) – vulnerable and emotionally unstable vs. calm and stable

There are various behavioral and linguistic implications of the Big 5 model. The Big 5 traits resonate with distinctive and defining behavioral characteristics. For example, it was found that people who scored high on Introversion and Neuroticism preferred written communication methods rather than face-to-face contexts (Hertel et al., 2008). The Big 5 model has also been considered in the context of organizational behavior. It has been shown that a CEO's high Extraversion and Agreeableness and low Neuroticism, Openness and Conscientiousness positively influenced the company's business performance (Wang and Chen, 2019). The Big 5 model has also been heavily applied in advertising and marketing because of the need for going beyond regular demographics to grasp deeper insights about the psychographic profiling of the consumers (Wells, 1975). Matz and Netzer (2017) have shown that ads and marketing materials tailored to personality styles lead to better targeting in consumers' favor (e.g. by leading them to eat healthier, or purchase things they really need) and towards their best interest (e.g. by persuading them against unhealthy habits). Matz et al. (2017) ran Facebook advertising campaigns with targeted messages and slogans for Extraversion and Openness. For a beauty product, they found that when Extraverted people were displayed Extraverted advertising messages (e.g. *Dance like no one's watching, but they totally are*), and when Introverted people were displayed Introverted messages (e.g. *Beauty does not have to shout*), the click and conversion rates drastically increased in just a few weeks. Moreover, another study demonstrated that Big 5-targeted marketing helps consumers express themselves better and become happier by purchasing personality-suited products. Introverts, who are characterized by reserved social skills, report more happiness when they spend money on products that match with introverted activities such as reading or gardening, whereas highly social Extraverts' well-being depend more on their social activities (Matz et al., 2016).

From the psycholinguistic perspective, it has been shown that each of the five dimensions is characterized by different styles in language usage. For example, Extraverts are found to talk more, louder, and more repetitively, have fewer pauses and hesitations, a lower type/token ratio, use more positive emotion words and a less formal language than Introverts (Mairesse et al., 2007; Furnham, 1990; Pennebaker and King, 1999; Gill and Oberlander, 2002; Scherer, 2003). Neurotics seem to use more first person pronouns, more negative emotions, and less positive emotion words (Pennebaker and King, 1999), as well as more concrete and frequent words (Gill and Oberlander, 2003). People with high Conscientiousness tend to avoid negations, and negative emotion words (Pennebaker and King, 1999). Some other linguistic cues of Conscientiousness, such as use of filler words and second person pronouns, has been shown to vary across gender (Mehl et al., 2006), introducing thus additional confounds in automatic personality detection. Avoidance of the past tense has been found to mark openness to experience, and use of swear words to mark disagreeableness (Pennebaker and King, 1999). It has also been noted that some

of the Big 5 dimensions, e.g. Extraversion, have produced more findings – in terms of their correlation with speaking and writing style – than others (Mairesse et al., 2007), questioning thus whether the full personality is possible to detect from texts, even using human judgements.

2.2 MBTI

The MBTI model bases upon the comprehensive theoretical work of Carl Jung (1921), and several decades of extensive practical use within the industrial and educational settings (Briggs-Myers and Myers, 1995). Jung originally came up with three personality dimensions which he defined as psychological functions that people prefer to use for their perception and judgment processes. Later on, Myers and Briggs added the fourth layer, and MBTI has become one of the most widely used non-clinical psychometric assessments.

MBTI lays out a binary classification based on four distinct functions, and draws the typology of the person according to the combination of those four values (e.g. INFP, ESTJ):

- **E**xtraversion/**I**ntroversion - preference for how people direct and receive their energy, based on the outer or inner world
- **S**ensing/I**N**tuition - preference for how people take in information, by five senses or interpretation and meanings
- **T**hinking/**F**eeling - preference for how people make decisions, by looking at logic or people and special circumstances
- **J**udgment /**P**erception - how people deal with the world, by organizing it or staying open for new information

MBTI types translate well into the behavioral context. For example, Extraverts prefer offline communication modes, due to socialization and physical closeness, whereas Introverts prefer online communication, due to the anonymity of online mode (Goby, 2006). In terms of occupational behavior, the two MBTI functions, Sensing/Intuition and Thinking/Feeling, are the most influential personality aspects as they both are highly related to preferences for information processing. For instance, a person who is a Sensing (focus on facts rather than possibilities) and Thinking (use of objective analysis rather than personal analysis) type could use their full potential at jobs dealing with facts and practical analysis such as finance, accounting, applied sciences, and law. Job satisfaction can be derived from MBTI personality types as well. For example, a Judgment type could enjoy the most working on individual projects due to their preference for orderly and organized work, whereas a Perception type would be more satisfied when they spontaneously work on multiple different projects. Employee turnover is also affected by the Extraversion/Introversion match. It has been shown that Extraverts stay the longest in high-stimulating and moving work environments, whereas Introverts prefer to work in environments where the ideas and the main activity take place quietly inside their heads (Briggs-Myers and Myers, 1995).

