Emerging Trends in Personality Identification Using Online Social Networks—A Literature Survey

VISHAL KAUSHAL and MANASI PATWARDHAN, Vishwakarma Institute of Technology

Personality is a combination of all the attributes—behavioral, temperamental, emotional, and mental—that characterizes a unique individual. Ability to identify personalities of people has always been of great interest to the researchers due to its importance. It continues to find highly useful applications in many domains. Owing to the increasing popularity of online social networks, researchers have started looking into the possibility of predicting a user's personality from his online social networking profile, which serves as a rich source of textual as well as non-textual content published by users. In the process of creating social networking profiles, users reveal a lot about themselves both in what they share and how they say it. Studies suggest that the online social networking websites are, in fact, a relevant and valid means of communicating personality. In this article, we review these various studies reported in literature toward identification of personality using online social networks. To the best of our knowledge, this is the first reported survey of its kind at the time of submission. We hope that our contribution, especially in summarizing the previous findings and in identifying the directions for future research in this area, would encourage researchers to do more work in this budding area.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Information systems** → **Data mining**; **Social networks**; • **Applied computing** → **Psychology**;

Additional Key Words and Phrases: Personality, personality prediction, online social network, Facebook, Twitter, mining social network

ACM Reference format:
Vishal Kaushal and Manasi Patwardhan. 2018. Emerging Trends in Personality Identification Using Online Social Networks—A Literature Survey. *ACM Trans. Knowl. Discov. Data.* 12, 2, Article 15 (January 2018), 30 pages.
http://dx.doi.org/10.1145/3070645

1 INTRODUCTION

Personality identification using online social networks aims at deducing the personality traits based on one's profile and/or behavior in online social networks. According to psychologists (DeYoung 2010) and neuroscientists (Adelstein et al. 2011), personality is defined as an affect processing system that describes persistent human behavioral responses to broad classes of environmental stimuli. It characterizes a unique individual and it is involved in communication processes and connected to how people interact with one another. Studies in psychology have demonstrated several personality types and traits that are characteristics of different individuals.

Authors' addresses: V. Kaushal and M. Patwardhan, Department of Computer Engineering, Vishwakarma Institute of Technology, 666, Upper Indiranagar, Bibwewadi, Pune - 411037, Maharashtra, INDIA; emails: {vishal.kaushal, manasi.patwardhan}@vit.edu.

http://dx.doi.org/10.1145/3070645

Ability to identify personalities of people has always been of great interest to the researchers due to its importance. It continues to find highly useful applications in many domains. For example, training systems would be more efficient if they could adapt to learner's personality, as personality traits affect academic motivation (Komarraju and Karau 2005). Dating websites can predict relationship success by trying to match personalities of individuals. Human Resources Department can benefit from predicting job satisfaction or making right decisions before hiring a potential employee. Recommender systems can improve their accuracy by recommending music, movies, books, or other items that are tailored to the user's personality profile. Knowledge of a user's personality also enables applications to personalize user interfaces.

Traditionally the only way personalities were identified was through questionnaire based personality tests that the subjects used to undergo. A personality test is a questionnaire or other standardized instrument designed to reveal aspects of an individual's character or psychological makeup. Such methods however, are costly, error-prone, and laborious. However, with the advancement in fields of computer science like image processing and artificial intelligence more automated techniques have become common. For example, researchers have tried to approach the problem of personality recognition through face recognition (Robin et al. 2011); signature, handwriting recognition (Kedar et al. 2015); linguistic cues from conversation, text, SMS, blogs, and so on (Mairesse and Walker 2006); fine emotion features extracted from essays (Mohammad et al. 2013); speech acts (Appling et al. 2013); smart phone data (Chittaranjan et al. 2011); and so on.

One of the major sources of identifying a user's personality is by analyzing the text produced by him. A central assumption of language psychology is that the words people use reflect who they are. Psychologists have documented the existence of personality cues in text by discovering correlations between a range of linguistic variables and personality traits, across a wide range of linguistic levels, including acoustic parameters (Smith et al. 1975; Scherer 1979), lexical categories (Pennebaker and King 1999; Pennebaker Mehl and Neiderhoffer 2003; Mehl Gosling and Pennebaker 2006; Fast and Funder 2007), *n*-grams (Oberlander and Gill 2006), and speech-act type (Vogel and Vogel 1986). Taking this a step further, Mairesse and Walker (2006) developed models for personality recognition based on user's language as revealed in his essays and conversation extracts. A recent publication (Agarwal 2014) presents a review of Personality Recognition from Text (PRT) techniques.

Social media in general and online social networks in particular, serve as a rich source of textual as well as non-textual content published by users (Chin and Wright 2014). Owing to the increasing popularity of online social networks, researchers have started looking into the possibility of predicting a user's personality based on his online social networking profile. Social networking sites are widely used as a primary medium for communication and networking (Boyd and Ellison 2007; Valkenburg and Peter 2009). They have become a central, virtually unavoidable medium for social interactions. In the process of creating social networking profiles, users reveal a lot about themselves both in what they share and how they say it. Through self-description, status updates, photos, interests, interactions with other people on the social networks like messages, likes, upvotes, replies, retweets, and so on, much of a user's personality comes out.

Consequently, researchers have made use of the rich textual content available in online social networking profiles in form of status updates to predict users' personalities using standard PRT techniques (Golbeck et al. 2011; Wald et al. 2012; Farnadi et al. 2013; Alam et al. 2013; Markovikj et al. 2013; Appling et al. 2013; Gou et al. 2013). In addition to textual content, researchers have also looked into the possibility of identifying a user's personality based on the non-textual content associated with the user's social networking profile. These include several structural and behavioral features that can be extracted to understand their correlations with personality and their utility in

predicting the user's personality. Examples of such features include, but are not limited to, number of friends, contact frequency, network density, centrality, and clustering measures, and so on.

A recent publication on personality computing (Vinciarelli and Mohammadi 2014) presents a survey of the technologies capable of dealing with human personality in general and it also aims at providing "a conceptual model underlying the three main problems addressed in the literature, namely Automatic Personality Recognition (inference of the true personality of an individual from behavioral evidence), Automatic Personality Perception (inference of personality others attribute to an individual based on her observable behavior) and Automatic Personality Synthesis (generation of artificial personalities via embodied agents)." As a complement to this work, Wright (2014) summarizes contemporary thinking and research in personality with potential directions for future research and in the response article, the original authors, Vinciarelli and Mohammadi (2014) elaborate on such directions from computing science point of view. These studies, however, are much broader and generic in nature and do not specifically address the topic of our work. In this article, we specifically review the state of the art in the emerging trends for personality identification using online social networks. To the best of our knowledge, this is the first reported work of its kind.

The organization of the rest of the article is as follows. In Section 2, we present a background on personality as an important dimension of humans, followed by the importance and applications of personality identification. We also provide a description of Big Five personality model, a de facto standard for studying personality, and a brief description of some other well-known models. We conclude Section 2 by providing references to past studies that have used language as a marker for personality. We begin Section 3 by talking about the ever-increasing usage of online social networks that make them appropriate sources of textual as well as non-textual information for various analyses. We present studies that reveal that online social networks act like extended environments, where users behave and function projecting their personality. We also present contrasting opinions about whether such projected personality is user's true personality or idealized personality. We conclude Section 3 by presenting studies that establish online social networks as indeed a true mirror of users' personality. The first step toward automated personality identification from online social networks is feature identification, and thus, in Section 4, we present a categorized summary and background knowledge of different types of features that can be extracted from online social networks for personality identification. In Section 5, we present a summary of various studies that establish correlations between these features and personality. Next, in Section 6, we report a survey of various studies involving automatic identification of personality from social network data categorized into supervised, unsupervised, and semi-supervised approaches. This is followed by Section 7, where we present the trait-wise summary of results of such studies. Finally, in Section 8, we present directions for future research in this area based on our survey.

2 BACKGROUND

2.1 Personality as an Important Dimension of Humans

The scientific term "personality" is conceptualized as the entire mental organization of a person's traits, where traits are defined as a cross-situational and temporally stable set of individual attributes. Personality is the complex of all the attributes—behavioral, temperamental, emotional, and mental—that characterizes a unique individual. Wikipedia defines personality as the particular combination of emotional, attitudinal, and behavioral response patterns of an individual. Personality traits are the most fundamental dimension of variation between humans.

2.2 Importance and Applications of Personality Identification

Ability to identify personalities of people has always been of great interest to the researchers due to its importance. Personality traits influence many aspects of user behavior, such as attitude toward machines (Sigurdsson 1991), overall job performance (Furnham Jackson and Miller 1999), academic ability and motivation (Komarraju and Karau 2005; Furnham and Mitchell 1991), psychological disorders (Saulsman and Page 2004), music preferences (Rawlings and Ciancarelli 1997; Rentfrow and Gosling 2003; Dollinger 1993; Hansen and Hansen 1991), evaluation of conversational agents (Reeves and Nass 1996; Cassell and Bickmore 2003), leadership ability (Hogan et al. 1994), sales ability (Furnham, Jackson and Miller 1999), teacher effectiveness (Rushton et al. 1987), marketing techniques (Odekerken-Schroder et al. 2003; Whelan and Davies 2006; Nov and Ye 2008), and so on. Consequently, personality identification continues to find interesting applications in these and related areas. Some examples include recommendation systems (Roshchina et al. 2011), deception detection (Enos et al. 2006), authorship attribution (Luyckx and Daelemans 2008; Reiter and Sripada 2004), dialog systems (Funder and Sneed 1993; McLarney et al. 2006; Mairesse and Walker 2007), training systems, dating websites for predicting successful relationships and checking compatibility (Donnellan et al. 2004; Zhao et al. 2008), assessing job candidates (Finder 2006), design of personalized user interfaces (Karsvall 2002; Hauser et al. 2009), analyzing meetings and the conversations of suspected terrorists (Hogan et al. 1994; Tucker & Whittaker 2004; Nunn 2005), targeted marketing (Nov and ye 2008), friend suggestion system on an online social network, customized website "skins" (Brinkman and Fine 2005); and so on.

2.3 Big Five and Other Models of Personality

Most work in this area has used Big Five model of personality (Digman 1990; McCrae and John 1992; Norman 1963), which is one of the most well-researched and well-regarded measures of personality structure. Big Five has over the years, emerged as a standard way of modeling personalities based on extensive research (Peabody and Raad 2002; Schmitt et al. 2007) and is replicated many times (Norman 1963; Peabody and Goldberg 1989). With work beginning over 50 years ago (Norman 1963) and journals dedicated to it, the Big Five model (also known as the Five Factor Model—FFM) is a well-accepted construct of personality (Digman 1990). Personality is formalized along five traits obtained from factor analysis of personality description questionnaires. The five bipolar personality traits, namely Extraversion, Emotional Stability, Agreeableness, Conscientiousness, and Openness, have been proposed by Costa and McCrae (1985). The five personality traits according to Big Five can be described as follows:

Openness is the propensity of individuals to display imagination, curiosity, originality, and open-mindedness. In contrast, low openness scores indicate people who are practical, traditional, and down-to-earth. Openness to experience represents an individual's willingness to consider alternative approaches, be intellectually curious, and enjoy artistic pursuits.

Extraversion is the extent to which individuals are outgoing, active, assertive, and talkative. In contrast, individuals with low levels of extraversion tend to be "introverted," reserved, serious, and prefer to be alone or stay within close circles. Extraversion reflects a tendency to be sociable and able to experience positive emotions.

Conscientiousness is the extent that an individual is dependable, careful, responsible, organized, and has a high will to achieve. It reflects the degree to which an individual is organized, diligent, and scrupulous.

Agreeable persons tend to be courteous, kind, flexible, trusting, forgiving, are inclined to cooperate but known to avoid conflict. It is another aspect of interpersonal behavior, reflecting a tendency to be trusting, sympathetic, and cooperative.

Neuroticism is the extent to which individuals experience and display negative affects like anxiety, sadness, embarrassment, depression, guilt, and is tied to the ability to cope with stress. Roughly speaking, it is the tendency to worry. It reflects a person's tendency to experience psychological distress and high levels of this trait are associated with a sensitivity to danger.

Barrick et al. (2001) described FFM as the most useful taxonomy in personality research, while Costa and McCrae (1992) consider it the most comprehensive and parsimonious model of personality.

