RESEARCH NOTE

Can online content indicate an individual's ‘real-life’ personality?

Alice Matthews & Andrew Hine

A study by data analyst Alice Matthews and entrepreneur Andrew Hine suggests that text data from a range of online sources can provide insights into an individual’s psychological traits in line with the IPIP 50[1] outcomes.

Introduction

In 2015, a seminal study by Youyou, Kosinski and Stillwell found that computers can judge personality more accurately than humans.[2] There are some other relevant academic studies[3] about inferring personality through online content, such as *Explainable personality prediction using answers to open-ended interview questions* by Dai et al.[4] and *Deep personality trait recognition: A survey* by Zhao et al.[5]

In our study, we conducted a data-driven investigation into the correlations between online content-based personality inference and psychometric survey-reported personality measurements. Our goal was to determine whether online content can better approximate more traditional psychometric surveys. To do this, we used several significant tools to assess personality. These included the Big 5 OCEAN traits model, which measures five

[1] The International Personality Item Pool which consists of 50 items.

[2] Youyou et al., 2015

[3] Christian et al, 2021; Clarke, 2016; Gill et al., 2009; Golbeck et al., 2011; Möllering and Guenther, 2010; Hirsh and Peterson, 2009; Mehta et al., 2020; Novikov et al., 2021; Schwartz et al., 2013; Yarkoni, 2010; Fast and Funder, 2008

[4] Dai et al., 2020

[5] Zhao et al., 2022

broad dimensions of personality: openness, conscientiousness, extraversion, agreeableness and neuroticism. We also used the IPIP 50 item personality questionnaire, a widely used measure of the Big Five personality traits. Finally, we used IBM Watson's Personality Insights, which uses linguistic analytics to infer personality characteristics from text.

Our results indicate that text content from a range of online sources can indeed reliably estimate the results of the IPIP 50 item psychometric survey.

Individuals who take the initiative to claim ownership of their online content for non-conventional uses, such as using it as part of a personality assessment process, may gain an advantage over others. This is because organisations that allow individuals to submit their online content as part of the assessment process may be able to save time and money by not having to administer traditional psychometric surveys. Additionally, analysing online content may offer a means of assessing candidates that is blind and impartial. This means the assessment process would be unbiased and not influenced by factors such as the candidate's appearance or background.

Methodology

We experimented with 348 volunteers who completed a standard 50 item Big 5 OCEAN questionnaire and were open to supplying their public profile handles on LinkedIn, Twitter, Stack Overflow and Reddit. Our volunteers were a diverse group from 41 countries aged between 18 and 65.

Volunteers applied from various online forums, groups and newsletters to receive a free multi-page report on how they may appear to others based on their written online content. After reading and signing a detailed description, disclaimer and ethics approval, they opted to provide links to all or none of their online profiles. If they chose to provide their email and at least one profile, they were sent a BIG 5 report generated just from their online profile(s) content. They also took a 50-question IPIP BIG 5 test to study the accuracy of their report. They could not provide any information they did not want (incomplete submissions were removed from our dataset).

We used our proprietary algorithm to collect each volunteer's public content only on LinkedIn, Twitter, Stack Overflow and Reddit. The IBM Watson Personality Insights tool was used to parse this content and assign trait scores to each user. While Personality Insights was trained on tweets only, we looked at various formal and informal online contexts.

Deriving insights from natural language processing (NLP) generally requires a large body of text to return significant results. Therefore, respondents with less than 1,000 words of content were removed from the analysis, and our final count (n) was 101. We then checked for a correlation between the NLP prediction of personality and the Big 5 OCEAN model prediction, using Spearman's rank correlation rs as our correlation coefficient.

We finally performed Ordinary Least Squares (OLS) regressions using the Stats Models Python package to determine the nature of the relationships between survey trait scores and the online content trait scores. An OLS regression can tell us how a variable of interest (the dependent variable) may be affected by changes in another variable (the independent variable).

In our OLS regressions, we set the dependent variables as IPIP 50-item survey results and the independent variables as 'Online content-based results'. First, we ran multivariate regressions with demographic variables to determine whether age, gender or country might influence the results. We then ran simple linear regressions on the five IPIP 50-item trait variables, using only the online content-derived personality scores as the independent variable.

Results

Our data models suggest that IBM's Personality Insights model, which analyses online content such as text from social media posts, can accurately estimate an individual's scores on three Big Five personality traits as measured by the IPIP 50-item psychometric survey. This relationship holds even when controlling for demographic factors such as age, gender and country.

Additionally, we found a moderate to strong correlation (Spearman's ratio of 50%-58%) between online content and psychometric survey measures of extraversion and conscientiousness. This suggests that an individual's online behaviour and communication can provide valuable insights into their personality traits.