Unlike the Big 5, there are not many studies investigating linguistic characteristics of different MBTI types. One general reason for this is that MBTI is fundamentally a qualitative approach that makes use of theoretical and professional contexts. Hence, the available data rarely refers to any linguistic context, but more to practical results of the questionnaires.

2.3 Big 5 vs. MBTI

Although both the Big 5 and MBTI are the two most popular personality models, they have contradictory features, which makes them both challenging and complementing to each other.

Theory-driven (MBTI) vs. Data-driven (Big 5). The MBTI originates from the seminal work of Jung that primarily included Jung's clinical observations and analysis of scholarly work. Although this provided a solid theoretical background to the MBTI personality test, the fact that the construction of the test did not include empirical studies consistently measuring the validity and reliability of the instrument

raises flags for psychometric limitations. The Big 5, in contrast, does not rely on any theoretical background. However, it has been validated by many studies, for several decades, in terms of its psychometric and predictive power (Furnham and Crump, 2005).

Type (MBTI) vs. Trait (Big 5). The biggest premise of the MBTI comes from the fact that it is able to profile individual characteristics and personality not only for clinicians but also for professionals, educators and laymen (McCrae and Costa, 1989). This is due to its typology system which makes the complex concept of personality more accessible and understandable. However, because of the binary classification, it loses information while assigning individuals only into discreet categories by ignoring their position in the scale. The Big 5 overcomes this kind of a problem by being a trait model explaining personality by using continuous scales, and providing an intensity-based score for each dimension. However, even with this approach, studies showed that some portion of variance in the results stems from measurement errors and rating biases (Anusic et al., 2009).

Approaches to Extraversion/Introversion. Jung was the first researcher in history ever creating the terms *extraversion* and *introversion*. In his original definition, as well as in the MBTI framework, these terms are described as a person's interest in the outer or inner world energies. Accordingly, Extraverted people get involved in their environment, whereas the inner world is the major energy source for the Introverts. The Big 5 dimension of Extraversion also includes the concept of energy, though rather than how the energy is handled, it focuses on the existence or non-existence of the energy in terms of socialness. An Extraverted person in the Big 5 framework is characterized by being energetic and enthusiastic, whereas an Introverted person is characterized as shy and reserved. Hence, although these two models seem to have at least one dimension in common, the operationalizations of the dimensions are different. Therefore, it is hard for NLP/CL models to use and test them comparably. For instance, the following post is detected as 'Extraverted personality' by both models: *You all are going to party at my funeral because mourning is for losers*. The way the person writes this post provides the idea that he/she is energized by the outer world (MBTI) and has an outgoing and enthusiastic nature (Big 5). The following post is an example of when two models contradict: *Really done with art and has nothing special going on anymore!* This post scores low on the Extraversion scale in the Big 5 model potentially due to the topic of art being of interest for Introverts. However, in the MBTI context, this post would be an Extraverted one as the person is directing his/her energy out with a certain linguistic style, e.g. intensifier, exclamation mark.

3 Use of Personality Detection in NLP Applications

People are more engaged in interacting with others who have similar personality profiles as it requires less information processing and cognitive load (Wu et al., 2017). This idea has been used in designing dialog systems and personal assistants emphasising the need for them to be able to detect and mimic the personality style of a user (Cassell and Bickmore, 2003; Ma et al., 2019; Seo-young Lee and Lee, 2019), in order to behave similarly to humans that instinctively adapt to the other person's personality (Funder and Sneed, 1993; McLarney-Vesotski et al., 2006). It has been shown that adding personality traits to virtual agents leads to significantly better perceived emotional intelligence of such systems (Ma et al., 2019). However, conversational agents so far have mostly exploited methods for adapting to user's emotions instead of deeper personality traits, probably due to insufficiently good performances of automatic personality detection systems on short utterances.

The psycholinguistic characteristics of the Big 5 personality traits (Furnham, 1990; Pennebaker and King, 1999) have been used for building PERSONAGE, the first fully automatic natural language generation system that can generate output controlling for the personality trait (Mairesse and Walker, 2007; Mairesse and Walker, 2008; Mairesse and Walker, 2010; Mairesse and Walker, 2011). The style of the generated output is controlled by many parameters: verbosity, restatement, content polarity, polarisation, concessions, concession polarity, positive context first, claim complexity, self-reference, claim polarity, hedge variation, hedge repetition, tag question insertion, etc. (Mairesse and Walker, 2007). The system showed promising results for four out of five dimensions (all but conscientiousness), being only slightly below the performances of a hand-crafted rule-based system.