Big Five is one of the examples of a trait theory. It is important to note the difference between personality type and personality trait. While type is a qualitative characterization of a person (Jung 1971), trait is more of a quantitative representation of a behavioral tendency. For example, according to type theories, introverts and extraverts are two fundamentally different categories of people. According to trait theories, however, introversion and extraversion are part of a continuous dimension, with many people in the middle. Type is discrete, while trait is continuous. HEXACO (Ashton et al. 2004) is another example of a trait theory. Examples of some type theories include, Type A Type B personality theory, MBTI (Myers-Briggs Type Indicator) (Myers and Briggs 1980, 1995) and Enneagram of personality. However, because personality test scores usually fall on a bell curve rather than in distinct categories, personality type theories have received considerable criticism among psychometric researchers.

A number of measures of the FFM of personality have been developed. For example, the NEO Personality Inventory (NEO-PI-R) is a psychological personality inventory, a 240 item measure of the Big Five personality traits. Other measures of the five traits of personality include (NEO-FFI) (Costa and McCrae 1992), the California Psychological Inventory (Gough 1987) and the Hogan Personality Inventory (Hogan and Hogan 1992), all of which are proprietary. While most research on personality and Facebook has relied on the NEO PI-R or NEO-FFI instrument, the IPIP (Goldberg et al. 2006; http://ipip.ori.org/) is very user friendly (i.e., non-proprietary and much shorter) and research has shown strong evidence of convergent and discriminant validity and interchangeability with the NEO-FFI (Lim and Ployhart 2006).

2.4 Use of Language in Personality Identification

One of the major sources of identifying a user's personality is by analyzing the text produced by him. A central assumption of language psychology is that the words people use reflect who they are. Psychologists have documented the existence of personality cues in text by discovering correlations between a range of linguistic variables and personality traits, across a wide range of linguistic levels, including acoustic parameters (Smith et al. 1975; Scherer 1979), lexical categories (Pennebaker and King 1999; Pennebaker et al. 2003; Mehl et al. 2006; Fast and Funder 2007), *n*-grams (Oberlander and Gill 2006), and speech-act type (Vogel and Vogel 1986).

Analysis of word frequencies have been used for different applications (called Linguistic Fingerprinting) like to distinguish letters written by soldiers in 1800s (Broehl and McGee 1981); to understand speaking styles of political leaders (Hart 1984); to establish the identities of authors of biblical and literary works; and to determine the anonymous author of a best-selling book surrounding the presidency of Bill Clinton (Foster 1996). One of the first and most successful word-based approaches was General Inquirer developed by Stone et al. (1966). In order to provide an efficient and effective method for studying the various emotional, cognitive, structural, and process components present in individuals' verbal and written speech samples, Pennebaker and Francis developed a word based text analysis application called Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al. 2001).

As a first step in personality identification from language, Pennebaker and King (1999) investigated whether language use reflects personality. They concluded that the ways people express

themselves in words are remarkably reliable across time and situations; despite the different ranges of writing topics, word use remained reliable; and language use may be thought of as an arena in which the impact of the person is unavoidable. Gill and Oberlander (2002) studied the use of more complex phrases as markers of personality. Argamon et al. (2005) worked on distinguishing high from low neuroticism and extraversion in authors of informal text based on self-reports of personality. Taking this a step further in their seminal first reported work, Mairesse and Walker (2006) developed models for personality recognition based on user's language as revealed in his essays and conversation extracts. They concluded that personality can be recognized by computers through language cues; extraversion, emotional stability, conscientiousness are easier to model; recognition models based on observed personality perform better than a baseline of average personality score as well as models using self-reports; and spoken language is easier to model than written text, since the conversation extracts are less formal than the essays, and personality may be best observed in the absence of behavioral constraints. Oberlander and Nowson (2006) classified extraversion, stability, agreeableness, and conscientiousness of blog authors using *n*-grams as features and Naïve Bayes (NB) as learning algorithm. Oberlander and Gill (2006) used content analysis tools and *n*-gram language models to identify markers in extravert and introvert emails. Mehl et al. (2006) showed that some linguistic cues vary greatly across gender and hence including gender in the models would produce better results. Mairesse et al. (2007) extended previous study by Mairesse and Walker (2006) to include ranking models in addition to classification and regression models. They systematically examined the use of different feature sets, suggested by psycholinguistic research. They predicted both personality scores and classes using Support Vector Machines (SVMs) and M5 trees, respectively. They also reported a long list of correlations between Big5 personality traits and two lexical resources they used.

A summary of features used for PRT (linguistic features) is presented in Section 4.1 and later, in Section 7, we present which features are most useful in personality recognition from online social networks.

3 ONLINE SOCIAL NETWORKS AND PERSONALITY

Almost half of the people in the world use Internet (http://www.internetworldstats.com/stats.htm). Not all, but most of them spend lots of time online. According to the survey of Pew Research Center, the widely taken actions on the Internet are sending or reading e-mails, searching for information (e.g., medical information, information about a hobby or interest, information about some people), getting news, blogging, using status update services (e.g., Facebook, Twitter), and so on (http://www.pewinternet.org/Trend-Data/Online-Activities-Total.aspx). Over the past 10 years, the internet usage has increased by 450%. In addition to this, 20% of current internet users do not use it as a read-only source; but also contribute data (http://www.pewinternet.org/Static-Pages/Trend-Data/Usage-Over-Time.aspx). One aspect of our daily existence where the Internet has introduced major changes is our social lives (Hamburger et al. 2002; Hamburger and Ben 2000). Social life on the Internet initially comprised social tools, such as chat forums and newsgroups. Today it has developed many additional components, such as blogs, fantasy environments, and social networks.

Social Network Sites (SNSs) are huge, virtually infinite, corpora where authors (users) and sentences (posts) are found together. Boyd and Ellison (2007) characterize SNSs as web-based services that allow individuals to (1) construct a web presence usually including a photo and descriptors like location, age, study concentration, and interests, (2) publicly display a list of other users with whom they share a connection, and (3) to traverse those list of connections to view the profiles of others within the system. By the second quarter of 2008, Forrester Research estimated 75% of Internet users were involved in some sort of "social media" (Kaplan and Haenlein 2010).

One of the most ubiquitous on-line environments, Facebook, is becoming an increasingly natural environment for a growing fraction of the world's population. In January 2005, a survey of social networking websites estimated that among all sites on the web there were roughly 115 million members (Golbeck 2005). In 2012, Facebook alone counted over 900 million active users, who can create personal profiles and communicate with friends and other users through private or public messages and receive information about their friends by means of a news feed timeline. To allay concerns about privacy, Facebook enables users to choose their own privacy settings and choose who can see specific parts of their profile. Only a user's name and profile picture are required to be accessible by everyone. The rest of the information on users' pages is by default visible only to friends or to friends-of-friends. Though it is quite challenging to extract data from it for its privacy policy, it is one of the largest and general purpose existing social networks online.

Through social media such as Facebook and Twitter, used regularly by more than 1/7th of the world's population (http://newsroom.fb.com), variation in mood has been tracked diurnally and across seasons (Golder and Macy 2011), used to predict the stock market (Bollen Mao Zeng 2011), and leveraged to estimate happiness across time (Kramer 2010; Dodds et al. 2011). Search patterns on Google detect influenza epidemics weeks before The Centers for Disease Control and Prevention (CDC) data confirm them (Ginsberg et al. 2009), and the digitization of books makes possible the quantitative tracking of cultural trends over decades (Michel et al. 2011).

Personality of an individual gets reflected in the environment around him. In the context of physical environments, like bedrooms and offices, Gosling et al. (2002) proposed two mechanisms by which an individual's personality can become expressed in an environment: identity claims and behavioral residue. Identity claims are the symbolic declarations that individuals make to themselves or others in an attempt to convey how they would like to be seen. Examples of identity claims range from subtle clues found in an individual's clothing choice to more direct claims, like bumper stickers or explicit verbal statements made about beliefs. Behavioral residue refers to the inadvertent clues left by one's behavior. For example, a neatly organized movie collection reflects an individual's tendency to organize, even if the organizing behavior was not performed to specifically to convey that information. It can be applied to other contexts of expression too, such as music preferences (Rentfrow and Gosling 2006), everyday behavior (Mehl et al. 2006), and personal websites (Vazire and Gosling 2004). Just as physical environments, personality has been seen to reflect in virtual environments as well, including blog entries, behavior on web and online social networks.

Thus, researchers have started looking into the possibility of predicting a user's personality based on his online social networking profile. Social networking sites are widely used as a primary medium for communication and networking (Boyd and Ellison 2007; Valkenburg and Peter 2009). They have become a central, virtually unavoidable medium for social interactions. In the process of creating social networking profiles, users reveal a lot about themselves both in what they share and how they say it. Through self-description, status updates, photos, and interests, much of a user's personality comes out through their profile. As people spend more and more time on online social networks like Facebook, Twitter, MySpace, LinkedIn, and so on, they leave behind, implicitly and explicitly, a lot of valuable information. Such information includes published interests, attributes, and social interactions in the form of structural features like personal network topology, degree, centrality, cohesion, and so on; behavioral features like frequency of use, temporal aspects, status messages, likes, reciprocity, attention, latency, and so on; and profile features like number of friends, preferences, language, profile picture, and so on.

Online social networks, for sure, is drawing more and more participation from users and is increasingly becoming an important source of large amounts of data. However, does this alone justify its use for personality identification? Does social network data indicate a user's personality?

Impression management (IM) refers to the attempt to control information in order to affect others' opinions of us. People on Facebook and other social networks do not lie as such, but rather stretch the truth (sometimes to its outer limits). As befits the slogan "tell me who your friends are and I will tell you who you are," people try to have the right list of friends in order to create their desired image. In addition, a widely held assumption, supported by content analyses, suggests that Online Social Networks (OSN) profiles are used to create and communicate idealized selves. According to this idealized virtual-identity hypothesis, profile owners display idealized characteristics that do not reflect their actual personalities. In such circumstances, even if identified, how trustworthy will such identification of personality from social networks be?

A contrasting view holds that OSNs may constitute an extended social context in which to express one's actual personality characteristics, thus fostering accurate interpersonal perceptions. OSNs integrate various sources of personal information that mirror those found in personal environments, private thoughts, facial images, and social behavior, all of which are known to contain valid information about personality (Vazire and Gosling 2004). Moreover, creating idealized identities should be hard to accomplish because (a) OSN profiles include information about one's reputation that is difficult to control (e.g., wall posts) and (b) friends provide accountability and subtle feedback on one's profile. Accordingly, the extended real-life hypothesis predicts that people use OSNs to communicate their real personality. Fortunately, researchers have carried out studies proving that personality is indeed revealed in users' actions online, thus making it possible to identify personality by studying online social networking profiles of users. Research has shown that distinctive characteristics of one's personality are more likely to manifest themselves in situations that satisfy individuals' basic psychological needs (Sherman et al. 2012). These needs are summarized as relatedness to others, competence, and autonomy, which social networking sites like Facebook and Twitter are well positioned to satisfy. As opposed to sites that invite more adversarial discussion like political blogs, sites like Facebook and Twitter allow one to create their own circle and communicate only with those in their circle.

Back et al. (2009) tested the two competing hypothesis and their results were consistent with the extended real-life hypothesis and contrary to the idealized virtual-identity hypothesis. In another study, Gosling et al. (2007) studied for the first time, how accurate are the impressions based on OSN profiles? Their results suggest that the online social networking websites are, in fact, a relevant and valid means of communicating personality (particularly for Extraversion). Profile authors did engage in some self-enhancement for the Big Five domains of Emotional Stability and Openness to Experience.

Work on personality recognition using social network analysis began in around 2008 when researchers started finding correlations between personality traits and social network characteristics. This led them to go a step further and try and develop classification or regression models by which, given the social network characteristics of a user, one would be able to predict his personality traits. In the following sections, we first present the various features that have played role in such experiments of past research followed by a review of early researches establishing correlations between these features and personality. After that we present a review of the emerging trends in automatic personality recognition through social network analysis.

4 FEATURES THAT CAN BE EXTRACTED FROM ONLINE SOCIAL NETWORKING PROFILES

For the purpose of studying correlations with personality or to automatically identify a user's personality using supervised or unsupervised learning approaches, one needs to begin with a set of features associated with an online social networking profile that can offer insights into a user's personality. These features can be categorized into linguistic features, extracted from the textual

content available in user's profile in the form of status updates, interests, about me blurbs, and so on, and non-linguistic features that are other structural, behavioral, or temporal attributes of a user's profile.

