Personality and Individual Differences 124 (2018) 150–159

Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com/locate/paid

Review

Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis

Danny Azucar, Davide Marengo*, Michele Settanni

Department of Psychology, University of Turin, 10124, Via Verdi 10, Turin, Italy

ARTICLE INFO

Keywords:
Social media
Digital footprint
Big 5 traits
Personality
Data mining
Predictive modeling

ABSTRACT

The growing use of social media among Internet users produces a vast and new source of user generated ecological data, such as textual posts and images, which can be collected for research purposes. The increasing convergence between social and computer sciences has led researchers to develop automated methods to extract and analyze these digital footprints to predict personality traits. These social media-based predictions can then be used for a variety of purposes, including tailoring online services to improve user experience, enhance recommender systems, and as a possible screening and implementation tool for public health. In this paper, we conduct a series of meta-analyses to determine the predictive power of digital footprints collected from social media over Big 5 personality traits. Further, we investigate the impact of different types of digital footprints on prediction accuracy. Results of analyses show that the predictive power of digital footprints over personality traits is in line with the standard "correlational upper-limit" for behavior to predict personality, with correlations ranging from 0.29 (Agreeableness) to 0.40 (Extraversion). Overall, our findings indicate that accuracy of predictions is consistent across Big 5 traits, and that accuracy improves when analyses include demographics and multiple types of digital footprints.

1. Introduction

1.1. Social media and digital footprints

Social media and social network sites have become increasingly popular; currently about 2 billion people worldwide have a Facebook account, and over 1250 million users access Facebook on a daily basis (Statista, 2017). Similarly, Twitter averages about 328 million active users (Statista, 2017), with about 100 million daily users (Aslam, 2017). Social media has revolutionized how people interact with each other, is a virtually unavoidable avenue for social interactions, and a place where users present themselves to the world by creating an online profile. Every day, millions of people express their immediate thoughts, emotions, and beliefs by writing, posting, and sharing content on social media, which is then viewable by the user's online social network. Evidence also suggests that content generated and shared on social media user profiles represents an extension of "one's self" and reflects the actual personality of its individual users rather than project their most desirable traits (Back et al., 2010; Seidman, 2013). Consequently, the interactive nature of social media coupled with its ever-increasing utilization results in a naturally occurring, immense, ecologically valid dataset of online human activity, or *digital footprints*, consisting of information shared by users on their social media profiles - e.g., personal information about age, gender orientation, place of residence, as well shared texts, pictures, and videos (Madden, Fox, Smith, & Vitax, 2007). These digital footprints can be recorded, and have been previously analyzed by researchers from diverse disciplines, including computer science, public health, and social sciences (e.g., De Choudhury, Counts, & Horvitz, 2013; De Choudhury, Counts, Horvitz, & Hoff, 2014; Eichstaedt et al., 2015; Gosling, Augustine, Vazire, Holtzman, & Gaddis, 2011; Matz & Netzer, 2017; Padrez et al., 2015; Settanni & Marengo, 2015). In particular, the human migration to social media has steered psychologists toward studying existing relationships between digital footprints and psychological characteristics (Kosinski, Matz, Gosling, Popov, & Stillwell, 2015). The emergence of, and access to, these large user data sets has reshaped the way social science researchers use content analysis to study psychological characteristics and has resulted in the convergence of social and computer sciences. This interdisciplinary work of social and computer sciences has allowed researchers to not only seek to *gain insights* from studying human behaviors on social media, but to also *predict* psychological characteristics and behaviors based on automated data mining and the analysis of digital footprints (Schwartz & Ungar, 2015).

* Corresponding author.
E-mail address: davide.marengo@unito.it (D. Marengo).

https://doi.org/10.1016/j.paid.2017.12.018
Received 1 September 2017; Received in revised form 9 December 2017; Accepted 11 December 2017