Apart from making dialog systems and personal assistants being perceived as more emotionally intelligent, automatic personality detection could be used to enhance various other areas of NLP. For example, inconsistencies in impressions of agreeableness traits across visual and acoustic mode are used as cues for deception detection by human judges (Heinrich and Borkenau, 1998), and therefore could be used to enhance systems for automatic deception detection. Komarraju and Karau (2005) indicated that tutoring systems might be more effective if they were able to adapt to the learner's personality, and Oberlander and Nowson (2006) suggested that opinion mining might be more efficient if personality information was used (Mairesse et al., 2007).

4 Approaches to Personality Detection

The first approaches to personality detection were manual, using personality questionnaires that were performed/analysed by trained psychologists. As they have shown a number of biases (Section 5), and at the same time, computational methods were gaining popularity in linguistics and psycholinguistics, personality detection shifted towards automatic methods that used some basic computations over texts (Section 4.2). Soon after that, CL and NLP communities started showing more interest in the topic, proposing machine learning models that used a number of psycholinguistic features automatically extracted from texts (Section 4.3). Those approaches were still highly connected with psychology theory, as the features were designed based on psychology studies on personality assessment. At the same time, powered by the computational advances, they were able to explore some new features for this task. Although many features showed strong correlations with the 'gold' personality labels obtained via questionnaires, the predictive power of the machine learning models that used those same features was quite low. With the expansion of social media, the focus of automatic personality detection shifted towards machine learning approaches that tried to predict personality from large numbers of Twitter posts, Facebook statuses, or users' behaviours on those social platforms (Section 4.4). However, those approaches did not achieve much higher performances than previous psycholinguistic approaches which used much less data.

4.1 Manual/Questionnaire Approach

Traditional personality assessment relies on self-reports collected via questionnaires and laboratory studies. The strength of this approach to personality assessment is that, since the instruments developed via these methods go through many validation processes, they sit on solid empirical evidences. However, self-reports are likely to suffer from several weaknesses. First, they require human assessment and training of the assessors. Second, they suffer from response biases usually in the form of social desirability bias, i.e. responding to questions in a way that makes people look more favorable to others (Krumpal, 2011). Third, they suffer from *the reference-group effect* (Heine et al., 2002), e.g. an objectively introverted person might perceive him/herself as extraverted if surrounded by a peer-group of even more introverted friends or colleagues (Wu et al., 2017). Finally, apart from those bias problems, it is always a matter of question whether having people answer questionnaires and attend laboratory studies is natural after all. To overcome those shortcomings and find implicit measurements rather than explicit self-reports, personality researchers sought new approaches that led to the utilization of digital methods (Stachl et al., 2019).

4.2 Computational Psychology Approaches

The shift from traditional modelling to digital approaches (i.e. computational psychology) has led to new research avenues bringing psychology theories and linguistic methods together. The early implementers of this combination, Pennebaker and King (1999), showed that linguistic style was the indicator of individual differences across time and different contexts by using a computer program for textual analysis called Linguistic Inquiry Word Count (LIWC) (Francis and Pennebaker, 1993; Pennebaker et al., 2015). Using the same program, various empirical studies demonstrated that function words (i.e. words that connect, shape and organize the written text such as pronouns, prepositions, and conjunctions) were more effective than content words (i.e. words that have a meaning labelling an object or action such as nouns, verbs, and adjectives) in terms of signaling individual differences (Pennebaker, 2011; Tausczik

and Pennebaker, 2010). These findings concreted the idea that *how* people communicate fundamentally provides more insights into their psychological world than *what* they communicate. However, those approaches had several limitations. First, they were constrained by computational limitations, e.g. the used linguistic methods based on word count are not able to capture irony, sarcasm or other contextual elements (Pennebaker et al., 2003). Second, they all just investigated correlations between the examined linguistic features and personality traits without attempting at automatically detecting personality from texts.