4.1 Linguistic Features

Several linguistic features can be extracted from status updates and other textual attributes of an online social networking profile, which serve as a rich source of text associated with a user.

LIWC Features: In order to provide an efficient and effective method for studying the various emotional, cognitive, structural, and process components present in individuals' verbal and written speech samples, Pennebaker and Francis developed a word based text analysis application called LIWC (Pennebaker et al. 2001). LIWC produces statistics on 81 different features of text in five categories. These include Standard Counts (word count, words longer than six letters, number of prepositions, etc.), Psychological Processes (emotional, cognitive, sensory, and social processes), Relativity (words about time, the past, the future), Personal Concerns (such as occupation, financial issues, health), and Other dimensions (counts of various types of punctuation, swear words).

MRC Features (http://www.psych.rl.ac.uk/MRC_Psych_Db_files/mrc2.html): MRC features can be computed using Medical Research Council's psycho linguistic database (Coltheart 1981) that is a list of over 150,000 words with linguistic and psycholinguistic features of each word. These include Kucera–Francis written frequency (Kucera and Francis 1967), number of categories, and number of samples; Brown verbal frequency, i.e., frequency of occurrence in verbal language derived from the London-Lund Corpus of English Conversation (Brown 1984); Familiarity rating; Meaningfulness via Colorado norms and via Paivio Norms; Concreteness; age of acquisition; Thorndike-Lorge written frequency; and the number of letters, phonemes, and syllables.

Speech acts: Speech acts are a "basic unit of human linguistic communication" (Searle 1975, 1976) and can be used for categorizing human conversational utterances. There have been five basic types of speech acts proposed: assertives, commissives, declaratives, directives, and expressives. For example, a piece of text can be analyzed for the following symptoms to extract the appropriate speech act—Ends with Period? Has a question word? Has a Copula Verb? Has an exclamation mark? Begins with Copula Verb? Has sentiment-laden words? Has a Question Mark? Contains emoticons? Alternate speech act categories have been proposed by Walker and Whittaker (1990) as command, prompt, question, and assertion.

Parts-of-Speech (POS) tags: POS tag features include the numbers and average numbers of words in specific grammar categories (e.g., adverbs, adjectives, verbs, pronouns).

H4LvD features: H4LvD features are computed using General Inquirer developed by Stone et al. (1966). General Inquirer is a tool for content analyses of textual data, which include 182 word tag categories merging four sources (www.wjh.harvard.edu/~inquirer/Home.html), including HIV-4 and Lasswell value dictionary. A word is classified using an intensity level scale, which is a combination of different valence categories, such as positive vs. negative, strong vs. weak, and active vs. passive. The H4Lvd categories span from affective words and motivation to socially-related and communication-specific ones. H4Lvd as compared to LIWC has the larger number of words per category and finer sophistication of subcategories.

Sentiment features: These features indicate positive or negative sentiment strength of a status update using Afinn words (Nielsen 2011), a list of 2,500 words, each annotated with emotional valence ranking in the range of −5 to 5.

Other linguistic features that can be extracted from texts include *n*-grams (following bag-of-words approach), computational stylometrics, and systemic functional grammar—like appraisal use, function words, and so on.

4.2 Non-linguistic Features

Several other structural, behavioral, and temporal features of an online social networking profile can be extracted to understand their correlations with personality and their utility in predicting the user's personality.

Structural features: These may include number of friends, density of network, i.e., what percentage of possible edges between friends exist, centrality and clustering measures, betweenness, brokerage (Hanneman and Riddle 2005)

Behavioral features: These may include personal network type, associations with groups, degree of revealing private information, number of different functionalities used, photos displayed on the individual profiles (as they constitute an important way to project the image we wish to present to others), number of pictures, presence of self-picture (a selfie), and a set of several other behavioral measures introduced by Adal and Golbeck (2012).

Temporal features: These may include contact frequency, frequency of accepting/rejecting friendship offer, difference in number of friends over a period of time, frequency of status updates per day, or number of status updates posted between, say, 6am to 9am in morning.

In addition, the profiles contain a lot of other information that could have interesting connections with personality. For example, gender, birth dates, marital status, home town, home neighborhood, political views, religious views, family members, activities, interests, activities, interests, favorite music, favorite TV shows, favorite movies, favorite books, favorite quotations, about me, contact information, education and work information.

Features can also be categorized as those that depend exclusively on a user's actions and those that depend on actions of friends or another user. The former include the number of published photos, events, and groups the user has uploaded or created and the number of objects the user has "liked." The latter includes aspects of the profile like the number of times a user has been tagged in photos, and the size and density of their friendship network.

5 STUDIES CORRELATING PERSONALITY AND SOCIAL NETWORKING FEATURES

As a first step toward identification of personality from online social networks, researchers started looking into the correlations of the big five personality traits with some of the online social networking profile features as mentioned above. When it comes to textual features, most work seems to include LIWC features in their study owing to the work of Pennebaker and King (1999) wherein they studied linguistic styles to understand language use as an individual difference. They found and published correlations between LIWC features and the five factor scores. Similarly, Mairesse et al. (2007) established correlations between MRC features and personality traits, thus motivating researchers to consider MRC features in their work of personality identification.

While some researchers studied the effect of personality on social networking features, thereby using personality as a predictor for them, others studied the role of such features in predicting personality. In the first category, Wehrli (2008) for the first time in 2008 explored how individual personality characteristics influence online social networking behavior. They regressed exogenous personality traits on behavioral and structural indicators. They applied maximum likelihood estimation and simple OLS regressions and concluded that personality seems to be predictive in how many persons someone meets, what position the person takes in the network, but not in the type of personality the person's interaction partners will have. In 2008 itself, Doeven-Eggens et al. (2008) for the first time studied the effect of personality on personal network type (defined as the set of ties surrounding individuals (Marsden 1990). They reported that extroverts tend to have more peer oriented personal network, autonomous people tend to have a mixed personal network and conscientious people tend to have primarily family oriented personal network.

Later, in 2011, Quercia et al. (2012) studied the relationship between Facebook popularity (number of contacts) and personality traits on a large number of subjects. They analyzed data from a Facebook (FB) application called myPersonality. They found that the predictor for number of friends in the real world (Extraversion) is also a predictor for number of Facebook contacts. They also tested whether people who have many social contacts on Facebook are the ones who are able to adapt themselves to new forms of communication, present themselves in likable ways, and have propensity to maintain superficial relationships. They showed that there is no statistical evidence to support such a conjecture. Their individual-to-individual correlations are all weak, yet, at population-level, they report, for the first time, a clear linear relationship between number of contacts and Extraversion. They have shown that, from personality scores, one is then able to partially predict the number of social contacts on Facebook based on the three variables Extraversion, Age, and Neuroticism. Through a similar work which they did on Twitter, they demonstrated relationship between personality and five types of twitter users—listeners (those who follow many users), popular (those who are followed by many), highly read (those who are often listed in others' reading lists), and two types of influential. In a parallel study by Gosling et al. (2011), they performed two descriptive exploratory studies to examine how traits are expressed on a range of self-reported activities on Facebook. Their results demonstrate that social and personality processes are alive and well in OSNs, and parallel the processes in non-virtual environments.

Falling in the second category, Ross et al. (2009) studied the connection between the personality of the individual users and their behavior on a social network. They found connection, but it was not strong. Unfortunately, this study was based on a relatively small ($n = 97$) and homogeneous sample (mostly female students of the same subject at the same university), which limited the power of their analyses and made it difficult to extrapolate their findings to a general population. Also, this study relied on participants' self-reports of their Facebook profile features, rather than direct observation. Consequently, Ross et al. were only able to present one significant correlation—between Extraversion and group membership, leaving all the remaining hypotheses unverified. Amichai-Hamburger et al. (2010) built upon the study of Ross et al. (2009) where the self-reports of subjects were replaced by more objective criteria, measurements of the user-information upload on Facebook. They found several significant relationships; however, their sample was still small ($n = 237$) and very homogeneous (Economics and Business Management students of an Israeli university). Moreover, some of their findings were opposite to those of Ross et al. For example, they find that Extraversion is positively correlated with the number of Facebook friends, but uncorrelated with the number of Facebook groups, whereas Ross et al. found that Extraversion has an effect on group membership, but not on the number of friends. Also, they find that high Neuroticism is positively correlated with users posting their own photo, but negatively correlated with uploading photos in general, while Ross et al. posit that high Neuroticism is negatively correlated with users posting their own photo.

Later, in 2012, Bachrach et al. (2012) also worked on examining correlations between users' personality and the properties of their Facebook profiles. They demonstrated how multivariate regression allows prediction of the personality traits of an individual user given their Facebook profile. They argue that the overlap between Facebook profile features that contain the actual personality cues and features used by people to form personality judgments does not have to be perfect. Their work is very similar to work by Gosling et al. (2011) (based on self-reported and observable Facebook profile information) but with a much larger sample size. While most previous work focused on correlating Facebook profile features with personality traits averaged over large groups, they were inaccurate on the individual level. Bachrach et al. however used their large sample to show that by combining signals from different Facebook features it is possible to reliably predict personality of individuals. Their dataset consists of the personality profiles and

Facebook profile data of 180,000 users. Due to the difficulty and cost of testing large samples using a laboratory approach, they used viral marketing to collect personality data using an application within the Facebook environment.

Another interesting methodology for finding effect of personality traits on language of social media is proposed by Schwartz et al. (2013). They use an "open vocabulary" technique where the data itself drives a comprehensive exploration of language that distinguishes people, finding connections that are not captured with traditional "closed-vocabulary" word-category analyses. Theirs is the largest known study, by an order of magnitude, of language, and personality. They collected 700 million words, phrases, and topic instances from the Facebook messages of 75,000 volunteers, who also took standard personality test. They extract a data-driven collection of words, phrases, and topics, in which the lexicon is based on the words of the text being analyzed. They use differential language analysis (DLA), their particular method of open-vocabulary analysis, to find language features across millions of Facebook messages that distinguish demographic and psychological attributes.

6 PERSONALITY IDENTIFICATION FROM ONLINE SOCIAL NETWORKING PROFILES

As discussed earlier, traditionally the only way personalities were identified was through questionnaire-based personality tests that the subjects used to undergo and such methods are costly, error-prone, and laborious. More recently social networks have been used as a rich source of identifying a user's personality. It has been recently established that computer-based personality judgments are more accurate than those made by humans (Youyou et al. 2015). This ability to identify personality from a user's online social network data comes as a welcome relief. However, the psychological experiments involved in such an approach also have their own bottlenecks. When experiment participants upload the self-report data, they could have reflected self-views rather than actual behavior. Other data collection methods such as observable information profile cost a lot of manual resources and are not desirable for large scale of dataset collection. This triggers the need of automatic approaches for personality identification.

First, in Table 1, we give an overview of different studies available in the literature that talk about automatic identification of personality from social networking profiles. We then present, in chronological order, the studies with respect to different features used and the machine learning approaches they employ to build prediction models. We end this section by presenting a discussion on special contributions or key results of some particular benchmarking studies.