In other words, it is possible to infer certain aspects of their personality by analysing the language used in an individual's online content, as well as the topics they discuss and the way they interact with others online. For example, an individual who frequently uses positive language and discusses social activities in their online content may score high on extraversion, while an individual who frequently discusses topics related to organisation and responsibility may score high on conscientiousness.

Discussion

Our results indicate that analysing online content may be a reasonable proxy for traditional psychometric survey methods to reveal Big Five personality traits. Continued research and development will deliver more sophisticated tools to allow individuals to leverage their online content to derive tangible benefits by improving their life situation. Individuals and organisations who move first to leverage this opportunity will gain a first-mover advantage in increasing the number of applicants they can assess, especially those from more diverse backgrounds, often lacking conventional references and reducing human biases. What is more, as governments and organisations roll out digital IDs, trust frameworks and metaverses, establishing trust and proof of reputation online will only become more critical for all members of digital society to function effectively.

Automated, accurate and cheap personality assessment tools could affect society in many ways: marketing messages could be tailored to users' personalities; recruiters could better match candidates with jobs based on their personality; products and services could adjust their behaviour to match their users' characters best and changing moods; and scientists could collect personality data without burdening participants with lengthy questionnaires. Furthermore, in the future, people might abandon their psychological judgments and rely on computers when making important life decisions, such as choosing activities, career paths, or even romantic partners. Such data-driven decisions may improve people's lives.

Last but not least, there are a few questions that we should be asking. For example, how do different platforms (e.g., social media, blogs, online forums) affect how people express their personalities online? Does the context of the platform influence people's behaviour and the content they share? How do cultural and linguistic differences affect the way people present themselves online and the content they share? Are specific cultural or linguistic markers more indicative of certain personality traits? Is it easier to game only content or hypothetical psychometric test answers via preference falsification to portray desired attributes? Or is revealed preference shown via tens, hundreds or even thousands of real-world online human-to-human interactions more accurate than conventional testing?

Theoretical and managerial implications

Online content can be used as a reliable source of information for inferring an individual's personality traits. This suggests that online behaviour and communication can provide valuable insights into an individual's personality, which can be helpful for researchers and practitioners in personality assessment.

From a managerial perspective, this finding could have practical applications in various domains such as recruitment, marketing and social media management. For example, employers could use online content to screen job applicants for certain personality traits relevant to the position. Marketers could use online content to tailor their advertising campaigns to target specific personality types. Social media managers could use online content to understand their audience better and create more engaging content.

Here are some use cases. The 2019 Global Talent Trends survey of 5,165 talent and managers from LinkedIn[6] found that soft skills are the top concern for employers, as are overwork flexibility, anti-harassment, and pay transparency. Ninety-two percent of respondents said that soft skills matter more than technical skills, and 89% reported that bad hires typically lack suitable soft skills. Complicating the issue is that soft skills are notoriously difficult to identify in candidates: only 41% of companies had a formal process to assess soft skills in job candidates, and 68% of respondents said that they rely mainly on social cues to judge candidates' soft skills. Unfortunately, these perceptions are not predictive; worse, they are often unconsciously biased.

This situation illustrates a significant gap in current recruitment processes: assessing soft skills is vital in finding the right talent, but we lack the means to do so quickly, easily and accurately. Online content analysis may offer a solution. For example, if a tech candidate has a 'Gold' reputation on Stack Overflow (a technical question and answer platform) and is in the top 2% of most active users on GitHub (a code hosting and sharing site), indeed, this indicates a passion for their field and an enthusiasm to participate in communities that many employers must value. As our research has shown, there is a link between an individual's online activity and their real-life personality traits.

The recruitment process is often impacted by the unconscious biases of those involved in the hiring process, which are hard to identify. These unconscious biases negatively affect many groups in society, including the LGBTQ+ community, recent migrants, religious minorities and international students. A study by the International Education Association of Australia[7] found that international students are a segment of the market that many employers do not fully understand or hesitate to recruit. A 2019 ABC investigation found that

[6] https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions/resources/pdfs/global-talent-trends-2019-old.pdf

[7] Tran and Bui, 2019

Asian Australians face a bamboo ceiling in the job market,[8] while a Monash University study found that 'One in four permanent skilled migrants work in a job beneath their skill level', costing the economy up to $1.25 billion annually.[9] This is a particular problem, as it has been found that cognitively diverse teams perform better.[10] The solution is to design the decision-making process to minimise the influence of human bias.