4.3 Computational Linguistic Approaches

Building on those previous studies that used LIWC program, Argamon et al. (2005) were the first to attempt at automatically detecting personality from texts. They attempted at binary classification task of detecting extraversion and neuroticism (emotional stability) from the essays corpus (Pennebaker and King, 1999). As features, they used relative frequencies of 675 function words and word categories. They reached an accuracy of 58% on both tasks. Oberlander and Nowson (2006) attempted at automatic personality detection based on four out of five Big 5 traits, using an approach that differ from that of Argamon et al. (2005) in several ways. First, instead of essays, they used a corpus of personal weblogs. Second, instead of relying on previous psycholinguistic studies to design their features (also known as *closed-vocabulary approach*), they used n-grams as features thus allowing themselves that, apart from attempting at building an automatic personality detection system, might discover some other linguistic signals for the four personality traits in question, signals that might have not been explored in the previous studies (also known as *open-vocabulary approach*). Finally, instead of approaching the problem only as a binary classification task, which is in contradiction with the original definition of the Big 5 model which models the traits on a continuous scale (Section 2.1), they further tried seven different ways of partitioning the original corpus into classes, thus approximating a continuous modelling approach (Mairesse et al., 2007). Most importantly, the study of Oberelander and Nowson (2006) showed that careful feature selection, which eliminates the noise of non-discriminating features, can improve the accuracy by large, from 54% to 93% for agreeableness. Although their selected set of features achieved very high accuracies on the original corpus of personal weblogs (up to 83% for emotional stability, and 93% for agreeableness on the binary classification task), it failed the generalisability test by achieving accuracies in the range between 55% and 65% on a different weblog corpus. On the five-level classification task, the highest accuracy (44.7%) was achieved for the extroversion detection task.

Mairesse et al. (2007) were first to model personality detection not only using the self-reports, as in the previous studies (Argamon et al., 2005; Oberlander and Nowson, 2006), but also using observer reports, as 'gold' labels. They used a long list of features, combining the LIWC and MRC (Coltheart, 1981) features with utterance type and prosodic features. They found that observer scores can be more accurately predicted (ranging from 57.0% for openness to experience, to 73.9% for emotional stability) than scores based on self-reports, which did not significantly outperform the majority-class baseline.

A recent study on automatic personality detection from audio data outperformed random guessing only for extraversion and agreeableness, reaching the F_1-scores of 0.56 and 0.58, respectively (Yu et al., 2019). A deep learning model that used a combination of textual, audio and video features reached the accuracy between 88% and 91%, depending on the trait, on the binary classification task (Kampman et al., 2018), showing that combining different modalities significantly improves the results.

4.4 Automatic Personality Detection from Social Media Data

The main objection to all above-mentioned studies is that they focus on small samples and/or closed-vocabulary investigations, and thus cannot generalise well, nor have statistical power of results (Iacobelli et al., 2011; Schwartz et al., 2013; Plank and Hovy, 2015). To overcome those limitations, several datasets have been compiled using large amounts of social media data.

Kosinski et al. (2013) attempted at modelling personality (Big 5) on a continuous scale, in short Facebook posts, relying solely on the Facebook Likes. Only the performance of the model for openness to experience achieved the Pearson's correlation scores close to those of the test-retest reliability. Park et al. (2015) used various linguistic features (words, phrases, topics) and achieved Pearson's correlation

score between 0.35 and 0.43 depending on the personality trait and the type of the gold label used. By using Facebook Likes, Wu et al. (2015) trained machine learning (ML) models to measure if they could outperform human judgments (i.e. judgment scores collected from the users' family and friends, a sample of 86,220 people) in personality detection. The correlation with users' own ratings (obtained via Big 5 questionnaire) were higher for the ML model (r=0.56) than for the human judgments (r=0.49). In another study, Facebook statuses (language-based social media data) outperformed both Facebook Likes (behavioral social media data) and self ratings (Big 5 questionnaire data) in predicting homophily (i.e. preference for interacting with similar others), which is a highly complex interpersonal construct (Wu et al., 2017). Kulkarni et al. (2018) also made use of Facebook statuses and developed a language-based personality construct by factor analyzing users' statuses, which showed high predictive validity for behavioral outcomes (e.g. income and IQ), and high test-retest scores. These studies indicated that social media data, which depicts personality on a fine-grained and natural level, can be a good source for automatic personality detection eliminating the biases that come from questionnaires.

The MBTI profiling of Twitter users based on their posts, where each personality aspect is modelled separately as a binary function, has recently attracted much attention and has been attempted for various languages. Plank and Hovy (2015) compiled a corpus of 1.2M English tweets annotated with MBTI type and gender. They tried to detect user's personality (MBTI type) by modelling four binary classifiers, one per each MBTI dimension (I-E, S-N, T-F, and J-P). They trained logistic regression classifier using a combination of binary n-grams, gender, and several discretized count-based meta-features, i.e. counts of tweets, followers, statuses, favorites, and listed counts. Despite such a large training dataset, their systems managed to outperform the majority-class baseline only on two dimensions, Introversion/Extraversion and Thinking/Feeling, achieving the accuracy of 72.5% and 61.5% on those binary tasks. Verhoeven et al. (2016) attempted at detecting MBTI from short posts on Twitter for six Western European languages using word and character n-grams. Their models achieved the F_1 measure between 0.47 and 0.79 on the binary classification tasks.Yamada et al. (2019) showed that textual features (extracted from users' posts) have better predictive power than the behavioural features in MBTI modelling from Twitter posts in Japanese. All those studies which tried to automatically detect MBTI personality label from tweets, despite large training datasets, barely manage to outperform the majority-based and random-guessing baselines.

