Analyzing Demographic Patterns of Vaccine Behaviors in Social Media

Michael Paul
University of Colorado Boulder

with Xiaolei Huang, Michael Smith, David Broniatowski, Mark Dredze, Karen Hilyard

Text as Data   |   October 15, 2016
Hip, Laid-Back Doctor Refers To Influenza As ‘The Flu’

NEWS IN PHOTOS
November 14, 2014
VOL 50 ISSUE 45
Local · Doctors · Disease · Healthcare

MORE FROM ONION, INC.

Infuriating: Man Licks Cone Instead Of Ice Cream (Presented by REDD’S Apple Ale)

Incoming Class Of Subway Trainees Spends Week Practicing On Sandwich Ca... The Onion
Overview

Focusing on influenza (flu), we’ll see how Twitter can be used to learn about:

• Disease incidence
• Disease awareness
• Disease prevention
  • Vaccination
Text as Data

United States

Influenza Rate

Week

Overview

Focusing on influenza (flu), we’ll see how Twitter can be used to learn about:

• Disease incidence
• Disease awareness
• Disease prevention
Flu Incidence: Motivation

Government (CDC) monitoring is the gold standard
- Data used for planning and forecasting
- But reports have a delay of 1-2 weeks

Text-driven systems can produce estimates immediately
- Internet data: we’ll focus on tweets
  - advantage: huge, public, free
Flu Incidence: Methods

Simple approach: keyword counting
Flu Incidence: Methods

Simple approach: keyword counting

Do Twitter users really describe colds this way?
Flu Incidence: Methods

Common approach: regression

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg¹, Matthew H. Mohebbi¹, Rajan S. Patel¹, Lynnette Brammer², Mark S. Smolinski¹ & Larry Brilliant¹

¹Google Inc. ²Centers for Disease Control and Prevention

\[
\text{logit}(P) = \beta_0 + \beta_1 \times \text{logit}(Q) + \varepsilon
\]

This is a scalar. Seems crazy if you know NLP!
Flu Incidence: Methods

Common approach: regression

Multivariate models have problems too:

\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_V x_{iV} \]

Flu rate in week \( i \) (given by CDC)

Count of word 2 in week \( i \)

Conclusion:
We need language understanding!
Flu Incidence: Methods

We only want to count tweets about the flu
  • Not about Christmas

We want to include only tweets that are experiential
  “think I’m coming down with the flu” vs “tired of hearing about the flu”

Our labeled data: Infection vs Awareness

What we’re trying to measure

Affected by panic, undue media attention
Flu Incidence: Results

Correlation with classifier: 0.990
Correlation with keywords: 0.977
Flu Incidence: Results

Correlation with classifier: 0.93
Correlation with keywords: 0.75
Flu Incidence: Results

Canada

England and Wales

United States

Australia

New Zealand

South Africa
Overview

Focusing on influenza (flu), we’ll see how Twitter can be used to learn about:

• Disease incidence
• Disease awareness
• Disease prevention
Flu Awareness: Motivation

• A population’s awareness of (and reaction to) a disease can affect its spread
• Officials need to manage public awareness in addition to managing the disease itself
Flu Awareness: Motivation

• A population’s awareness of (and reaction to) a disease can affect its spread

• Officials need to manage public awareness in addition to managing the disease itself

• Unlike flu incidence, **there are no standard surveillance systems for flu awareness**
Flu Awareness: Methods

We can use the same classifier as before.

Our labeled data: **Infection** vs **Awareness**

Also important to study.
Flu Awareness: Results
Overview

Focusing on influenza (flu), we’ll see how Twitter can be used to learn about:

• Disease incidence
• Disease awareness
• Disease prevention
Flu Vaccination: Motivation

• Vaccination one of biggest successes of public health
Flu Vaccination: Motivation

• Vaccination one of biggest successes of public health
• CDC wants flu vaccine coverage for 70% of Americans by 2020, but current numbers only 40-50% each year
• Critical to understand why people receive or refuse vaccines
  • And importantly, how this varies by geography and demographic group
Flu Vaccination: Methods

• Filtered for tweets matching relevant keywords
  • flu, influenza
  • shot(s), vaccine(s), vaccination(s)

• Dataset: about 1 million tweets since 2013

• Classified tweets according to the following criterion:
  • Does this tweet indicate that someone received or intends to receive a flu vaccine?
  • About 30% of the original tweets
  • Classifier F1 of .82 (similar to infection classifier)
Flu Vaccination: Methods

• Geolocated tweets using Carmen


• Inferred gender using Demographer

Rebecca Knowles, Josh Carroll, Mark Dredze. Demographer: Extremely Simple Name Demographics. EMNLP Workshop on Natural Language Processing and Computational Social Science, 2016.
Flu Vaccination: Results

The graph shows the number of tweets mentioning vaccine intention/receipt and other vaccine mentions over time. Peaks can be observed around August and September, likely corresponding to flu vaccine season.
Flu Vaccination: Results

Correlation for intention/receipt: 0.90
Correlation for other mentions: 0.82
Flu Vaccination: Results

By geographic region:

- Correlation for intention/receipt: .71
- Correlation for other mentions: .31
Flu Vaccination: Results

Vaccine coverage according to CDC:
Female: 42.0%  Male: 35.4%
Future Plans

• More analysis of our current results
• More demographic attributes
• Improved classifiers
• Sentiment classification
  • Goal: understand beliefs & attitudes
• How to integrate with traditional data sources
• Beyond validation: new findings

Thank you!