Without references or ratings, engaging the first client who can give you a reference or rating is challenging. This is a classic catch-22, coined the 'cold start problem'[11] that many gig economy workers are sadly familiar with. A way of leveraging online presence and content to prove conscientiousness, reliability and other social and technical skills would help workers in the gig economy get their businesses off the ground.

However, it is essential to note that the use of online content for personality assessment should be done with caution and in an ethical manner. Individuals have the right to privacy, and their online content should not be used without consent.

References

Bogoda Arachchige, P. (2021), 'Billion-dollar hit: The barriers skilled migrants face in finding jobs at their full capacity, and the economic cost', *Lens*, 20 May, https://lens.monash.edu/@politics-society/2021/05/20/1383170/billion-dollar-hit-the-barriers-skilled-migrants-face-in-finding-jobs-at-their-full-capacity-and-the-economic-cost, accessed 18 July 2024

Chen, A. (2022), *Cold start problem: How to start and scale network effects*, Harper Collins

Christian, H., Suhartono, D., Chowanda, A. et al. (2021), 'Text-based personality prediction from multiple social media data sources using pre-trained language model and model averaging', *Journal of Big Data*, vol. 8, p. 68

Clarke, M. (2016), 'Addressing the soft skills crisis', *Strategic HR Review*, vol. 15, no. 3, pp. 137–139

Dai, Y., Jayaratne, M. and Jayatilleke, B. (2020), 'Explainable personality prediction using answers to open-ended interview questions', *Frontiers in Psychology*, vol. 13

Fast, L.A. and Funder, D.C. (2008), 'Personality as manifest in word use: Correlations with self-report, acquaintance report, and behavior', *Journal of Personality and Social Psychology*, vol. 94, no. 2, pp. 334–346

Gill, A. J., Nowson, S. and Oberlander, J. (2009), 'What are they blogging about? Personality, topic and motivation in blogs', *AAAI Publications*, Third International AAAI Conference on Weblogs and Social Media, pp. 18–25

Golbeck, J., Robles, C., Edmondson, M. and Turner, K. (2011), 'Predicting personality from Twitter', *Proceedings of IEEE International Conference on Social Computing*

Hirsh, J.B. and Peterson, J.B. (2009), 'Personality and language use in self-narratives', *Journal of Research in Personality*, vol. 43, pp. 524–527

Mehta, Y., Fatehi, S., Kazameini, A., Stachl, C., Cambria, E. and Eetemadi, S. (2020), 'Bottom-up and top-down: Predicting personality with psycholinguistic and language model features', *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 1184–1189

[8] Xiao and Handley, 2019

[9] Bogoda Arachchige, 2021

[10] Reynolds and Lewis, 2017

[11] Chen, 2022

Möllering, G. and Guenther, T. (2010), 'A framework for studying the problem of trust in online settings', *International Journal of Dependable and Trustworthy Information Systems* (IJDTIS), vol. 1, no. 3

Novikov, P., Mararitsa, L.V. and Nozdrachev, V. (2021), 'Infrared vs traditional personality assessment: Are we predicting the same thing?' ArXiv, abs/2103.09632

Reynolds, A. and Lewis, D. (2017), 'Teams solve problems faster when they're more cognitively diverse', *Harvard Business Review*, 30 March, https://hbr.org/2017/03/teams-solve-problems-faster-when-theyre-more-cognitively-diverse, accessed 18 July 2024

Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., et al. (2013), 'Personality, gender, and age in the language of social media: The open-vocabulary approach', *PLOS ONE*, vol. 8, no. 9

Tran, L.T. and Bui, N.T.N. (2019), 'International graduates navigating the host and home labour markets', *International Education Association of Australia (IEAA)*, p. 7

Yarkoni, T. (2010), 'Personality in 100,000 words: A large-scale analysis of personality and word usage among bloggers', *Journal of Research in Personality*, vol. 44, no. 3, pp. 363–373

Youyou, W., Kosinski, M. and Stillwell, D. (2015), 'Computer-based personality judgments are more accurate than those made by humans', *Proceedings of the National Academy of Sciences*, vol. 112, no. 4, pp. 1036–1040

Xiao, B. and Handley, E. (2019), 'How Asian-Australians are struggling to break through the "bamboo ceiling" ', *ABC News*, 2 November, https://www.abc.net.au/news/2019-11-02/asian-australians-struggling-to-break-bamboo-ceiling/11665288, accessed 18 July 2024

Zhao, X., Tang, Z. and Zhang, S. (2022), 'Deep personality trait recognition: A survey', *Frontiers in Psychology*, vol. 13

JOURNAL OF BEHAVIOURAL ECONOMICS AND SOCIAL SYSTEMS

Volume 6, Number 1, 2024

gap | TCG