6.1 Features

In terms of features used, as presented in Table 1, several studies have employed the standard textual features like LIWC, MRC, bag of words, POS tags, and so on and standard non-textual features like personal information, demographic data, number of friends, and so on. However, certain studies have explored the possibility of other features. On YouTube personality dataset (https://www.idiap.ch/dataset/youtube-personality), Sarkar et al. (2014) have used logistic regression model with a ridge estimator for the classification and experimented with audio-visual features, bag of word features, sentiment based, and demographic features and have provided important insights about the significance of different feature types for personality classification task. Celli et al. used bag-of-visual-words technique to extract features from Facebook profile pictures (Celli et al. 2014). Quercia et al. (2011), however, made use of publicly available counts of—what Twitter calls—"following," "followers," and "listed" counts, to identify three types of users: listeners (those who follow many users), popular (those who are followed by many), and highly read (those who are often listed in others' reading lists). Bai et al. (2012) added emotion and time-related

Table 1. Overview of Various Studies for Personality Recognition from Online Social Networks

Paper	Quick summary	Data	Features used and feature selection	Machine learning algorithms
Golbeck et al. (2011)	First work on predicting personality using FB profiles	Administered 45-question version of the Big Five Personality Inventory to 279 subjects through a Facebook application	A total of 161 statistics in following categories—structural, personal, activities and preferences, language features (LIWC), internal FB stats. A number of features were reduced to 74 per user after pre-processing.	Regression analysis in Weka with a 10-fold cross-validation with 10 iterations using two algorithms: M5′ Rules, a rule-based variation of the M5′ algorithm, and Gaussian Processes
Quercia et al. (2011)	First work on Twitter. Studied relationship between personality and five types of twitter users and suggested a way to predict personality using three counts in Twitter—listeners, popular, highly-read, and two types of influential (one is based on Klout score—klout.com—and the other is based on a measure used by TIME magazine to rank public figures)	335 Twitter users who also had FB accounts. Collected personality data from myPersonality FB app.	Studied the Pearson product-moment correlation between the logarithm of the five user characteristics and each of the (big) five personality traits, plus two additional attributes, namely age and sex.	Regression analysis with a 10-fold cross-validation with 10 iterations using M5 Rules
Daniel Chapsky (2011)	Presented initial work done on generating a probabilistic model of personality, which uses representations of people's connections to other people, places, cultures, and ideas, as expressed through Facebook.	Facebook data of 615 FB users combined with external data sources to allow further inference. Used FB App called Bayesian Personality for collecting user data. They used the NEO Personality Inventory for personality scores.	Also included home town data, movie interests, and music interests as features	Bayesian Network
Golbeck et al. (2011)	Predicting personality from publicly available information on their Twitter profile	Administered the Big Five Personality Inventory to 279 subjects through a Twitter application. Gathered their 2,000 most recent public Twitter posts (tweets).	Used MRC in addition to LIWC for language features. They also performed a word by word sentiment analysis of each user's tweets using General Inquirer dataset	Two regression algorithms: Gaussian Process and ZeroR, each with a 10-fold cross-validation with 10 iterations

(Continued)

Table 1. Continued

Paper	Quick summary	Data	Features used and feature selection	Machine learning algorithms
Bai et al. (2012)	Prediction of personality from RenRen, a Chinese SNS	Used the 44 questions Big Five Inventory (BFI), designed by Berkeley Personality Lab	Used 41 features—basic info, SNS usage, time related usage, emotion related usage, time & emotion related usage. Used C4.5 decision tree which uses Gain Ratio to extract features. Used their Naïve Bayesian method based emotion predictor to classify an article into different emotion according to its content.	Many classification algorithms, such as Naive Bayesian (NB), Support Vector Machine (SVM), Decision Tree, and so on.
Adal and Golbeck (2012)	They examine to which degree behavioral measures can be used to predict personality.	Used 44-question version of the Big Five inventory	Introduced a number of measures that are based on the intensity and number of social interactions one has with friends along a number of dimensions such as reciprocity and priority	Two regression algorithms: Gaussian Process and ZeroR
Bachrach et al. (2012)	Used large sample to show that by combining signals from different Facebook features it is possible to reliably predict personality of individuals	Personality profiles and Facebook profile data of 180,000 users collected through myPersonality FB application	Two broad categories—aspects of the profile that depend exclusively on a user's actions (e.g., no. of photos uploaded) and aspects of the profile that depend on the actions of a user and their friends (e.g., no. of times user has been tagged in photos)	Multi-variate linear regression. Also, applied several more sophisticated machine learning methods for predicting traits, including tree based rule-sets, support vector machines, and decision stumps.
Wald et al. (2012)	Predict users' personality traits using only demographic and text-based attributes extracted from their profiles	Collected personality data using the Big-Five Personality Index. Their study was based on 537 FB users.	Included 31 demographic attributes for each individual and 80 text based attributes using LIWC	Linear regression, REPTree, and decision tables
Kosinski et al. (2013)	Showed that Facebook Likes, can be used to automatically and accurately predict a range of highly sensitive personal attributes	Dataset of over 58,000 volunteers who provided their Facebook Likes, detailed demographic profiles, and the results of several psychometric tests.	Facebook "Likes" information of each user	Numeric variables were predicted using a linear regression model, whereas dichotomous variables such as gender or sexual orientation were predicted using logistic regression

(Continued)

Table 1. Continued

Paper	Quick summary	Data	Features used and feature selection	Machine learning algorithms
Verhoeven et al. (2013)	Report a proof of concept of using ensemble learning as a way to alleviate the data acquisition problem	Used myPersonality dataset of FB profiles and essays labeled with personality scores	Used 2,000 frequent trigrams as initial features taken from the Essays corpus, then exploited ensemble methods based on meta-learning to generalize features across genres and trained SVMs classifiers	Ensemble methods to improve the accuracy by combining the predictions of different component classifiers
Farnadi et al. (2013)	Proposed F-measure weighted by class-size average as evaluation measure and tested cross-domain learning	Used myPersonality dataset of FB profiles and essays labeled with personality scores	Used four different sets of features: lexical (LIWC), network measures (social), status update timestamps (time), and other measures like posts per user, capital letters, repeated words (others)	Three different learning algorithms: NB, SVMs, and Nearest Neighbors (kNN)
Markovikj et al. (2013)	Exploited rich linguistic patterns and used large feature space for personality prediction.	Used myPersonality dataset of FB profiles	Used a very large feature space (725 features), including social and demographic info, lexical resources, Part Of Speech Tags, word emotional values (AFINN), and word intensity scale (H4Lvd). Used ranking algorithms for feature selection.	SVMs and Boosting (MultiBoostAB and AdaBoostM1)
Alam et al. (2013)	Followed bag of words approach to predict personality	Used myPersonality dataset of FB profiles	Followed bag of words approach and used unigrams as features	SVMs (SMO with linear kernel), Bayesian Logistic Regression (BLR), and Multinomial Naïve Bayes (mNB) sparse model

(Continued)

Table 1. Continued

Paper	Quick summary	Data	Features used and feature selection	Machine learning algorithms
Appling et al. (2013)	Used speech acts in status updates to predict personality	Used myPersonality dataset of FB profiles	Focused on simple speech act features in text like Ends with Period? Has a question word? Has a Copula Verb? Has an exclamation mark? Begins with Copula Verb? Has sentiment-laden words? Has a Question Mark? Contains emoticons?	Multiple regression analysis in which each personality trait was individually regressed onto all speech acts at once

features. Their time-related features included the features that correlated with the recent psychological state such as status or blog publishing count during recent one month; emotion related features included features that related with the emotion distribution (angry, funny, surprised, and moving) of the user such as the top emotion count of all her blog; and time and emotion related features included recent emotion tendency of the user such as the status emotion of the newest status and its emotion length (time sustained of the recent emotion). Adal and Golbeck's (2012) work is noteworthy in the sense that they introduced a number of new behavioral measures that are based on the intensity and number of social interactions one has with friends along a number of dimensions, such as reciprocity and priority. In using speech acts of status updates as features, Appling et al. (2013) suggested a meta-level approach as against the content-level approach used by other studies. Content approaches make use of the particular topics and the description of those topics through words associated with specific psychological phenomena. Such approaches require human judges to rate whether or not content should be included in a particular dictionary. This process, though producing high quality association lists, requires continual maintenance and versioning from human experts as language use changes over time and new words and expressions come into and out of use. Appling's approach however, makes use of meta-level indicators of speech act phenomena to predict personality dimensions.

6.2 Supervised Approaches

In terms of machine learning algorithms used and results obtained, most studies have used different classification and/or regression models to predict personality. Golbeck et al. (2011) reported accuracy within 11% of actual values with M5′ Rules performing better than Gaussian Processes. It is important to note, however, that their sample ($n = 167$) was very small, especially given the number of features used in prediction ($m = 74$), which limits the reliability and generalizability of their results. Working on Twitter data, Quercia et al. (2011) showed that, based on three publicly available counts, they can accurately predict users' personality as good as state-of-the-art recommender systems predict user ratings for movies. Golbeck et al. (2011) reported similar performance of both Gaussian Process and ZeroR when working on Twitter data. While working with time and emotion related features on a Chinese online social network, Bai et al. (2012) found out that C4.5 Decision Tree can get the best results.

In their work focused on behavioral measures, Adal and Golbeck (2012) report that timing between messages, text length, and propagations appear to be the most informative features for

personality. Also, they show that studying friends' behavior in general provides with many useful features for understanding the person's personality from a normative perspective. However, performance was unacceptably low for some traits, especially Neuroticism and Extraversion. Results of Bachrach et al. (2012) showed that accuracy of prediction for Extraversion is greater than Neuroticism, which is greater than Conscientiousness which in turn is greater than Openness and Agreeableness which turned out to be the poorest trait to predict. Interestingly, they also report that, for all of the personality traits, RMSE values change very little when using more sophisticated machine learning methods. While working with only demographic and text-based attributes, Wald et al. (2012) found REPTree model to perform best among linear regression, REPTree, and decision tables.

Chapsky (2011), however, took a different approach and took a first step toward a more holistic approach toward personality and its effects. He argued that a model predicting personality based on online data which cannot be extrapolated to "real world" situations is of limited utility for researchers. Thus, he presented initial work done on generating a probabilistic model of personality, which uses representations of people's connections to other people, places, cultures, and ideas, as expressed through Facebook using Bayesian Network. His main motivation behind modeling Bayesian Network was that knowing how a prediction of personality is made can be as interesting as the prediction itself.

6.3 Unsupervised and Semi-Supervised Approaches

All of the above studies have essentially used supervised approach for personality identification. The main problems, however, with supervised approach are the limitations in data annotation and language dependency. Data labeled with personality types are usually costly and time consuming to collect. Driven by these limitations, for the first time, Celli (2011, 2012) proposed an unsupervised approach for personality prediction. Given as input a database of users and posts, the system outputs one personality model for each user and gives accuracy (measure of reliability of personality model) and validity (measure of variability of personality traits within each user—how much the model is valid across posts of the same user) as evaluation measures. This is based on the assumption that one user has one and only one complex personality, and that this personality emerges at various levels from written text as well as from other extralinguistic cues. They worked on a SNS called FriendFeed and focused their study on 156 users. They used a subset of 12 features from Mairesse et al. (2007) along with its reported correlations with personality traits. For each dimension the possible labels are yes, no, and balanced. If a sentence shows a feature whose frequency is higher than mean + standard deviation and it correlates positively with one personality trait, add +0.1 to that personality trait. If it correlates negatively, add −0.1 to that personality trait. Then they convert numerical value into nominal by the following scheme: >0 means y, <0 means n, and 0 means o. Majority class of each personality trait is then calculated for each user and that is the user's personality. Evaluation is carried out comparing many posts of the same user under the assumptions that one user has one and only one personality and this personality emerges at different degrees from user's posts. Celli proposed two measures of performance, accuracy and validity as follows.

Accuracy (a) = (tp + tn) / (tp + tn + fp + fn) and,
Validity = 1 − (a/P) where,
P = no. of posts of one user
tp = sum of each post personality attribute matching within the same user (y–y, n–n, o–o)
tn = sum of opposite attributes within the same user (y–n and n–y)
fp = sum of possible attributes turned to the balance value within the same user (y to o and n to o) calculating the majority class for each attribute

fn = sum of possible attributes turned to positive or negative (o to y or o to n) in order to calculate majority class for each attribute

They reported mean accuracy of 0.631 and mean validity is 0.729. They conclude that accuracy is in line with the classification accuracies reported by Mairesse et al. for observer ratings evaluation thus establishing the feasibility of such an unsupervised approach. In another work on twitter, Celli (2011) used this approach for automatically annotating 25,700 tweets with personality labels.

We came across one semi-supervised approach proposed by Lima and Castro (2014), and it is different from others in that it works with groups of texts, instead of single texts, and does not take users' profiles into account. Also, their approach extracts meta-attributes from texts and does not work directly with the content of the message. They have reported a prediction accuracy of 83%.

6.4 Studies Toward Providing a Benchmark

In 2013, Workshop on Computational Personality Recognition (Celli et al. 2013), a shared task was organized by Celli et al. to provide a benchmark for discovering which feature sets, resources, and learning techniques are useful in the extraction of personality from text and from social network data. They released two datasets, different in size and domain, annotated with gold standard personality labels. The dataset used for this workshop is a subset (250 users and about 9,900 status updates) of the myPersonality sample. Eight teams participated in the workshop and overall results can be summarized as follows: (a) feature selection with ranking algorithms over a large initial feature space is very effective (Markovikj et al. 2013); (b) bottom-up approaches based solely on words (unigrams, bigram, trigrams) are not very effective, and seem to work best with probabilistic algorithms, like NB (Alam et al. 2013); (c) top-down approaches, based on lexical resources (including the ones for sentiment analysis) and social info, in general seem to help personality recognition more than bottom-up approaches, based on words or *n*-grams; (d) ensemble methods are effective (Verhoeven et al. 2013); and (e) cross domain learning is possible (Farnadi et al. 2013).