Big 5 models trained on Facebook data seem to do a better job in personality prediction than untrained humans (Kosinski et al., 2013; Wu et al., 2015; Wu et al., 2017). The MBTI models trained on Twitter data, in contrast, seem to barely outperform majority-class baselines, and only for certain traits (Plank and Hovy, 2015; Verhoeven et al., 2016). The first explanation that comes to mind is that the potential reason for this would be the data-driven nature of the Big 5 model. The MBTI framework, in contrast, covers more complex and deeper characteristics that come from cognitive and information processing hypotheses, and thus might require a more elaborate modelling which can detect and aggregate different layers of personality. However, here is important to note that the previous studies that model the Big 5 and those that model the MBTI from social media data have several important methodological differences, and thus cannot be fairly compared. First, the Big 5 models are trained on Facebook data, and the MBTI models on Twitter data. It is likely that Facebook data contains more personal statements and is thus more suited for automatic personality detection. Second, the Big 5 models are compared to the performance of untrained humans, and not to majority-based or random-guessing baselines, which was the case for the MBTI models. The methodological issues in psychology-based modelling of Big 5 personality traits, i.e. the absence of any kind of baseline systems (majority-based or random-guessing baselines) and comparison to the performances of trained human assessors, make it difficult to assess the real potential of the proposed models to replace traditional questionnaire-based personality assessment.

To more objectively compare the performances of the Big 5 and MBTI models from the computational perspective, Celli and Lepri (2018) compared them on Twitter data. They extracted several types of features: 1000 character bi-grams and tri-grams with a minimum frequency of 3, LIWC match ratio (68 features), and ten metadata features (the followers/following ratio, hashtags/words ratio, background colour, text colour, etc.) and used the Support Vector Machines (SVM) classifier with feature standardi-

sation for training the binary models in all nine tasks (four from MBTI and five from Big 5). They found that modelling MBTI leads to better performances than modelling the Big 5 in this setup, achieving the average accuracy of 65% on the MBTI tasks (ranging from 60% to 69% depending on the dimension) and 61% on the Big 5 tasks (ranging from 60% to 66% depending on the dimension) in English. By using the AutoWeka, a meta-classifier that automatically finds the best algorithm and settings for the task, they found that the type of architecture used for modelling has greater influence on the success of modelling the Big 5 than the MBTI personality traits. The choice of algorithm and settings led to a 15% and 12% improvement on the binary tasks of assessing emotional stability (neuroticism) and agreeableness, respectively, while the influence on the task performance was smaller for the other three dimensions of the Big 5 model. The choice of architecture and settings used for modelling the four MBTI dimensions, in contrast, only led to 0-2% improvement in accuracy, depending on the task. These results suggest that either Twitter data simply does not contain sufficient amount of linguistic and behavioural (through metadata) signals for successful MBTI personality assessment, or the complex theoretically-based MBTI model simply cannot be inducted from purely textual data, but rather requires insights into long-term behavioural habits and preferences in various spheres of the person's life.

5 MBTI Signals in Twitter Data

We argue that Twitter data simply do not make for a good dataset for the MBTI personality detection, and even more, that purely textual data do not exhibit clear linguistic (either content-based or stylistic) signals for it. The Sensing/Intuitive (S-N), and Judging/Perceiving (J-P) dimensions depend on behavioral signals rather than linguistic. The S-N typology is the result of a preference for information processing, which in fact is the product of cognitive processing and requires cognitive methodologies to understand fully (e.g. abstract reasoning test). For example, the sentence *This one contains the raw data, but it isn't available to the public*, although it mentions *data*, does not provide sufficient information to determine if the user is interested in the mentioned data in a numerical and sensing way, or in an interpretative and intuitive way. The J-P dimension includes different sets of behavioral information about a person, such as planning and organization skills. Short Twitter posts usually do not include information about planning and organization.

