Towards Real-Time Measurement of Public Epidemic Awareness

Monitoring Influenza Awareness through Twitter

Michael C. Smith
David A. Broniatowski
The George Washington University

Michael J. Paul
University of Colorado, Boulder

Mark Dredze
The Johns Hopkins University
Goal of Research

• Characterize influenza awareness signal (public concern) to determine which factors most influence disease awareness in a population

  • Influenza incidence rate
  • Infection surveillance using Twitter
  • News media regarding influenza
Agenda

• Introduction / Literature Review:
  • Why disease awareness?
  • How disease awareness?

• Data
  • Operationalize!

• Analysis
  • What drives disease awareness?

• Discussion and Conclusions
Why Disease Awareness?

- Studies have shown that the public’s awareness and their reaction to it may affect disease spread
  - Funk et al. (2009); Jones and Salathe (2009); Granell, Gomez, and Arenas (2013)

- Public health officials often manage and monitor disease awareness during epidemics or threats
  - Specific context: influenza

- Problem: no effective method exists for efficient, current awareness surveillance
  - Such systems exist for flu incidence

- We don’t know what drives awareness!
Agenda

• Introduction / Literature Review:
  • Why disease awareness?
  • How disease awareness?

• Data
  • Operationalize!

• Analysis
  • What drives disease awareness?

• Discussion and Conclusions
How Disease Awareness?

• Web and social media data!
• Numerous studies on disease surveillance showing benefits: cheap, real-time
  • Search queries:
    • Google Flu Trends (GFT): Ginsberg et al. (2009)
    • Yuan et al. (2013); Santillana et al. (2014); Preis and Moat (2014)
  • Social media
    • Culotta (2010); Aramaki, Maskawa, and Morita (2011); Lampos and Cristianini (2012)
    • Combining multiple sources - state-of-the-art
      • Santillana et al. (2015)
• Gov’t surveillance systems (tracking hospital visits) may not capture awareness (nor might other systems built to emulate)
• Social media might: people discuss and share concern
How Disease Awareness?

• Previous work separated infection from awareness

  • Lamb, Paul, and Dredze (2013); Broniatowski, Paul, Dredze (2013)

  • “I have the flu” vs. “tired of hearing about the flu”

  • “I’m sick with the flu” vs. “it’s flu season – don’t get sick”

  • “My kids gave me the flu” vs. “wash your hands, don’t get the flu”
Twitter flu prediction

Our current system uses a cascade of 3 MaxEnt classifiers:
• about health vs not about health
• about flu vs not about flu
• flu infection vs flu awareness

Estimated weekly flu rate:

\[
\frac{\text{# tweets about flu infection that week}}{\text{# of all tweets that week}}
\]
Twitter flu prediction

Features:

- Stylometry
  - Retweets, user mentions, URLs, emoticons
- 8 manually created word classes

<table>
<thead>
<tr>
<th>Infection</th>
<th>getting, got, recovered, have, having, had, has, catching, catch, cured, infected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease</td>
<td>bird, flu, sick, epidemic</td>
</tr>
<tr>
<td>Concern</td>
<td>afraid, worried, scared, fear, worry, nervous, dread, dreaded, terrified</td>
</tr>
<tr>
<td>Treatment/Prevention</td>
<td>vaccine, vaccines, shot, shots, mist, tamiflu, jab, nasal spray</td>
</tr>
</tbody>
</table>

...
Twitter flu prediction

Features:

• Part of speech templates
  • (subject, verb, object) tuples
    • always a good feature, IMO
  • numeric references
    • “100 more cases of swine flu”
• whether “flu” is a noun or adjective
  • “tired of the flu” vs “tired of the flu hype”
• whether “flu” is the subject or object
  • “I have the flu” vs “the flu is going around”
• … and others
How Disease Awareness?

• Previously focused on social media infection signal
• Now focus on social media awareness signal
  • 2012-2013 flu season

• Our work builds upon this by analyzing awareness
  • News media potential driver, can lead to overestimates in infection
    • ‘12–’13 GFT

• Studying flu awareness:
  • Yields trends for public health officials
  • Determines drivers for awareness in a population
Agenda

• Introduction / Literature Review:
  • Why disease awareness?
  • How disease awareness?

• Data
  • Operationalize!

• Analysis
  • What drives disease awareness?

• Discussion and Conclusions
Data

- 2012-2013 flu season
  - September 30, 2012 to May 25, 2013
- National and regional levels

Map of Health and Human Services regions
Source:
Data – Twitter Awareness

• Healthtweets.org (Dredze et al. 2014)

• Lamb et al. (2013); Broniatowski et al. (2013)
  • Normalized by public tweet counts
  • US tweets, and state level tweets
  • Also downloaded Twitter infection signal

Example trends from HealthTweets.org: weekly United States influenza counts
Data – Government Flu Incidence

• Center for Disease Control and Prevention (CDC)
  • US Outpatient Influenza-like Illness Surveillance Network (ILINet) yields Influenza-like Illness rates (ILI)
• Publicly available on the CDC’s flu dashboard:

![Screenshot of the CDC’s flu dashboard. Source: http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html](http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html)
Data – News Media

Data – Summary

• Twitter awareness

• Infection signals:
  • Twitter infection
  • CDC ILI

• News Media
Agenda

• Introduction / Literature Review:
  • Why disease awareness?
  • How disease awareness?

• Data
  • Operationalize!

• Analysis
  • What drives disease awareness?

• Discussion and Conclusions
Analysis – Overview

• Compare ILI to Twitter awareness, Twitter infection via Pearson correlations
  • How related are they?

• Plot weekly regional awareness, Twitter infection
  • How similar are the trends?

• Plot weekly awareness, ILI, infection, media
  • What can we glean?