Between the two approaches—top-down (making heavy use of external resources, such as LIWC and MRC, and testing the correlations between those resources and personality traits) and bottom-up (starting from the data and seeking for linguistic cues associated to personality traits), the former approach seems to achieve the best improvement from the baselines but it is more prone to over-fitting and should be done on very large corpora, while the latter is more robust but yields smaller improvements. However, the bottom-up approaches are better suited for handing internet slang, contractions, abbreviations, and emoticons which are increasingly becoming common in online social media language (unless the dictionaries used in top-down approaches are made to include such words).

7 TRAIT-WISE SUMMARY OF RESULTS OF CORRELATION AND PREDICTION STUDIES

Next, we present a trait-wise summary of results of various correlation and prediction studies. Extraversion/introversion dimension has received most attention as it is the most important one for discriminating between people (Peabody and Goldberg 1989) and hence maximum results have been reported for Extraversion. However, other traits have also received a fair amount of experimentation and the reported results are useful.

As far as "Openness to Experience" trait is concerned, it has been reported that it is positively correlated with preference for longer words and words expressing tentativity (e.g., perhaps and maybe) (Pennebaker and King 1999); the use of words related to insight (Mehl et al. 2006); articles, second person pronouns (You), and long words (Sixltr) (Mairesse et al. 2007); number of features

used from personal information section of FB (Hamburger et al. 2010); providing a URL to personal website (Golbeck et al. 2011); amount of personal information revealed on FB profile (Hamburger et al. 2010); frequency of status updates (Bai et al. 2012); number of users' likes, group associations and status updates (effect being strong for lower numbers of likes and saturates as the number of likes increases) (Bachrach et al. 2012); use of dictionary words, social interaction words, affective processes, cues associated with hearing, 2nd person singular, and 3rd person plural pronouns for updating statuses (Farnadi et al. 2013); and use of words from an artistic domain (e.g., soul, dreams, universe, music), use of more pronoun-type words (e.g., we're, you're, I've, I'll) (Schwartz et al. 2013).

"Openness to Experience" has been found not related to network characteristics (Wehrli 2008), strongly related to prosodic features (Mairesse and Walker 2006) and negatively correlated with high use of shorthand language (e.g., wat, ur, 2day), misspellings, and contractions (e.g., dont vs. don't) (Schwartz et al. 2013); use of 1st person singular pronouns and present tense forms (Pennebaker and King 1999); use of past tense (Mehl et al. 2006); density (Golbeck et al. 2011); and use of words about their occupations (Occup, Home, and School) and themselves (Self) (Mairesse et al. 2007). According to Markovikj et al. (2013), best predictors for "Openness to Experience" have been found to be geo-location, number of groups, punctuation, H4Lvd (weakness, wealth, enlightenment participants, aesthetic skills).

For "Conscientiousness," it has been reported that it is positively correlated with use of longer words (Mairesse and Walker 2006); words related to communication (e.g., talk and share) (Mairesse et al. 2007); words about work, insight words (e.g., think and know), words expressing positive feelings like happy and love (Mairesse et al. 2007); number of friends (Hamburger et al. 2010); words surrounding social processes, as well as the subset of words that describe people (Golbeck et al. 2011); frequency of using guestbook calling for help from others (Bai et al. 2012); number of uploaded photos (Bachrach et al. 2012); network size (Farnadi et al. 2013); and words reflecting achievement, school, and work (e.g., success, finals, to_work, work_tomorrow, long_day), and activities that support relaxation and balance (e.g., weekend, family, workout, vacation, day_off, lunch_with), and general enjoyment (e.g., much_fun, blessed, enjoying, wonderful) (Schwartz et al. 2013).

Conscientiousness is shown to be correlated with MRC features (Mairesse & Walker 2006); personal network structure (McCarty and Green 2005); and average valence of words (Markovikj et al. 2013). When regressed onto speech acts, it negatively predicted Assertives (Appling et al. 2013). It has been found to be negatively correlated with use of swear words and content related to sexuality (Mairesse and Walker 2006; Mairesse et al. 2007; Schwartz et al. 2013); use of negations, negative emotion words, and words reflecting discrepancies (e.g., should and would) (Pennebaker and King 1999; Mairesse et al. 2007); use of pronouns (Mairesse et al. 2007); frequency and duration of visiting social networking websites (Wehrli 2008; Gosling et al. 2011); use of picture upload feature (Hamburger et al. 2010); number of likes and group membership (Bachrach et al. 2012); and network density, updating of statuses between 00am and 11am, use of verbs, 1st person singular pronouns, or negative emotions (Farnadi et al. 2013). Those high on Conscientiousness are found to have family oriented personal networks (Doeven-Eggens et al. 2008). Moreover, according to Markovikj et al. (2013), best predictors for Conscientiousness are H4Lvd (reaping affect, submission to authority, dependence of others, vulnerability to others, social relations adjectives).

As far as "Extraversion" trait is concerned, it has been reported that it is positively correlated with use of positive emotion words and informal style (Pennebaker and King 1999); less formal phrases such as "Take care" and "Hi" (Oberlander and Gill 2006); number of friends (Hamburger et al. 2010; DeYoung 2010; Harrington et al. 2003; McCrae and Costa 1999; Golbeck et al. 2011, Quercia et al. 2012; Sigurdsson 1991; Golbeck and Hansen 2011; Wehrli 2008); centrality measures

(Wehrli 2008); number of groups (Ross et al. 2009); ease of use of social media (Roshchina et al. 2011); commenting on another's page (Gosling et al. 2011); use of emoticons (Bai et al. 2012); length of text (Adal and Golbeck 2012); degree of interaction by more sharing and likes (Bachrach et al. 2012); use of dictionary words, 2nd person and 3rd person singular pronouns, past tense verbs, social interaction words, cues associated with the five senses, health related words (Farnadi et al. 2013); and use of positive emotions words (e.g., ;), excited), and social words and phrases such as party, girls, and can't_wait (Schwartz et al. 2013).

LIWC features (use of social words, emotion words, first person pronouns, and present tense verbs), MRC features and prosodic features were found good to model Extraversion (Mairesse and Walker 2006). It was also found to be correlated with average valence of words (Markovikj et al. 2013). Extraversion is found to be negatively correlated with first person singular pronouns (e.g., I don't) and formal greetings (e.g., Hello) (Oberlander and Gill 2006); concreteness (Gill and Oberlander 2002); use of personal information, number of groups (Hamburger et al. 2010); transitivity, use of swear words (Farnadi et al. 2013); and words related to isolation and a focus on computer-related activities such as internet and reading (Schwartz et al. 2013). Extraverts are found to have peer oriented personal networks (Doeven-Eggens et al. 2008). Further, according to Markovikj et al. (Markovikj et al. 2013) the best predictors for Extraversion are brokerage, network size, punctuation, adjectives, verbs, H4Lvd (reaping affect, decrease as process, other relations).

For "Agreeableness" it has been reported that it is positively correlated with words using positive emotions (Penebaker and King 1999; Mairesse et al. 2007); length of words, tentative words like may be or perhaps (Mairesse et al. 2007); no. of pictures uploaded (but only among females) (Hamburger et al. 2010); use of affective process words (words describing feelings) in general, and positive emotion words in particular (Golbeck et al. 2011); chat activity, use of emoticons (Bai et al. 2012); number of tags (for users with more than 50 tags) (Bachrach et al. 2012); use of sexual words (Farnadi et al. 2013); and; use of religious words (e.g., prayer, church, god_bless), well-being (e.g., excited, wonderful, amazing, blessed), and positive social relationships (e.g., love_you_all, thank_you, friends_and_families) (Schwartz et al. 2013).

Agreeableness is also found to be correlated with personal network structure (McCarty and Green 2005) and average valence of words (Markovikj et al. 2013). It is negatively correlated with words using negative emotions, articles (Penebaker and King 1999; Mairesse et al. 2007); swearing words (Mehl et al. 2006; Mairesse et al. 2007; Schwartz et al. 2013); anger words (Mairesse et al. 2007); no. of page features used (Hamburger et al. 2010); number of likes (Bachrach et al. 2012); transitivity (Farnadi et al. 2013; Markovikj et al. 2013); use of assertives (Appling et al. 2013); and words reflecting aggressiveness, substance abuse, and other words reflecting a hostile approach to the world (e.g., kill, punch, knife, drunk, i_hate, racist, idiots). Speech acts are best to model agreeableness, among all traits (Mairesse and Walker 2006). Agreeableness has also been found to be unrelated with network characteristics (Wehrli 2008). Moreover, according to Markovikj et al. (2013), best predictors for Agreeableness are geo-location, Afinn number of words with valence +3/+5, transitivity, punctuation, "to," H4Lvd (qualities, senses-detectable degrees of qualities, shame).

As far as "Neuroticism" trait is concerned, it has been reported that it is positively correlated with use of negative appraisal words (like bad, ugly, evil) (Argamon et al. 2005); concreteness and use of more frequent words (Gill and Oberlander 2002); use of 1st person singular pronouns, negative emotion words (Pennebaker and King 1999); self-references (Pennebaker and King 1999; Mehl et al. 2006); frequency and duration of visiting social networking websites (Wehrli 2008); use of self-pictures in profile (Hamburger et al. 2010); frequency of words that express anxiety, character length of subject's last name (Golbeck et al. 2011); posting private information (Hamburger et al. 2010); content that makes people angry (Bai et al. 2012); standard deviation of text length and response time (Adal and Golbeck 2012); number of Facebook likes (Bachrach et al. 2012); number

of friends (when less than 200) (Bachrach et al. 2012); transitivity (Farnadi et al. 2013); use of anger words in updating statuses (Farnadi et al. 2013); and use of swear words reflecting depression, loneliness, worry, and psychosomatic symptom words (e.g., depressed, lonely, scared, headache) (Schwartz et al. 2013).

Neuroticism is also found to be correlated with LIWC features (Mairesse & Walker 2006) and average valence of words (Markovikj et al. 2013). It is found to be negatively correlated with use of positive appraisal words (like good, beautiful, nice) (Argamon et al. 2005); use of positive emotion words (Pennebaker and King 1999); uploading other pictures (Hamburger et al. 2010); number of Facebook contacts (Quercia et al. 2012); propagation (Adal and Golbeck 2012); number of friends (when greater than 200) (Bachrach et al. 2012); use of social interaction words, positive emotions or/and prepositions (Farnadi et al. 2013); use of commissives (Appling et al. 2013); and use of words reflecting positive social relationships (e.g., team, game, success), activities that could build life balance (e.g., blessed, beach, sports), and sport-related words (e.g., lakers, basketball, soccer) (Schwartz et al. 2013). Further, according to the study of Markovikj et al. (2013), best predictors for neuroticism are gender, network size, number of groups, number of tags, "to," H4Lvd (positive feelings, acceptance, appreciation, emotional support, enjoyment of a feeling, confidence, interest, and commitment).

8 DIRECTIONS FOR FUTURE RESEARCH

In Table 1, we provided an overview of the various studies that have been carried out for automatic identification of personality from social networking profiles. For each of the studies, we have listed the datasets used, features used, and machine learning algorithm employed, followed by a discussion on their major results. From the table, we can observe that many studies have been undertaken with little consensus on the choice of features and results obtained. Most studies use a different dataset, different set of features, and different data mining models. Surprisingly, few researchers compare their newly proposed techniques with prior studies. Also, it is not clear whether the accuracies are high enough to be useful in the context of a particular application domain. These observations lead to a number of directions for future research in this area, which can be categorized in the following groups, as discussed next.

Notion of personality and purpose of research

Most studies in personality identification using online social networking profiles have considered Big Five as the de facto standard model for personality modeling. However, Big Five has its own limitations (Eysenck 1991; Paunonen and Jackson 2000). Though there is a general agreement on the number of traits, there is no full agreement on their meaning, since some traits are vague. For example, there is some disagreement about how to interpret the openness factor, which is sometimes called "intellect" rather than openness to experience. One may also consider the correlation of these five personality dimensions, since these five dimensions are not absolutely orthogonal. Thus, future work should consider evaluating other models as alternatives or predicting additional attributes of personality like basic human values and fundamental needs in addition to the big five traits (Gou et al. 2014).

