Researchers use Facebook data to predict users' age, gender and personality traits

September 26 2013

Word clouds that compare the language that extraverts (top) and introverts (bottom) used in their status messages.

In the age of social media, people's inner lives are increasingly recorded through the language they use online. With this in mind, an interdisciplinary group of University of Pennsylvania researchers is

interested in whether a computational analysis of this language can provide as much, or more, insight into their personalities as traditional methods used by psychologists, such as self-reported surveys and questionnaires.

In a recent study, published in the journal *PLOS ONE*, 75,000 people voluntarily completed a common personality questionnaire through a Facebook application and made their Facebook status updates available for research purposes. The researchers then looked for overall linguistic patterns in the volunteers' language.

Their analysis allowed them to generate computer models that were able to predict the individuals' age, gender and their responses on the personality questionnaires they took. These prediction models were surprisingly accurate. For example, the researchers were correct 92 percent of the time when predicting users' gender based only on the language of their status updates.

The success of this "open" approach suggests new ways of researching connections between personality traits and behaviors and measuring the effectiveness of psychological interventions.

The study is part of the World Well-Being Project, an interdisciplinary effort with members of the Computer and Information Science Department in Penn's School of Engineering and Applied Science and the Department of Psychology and its Positive Psychology Center in the School of Arts and Sciences.

It was led by H. Andrew Schwartz, a postdoctoral fellow in computer and information science and the Positive Psychology Center, and included graduate student Johannes Eichstaedt, postdoctoral fellow Margaret Kern and director Martin Seligman, all of the Positive Psychology Center, as well as professor Lyle Ungar of Computer and

Information Science.

The Penn team collaborated with Michal Kosinski and David Stillwell of The Psychometrics Centre at the University of Cambridge, who originally collected the data from Facebook users.

The researchers' study draws on a long history of studying the words people use as a way of understanding their feelings and mental states, but took an "open" rather than "closed" approach to analyzing the data at its core.

"In a 'closed vocabulary' approach," Kern said, "psychologists might pick a list of words they think signal positive emotion, like 'contented,' 'enthusiastic' or 'wonderful' and then look at the frequency of a person's use of these words as a way to measure how happy that person is. However, closed vocabulary approaches have several limitations, including that they do not always measure what they intend to measure."

"For example," Ungar said, "one might find the energy sector uses more negative emotion words, simply because they use the word 'crude' more. But this points to the need to use multi-word expressions to understand the intended meaning. 'Crude oil' is different than 'crude,' and, likewise, being 'sick of' is different from merely being 'sick.'"

Another inherent limitation to the closed vocabulary approach is that it relies upon a preconceived, fixed set of words. Such a study might be able to confirm that depressed people do indeed use expected words (like "sad") more frequently but cannot generate new insights (that they talk less about sports or social activities than happy people, for example.)

Past psychological language studies have necessarily relied on closed vocabulary approaches as their small sample sizes made open approaches impractical. The emergence of massive language datasets afforded by

social media now allows for qualitatively different analyses.

"Most words occur rarely—any sample of writing, including Facebook status updates, only contains a small portion of the average vocabulary," Schwartz said. "This means that, for all but the most common words, you need writing samples from many people in order to make connections with psychological traits. Traditional studies have found interesting connections with pre-chosen categories of words such as 'positive emotion' or 'function words.' However, the billions of word instances available in social media allow us to find patterns at a much richer level."

The open-vocabulary approach, by contrast, derives important words and phrases from the sample itself. With more than 700 million words, phrases and topics drilled out of this study's sample of Facebook status messages, there was enough data to dig past the hundreds of common words and phrases and to find open-ended language that more meaningfully correlates with specific characteristics.

This large data size was critical to the specific technique the team used, known as differential language analysis, or DLA. The researchers used DLA to isolate the words and phrases that clustered around the various characteristics self-reported in the volunteers' questionnaires: age, gender and scores for the "Big Five" personality traits, which are extraversion, agreeableness, conscientiousness, neuroticism and openness. The Big Five model was chosen as it is a common and well-studied way of quantifying personality traits, but the researchers' method could be applied to models that measure other characteristics, including depression or happiness.

To visualize their results, the researchers created word clouds that summarized the language that statistically predicted a given trait, with the correlation strength of a word in a given cluster being represented by its size. For example, a word cloud that shows language used by

extraverts prominently features words and phrases like "party," "great night" and "hit me up," while a word cloud for introverts features many references to Japanese media and emoticons.

"It may seem obvious that a super extraverted person would talk a lot about parties," Eichstaedt said, "but taken all together, these word clouds provide an unprecedented window into the psychological world of people with a given trait. Many things seem obvious after the fact and each item makes sense, but would you have thought of them all, or even most of them?"

"When I ask myself," Seligman said, "'What's it like to be an extrovert?' 'What's it like to be a teenage girl?' 'What's it like to be schizophrenic or neurotic?' or 'What's it like to be 70 years old?' these word clouds come much closer to the heart of the matter than do all the questionnaires in existence."

To test how accurately they were capturing people's traits through their open-vocabulary approach, the researchers split the volunteers into two groups and saw if a statistical model gleaned from one group could be used to infer the traits of the other. For three-quarters of the volunteers, the researchers used machine-learning techniques to build a model of the words and phrases that predict questionnaire responses. They then used this model to predict the age, gender and personalities for the remaining quarter based on their Facebook posts.

"The model was 92 percent accurate in predicting a volunteer's gender from their language usage," Schwartz said, "and we could predict a person's age within three years more than half the time. "Our personality predictions are inherently less accurate but are nearly as good as using a person's questionnaire results from one day to predict their answers to the same questionnaire on another day."

With the open-vocabulary approach shown to be equally or more predictive than closed approaches, the researchers used the word clouds to generate new insights into relationships between words and traits. For example, participants who scored low on the neurotic scale (i.e., those with the most emotional stability) used a greater number of words that referred to active, social pursuits, such as "snowboarding," "meeting" or "basketball."

"This doesn't guarantee that doing sports will make you less neurotic; it could be that neuroticism causes people to avoid sports," Ungar said. "But it does suggest that we should explore the possibility that neurotic individuals would become more emotionally stable if they played more sports."

By building a predictive model of personality based on the language of social media, researchers can now more easily approach such questions. Instead of asking millions of people to fill out surveys, future studies may be conducted by having volunteers submit their Facebook or Twitter feeds for anonymized study.

"Researchers have studied these personality traits for many decades theoretically," Eichstaedt said, "but now they have a simple window into how they shape modern lives in the age of Facebook."

More information: dx.plos.org/10.1371/journal.pone.0073791

Provided by University of Pennsylvania