To test our hypothesis that Twitter data does not contain sufficient information for MBTI personality detection, we conducted the following annotation experiment. We hired two people, one coming from psychology background and the other from computational linguistic background, both with annotation experience and thorough knowledge of the MBTI model. We randomly selected 96 user data from the MBTI-Twitter dataset (Plank and Hovy, 2015), controlling for having equal number (six) of users from each of the 16 categories, and controlling for gender (three male and three female in each category), as we wanted to have a representative sample with equal number of user-data for each polar in each dimension. The dataset contained 50 concatenated tweets for each user/instance.[1] We asked the annotators to assign to each user (after reading all 50 tweets), for each of the MBTI dimensions one of the following four labels: either of the two polars (e.g. E or I for the E-I dimension, or S or N for the S-N dimension, etc.), *unsure* (in case that they saw many signals from both polars and are not confident to make a binary decision), or *not enough signal* (in the case that they did not find any signal for any of the two polars). The annotators were instructed to have enough breaks to avoid the *fatigue* effect.

We found that, as expected, J-P and S-N were the two dimensions for which the annotators did not find any signal in many instances (Table 1). Here is interesting to note that the Annotator A is the one with the computational linguistics background, and the Annotator B is the one with the psychology background. A feedback interview with the annotators revealed that the annotator with the psychology background was focusing more on the overall impression about the user, while the other annotator focused more on the linguistic clues, the content words and style. Interestingly, the percentage of instances on which both annotators felt confident to assign one of the classes was extremely low for the J-P and S-N tasks (15%

[1]We chose the dataset with 50 tweets of each user, as opposed to those with 100 or 200 tweets, to make the task manageable for human assessors. We believe that having more tweets per user would lead to faster fatigue effect in annotators, while not leading to significantly different overall findings.

Annotator	Statistic	E-I	S-N	T-F	J-P
A (Computational Linguist)	Not enough signal	0	30	8	57
	Unsure	13	13	22	7
	Confident (class assigned)	83	53	68	32
B (Psychologist)	Not enough signal	10	16	10	32
	Unsure	26	16	18	24
	Confident (class assigned)	60	64	68	40

Table 1: Human annotation statistics (in % of instances) on the subset of the MBTI-Twitter dataset.

and 30%, respectively), and somewhat higher for the T-F and E-I tasks (43% and 53%, respectively). The percentage of cases (out of those for which the annotator was confident enough to assign the class) in which each annotator agreed with the 'gold label' varied from 44% to 77%.

Although to be taken only as preliminary due to the small number of annotators and instances involved, these results confirm the raised doubts in possibility to use Twitter posts as training datasets for automatic detection of MBTI personalities. They show that large amounts of Twitter posts do not seem to contain any signals about MBTI dimensions, and that often Twitter posts by the same user show a mixture of signals, thus making Twitter data very noisy and unreliable for the tasks. Furthermore, the annotators had low agreements with 'gold labels', as well as among themselves, even on the instances for which both annotators expressed their confidence. These results raise question if the MBTI constructs from traditional questionnaire-based assessments show expected linguistic patterns on textual data.

6 Ethical Concerns in Automatic Personality Detection

Apart from concerns about whether social media data make for good training data for building reliable personality models, the approaches based on social media data carry some ethical concerns given that the data comes from social media profiles, a construct that is considered as highly private. The biggest issue is related to misuse of social media data without the knowledge or consent of the users. Especially after the infamous Cambridge Analytica study, the issue has started gaining more attention and awareness, and public's opinion on digital targeting and modelling has deteriorated drastically (Schneble et al., 2018). Alongside the consent problem, algorithmic bias could be another issue in this context, making certain groups susceptible to negative effects of the psychological targeting. Various research has suggested that detecting private characteristics of users can lead to biases (Bolukbasi et al., 2016), and put certain groups in unfair situations. For instance, an algorithm that detects a Neurotic user might recommend jobs that do not include any human interaction by ignoring the user's educational and occupational background only because data indicates that Neurotic people do not engage in social interactions as much.

The use of personality detection via psychographic targeting in advertising and marketing also brings ethical concerns. It enables companies to easily target their consumer segments on large scales and increase their sales and conversion rates. However, psychographic targeting can be particularly harmful for vulnerable groups who engage with risky behaviors, such as targeting addicted people with online gambling advertising (Matz et al., 2017; Gladstone et al., 2019). At the same time, when handled properly, the same targeting approach could make consumers more satisfied by directing them to spend their money only on products which are compatible with their personality, and avoiding unnecessary and impulse purchasing behaviors (Matz et al., 2016). Hence, as any research, personality detection can be used for good and for bad.

7 Conclusions

We presented the trajectory of automatic personality detection, pinpointing the main strengths and weaknesses of each of the approaches along the way. Psychology research seems to use more adequate datasets for personality detection, but does not evaluate its models with the same rigour as the NLP/CL approaches. The NLP/CL approaches, in contrast, seem to not pay much attention to the psychology theories that set grounds for the personality modelling and thus set unrealistic expectations that certain

personality signals can be found in any kind of textual data. This is particularly important for the MBTI modelling since it covers a theoretical framework that does not originate from (linguistic) data.