Further, as there are discrepancies between markers of self-assessed and observed personality (Mairesse and Walker 2006; Mairesse et al. 2007), another issue is the identification of the most appropriate model given a specific application. A hypothesis that remains to be tested is that traits with a high visibility (e.g., extraversion) are more accurately assessed using observer reports, as they tend to yield a higher inter-judge agreement, while low visibility traits (e.g., emotional stability) are better assessed by oneself. A personality recognizer aiming to estimate the true personality

would therefore have to switch from observer models to self-report models, depending on the trait under assessment.

On similar lines, it is important to ask whether a personality identification system claims to identify a user's identity or his reputation. Hogan (1982) introduced the distinction between the agent's and the observer's perspective in personality assessment. While the agent's perspective conceptually taps into a person's identity (or "personality from the inside"), the observer's perspective in contrast taps into a person's reputation (or "personality from the outside"). Both facets of personality have important psychological implications. A person's identity shapes the way the person experiences the world. A person's reputation, however, is psychologically not less important: it determines whether people get hired or fired (e.g., reputation of honesty), get married or divorced, get adored or stigmatized. Because it is harder to assess, this observer's perspective has received comparatively little attention in psychology. Given that in everyday life people act as observers of other people's behaviors most of the time, the external perspective naturally has both high theoretical importance and social relevance (Hogan 1982).

A fundamental assumption is worth investigating—does style of expression reveal personality disposition or mere verbal ability? A person's linguistic skills or verbal ability need not necessarily be due to his personality disposition (for example, if he has received a particular training in communication skills), in which case these two may not be correlated. The correlations revealed in the literature are rather weak and may need further investigation.

As we saw earlier (in Section 2), personality identification is a very important area and is increasingly becoming prominent owing to the wide variety of its applications. However, typically, information of a user's personality is useful only to the extent it can predict a specific behavior in a particular context. For example, ability to predict a user's shopping habits or his preference for user interfaces or his music preferences. This leads one to think that, instead of trying to predict "personality" per se, which is a complex attribute to predict accurately, one might as well try to predict these specific behaviors for a given context which might be simpler and more accurate to predict, for example, the ability to predict mental disorders or ability to identify suicidal tendencies, given an online social networking profile of a user. Future work can explore this line of thought.

Automatic personality recognition researches can also be extended to studying community characteristics, particularly from the perspective of participating personalities if the relationship between individual personalities and community structures can be better understood.

Use of Online Social Networks

Difference between online social networking platforms has to be recognized. For example, Facebook is a social networking site that generally connects people who already know each other (e.g., friends, family and co-workers)—the very default is that two individuals need to be mutual friends on Facebook to fully share what they have been up to. Instead, Twitter is a social media site on which users can see just about anything about anybody, unless they protect their updates, which only a very tiny minority of active users do (Meeder et al. 2010). It is yet to be understood if personality markers change across different social networks as a result of such differences.

Another perspective of using various online social networks is the possibility of profiling for personality recognition across various online social networks that is worth investigating. This may require coming up with ways to measure relevant features irrespective of the differences in details across various OSNs. There seems to be some work done on these lines (Song et al. 2015).

Challenges in data sampling

Collecting data from online social networking sites is a challenge—there are privacy issues, authentication issues, limitations of snowball crawling approach, and so on. This puts heavy restrictions

on data sampling. This challenge is yet to be addressed and is one of the biggest limitations restricting researchers from making a major contribution in this domain. This also creates a need of more published, well-defined datasets that can be used by many researchers and would provide a strong foundation for comparing results.

PRT in online social networks is a really challenging task. Posts are often very short and noisy, and normal tools for Natural Language Processing (NLP) often perform bad online (Maynard et al. 2012). One of the ways this problem is addressed in the literature is by aggregating many tweets or status updates from a user that gives more information (i.e., makes it statistically significant for lexical analysis). However, the resulting document per user is more of a series of disconnected statements rather than a coherent document as was used in other studies for PRT. The validity of such combining remains to be tested.

Further, the validity of any dataset is worth giving a thought. Any sample collected via a FB or Twitter application is a convenience sample of self-selected users interested in their personality, and so the results may not represent the Facebook population as a whole.

Another connected issue of data quality is the presence of fake and spam accounts on online social networks. A thorough study should eliminate possibility of a profile being fake or spam (Xiao et al. 2015; Cao et al. 2012)

Use of different feature sets

Use of LIWC does not come without its own set of limitations. LIWC was designed to study the language people use when writing about traumatic experiences. Hence, its application in any general context is questionable. Also, LIWC categories were somewhat arbitrarily created by the authors and their colleagues.

Studies can be performed on other textual contents of a user's profile on the social network, to extract some more previously unknown features to see their effect on personality identification. New set of features can be extracted and evaluated in their utility in identifying a user's personality. For example, change in number of friends over time or examining what has been "liked" by the user instead of just counting the number of "likes." The standard personality questionnaires or inventories can also serve as valuable insights in coming up with new features that can potentially be extracted from a user's online social networking profile.

Eliminating the interplay of other factors

Researchers can analyze if user's background information such as his gender, ethnicity, cultural and financial background, family size, and the extent to which he belongs to different SNSs has any effect on his personality so as to evaluate the consistency of features across these parameters and their effect on the personality identification. For example, Quercia et al. (2012) found negative correlations between age and number of Facebook contacts (i.e., younger people tend to have more social contacts on Facebook). However, by contrast, the opposite holds in Twitter—where a positive correlation between number of followers/following and age has been found (Quercia et al. 2011). These contrasting results reflect the different user demographics in the two platforms and hence indicate the importance of controlling for age in the experiments. Further, inclusion of gender as a feature would produce better models, as the actual language correlates of perceived personality were shown to depend on the gender of the speaker (Mehl et al. 2006).

Scherer (1979) showed that extraverts are perceived as talking louder and with a more nasal voice, and that American extraverts tend to make fewer pauses, while German extraverts produce more pauses than introverts. Thus, personality markers are culture-dependent, even among western societies. Corpora in other languages need to be included to study cross-lingual and cross-cultural effects on personality.

Another interesting dimension comes from the fact that it is important to study how people design and change their online social networking profiles over a significant period of time. Such a study, say over the period of a year, would enable us to increase our understanding of the long-term interaction between personality and the online social network dynamic. Effect of time on personality and how such a change is reflected in social network characteristics needs to be better understood. Studies have shown that while certain aspects of personality remain fixed (Roberts et al. 2000), certain other aspects do change over time (Roberts et al. 1997; Caspi and Bren 2001; Helson et al. 2002).

Also, in the time of unrest, uncertainty and risk, the expression of personality by behavior may display very different characteristics. Study of such inter-dependencies is certainly a topic of future research.

9 CONCLUSION

Personality identification using social network analysis is a relatively new domain within machine learning research. Since its introduction, however, it has drawn increasing attention by the research community with applications in wide variety of domains. Traditionally the only way personalities were identified was through questionnaire based personality tests that the subjects used to undergo. The surveyed techniques for automatic identification from online social networking profiles have yielded promising outcomes. Yet, many challenges and opportunities exist. Surveying this topic, we listed some challenges and insights that constitute promising research directions.

REFERENCES

S. Adal and J. Golbeck. 2012. Predicting personality with social behavior. In *Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*.

J. S. Adelstein, Z. Shehzad, M. Mennes, C. G. DeYoung, X.-N. Zuo, C. Kelly, D. S. Margulies, A. Bloomfield, J. R. Gray, X. F. Castellanos, M. P. Milham. 2011. Personality is reflected in the brain's intrinsic functional architecture. *PLoS ONE* 6, 11 (2011), 1–12.

B. Agarwal. 2014. Personality detection from text: A review. *International Journal of Computer System* 1, 1, (Sep. 2014).

F. Alam, E. Stepanov, G. Riccardi. 2013. Personality traits recognition on social network - Facebook. In *Proceedings of Workshop on Computational Personality Recognition (WCPR'13)*. AAAI Press, Melon Park, CA, 6–9.

G. W. Allport and H. S. Odbert. 1936. Trait names: A psycho-lexical study. *Psychological Monographs* 47, 1, Whole 211 (1936), 171–220.

Appling D. Scott, Erica J. Briscoe, Heather Hayes, and Rudolph L. Mappus. 2013. Towards automated personality identification using speech acts. In *Proceedings of Workshop on Computational Personality Recognition (WCPR'13)*. AAAI Press

S. Argamon, S. Dhawle, M. Koppel, and J. W. Pennebaker. 2005. Lexical predictors of personality type. In *Proceedings of 2005 Joint Annual Meeting of the Interface and the Classification Society of North America*.

M. Arrington. 2005. 85% of College Students Use Facebook Techcrunch. (September 2005). Retrieved March 17, 2006 from http://www.techcrunch.com/2005/09/07/85-of-collegestudents-use-facebook.

Michael C. Ashton, Kibeom Lee, Marco Perugini, Piotr Szarota, Reinout E. de Vries, Lisa Di Blas, Kathleen Boies, and Boele De Raad. 2004. A six-factor structure of personality-descriptive adjectives: Solutions from psycholexical studies in seven languages. *Journal of Personality and Social Psychology* 86, 2 (2004), 356–366.

M. C. Ashton and K. Lee. 2007. Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review* 11, 2 (2007), 150–66.

Y. Bachrach, M. Kosinski, T. Graepel, P. Kohli, and D. Stillwell. 2012. Personality and patterns of Facebook usage. In *Proceedings of ACM Conference on Web Sciences, ACM Web Sciences*.

Back Mitja, D. Stopfer, M. Juliane, S. Vazire, S. Gaddis, C. Schmukle Stefan, B. Egloff, and S. D. Gosling. 2009. Facebook profiles reflect actual personality. *Psychological Science, Not Self-Idealization* 21, 3 (2009), 372–374.

S. Bai, Tingshao Zhu, and Li Cheng. 2012. Big-five personality prediction based on user behaviors at social network sites. *arXiv preprint arXiv:1204.4809*.

M. Barrick, M. Mount, and T. Judge. 2001. Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment* 9, 1&2 (2001), 9–30.

J. Bollen, H. Mao, and X Zeng. 2011 Twitter mood predicts the stock market. *Journal of Computational Science* 2 (2011), 1–8.

D. M. Boyd and N. B. Ellison. 2007. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*,13, 210–230.

W.-P. Brinkman and N. Fine. 2005. Towards customized emotional design: An explorative study of user personality and user interface skin preferences. In *Proceedings of the 2005 Annual Conference on European Association of Cognitive Ergonomics (EACE'05)*. University of Athens, 107–114.

W. G. Broehl and V. W. McGee. 1981. Content analysis in psychohistory: A study of three lieutenants in the Indian Mutiny - 1857-1858. *Journal of Psychohistory* 8 (1981), 281–306.

G. D. A. Brown. 1984. A frequency count of 190,000 words in the London-Lund Corpus of English Conversation. *Behavioural Research Methods Instrumentation and Compu taters* 16, 6 (1984), 502–532.

E. Brunswik. 1956. *Perception and the Representative Design of Psychological Experiments*. University of California Press, Berkeley, CA.

Cao Qiang. 2012. Aiding the detection of fake accounts in large scale social online services. In *Proceeding of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI'12)*. 2012.

Caspi Avshalom and Brent W. Roberts. 2001. Personality development across the life course: The argument for change and continuity. *Psychological Inquiry* 12, 2(2001) 49–66.

A. Caspi, H. Harrington, B. Milne, J. W. Amell, R. F. Theodore, and T. E. Moffitt. 2003. Children's behavioral style at age 3 are linked to their adult personality traits at age 26. *Journal of Personality* 71, 4 (2003), 495–514.

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

F. Celli. 2011. *Unsupervised Recognition of Personality From Linguistic Features*. Technical Report available at http://clic.cimec.unitn.it/fabio, 2011.

F. Celli and L. Rossi. 2012. The role of emotional stability in twitter conversations. In *Proceedings of Workshop on Semantic Analysis in Social Media - EACL*. 10–17.

F. Celli. 2012. Unsupervised personality recognition for social network sites. In *Proceedings of the 6th International Conference on Digital Society (ICDS'12)*.

F. Celli, Fabio Pianesi, David Stillwell, and Michal Kosinski. 2013. *Workshop on Computational Personality Recognition: Shared Task*. AAAI Technical Report.

