Personality and Individual Differences 102 (2016) 229–233

Contents lists available at ScienceDirect

Personality and Individual Differences

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

Assessing the Big Five personality traits with latent semantic analysis

Peter J. Kwantes [a,*], Natalia Derbentseva [a], Quan Lam [a], Oshin Vartanian [a,b], Harvey H.C. Marmurek [c]

[a] Defence Research and Development Canada, Canada
[b] University of Toronto at Scarborough, Canada
[c] University of Guelph, Canada

ARTICLE INFO

Available online 18 July 2016

Keywords:
Personality
Big Five traits
Latent semantic analysis

ABSTRACT

We tested whether the characteristics of a person's personality can be assessed by an automated analysis of the semantic content of a person's written text. Participants completed a questionnaire measuring the so-called Big Five personality traits. They also composed five short essays in which they were asked to describe what they would do and how they would feel in each of five scenarios designed to invoke the creation of narrative relevant to the Big Five personality traits. Participants' essays were processed for content by Latent Semantic Analysis (LSA; T. Landauer & S. Dumais, 1997), a model of lexical semantics. We found that LSA could assess individuals on three of the Big Five traits, and we discuss ways to improve such techniques in future work.

1. Introduction

In this article we build on classic work (e.g., Allport & Odbert, 1936; Cattell, 1943) that explores the role that words play in the description of personality. Specifically, we tested whether a collection of words that describes a trait can be used in an automated tool to assess a person's personality from information contained in his or her written text.

Current techniques of personality analysis from text samples are dominated by algorithms that tally and track word usage and map the patterns of usage across word categories onto personality traits. The basic idea of the text analytic approach is that personality influences word choice in one's speech or writing behavior. To the extent that we can characterize the word usage patterns common to the different personality types, we should be able to assess personality based on an examination of language samples generated by a speaker/author. Among available software-driven text analytic techniques, the most widely cited is Pennebaker's Linguistic Inquiry and Word Count (LIWC; Pennebaker & King, 1999). LIWC is essentially a word frequency counter that tallies an author's use of words that are yoked to linguistic (e.g., prepositions, articles, numbers), psychological (e.g., optimism, anger, insight), or physical (e.g., work, sleep, sexuality) categories. The LIWC yields a profile of the speaker's words across the categories, and the pattern with which they are distributed across categories is believed to be driven in part by the author or speaker's psychological traits or states at the time.

Most of the language-based work focuses on personality as expressed by the so-called Big Five personality traits (McCrae & John, 1992). The Big Five personality traits comprise the following: *extraversion* (described as being friendly, assertive and sociable), *conscientiousness* (described as being organized, dependable, and motivated), *agreeableness* (described as being cooperative, trusting, and helpful), *openness to new experience* (described as being emotional, curious, creative, imaginative, and hereafter referred to as, *Openness*), and *neuroticism* (associated with easily being made to feel upset, angry, anxious, or depressed). Pennebaker and King (1999) and Yarkoni (2010), for example, reported that authors who scored high on Extraversion tended to use fewer negative emotion words and more social words (e.g., restaurant, meet) than introverts. By contrast, Yarkoni also reported that people high on the neuroticism trait tended to use more negative emotion words than people low on neuroticism. In sum, the language we use to express ourselves seems to provide a window into how we feel, how we think, and how we are built psychologically. Current techniques to examine personality from language focus on the classification of word usage. The purpose of the current study was to augment such analyses of language by analyzing the semantic content of a speaker or author's text.

Analyzing text for word usage represents a categorical characterization of an author's text. Another way to characterize text is by the semantic content it carries. In this article, we examine whether a formal analysis of word usage, like that provided by the LIWC, can be complemented with a meaning-based analysis of the text generated by an author. Over the past two decades, computational models have been developed to create semantic representations for words encountered in text. One such model is Latent Semantic Analysis (LSA; Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998). LSA is a computational model that works on the notion that words with similar meanings tend to appear in similar contexts. It creates semantic representations for words by analyzing the pattern with which words occur together in documents across

* Corresponding author at: DRDC Toronto Research Centre, 1133 Sheppard Ave W., Toronto, ON M3K 2C9, Canada.
E-mail address: peter.kwantes@drdc-rddc.gc.ca (P.J. Kwantes).

http://dx.doi.org/10.1016/j.paid.2016.07.010
0191-8869/

thousands of text samples provided to it in a training corpus. Then, from an analysis of the words that do and do not co-occur in the corpus, the model estimates what words should occur in similar documents (i.e., contexts) and are, therefore, close to each other in semantic space.

Using LSA to explore aspects of an author's psychological make-up is not entirely new. For example, Campbell and Pennebaker (2003) reported results from participants who were asked, over several sessions, to write about an emotional time in their life or about some emotional experiences. They used LSA to characterize changes in writing content and style over time and measured the extent to which changes were related to improved wellness. They found no evidence that the semantic content of patients' written text over the time spent in treatment was in any way correlated with changes in well-being. They did, however, find that changes in writing style over time, especially the use of pronouns, were related to changes in patients' well-being.