References

Ivana Anusic, Ulrich Schimmack, Rebecca T. Pinkus, and Penelope Lockwood. 2009. The nature and structure of correlations among Big Five ratings: The halo-alpha-beta model. *Journal of Personality and Social Psychology*, 97(6):1142–1156.

Shlomo Argamon, Sushant Dhawle, Moshe. Koppel, and James W. Pennebaker. 2005. Lexical predictors of personality type. In *Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America*.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *In Advances in neural information processing systems*, pages 4349–4357.

Isabel Briggs-Myers and Peter B. Myers. 1995. *Gifts differing: Understanding personality type*. Davies-Black Publishing.

Justine Cassell and Timothy Bickmore. 2003. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-Adapted Interaction*, 13:89–132.

Raymond B. Cattell. 1946. *The description and measurement of personality*. Yonkers-on-Hudson.

Fabio Celli and Bruno Lepri. 2018. Is Big Five Better than MBTI? A Personality Computing Challenge Using Twitter Data. In *CLiC-it*.

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

Paul T. Costa, Jr and Robert R. McCrae. 1992. *Revised NEO Personality Inventory (Neo-PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional manual*. Psychological Assessment Resources.

Martha E. Francis and James W. Pennebaker. 1993. *Linguistic inquiry and word count*. Dallas, TX: Southern Methodist University.

David C. Funder and Carl D. Sneed. 1993. Behavioral manifestations of personality: An ecological approach to judgmental accuracy. *Journal of Personality and Social Psychology*, 64(3):479—-490.

Adrian Furnham and John Crump. 2005. Personality traits, types, and disorders: an examination of the relationship between three self-report measures. *European Journal of Personality*, 19(3).

Adrian Furnham, 1990. *Handbook of Language and Social Psychology*, chapter Language and personality. Winley.

Alastair J. Gill and Jon Oberlander. 2002. Taking care of the linguistic features of extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pages 363—-368.

Alastair J. Gill and Jon Oberlander. 2003. Perception of e-mail personality at zero-acquaintance: Extraversion takes care of itself; neuroticism is a worry. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, pages 456—-461.

Joe J. Gladstone, Sandra C. Matz, and Alain Lemaire. 2019. Can Psychological Traits Be Inferred From Spending? Evidence From Transaction Data." Psychological science. *Psychological Science*, 30(7):1087–1096.

Valerie Priscilla Goby. 2006. Personality and online/offline choices: MBTI profiles and favored communication modes in a Singapore study. *CyberPsychology Behavior*, 9:5–13.

Lewis R. Goldberg. 1982. From Ace to Zombie: Some explorations in the language of personality. *Advances in personality assessment*, 1:203–234.

Steven J. Heine, Darrin. R. Lehman, Kaiping Peng, and Joe Greenholtz. 2002. What's wrong with cross-cultural comparisons of subjective likert scales?: The reference-group effect. *Journal of Personality and Social Psychology*, 82(6):903—-918.

Christina U. Heinrich and Peter Borkenau. 1998. Deception and deception detection: The role of cross-modal inconsistency. *Journal of Personality*, 66(5):687—-712.

Guido Hertel, Joachim Schroer, Bernad Batinic, and Sonja Naumann. 2008. Do shy people prefer to send e-mail? Personality effects on communication media preferences in threatening and nonthreatening situations. *Social Psychology*, 39(4):231–243.

Francisco Iacobelli, Alastair J. Gill, Scott Nowson, and Jon Oberlander. 2011. Large scale personality classification of bloggers. In *Proceedings of the 4th international conference on affective computing and intelligent interaction*, pages 568–577.

Carl G. Jung. 1921. *Psychological Types: Volume 6*. Routledge.

Onno Kampman, Elham J. Barezi, Dario Bertero, and Pascale Fung. 2018. Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 606–611.

Meera Komarraju and Steven J. Karau. 2005. The relationship between the big five personality traits and academic motivation. *Personality and Individual Differences*, 39:557—-567.

Michal Kosinski, David Stillwell, and Thore Graepell. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 15:5802–5805.

Ivar Krumpal. 2011. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality Quantity*, 47(4).

Vivek Kulkarni, Margaret L. Kern, David Stillwell, Michal Kosinski, Sandra Matz, Lyle Ungar, Steven Skiena, and H. Andrew Schwartz. 2018. Latent human traits in the language of social media: An open-vocabulary approach. *PloS one*, 13(11).

Xiaojuan Ma, Emily P. Yang, and Pascale Fung. 2019. Exploring perceived emotional intelligence of personality-driven virtual agents in handling user challenges. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 1222–1233.