F. Celli. 2011. *Mining User Personality in Twitter, Language, Interaction And Computation CLIC, University of Trento*. (September 2008). Retrieved October 4, 2013 from http://clic.cimec.unitn.it/.

F. Celli, E. Bruni, and B. Lepri. Automatic personality and interaction style recognition from Facebook profile pictures. In *Proceedings of the 22nd ACM International Conference on Multimedia Pages*. 1101–1104.

D. Chapsky. 2011. Leveraging online social networks and external data sources to predict personality. In *Proceedings of International Conference on Advances in Social Networks Analysis and Mining*. 2011.

N. Chin David, William, and R. Wright. 2014. Social media sources for personality profiling. In *Proceedings of UMAP Workshops*. 2014

G. Chittaranjan, Jan Blom, and Daniel Gatica-Perez. 2011. *Mining Large-Scale Smartphone Data For Personality Studies*. Personal and Ubiquitous Computing. Springer, Page 1.

M. Coltheart. 1981. The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology* 33A (1981), 497–505.

P. T. Costa and R. R. Jr. McCrae. 1985. *The NEO Personality Inventory Manual*. Psychological Assessment Resources, 5–13.

Costa P. T. Jr. and R. R. McCrae. 1992. *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) Manual*. Psychological Assessment Resources, Odessa, FL.

C. G. DeYoung. 2010. Toward a theory of the big five. *Psychological Inquiry* 21 (2010), 26–33.

J. M. Digman. 1990. Personality structure: Emergence of the five-factor model. *Annual Review of Psychology* 41 (1990), 417–440.

P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE* 6 (2011), 26.

L. Doeven-Eggens, A. A. Filip De Fruyt, Jolijn Hendriks, Roel J. Bosker, Margaretha P. C. Van der Werf. 2008. Personality and personal network type. *Personality and Individual Differences* 45, 7 (2008), 689–693.

S. Dollinger. 1993. Research note: Personality and music preference: Extraversion and excitement seeking or openness to experience? *Psychology of Music* 21, 1 (1993), 73–77.

M. B. Donnellan, R. D. Conger, and C. M. Bryant. 2004. The big five and enduring marriages. *Journal of Research in Personality* 38 (2004), 481–504.

T. DuBois, J. Golbeck, J. Kleint, and A. Srinivasan. 2009. Improving recommendation accuracy by clustering social networks with trust. In *Proceedings of Recommender Systems & the Social Web*. 2009.

F. Enos, S. Benus, R. L. Cautin, M. Graciarena, J. Hirschberg, and E. Shriberg. 2006. Personality factors in human deception detection: Comparing human to machine performance. In *Proceedings of INTERSPEECH - ICSLP*. 813–816.

D. C. Evans, S. D. Gosling, and A. Carroll. 2008. What elements of an online social networking profile predict target-rater agreement in personality impressions. In *Proceedings of International Conference on Weblogs and Social Media*, 2008.

H. J. Eysenck. 1991. Dimensions of personality: 16, 5 or 3? Criteria for a taxonomic paradigm. *Personality and Individual Differences* 12, 8 (1991), 773–790.

G. Farnadi, S. Zoghbi, M. F. Moens, and M. De Cock. 2013. Recognising personality traits using facebook status updates. In *Proceedings of Workshop on Computational Personality Recognition (WCPR'13)*.

L. A. Fast and D. C. Funder. 2007. Personality as manifest in word use: Correlations with self-report, acquaintance-report, and behavior. *Journal of Personality and Social Psychology* 94, 2 (2008), 334.

A. Finder. 2006. For some, online persona undermines a réesumé. *New York Times* 11 (2006).

D. Foster. 1996. Primary culprit: Who is anonymous? *New York Magazine* 29 (1996), 50–57.

Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. 1998. An efficient boosting algorithm for combining preferences. In *Proceedings of the 15th International Conference on Machine Learning*. 170–178.

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

A. Furnham and J. Mitchell. 1991. Personality, needs, social skills and academic achievement: A longitudinal study. *Personality and Individual Differences* 12 (1991), 1067–1073.

A. Furnham, C. J. Jackson, and T. Miller. 1999. Personality, learning style and work performance. *Personality and Individual Differences* 27 (1999), 1113–1122.

A. J. Gill and J. Oberlander. 2002. Taking care of the linguistic features of extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*. 363–368.

J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. 2009. Detecting inuenza epidemics using search engine query data. *Nature* 457 (2009), 1012–1014.

J. Golbeck, Cristina Robles, and Karen Turner. Predicting personality with social media. In *Proceedings of Extended Abstracts on Human Factors in Computing Systems (CHI EA'11)*. ACM, New York, NY, 253–262. doi:10.1145/1979742.1979614

J. Golbeck. 2005. *Computing and Applying Trust in Web-Based Social Networks*. Ph.D. Thesis. University of Maryland, College Park, MD.

J. Golbeck and D. L. Hansen. 2011. Computing political preference among twitter followers. In *Proceedings of CHI 2011*. 1105–1108.

J. Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011. Predicting personality from Twitter. In *Proceedings of IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing*, 2011.

L. R. Goldberg. 1993. The structure of phenotypic personality traits. *American Psychologist* 48, 1 (1993), 26–34.

L. R. Goldberg, J. A. Johnson, H. W. Eber, R. Hogan, M. C. Ashton, C. R. Cloninger, and H. C. Gough. 2006. The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality* 40 (2006), 84–96.

J. Goldenberg, S. Han, D. R. Lehmann, and J. W. Hong. 2009. The role of hubs in the adoption processes. *Journal of Marketing* 73, 2 (2009), 1–13.

S. Golder and M. Macy. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* 333 (2011), 1878–1881.

S. D. Gosling, S. J. Ko, T. Mannarelli, and M. E. Morris. 2002. A room with a cue: Personality judgments based on offices and bedrooms. *Journal of Personality and Social Psychology* 83, 3 (2002), 379–398.

S. D. Gosling, P. J. Rentfrow, and W. B. Swann Jr. 2003. A very brief measure of the big-five personality domains. *Journal of Research in Personality* 37 (2003), 504–528.

Samuel D. Gosling, S. Gaddis, S. Vazire. 2007. Personality impressions based on facebook profiles. In *Proceedings of ICWSM, Vol.* 7. 1–4.

S. D. Gosling, Adam A. Augustine, Simine Vazire, Nicholas Holtzman, and Sam Gaddis. 2011. Manifestations of personality in online social networks: Self-reported Facebook-related behaviors and observable profile information. *Cyberpsychology, Behavior, and Social Networking* 14, 9 (September 2011), 483–488. doi: 10.1089/cyber.2010.0087

Gou Liang. 2013. Personalityviz: A visualization tool to analyze people's personality with social media. *Proceedings of the Companion Publication of the 2013 International Conference on Intelligent User Interfaces Companion*. ACM.

Gou Liang, Michelle X. Zhou, and Huahai Yang. 2014. Knowme and shareme: Understanding automatically discovered personality traits from social media and user sharing preferences. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM.

H. G. Gough. 1987. *California Psychological Inventory Administrator's Guide*. Consulting Psychologists Press, Palo Alto, CA.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11, 1 (2009), 10–18.

Y. A. Hamburger and E. Ben-Artzi. 2000. The relationship between extraversion and neuroticism and the different uses of the Internet. *Computers in Human Behavior* 16 (2000), 441–449.

Y. A. Hamburger. 2002. Internet and personality. *Computers in Human Behavior* 18 (2002), 1–10.

Y. A. Hamburger, G. Wainapel, and S. Fox. 2002. On the internet no one knows i'm an introvert: Extroversion, neuroticism, and internet interaction. *Cyberpsychology & Behavior* 2 (2002), 125–128.

Y. A. Hamburger, H. Kaplan, and N. Dorpatcheon. 2008. Click to the past: The impact of extroversion by users of nostalgic website on the use of Internet social services. *Computers in Human Behavior* 24 (2008), 1907–1912.

Y. A. Hamburger and Gideon Vinitzky. 2010. Social network use and personality, *Computers in Human Behavior*, 26, 6, (2010), 1289–1295. DOI: http://dx.doi.org/10.1016/j.chb.2010.03.018

Robert A. Hanneman and Mark Riddle. 2005. *Introduction to Social Network Methods*. University of California, Riverside, CA (published in digital form at http://faculty.ucr.edu/~hanneman/).

C. Hansen and R. Hansen. 1991. Constructing personality and social reality through music: Individual differences among fans of punk and heavy metal music. *Journal of Broadcasting & Electronic Media*, 35, 3 (1991), 335–350.

R. P. Hart. 1984. *Verbal Style and The Presidency: A Computer Based Analysis*. Academic Press, New York, NY.

J. R. Hauser, G. Urban, G. Liberali, and M. Braun. 2009. Website morphing. *Marketing Science* 28, 2 (2009) 202–223.

Ravenna Helson, Constance Jones, and Virginia S. Y. Kwan. 2002. Personality change over 40 years of adulthood: Hierarchical linear modeling analyses of two longitudinal samples. *Journal of Personality and Social Psychology* 83, 3 (2002), 752–766.

R. Hogan. 1982. A socioanalytic theory of personality. *Nebraska Symposium of Motivation* 30 (1982), 55–89.

R. Hogan and J. Hogan. 1992. *Manual for the Hogan Personality Inventory*. Hogan Assessment Systems, Tulsa, OK.

R. Hogan, G. J. Curphy, and J. Hogan. 1994. What we know about leadership: Effectiveness and personality. *American Psychologist* 49, 6 (1994), 493–504.

O. P. John, E. M. Donahue, and R. L. Kentle. 1991. The "big five" inventory: Versions 4a and 5b. *Technical report, Berkeley: University of California, Institute of Personality and Social Research.*

Jung Carl Gustav (1971). *Psychological Types. Collected Works of C.G. Jung*, Volume 6. Princeton University Press.

A. Kaplan and M. Haenlein. 2010. Users of the world, unite! The challenges and opportunities of social media. *Business Horizons* 53, 1 (2010), 59–68.

A. Karsvall. 2002. Personality preferences in graphical interface design. In *Proceedings of the 2nd Nordic Conference on Human-Computer Interaction (NordiCHI'02)*. ACM, New York, NY, 217–218.

S. Kedar, V. Nair, and S. Kulkarni. 2015. Personality identification through handwriting analysis: A review. *International Journal of Advanced Research in Computer Science and Software Engineering* 5, 1 (2015).

M. Komarraju and S. J. Karau. 2005. The relationship betweenthe big five personality traits and academic motivation. *Personality and Individual Differences* 39 (2005), 557–567.

M. Kosinski, D. J. Stillwell, and T. Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110, 15 (2013), 5802–5805.

A. Kramer. 2010. An unobtrusive behavioral model of gross national happiness. In *Proceeding of the 28th International Conference on Human Factors in Computing Systems*. ACM, 287–290.

J. Kratzer and C. Lettl. 2009. Distinctive roles of lead users and opinion leaders in the social networks of schoolchildren. *Journal of Consumer Research* 36, 4 (2009), 646–659.

Kucera and W. N. Francis. 1967. *Computational Analysis of Present-day American English*. Brown University Press, Providence.

D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, and T. Jebara. 2009. Life in the network: The coming age of computational social science. *Science* 323, 5915 (2009), 721–723.

B.-C. Lim and R. E. Ployhart. 2006. Assessing the convergent and discriminant validity of Goldberg's international personality item pool. *Organizational Research Methods* 9, 1 (2006), 29–54.

Ana Lima, E. S. Carolina, and Leandro Nunes De Castro. 2014. A multi-label, semi-supervised classification approach applied to personality prediction in social media. *Neural Networks* 58 (2014), 122–130.

K. Luyckx and W. Daelemans. 2008. Personae: A corpus for author and personality prediction from text. In *Proceedings of the 6th International Language Resources and Evaluation Conference (LREC'08)*. 2981–2987.

F. Mairesse and M. Walker. 2006. Words mark the nerds: Computational models of personality recognition through language. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*. Vancouver.

F. Mairesse and M. Walker. 2007. PERSONAGE: Personality generation for dialogue. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*. 496–503.

F. Mairesse, M. Walker, M. Mehl, and R. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research* 30 (2007), 457–500.

D. Markovikj, S. Gievska, M. Kosinski, and D. S. Stillwell. 2013. Mining facebook data for predictive personality modeling. In *Proceedings of the 7th International AAAI Conference On Weblogs And Social Media (ICWSM 2013), Proceedings of Workshop on Computational Personality Recognition (WCPR'13)* (2nd. ed.). Guilford Press, New York, 139–153.