1.2. Personality prediction from social media

Personality has been regarded as one of the most important topics in psychological research (Li, Li, Hao, Guan, & Zhu, 2014; Ozer & Benet-Martinez, 2006). Research has shown that personality may be predictive of many aspects of life, including academic success (e.g., Komarraju, Karau, & Schmeck, 2009), job performance (e.g., Judge, Higgins, Thoresen, & Barrick, 1999; Neal, Yeo, Koy, & Xiao, 2012), social status (e.g., Anderson, John, Keltner, & Kring, 2001), health (e.g., Soldz & Vaillant, 1999), success in romantic relationships (e.g., Donnellan, Conger, & Bryant, 2004; Donnellan, Larsen-Rife, & Conger, 2005), political attitudes (e.g., Gerber, Huber, Doherty, Dowling, & Ha, 2010), subjective well-being (e.g., Hayes & Joseph, 2003), and online behaviors (e.g., Wang, 2013). While several models to describe personality exist, one of the most well researched, well regarded, and widely accepted theoretical frameworks of personality is the five-factor (or Big 5) model, comprised of openness to new experiences, conscientiousness, extraversion, agreeableness and neuroticism (McCrae & Costa, 1987; McCrae & John, 1992). Big 5 traits have been shown to be significantly associated with users' behaviors on social media. For example, individuals with high extraversion have been characterized by higher levels of activity on social media (e.g., Blackwell, Leaman, Tramposch, Osborne, & Liss, 2017; Kuss & Griffiths, 2011), and have a greater number of friends (Kosinski, Bachrach, Kohli, Stillwell, & Graepel, 2014) than introverted individuals. Individuals with high neuroticism are more prone to self-disclose hidden aspects of themselves, use social media as a passive way to learn about others (Seidman, 2013), and use more negative words in their posts, or 'status updates' (Schwartz et al., 2013). On the other hand, agreeable individuals tend to use fewer swear words and express positive emotions more frequently in their posts (Schwartz et al., 2013), and are more likely to post pictures expressing a positive mood (Liu, Preotiuc-Pietro, Samani, Moghaddam, & Ungar, 2016). Individuals with high conscientiousness appear to be cautious in managing their social media profiles; they tend to post fewer pictures (Amichai-Hamburger & Vinitzky, 2010), express less "Likes", and engage in less group activity on social media (Kosinski et al., 2014). Furthermore, individuals with high openness tend to have larger networks (Quercia, Lambiotte, Stillwell, Kosinski, & Crowcroft, 2012), and "Like" more content found on social media (Bachrach, Kosinski, Graepel, Kohli, & Stillwell, 2012) than individuals low on the trait. Driven by increasing evidence of the presence of links between personality and online behaviors, researchers have begun exploring the use of digital footprints left by people on social media to infer the Big 5 traits. Researchers in this field have generally employed a common research design consisting of, 1. The administration of self-report questionnaires to assess personality traits of social media users, 2. The collection of digital footprints from users' social media profiles, 3. The processing of these digital footprints to extract single or multiple features to be employed in predictive models, and 4. The evaluation of accuracy of personality predictions based on these features. However, studies vary in terms of type of digital footprints (e.g., text, pictures, Likes, user activity, which may be examined separately or in combination), and social media platforms (e.g., Facebook, Twitter, Instagram, Youtube) examined. For instance, Schwartz et al. (2013) investigated the feasibility of predicting personality traits based on textual features extracted from Facebook status updates using topic-modeling techniques. Similarly, Liu et al. (2016) and Qiu, Lin, Ramsay, and Yang (2012) both analyzed language/text used on Twitter to build predictive models for the Big 5 traits. While Gao et al. (2013), Li et al. (2014), and Wei et al. (2017) inferred the Big 5 traits using samples from the Sina Weibo micro blog albeit using different combinations of digital footprints (activity vs. activity + language vs. activity + language + pictures) in their analysis. Additionally, Kosinski, Stillwell, and Graepel (2013) and Youyou, Kosinski, and Stillwell (2015) explored Big 5 personality predictions based on Facebook Likes. Findings emerging from these studies are heterogeneous with respect to the accuracy of prediction for each personality trait. For instance, using "Likes" data extracted from Facebook, Kosinski et al. (2013) found prediction accuracy to vary significantly across traits, with openness being the easiest to predict. Conversely, Li et al. (2014) analyzed user activity statistics from the Sina Weibo microblog and achieved similar prediction accuracy among all Big 5 Personality traits, and Skowron, Tkalčič, Ferwerda, and Schedl (2016) analyzed language + user features from users of both Twitter and Instagram and found a high prediction accuracy for conscientiousness, but a relatively low prediction accuracy for agreeableness. Even though many studies have been conducted on the subject, this area of psychological research is still quite young, which in part explains the reason for the lack of uniformity in the employed research methods. For example, studies vary largely on sample sizes, type of digital footprints analyzed, and social media platform used for data collection. Given these circumstances with psychological research conducted on social media, there is a need to synthesize and summarize the existing literature in order to evaluate their accuracy, and recommend best methods for personality prediction from social media.

The ability to use digital footprints to accurately predict personality traits may represent a rapid, cost-effective alternative to surveys and reach larger populations, which can be beneficial for academic, health-related, and commercial purposes. With respect to academic research, the development of automated procedures to measure personality would permit to reach larger samples, and obtain measures potentially less prone to social-desirability bias. Furthermore, personality traits have also been shown to act as potential risk and protective factors for many health-related outcomes (Booth-Kewley & Vickers, 1994; Raynor & Levine, 2009; Widiger & Oltmanns, 2017), and to influence beliefs about health (e.g., Hill & Gick, 2011). Therefore, the ability to distinguish online users based on their personality profiles could be leveraged in order to tailor techniques aimed at improving the efficacy of health related messages (Gale, Deary, Wardle, Zaninotto, & Batty, 2015; Lawson, Bundy, & Harvey, 2007; Neeme, Aavik, Aavik, & Punab, 2015; Rimer & Kreuter, 2006) and individual interventions (Chapman, Hampson, & Clarkin, 2014; Franks, Chapman, Duberstein, & Jerant, 2009) directed at online populations, and thus assist in the effective implementation of public health policies (Chapman, Roberts, & Duberstein, 2011; Hengartner, Kawohl, Haker, Rössler, & Ajdacic-Gross, 2016). With respect to commercial applications, knowledge about individuals' personalities can allow for the enhancement and personalization of recommender systems in order to improve user experiences (Bachrach et al., 2012; Farnadi et al., 2016). Also, social media sites, online advertisers, e-commerce retailers, and e-learning websites may be tailored based on individual personality and present information in ways that will be better received by users (Bachrach et al., 2012; Gao et al., 2013; Golbeck, Robles, & Turner, 2011; Kosinski et al., 2013; Markovikj, Gievska, Kosinski, & Stillwell, 2013).

1.3. Aims

The aim of the current study is to conduct a series of meta-analyses to estimate the mean predictive value of digital footprints on each of the Big 5 Personality Traits. Further, we aim to study if the use of different types of digital footprints influence the accuracy of personality prediction, and if data from different social media platforms lead to different results. Lastly, we will check for possible bias in effect size estimates due to study quality.

2. Methods

2.1. Literature search

To identify relevant studies on the relationships between Big 5 personality traits and digital footprints, we followed the literature search strategies proposed by Durlak and Lipsey (1991). We conducted

https://daneshyari.com/article/7249047