Little work has been done to test LSA as a tool in the evaluation of authors' or speakers' personalities. To fill that gap in research, we examined how strongly the semantic content of authors' text is driven by personality and ask whether we can use LSA to measure aspects of it. To do so, participants were presented with five scenarios, and for each, they were asked to describe how they would feel and what they would do. The five scenarios were designed to invoke the production of narratives relevant to each of the five dominant, personality traits (McCrae & John, 1992). We postulate that when participants write about themselves in the scenarios, they will use terms that express their status on each of the Big Five traits. We hypothesize further that the more strongly a participant identifies with a trait, the more his or her narrative will contain text relevant to the trait, and that such differences can be detected using models like LSA.

2. Methods

2.1. Participants

One hundred and fifteen first-year (19 male) undergraduate students in an introductory psychology course at the University of Guelph participated in the study for course credit. Average age of participants was 19 years old (range 18–23 years). All but 12 of the participants reported English as their first language.

2.2. Materials

2.2.1. Testing materials

Five scenarios were developed to describe situations in which participants were to imagine themselves. For each scenario, participants were prompted to ponder how they would feel and what they would do. Each of the five scenarios, reproduced in Appendix A, was designed to be relevant to one of the Big Five personality traits. The scenarios were devised by a focus group of three researchers at Defence Research and Development Canada (DRDC) Toronto. Validation and fine-tuning of the scenarios was then done using feedback from a separate sample of three researchers at DRDC Toronto.

The Big Five Inventory (BFI; John, Donahue, & Kentle, 1991; John, Naumann, & Soto, 2008) was used in the study. The BFI is a 44-item test wherein respondents indicate their agreement with statements about themselves on a five-point scale. John and Srivastava (1999) reported alpha reliabilities for the five scales of between 0.75 and 0.90 and test–retest reliability between 0.80 and 0.90. They also report strong agreement ($M = 0.87$) between the BFI and other tests like the Costa and McCrae's (1985) NEO Five Factor Inventory and Trait Descriptive Adjectives (Goldberg, 1992), which also assess the Big Five personality traits.

2.2.2. LSA

As mentioned above, LSA is an algorithm that generates a *semantic space* from a statistical analysis of the frequencies with which words co-occur in a large collection of documents (i.e., contexts). The process by which LSA builds a semantic space from the document collection is called 'training'. After training, the semantic space comprises a set of vectors containing the semantic features for each word encountered in the document collection. We refer to the vectors in the semantic space as, *semantic vectors*. Generally speaking, the more documents contained in a training corpus, the more contextual information the system has to semantically differentiate or align words. We used different corpus sizes to ensure that if LSA failed to detect differences in authors' personalities, the results might suggest whether it was because of an insufficient number of documents during the training phase.

The other important aspect to consider when building LSA's semantic space is the choice of training corpus. LSA builds its semantic knowledge by exploiting the associations among words within the thousands of training documents. As a consequence, how the training corpus uses language and how words are associated in the training corpus will drive the system's interpretation of a word. For example, if LSA was trained on a document collection dominated by sports-related articles, its semantic representation for the word *play* would have different close associates than if the collection were dominated by, say, theatre-related articles. We trained LSA on two types of corpora. For one version, we used a random collection of articles from Wikipedia. For the other, we trained LSA on Wikipedia articles relevant to the Big Five personality traits. Done this way, LSA's understanding of the words in the collection was in the context of materials related to the personality traits and might therefore amplify the extent to which terms a person uses are considered to be related to the five personality traits.

We trained LSA separately on seven training corpora. The corpora were constructed by creating collections of varying sizes using different criteria for selecting documents. For the first three corpora, we trained the system on the first 200 words of randomly selected articles taken from Wikipedia. The three corpora differed in size. One corpus contained 12,000 articles, another 30,000, and the final one 50,000 articles. For the next three corpora, we used the Lucene (lucene.apache.org) indexer to search the Wikipedia corpus for terms relevant to the Big Five personality traits. Articles were selected by forming a query from terms (mainly adjectives) that describe one extreme on the trait's continuum. The terms were taken from the traits' definitions as reported in Wikipedia and those contained in the BFI. Adjectives from reverse-keyed items on the BFI were changed to their antonyms. The terms in the query are listed in Table 1. Again, we selected three different corpus sizes with collections of the 5000, 10,000, and 15,000 most relevant documents to train the system. The seventh corpus was also one which only contained documents relevant to the Big Five but was constructed slightly differently. Instead of extracting documents from Wikipedia using a query that included search terms relevant to all five personality traits, we extracted five 1000-document collections, each relevant to a single trait, and combined the results to create a 5000-document corpus containing document about all five traits. The rationale for creating the final corpus the way we did was to ensure that the proportion of documents relevant to each of the five traits was equal across them.

Once a semantic space is created from a training corpus, the semantic vectors it creates for words can be used to create semantic vectors for new documents. Creating a document vector is straightforward and involves summing the semantic vectors of a document's content words. Once created, pairs of documents can be compared by calculating the cosine between their vectors. The cosine behaves much like a correlation in that a cosine of 0 indicates that two vectors are orthogonal and a cosine of 1 indicates that they have identical projections in high-dimensional space. Words and documents with high cosines project in similar directions in semantic space and are therefore considered by LSA to be semantically related. Likewise, the lower the cosine between two vectors, the less related they are considered to be.