P. V. Marsden. 1990. Network data and measurement. *Annual Review of Sociology* 16 (1990), 435–463.

D. Maynard, K. Bontcheva, and D. Rout. 2012. Challenges in developing opinion mining tools for social media. In *Proceedings of @NLP can u tag usergeneratedcontent?! Workshop at LREC 2012.* 15–22.

C. McCarty and H. D. Green. 2005. Personality and personal networks. In *Proceedings of Sunbelt XXV, Conference Contribution.*

R. M. McCrae and O. P. John. 1992. An introduction to the five-factor model and its applications. *Journal of Personality* 60, 2 (1992), 175–215. doi:10.1111/j.1467-6494.1992.tb00970.x. PMID 1635039.

R. R. McCrae and P. T. Costa. 1999. A five-factor theory of personality. In *Handbook of Personality: Theory and Research* (2nd. ed.). L. A. Pervin & O. P. John (Eds.), Guilford Press, New York, NY, 139–153.

A. R. McLarney-Vesotski, F. Bernieri, and D. Rempala. 2006. Personality perception: A developmental study. *Journal of Research in Personality* 40, 5 (2006), 652–674.

B. Meeder, J. Tam, P. Kelley, and L. F. Cranor. 2010. RT@ Iwantprivacy: Widespread violation of privacy settings in the twitter social network. In *Proceedings of the Web 2.0 Privacy and Security Workshop Co-located with IEEE Symposium on Security and Privacy.*

M. R. Mehl, S. D. Gosling, and J. W. Pennebaker. 2006. Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology* 90, 5 (2006), 862–877.

J. B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, and M. K. Gray. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331 (2011), 176–182.

G. Miller. 2011. Social scientists wade into the tweet stream. *Science* 333 (2011), 1814–1815.

Saif M. Mohammad and S. Kiritchenko. 2013. Using nuances of emotion to identify personality. In *Proceedings of the ICWSM Workshop on Computational Personality Recognition.* Boston

Myers, Isabel Briggs with Peter B. Myers (1980, 1995). *Gifts Differing: Understanding Personality Type.* Davies-Black

D. Nettle. 2007. *Personality: What Makes You the Way You Are.* Oxford University Press.

F. Å. Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big Things Come in Small Packages 718 in CEUR Workshop.* 93–98

W. T. Norman. 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality rating. *Journal of Abnormal and Social Psychology* 66 (1963), 574–583.

O. Nov and C. Ye. 2008. Personality and technology acceptance: The case for personal innovativeness in IT, openness and resistance to change. In *Proceedings of the 41st Hawaii International Conference on System Sciences (HICSS'41).* IEEE Press, Hawaii.

S. Nunn. 2005. Preventing the next terrorist attack: The theory and practice of homeland security information systems. *Journal of Homeland Security and Emergency Management* 2, 3 (2005), 1547–7355.

J. Oberlander and A. J. Gill. 2006. Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes* 42 (2006), 239–270.

J. Oberlander and S. Nowson. 2006. Whose thumb is it anyway?: Classifying author personality from weblog text. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions (COLING-ACL'06).* Association for Computational Linguistics, Stroudsburg, PA, 627–634.

G. Odekerken-Schroder, K. D. Wulf, and P. Schumacher. 2003. Strengthening outcomes of retailer-consumer relationships: The dual impact of relationship marketing tactics and consumer personality. *Journal of Business Research* 56, 3 (March 2003), 177–190.

S. V. Paunonen and D. N. Jackson. 2000. What is beyond the big five? plenty! *Journal of Personality* 68, 5 (2000), 821–836.

D. Peabody and B. De Raad. 2002. The substantive nature of psycholexical personality factors: A comparison across languages. *Journal of Personality and Social Psychology* 83, 4 (2002), 983–997.

D. Peabody and L. R. Goldberg. 1989. Some determinants of factor structures from personality-trait descriptor. *Journal of Personality and Social Psychology* 57, 3 (1989), 552–567.

J. W. Pennebaker, M. E. Francis, and R. J. Booth. 2001. *Linguistic Inquiry and Word Count: LIWC2001.* Erlbaum, Mahwah, NJ (www.erlbaum.com).

J. W. Pennebaker and L. A. King. 1999. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology* 77 (1999), 1296–1312.

J. W. Pennebaker, M. Mehl, and K. Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology* 54 (2003), 547–577.

D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft. 2011. Our twitter profiles, our selves: Predicting personality with twitter. In *Proceedings of the 3rd IEEE Conference on Social Computing (SocialCom).*

D. Quercia, Renaud Lambiotte, David Stillwell, Michal Kosinski, Jon Crowcroft. 2012. The personality of popular facebook users. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW'12)*. ACM, New York, NY, 955–964. doi:10.1145/2145204.2145346

D. Rawlings and V. Ciancarelli. 1997. Music preference and the five-factor model of the NEO personality inventory. *Psychology of Music* 25, 2 (October 1997), 120–132.

B. Reeves and C. Nass. 1996. *The Media Equation*. University of Chicago Press.

E. Reiter and S. G. Sripada. 2004. Contextual influences on near-synonym choice. In *Proceedings of the International Natural Language Generation Conference*. 161–170.

P. Rentfrow and S. Gosling. 2003. The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology* 84, 6 (2003), 1236–1256.

P. J. Rentfrow and S. D. Gosling. 2006. Message in a ballad: The role of music preferences in interpersonal perception. *Psychological Science* 17, 3 (2006), 236–242.

Brent W. Roberts and Ravenna Helson. 1997. Changes in culture, changes in personality: The influence of individualism in a longitudinal study of women. *Journal of Personality and Social Psychology* 72, 3 (1997), 641–651.

Brent W. Roberts, DelVecchio, and F. Wendy. 2000. The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin* 126, 1 (2000), 3–25.

S. S. Robin and E. KramerJames. 2011. King robert ward, identifying personality from the static, non-expressive face in humans andchimpanzees: Evidence of a shared system for signaling personality. *Evolution and Human Behavior* 32 (2011), 179–185.

A. Roshchina, J. Cardiff, and P. Rosso. 2011. A comparative evaluation of personality estimation algorithms for the TWIN recommender system. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*. 11–18.

C. Ross, E. S. Orr, M. Sisic, J. M. Arseneault, M. J. Simmering, and R. R. Orr. 2009. Personality and motivations associated with facebook use. *Computers in HumanBehavior* 25, 578–586.

J. P. Rushton, H. G. Murray, and S. Erdle. 1987. Combining trait consistency and learning specificity approaches to personality, with illustrative data on faculty teaching performance. *Personality and Individual Differences* 8 (1987), 59–66.

M. Saleem. 2010. By the numbers: Facebook vs the United States. Retrieved April 5, 2010 from <http://mashable.com/2010/04/05/facebook-us-infographic/>.

C. Sarkar, S. Bhatia, A. Agarwal, and Li Juan. 2014. Feature analysis for computational personality recognition using youtube personality data set. In *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition*. 11–14.

L. Saulsman and A. Page. 2004. The five-factor model and personality disorder empirical literature: A meta-analytic review. *Clinical Psychology Review* 23, 8 (2004), 1055–1085.

K. R. Scherer. 1979. Personality markers in speech. In *Social Markers in Speech*. K. R. Scherer, & H. Giles (Eds.), Cambridge University Press, 147–209.

D. Schmitt, J. Allik, R. McCrae, and V. Benet-Martinez. 2007. The geographic distribution of big five personality traits: Patterns and profiles of human self-description across 56 nations. *Journal of Cross-Cultural Psychology* 38, 2 (2007), 173.

H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, and S. M. Ramones. 2013 Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE* 8, 9 (2013), e73791. doi:10.1371/journal.pone.0073791

J. Searle. 1975. Indirect speech acts. In *Syntax and Semantics 3: Speech Acts*, 59–82.

J. Searle. 1976. The classification of illocutionary acts. *Language in Society* 5 (1976) 1–23.

R. A. Sherman, C. S. Nave, and D. C. Funder. 2012. Properties of persons and situations related to overall and distinctive personality-behavior congruence. *Journal of Research in Personality* 46, 1 (2012), 87–101.

J. F. Sigurdsson. 1991. Computer experience, attitudes toward computers and personality characteristics in psychology undergraduates. *Personality and Individual Differences* 12, 6 (1991), 617–624.

B. L. Smith, B. L. Brown, W. J. Strong, and A. C. Rencher. 1975. Effects of speech rate on personality perception. *Language and Speech* 18 (1975), 145–152.

Song Xuemeng et al. 2015. Multiple social network learning and its application in volunteerism tendency prediction. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

A. Srinivasan. 2009a. Ten Interesting Facts About Facebook. Retrieved December 4, 2009 from <http://techcrunchies.com/ten-interesting-facts-about-facebook/>.

Susan J. Stabile. 2002. The use of personality tests as a hiring tool: Is the benefit worth the cost? *Journal of Labor and Employment Law* 4 (2001), 279.

P. J. Stone, D. C. Dunphy, M. S. Smith, and D. M. Ogilvy. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, MA.

J. J. Tickle, T. F. Heatherton, and L. G. Wittenberg. 2001. Can personality change? In *Handbook of Personality Disorders*: Theory, Research and Treatment. J. W. Livesley (Ed.), Guilford Press, New York, 242–259.

S. Tucker and S. Whittaker. 2004. Accessing multimodal meeting data: Systems, problems and possibilities. *Lecture Notes in Computer Science, Machine Learning for Multimodal Interaction* 3361 (2004), 1–11.

P. M. Valkenburg and J. Peter. 2009. Social consequences of the Internet for adolescents: A decade of research. *Current Directionsin Psychological Science* 18 (2009), 1–5.

S. Vazire and S. D. Gosling. 2004. e-Perceptions: Personality impressions based on personal websites. *Journal of Personality and Social Psychology* 87 (2004), 123–132.

B. Verhoeven, Walter Daelemans, and Tom De Smedt. 2013. Ensemble methods for personality recognition. In *Proceedings of Workshop on Computational Personality Recognition (WCPR'13).*

A. Vinciarelli and G. Mohammadi. 2014. A survey of personality computing. *IEEE Transactions on* affective computing 5, 3 (2014), 273–291.

A. Vinciarelli and G. Mohammadi. More personality in personality computing. *IEEE Transactions on Affective Computing*, 5, 3 (July–September 2014), 297–300. DOI: 10.1109/TAFFC.2014.2341252

K. Vogel and S. Vogel. 1986. L'interlangue et la personalit´e de l'apprenant. *International Journal of Applied Linguistics.* 24, 1 (1986), 48–68.

R. Wald, T. Khoshgoftaar, and C. Sumner. 2012. Machine prediction of personality from facebook profiles. In *Proceedings of the 2012 IEEE 13th International Conference on Information Reuse and Integration (IRI'12).* 109–115. DOI: 10.1109/IRI.2012.6302998

M. Walker and S. Whittaker. 1990. Mixed initiative in dialogue: An investigation into discourse segmentation. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics.* 70–78.

Wehrli Stefan. 2008. Personality on social network sites: An application of the five factor model. In *Proceedings of ETH Zurich Sociology Working Papers No. 7.* Chair of Sociology. http://EconPapers.repec.org/RePEc:ets:wpaper:7.

S. Weinberger. 2011. Web of war: Can computational social science help to prevent or win wars? the pentagon is betting millions of dollars on the hope that it will. *Nature* 471 (2011), 566–568.

S. Whelan and G. Davies. 2006. Profiling consumers of own brands and national brands using human personality. *Journal of Retailing and Consumer Services* 13, 6 (2006), 393–402.

A. G. C. Wright. 2014. Current directions in personality science and the potential for advances through computing. *IEEE Transactions on Affective Computing* 5, 3 (July–September 2014), 292–296.

Xiao Cao, David Mandell Freeman, and Theodore Hwa. 2015. Detecting clusters of fake accounts in online social networks. In *Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security.* ACM.

W. Youyou, M. Kosinski, and D. Stillwell. 2015. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences of the United States of America* 112, 4 (January 2015), 1036–1040. DOI: http://www.pnas.org/cgi/doi/10.1073/pnas.1418680112

S. Zhao, S. Grasmuck, and J. Martin. 2008. Identity construction on facebook: Digital empowerment in anchored relationships. *Computers in Human Behavior* 24, 5 (2008), 1816–1836.

Received June 2015; revised November 2016; accepted March 2017