Geotagged Tweets as Predictors of County-Level Health Outcomes

Quynh Nguyen, PhD
University of Utah, College of Health
BD2K All Hands Meeting
Nov 30, 2016

Funding provided by NIH’s Big Data to Knowledge Initiative: 5K01ES025433 (PI: Dr. Nguyen)
Here Are the Most Popular Foods on Twitter | TIME

HEALTH | Tue Oct 18, 2016 | 11:53pm IST

What do tweets say about our health?

New Study Shows We Are What We Tweet When It Comes To Health And Happiness

Researchers mapped the happiest, healthiest tweets around the nation.

How Twitter shapes the food trends we follow

Study shows people take to social media to talk about what's healthy and how it makes them happy
What Factors Determine Our Health?

Family Health History

Environment

Behaviors/Lifestyles

Health care

Sources: CDC, WHO
What is the Built Environment?
Ways Social Processes and networks can affect health

1. **Norms** around healthy behaviors via informal social control

2. **Stimulation of new interests** such as a new sport or exercise

3. **Political advocacy** for access to neighborhood amenities and protection against stressors and toxic agents

4. **Emotional support**

5. The **dispersal of knowledge** about health promotion practices
Data challenges

• Consistently constructed neighborhood indicators across geographies

• Neighborhood surveys are costly and time-intensive
Twitter data collection

• April 2015– March 2016

• 80 million geotagged tweets from 603,363 unique Twitter users across the contiguous United States
HashtagHealth

- Uses geotagged social media data to characterize community characteristics
  - Happiness
  - Diet
  - Physical activity
  - Substance use
<table>
<thead>
<tr>
<th>Table 1. Descriptive statistics, county level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>% tweets that are happy</td>
</tr>
<tr>
<td>Calories density</td>
</tr>
<tr>
<td>% tweets about food</td>
</tr>
<tr>
<td>% tweets about healthy foods</td>
</tr>
<tr>
<td>% tweets about fast food</td>
</tr>
<tr>
<td>% tweets about physical activity</td>
</tr>
<tr>
<td>% tweets about alcohol</td>
</tr>
</tbody>
</table>

Twitter data collection period: April 2015– March 2016. County summaries were derived from 80 million tweets from the contiguous United States.
<table>
<thead>
<tr>
<th></th>
<th>Beta (95% CI)</th>
<th>Beta (95% CI)</th>
<th>Beta (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent food</td>
<td>-0.81 (-1.10, -0.51)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tweets</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent healthy</td>
<td>-0.40 (-0.67, -0.13)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>food tweets</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caloric density</td>
<td>0.35 (0.13, 0.56)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>of food mentions</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Covariates**

<table>
<thead>
<tr>
<th></th>
<th>Beta (95% CI)</th>
<th>Beta (95% CI)</th>
<th>Beta (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median age</td>
<td>-0.58 (-1.03, -0.12)*</td>
<td>-0.53 (-1.00, -0.05)*</td>
<td>-0.59 (-1.03, -0.14)*</td>
</tr>
<tr>
<td>Percent non-Hispanic white</td>
<td>0.16 (-0.79, 1.11)</td>
<td>0.12 (-0.83, 1.07)</td>
<td>0.14 (-0.81, 1.09)</td>
</tr>
<tr>
<td>Median household</td>
<td>-2.32 (-2.89, -1.76)*</td>
<td>-2.52 (-3.13, -1.91)*</td>
<td>-2.49 (-3.09, -1.89)*</td>
</tr>
<tr>
<td>income</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R-squared</td>
<td>0.32</td>
<td>0.30</td>
<td>0.30</td>
</tr>
<tr>
<td>N</td>
<td>3057</td>
<td>2899</td>
<td>3057</td>
</tr>
</tbody>
</table>

---

*2011-2013 Behavioral Risk Factor Surveillance System on adults 20 years and older

*Twitter variables were standardized to have a mean of 0 and standard deviation of 1. Adjusted linear regression models were run for each outcome separately. Models controlled for county-level demographics: median age, % non-Hispanic white, median household income. Standard errors accounted for clustering of county values at the state level

*p<0.05
Results at County Level

• 1 standard deviation increase in happy tweets was related to
  ▫ Lower obesity (-0.67%)
  ▫ Lower physical inactivity (-0.75%)

• 1 standard deviation increase food tweets related to
  ▫ Lower premature mortality (-339 per 100,000)
  ▫ Lower obesity (-0.81%)

• 1 sd increase in physical activity tweets related to
  ▫ Lower premature mortality (-204 per 100,000)
  ▫ Lower physical inactivity (-0.81%)
Results at County Level

• 1 standard deviation increase in alcohol tweets
  ▫ Higher excessive drinking (+0.48%)
  ▫ Higher percent drinking deaths that are alcohol related (+1.08%)
In Summary

• Twitter indicators of happiness, food, and physical activity were associated with lower premature mortality, obesity and diabetes at the county level.

• Social media represents a new data resource to conduct population assessments
Acknowledgments

Faculty
- Dr. Quynh Nguyen (Principal Investigator), Assistant Professor, College of Health
- Dr. Ken R. Smith, Professor, Department of Family and Consumer Studies
- Dr. James A. VanDerslice, Research Associate Professor, Department of Family and Preventive Medicine
- Dr. Ming Wen, Professor, Department of Sociology
- Dr. Feifei Li, Associate Professor, School of Computing

Research Assistants
- Hsien-Wen (Sherry) Meng, Ph.D. Student, College of Health
- Matt McCullough, Ph.D. Student, Department of Geography
- Debjoyti Paul, Ph.D. student, School of Computing

Alumni
- Suraj Kath, Software Engineer, Google
- Dapeng Li, Postdoctoral Researcher, Michigan State University
THANK YOU