Francois Mairesse and Marilyn Walker. 2007. Personage: Personality generation for dialogue. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Francois Mairesse and Marilyn Walker. 2008. Trainable generation of big-five personality styles through data-driven parameter estimation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Francois Mairesse and Marilyn Walker. 2010. Towards personality-based user adaptation: Psychologically informed stylistic language generation. *User Modeling and User-Adapted Interaction*, 20(3):227–278.

Francois Mairesse and Marilyn Walker. 2011. Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics*, 37(3).

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

Sandra C. Matz and Oded Netzer. 2017. Using big data as a window into consumers' psychology. *Current opinion in behavioral sciences*, 18:7–12.

Sandra C. Matz, Joe J. Gladstone, and David Stillwell. 2016. Money buys happiness when spending fits our personality. *Psychological science*, 27(5):715–725.

Sandra C. Matz, Michal Kosinski, Gideon Nave, and David J. Stillwell. 2017. Psychological targeting as an effective approach to digital mass persuasion. *PNAS*, 114:12714–12719.

Robert R. McCrae and Paul T. Jr. Costa. 1989. Reinterpreting the Myers-Briggs type indicator from the perspective of the five-factor model of personality. *Journal of personality 57*, 57(1):17–40.

Amber R. McLarney-Vesotski, Frank Bernieri, and Daniel Rempala. 2006. Personality perception: A developmental study. *Journal of Research in Personality*, 40(5):652—-674.

Matthias R. Mehl, Samuel D. Gosling, and James W. Pennebaker. 2006. Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Personality and Social Psychology*, 90:862—-877.

Jon Oberlander and Scott Nowson. 2006. Whose thumb is it anyway? classifying author personality from weblog text. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Gregory Park, H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Michal Kosinski, David J. Stillwell, Lyle H. Ungar, and Martin E.P. Seligman. 2015. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108:934–952.

James W. Pennebaker and Laura A. King. 1999. Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296–1312.

James W. Pennebaker, Matthias R. Mehl, and Niederhoffer Kate G. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577.

James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Environment and Planning D: Society and Space.

James W. Pennebaker. 2011. *The secret life of pronouns: What our words say about us*. Bloomsbury Press.

Barbara Plank and Dirk Hovy. 2015. Personality traits on twitter—or—how to get 1,500 personality tests in a week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–98, Lisbon, Portugal, September. Association for Computational Linguistics.

Brent W. Roberts and Daniel Mroczek. 2008. Personality trait change in adulthood. *Current directions in psychological science*, 17:31–35.

Klaus R. Scherer. 2003. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40 (1-2):227—-256.

Christophe Olivier Schneble, Bernice Simone Elger, and David Shaw. 2018. The Cambridge Analytica affair and Internet-mediated research. *EMBO reports*, 19(8).

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stilwell, Martin E. Seligman, and Lyle H. Ungar. 2013. Personality, gender, and age in the language of social media: The open vocabulary approach. *PLOS ONE*, 8.

Soomin Kim Seo-young Lee, Gyuho Lee and Joonhwan Lee. 2019. Expressing personalities of conversational agents through visual and verbal feedback. *Electronics*, 8.

Clemens Stachl, Florian Pargent, Sven Hilbert, Gabriella M. Harari, Ramona Schoedel, Sumer Vaid, Sam Gosling, and Bühner Markus. 2019. Personality Research and Assessment in the Era of Machine Learning.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Ernest C. Tupes and Raymond E. Christal. 1961. Recurrent personality factors based on trait ratings. *USAF ASD Technical Report*, pages 61–97.

Ben Verhoeven, Walter Daelemans, and Barbara Plank. 2016. Twisty: a multilingual twitter stylometry corpus for gender and personality profiling. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1632–1637, Portoroz, Slovenia, May. European Language Resources Association (ELRA).

Shichao Wang and Xi Chen. 2019. Recognizing CEO personality and its impact on business performance: Mining linguistic cues from social media. *Information Management*.

William D. Wells. 1975. Psychographics: A critical review. *Journal of marketing research* , 12:196–213.

Youyou Wu, Michal Kosinski, and David Stillwell. 2015. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040.

Youyou Wu, David Stillwell, H. Andrew Schwartz, and Michal Kosinski. 2017. Birds of a feather do flock together: behavior-based personality-assessment method reveals personality similarity among couples and friends. *Psychological Science*, 28:276–284.

Kosuke Yamada, Ryohei Sasano, and Koichi Takeda. 2019. Incorporating textual information on user behavior for personality prediction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 177–182, Florence, Italy, July. Association for Computational Linguistics.

Mingzhi Yu, Emer Gilmartin, and Diane J. Litman. 2019. Identifying personality traits using overlap dynamics in multiparty dialogue. *ArXiv*, abs/1909.00876.