Bringing state-of-the-art text analysis techniques to the social sciences

November 4 2015, by Laura Paisley

Morteza Dehghani and a team of psychology researchers and computer scientists at USC Dornsife have developed a new text analysis software program for social scientists.

According to a 2015 Global Web Index report, today's average adult spends 6.15 hours a day online. More than a quarter of this time is spent on social networking sites. Life in the digital age means an abundance—arguably an overabundance—of online information and communication.

Among social scientists, there is growing interest in using the vast

amount of text that is written online every day to explore different research questions. Certain tools are already available for very specific types of text analysis—looking at the content and structure of messages within a text—but a unified, open-source software system for gathering, managing and analyzing text hasn't existed before now.

Morteza Dehghani, assistant professor of psychology and computer science, has overseen the development at USC of TACIT (Text Analysis, Crawling and Interpretation Tool). The program aims to make sophisticated and highly customizable text analysis available to social science researchers.

"Currently [text analysis] techniques are available as independent programs or software, but they require a lot of expertise and because social scientists often don't have the programming background, they don't use them," Dehghani said. "So we've created a very researcher-friendly environment where they can easily access and use these methods. And if they want more, anyone can write their own plugins for the system."

TACIT offers a suite of techniques that can be used for a variety of textual analyses. The software is groundbreaking for its open-source, plugin architecture, a characteristic that will help ensure its continued growth and adaptation amid rapidly changing technology and scientific needs.

Bridging two disciplines

"We wanted to bring state-of-the-art text analysis techniques to the social sciences and bridge the two fields of psychology and computer science," Dehghani said.

Measuring psychological and demographic properties using

computational text analysis, he added, is becoming a field norm.

For example, a political scientist might be interested in analyzing political rhetoric. If researchers want to see the type of language applied by conservatives versus liberals regarding the death penalty, explained Dehghani, they can gather text through forums and newspapers and do text analyses to better understand the word choices surrounding these issues.

In terms of health psychology, Dehghani explained, researchers can explore a variety of questions: Can we better understand people's and groups' moral intuitions about different issues just by looking at what and how they write? Can we detect depression automatically through social media activity? Is there any way we can design intervention techniques if we can detect early signs of suicide?

TACIT's plugin architecture has three primary components to help users glean insights: a crawling plugin to allow for automated text collection from online sources like Twitter and Reddit in addition to content such as United States Supreme Court speech transcriptions; a corpus management plugin to allow processing and storing of formal bodies of text such as The New York Times corpus; and analysis plugins to do things like count instances or ratios of specific words, and otherwise identify and classify text related to research topics.

Reaching across the university

Dehghani's development team for TACIT is a collaboration between USC Dornsife and USC Viterbi School of Engineering. The group includes psychology doctoral students Kate M. Johnson and Joe Hoover as well as eight computer scientists from USC Viterbi's computer science department.

"It's an army, a team of people in Dornsife and Viterbi," Dehghani laughed. "I try to lead the effort."

The team has been working on the software for about a year and a half, with a beta 2 version set to release in December. The final release is expected in March 2016.

"In the first week the program launched," Dehgani said, "there were over 2,500 hits to our website. We had people from Kenya to Vietnam, from Uruguay to Estonia and from Hawaii to Maine downloading the software."

Though researchers have known since the era of Sigmund Freud that language affects cognition, psychologists became particularly interested in the subject of text analysis in the early 1990s. Some psychologists and social scientists started to develop individual techniques for text analysis, which proved very helpful, but researchers were still limited to one specific technique in any given software program.

During the same period, the fields of computational linguistics and computer science really took off, Dehghani said. This is what eventually gave rise to modern innovations like Siri, predictive typing, voice recognition and automatic language translation.

Dehghani and his team hope that TACIT can facilitate and encourage the use of advancements in computational linguistics in psychological research, and by doing so, help researchers make use of the ever-growing documents of our social discourse in ways that have previously not been possible.

Provided by University of Southern California