• Regressions: drivers of awareness
  • Infection signals, news media
  • Regional, national
Analysis – Compare to Gold Standard

- Compare ILI to both awareness, Twitter infection
- Awareness significantly lower than Twitter infection (p=.029) nationally

<table>
<thead>
<tr>
<th>Region</th>
<th>Infection</th>
<th>Awareness</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.802</td>
<td>.588</td>
</tr>
<tr>
<td>2</td>
<td>.804</td>
<td>.620</td>
</tr>
<tr>
<td>3</td>
<td>.815</td>
<td>.575</td>
</tr>
<tr>
<td>4</td>
<td>.812</td>
<td>.489</td>
</tr>
<tr>
<td>5</td>
<td>.818</td>
<td>.547</td>
</tr>
<tr>
<td>6</td>
<td>.868</td>
<td>.633</td>
</tr>
<tr>
<td>7</td>
<td>.885</td>
<td>.626</td>
</tr>
<tr>
<td>8</td>
<td>.869</td>
<td>.667</td>
</tr>
<tr>
<td>9</td>
<td>.778</td>
<td>.548</td>
</tr>
<tr>
<td>10</td>
<td>.846</td>
<td>.658</td>
</tr>
<tr>
<td>National</td>
<td>.827</td>
<td>.555</td>
</tr>
</tbody>
</table>

Table 1: Correlations between the Twitter infection and awareness data and the CDC’s ILINet influenza prevalence data.
Analysis – Awareness Signal

• Awareness more similar than Twitter infection
• Differences in peak activity, off-peak activity
Analysis – Awareness Signal

US data

Values (z-scores)

Week (Number, Year)

-2

-1

0

1

2

3

4

5

CDC ILI

Awareness

Infection

Media
Analysis – Effect of News Media

<table>
<thead>
<tr>
<th>Awareness Correlation with:</th>
<th>Mean Regional Correlation</th>
<th>National Correlation:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Media</td>
<td>.905 (SD .044)</td>
<td>.940</td>
</tr>
<tr>
<td>Twitter Infection</td>
<td>.907 (SD .026)</td>
<td>.900</td>
</tr>
</tbody>
</table>

• Can we better define the relationship?
Analysis – Effect of News Media

• Bivariate linear regression model at the regional level:

\[ \text{awareness}_{rw} = \beta_{r0} + \beta_{r1}\text{infection}_{rw} + \beta_{r2}\text{media}_{rw} + \epsilon_{rw} \]

<table>
<thead>
<tr>
<th>Region</th>
<th>Infection</th>
<th>Media</th>
<th>Infection</th>
<th>Media</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.898</td>
<td>0.034</td>
<td>0.098</td>
<td>0.761</td>
</tr>
<tr>
<td>2</td>
<td>0.484</td>
<td>0.514</td>
<td>0.143</td>
<td>0.873</td>
</tr>
<tr>
<td>3</td>
<td>0.614</td>
<td>0.359</td>
<td>0.196</td>
<td>0.776</td>
</tr>
<tr>
<td>4</td>
<td>0.386</td>
<td>0.652</td>
<td>0.192</td>
<td>0.869</td>
</tr>
<tr>
<td>5</td>
<td>0.547</td>
<td>0.431</td>
<td>0.073</td>
<td>0.847</td>
</tr>
<tr>
<td>6</td>
<td>0.173</td>
<td>0.818</td>
<td>-0.003</td>
<td>0.978</td>
</tr>
<tr>
<td>7</td>
<td>0.341</td>
<td>0.645</td>
<td>0.119</td>
<td>0.852</td>
</tr>
<tr>
<td>8</td>
<td>0.580</td>
<td>0.401</td>
<td>0.161</td>
<td>0.785</td>
</tr>
<tr>
<td>9</td>
<td>0.490</td>
<td>0.531</td>
<td>0.186</td>
<td>0.834</td>
</tr>
<tr>
<td>10</td>
<td>0.561</td>
<td>0.435</td>
<td>0.228</td>
<td>0.741</td>
</tr>
<tr>
<td>National</td>
<td>0.340</td>
<td>0.645</td>
<td>0.0281</td>
<td>0.924</td>
</tr>
</tbody>
</table>

Table 2: Coefficients learned from two bivariate regression models that estimate each week’s flu awareness level (as measured from Twitter) as a linear combination of the week’s flu infection level and the week’s level of media attention (as measured by newspaper volume). The first model uses the Twitter-based estimate of flu infection, while the second model uses the CDC’s ILINet estimate.
Analysis – National / Regional Media

• What about national media?
• Similar regression model:
  • Weekly regional awareness as regional media and national media
  • Generated 10 regional models with coefficients for regional media and national media

<table>
<thead>
<tr>
<th>Mean Regional Media:</th>
<th>Mean National Media:</th>
</tr>
</thead>
<tbody>
<tr>
<td>-.088 (SD .582)</td>
<td>1.026 (SD.561)</td>
</tr>
</tbody>
</table>

Mean regression coefficients across all ten HHS regions

• National media explains much more than regional media!
Agenda

• Introduction / Literature Review:
  • Why disease awareness?
  • How disease awareness?

• Data
  • Operationalize!

• Analysis
  • What drives disease awareness?

• Discussion and Conclusions
Discussion

• Distinction between awareness and infection important for flu surveillance

• Awareness more a function of media than infection
  • National media more than regional media

• Little variation in regional awareness

• Awareness does not rise until the flu becomes severe
  • Drops sharply after peak, though infection still high
Conclusions

• Awareness more function of media than infection

• Opportunity: target only certain national distribution channels
  • National media levels contribute more than regional media

• Additional study:
  • Relationship more complex than regression model
  • News media

• Future work needed: generalize to other flu seasons

• Thoughts? Feedback?
  • mikesmith@gwu.edu
References