MedicalXpress

# Estimating county health statistics by looking at tweets

March 27 2014

A researcher at Illinois Institute of Technology (IIT) has found that Twitter knows if you're obese—or at least, if your county is. Tweets can accurately predict a county's rates of obesity, diabetes, teen births, health insurance coverage, and access to health foods, according to Aron Culotta, assistant professor of computer science and director of the Text Analysis in the Public Interest Lab. As a result, Twitter and other social media may complement other data sources for public health officials to identify at-risk communities and offer support. Culotta will report his findings in a paper, "Estimating County Health Statistics with Twitter," to be given at CHI 2014, the ACM (Association for Computing Machinery) CHI Conference on Human Factors in Computing Systems, April 26-May 1 in Toronto.

For each of the 100 most populous counties in the U.S., Culotta collected 27 health-related statistics. He also collected more than 1.4 million Twitter user profiles and 4.3 million Tweets over a nine-month span from the same 100 counties. He then performed a statistical analysis to identify how accurately the health outcomes can be predicted from the Twitter data and which linguistic markers are most predictive of each statistic.

Among other things, Culotta found the Tweets predicted county-level health statistics for 6 of 27 statistics, including obesity, diabetes, teen births, health insurance coverage, and access to healthy foods. Models that augmented demographic variables (race, age, gender, income) with linguistic variables (from Twitter) were more accurate than models using

demographic variables alone for 20 of the 27 health statistics considered. That is, the Twitter data helped to make the traditional models more accurate, suggesting that this new methodology can complement existing approaches. For two statistics—limited access to health foods and prevalence of fast foods—the Twitter model alone was actually more accurate than the demographic variable model.

Analysis of social media for most health concerns such as influenza focus on detecting specific mentions of a symptom of interest—e.g., "Staying home from work today with a sore throat." But Culotta investigated more nuanced linguistic cues that correlate with the overall health of a population. He identified the linguistic indicators that are most predictive of each outcome. For example, references to religion and certain pronouns ("we", "her") correlate with better socio-emotional support. References to money and inhibition correlate with lower unemployment. References to family and love correlate with higher rates of teen births. For obesity, indicators include what are known as Negative Engagement words (e.g., "tired", "bored", "sleepy"), as well as profanity.

"Twitter activity provides a more fine-grained representation of a community's health than demographics alone," Culotta said. "The reason for this appears to come from the insights Twitter provides into personality, attitudes, and behavior, which in turn correlate health outcomes.

The U.S. Centers for Disease Control and Prevention lead community health data collection and intervention efforts such as the Behavioral Risk Factor Surveillance System to identify vulnerable populations to better target intervention strategies. But such programs take considerable time and often are limited in sample size or geographic specificity. Culotta's research suggests that social media could be a complementary data source to identify at-risk communities.

Culotta said, "While this new methodology requires further experimentation, we believe it can aid public health researchers by providing (1) a more nuanced alternative to demographic profiles for identifying at-risk populations; (2) a low-cost method to measure risk across different subpopulations; (3) a process to help formulate new hypotheses about the relationship between environment, behaviors, and health outcomes, which can then be tested in a more controlled setting."

Provided by Illinois Institute of Technology

Citation: Estimating county health statistics by looking at tweets (2014, March 27) retrieved 23 April 2024 from https://medicalxpress.com/news/2014-03-county-health-statistics-tweets.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.