METHODS FOR IDENTIFYING PUBLIC HEALTH TRENDS

Mark Dredze
Department of Computer Science
Johns Hopkins University
Public Health

The prevention of disease, prolonging of life and promotion of health

- Disease surveillance
- Education
- Self medicating
- Tobacco use
- Vaccination
- Drug use
PUBLIC HEALTH CYCLE

Surveillance

REQUIRES:
DATA ON THE POPULATION

Population  Intervention  Doctors
WEB DATA

Twitter

Google Trends

Facebook

Forum

Bing Trends
PUBLIC HEALTH CYCLE

Population → Surveillance → Intervention → Doctors
PUBLIC HEALTH CYCLE

Surveillance

Find health trends

Data driven trend discovery

Measuring known trends
• Examples from public health
• Methods from computer science
PUBLIC AWARENESS

APRIL 10 is
#NYHAAD @YOUTHAIDSDAY

LEADING CAUSES OF EYE INJURIES
SPORTS ACCIDENTS CONSUMER FIREWORKS HOUSEHOLD CHEMICALS

APRIL is ALCOHOL AWARENESS MONTH
Help for Today Hope for Tomorrow

LEADERSHIP IN PREVENTION MONTH
1,000 EYE INJURIES OCCUR IN AMERICAN WORKPLACES

NATIONAL YOUTH HIV/AIDS AWARENESS DAY

INTERNATIONAL FOUNDATION
APRIL IS EYE INJURY AWARENESS MONTH
SEPTEMBER IS SPORTS EYE INJURY

DON'T TRY TO TREAT AN EYE INJURY YOURSELF. CONTACT YOUR EYE DOCTOR OR EMERGENCY ROOM IMMEDIATELY FOR HELP.

Image Credit: allshares.net
DO AWARENESS CAMPAIGNS WORK?
HOW MANY PEOPLE QUIT SMOKING?
MEASURE ONLINE BEHAVIOR

• Media covers awareness event
• People go online to find information about tobacco cessation
• Use these digital signals to quantify impact of GASO
ONLINE DATA SOURCES

- News articles
- Social media: Twitter
- Google searches
- Wikipedia page views

Google Trends

NEWS
AWARENESS EVENTS

• Not all awareness events are planned
ONLINE DATA SOURCES

• News articles (Bloomberg Terminal)
  • Sheen and HIV

• Google searches
  • Searches for HIV, condom, symptom, testing
Daily news stories and searches for "HIV"
C. Hourly trends for queries regarding condoms

D. Hourly trends for queries regarding HIV symptoms
honored, humbled, inspired, hopeful...

bit.ly/1VzvcRO
Brazilians not buying Zika excuse for babies with shrunken brains

Jan 25 2016

http://82.221.129.208/ifyouareinamericanouprobablycantseethisj9.html

Jim Stone

Brazilians not buying Zika excuse for babies with shrunken brains

Over 4,000 babies have now been born in Brazil with shrunken brains since November 1 2015. Brazil normally gets approximately 150 cases of this type of birth defect per year, which means that if this all happened in less than a three month time window, abnormal births of this type have increased by approximately 13,000 percent. HERE IS A KEY REPORT...
PSEUDO SCIENTIFIC BELIEFS

• “GMO Mosquitoes are the cause of the zika virus.”

• “#Zika may help accelerate Sterilization in the US, and with the use of GMO Mosquitoes sterility will be delivered to you, #Depopulation#NATO”

• 0.1 babies had zika, 100% had DTAP given to mother during pregnancy? Wonder which caused this?

• Factors: Those pregnant women were #Vaxxed=dtap,GMO mozzies released, pesticides put in drinking water so blame #Zika
All Zika vaccines  Pseudo-scientific claims
• Examples from public health
• Methods from computer science
TREND IDENTIFICATION

• Identify health trends in social media

  • Organize data into themes or topics

  • Identify patterns in topics over time, location, sub-population
APPROACH

• Topic modeling for Twitter
  • Probabilistic models of text, e.g. latent Dirichlet allocation
  • Identify major themes in corpus
  • Yields human interpretable topics
  • Identify/correlate topics with survey responses
  • Unsupervised: no tweet level supervision
SUPERVISION OF TOPIC MODELS

• Standard topic models: unsupervised

• Supervised LDA: Document labels

• What do we have?
  • Supervision on aggregated documents
  • e.g. 60% of people in this location think X
COLLECTIVE SUPERVISION

• Train models at the document level to make predictions at the population level

• Data

  • Telephone survey results

  • Prediction: estimating survey values for populations from social media features

  • Topic models can learn low-dimensional, generalizable features that can be used in predictive models

  • Analysis: Topic models are interpretable: we can better understand trends by viewing topics
APPROACH

- Modify topic models to incorporate collective supervision
- Extend different types of topic models in different ways, and compare
- Evaluate effectiveness at predicting public health telephone surveys
Latent Dirichlet Allocation (LDA)

\[ \tilde{\theta}_{mk} = \exp(b_k); \theta_m \sim \text{Dirichlet}(\tilde{\theta}_m) \]
\[ \tilde{\phi}_{kv} = \exp(b_v); \phi_k \sim \text{Dirichlet}(\tilde{\phi}_k) \]
\[ z_{mn} \sim \theta_m; w_{mn} \sim \phi_{z_{mn}} \]
Supervised LDA (sLDA) [2]

\( \bar{z}_{mk} \) is the average proportion of topic \( k \) in document \( m \)

\( y_m \sim \mathcal{N}(\eta_b + \eta^T \bar{z}_m, \sigma_y^2) \)
Let $\bar{z}_{jk}$ be the average proportion of topic $k$ in collection $j$

$y_j \sim \mathcal{N}(\eta_b + \eta^T \bar{z_j}, \sigma_y^2)$

Supervised LDA is a special case of this, where each document has its own unique collection ID.
Dirichlet-multinomial regression (DMR) [3]

- $\alpha_m = y_{cm}$, feature value associated with document’s collection $c_m$
- $\tilde{\theta}_{mk} = \exp(b_k + \alpha_m \eta_k); \theta_m \sim \text{Dirichlet}(\tilde{\theta}_m)$
- $\tilde{\phi}_{kv} = \exp(b_v); \phi_k \sim \text{Dirichlet}(\tilde{\phi}_k)$
DMR with adaptive supervision (Upstream-ada)

\[ \alpha_m \sim \mathcal{N}(y_{cm}, \sigma^2_\alpha) \]

\[ \tilde{\theta}_{mk} = \exp(b_k + \alpha_m \eta_k) \]

\[ \tilde{\phi}_{kv} = \exp(b_v); \phi_k \sim \text{Dirichlet}(\tilde{\phi}_k) \]

Document value can deviate from given input – can help infer likely values when supervision is noisy or missing.
Supervision affects priors over words. Extension to DMR known as SPRITE [7].

- $\alpha_m = y_{cm}$
- $\tilde{\theta}_{mk} = \exp(b_k + \alpha_m \eta_k)$
- $\tilde{\phi}_{kv} = \exp(b_v + \omega_v \eta_k)$
DMR + adaptive + word prior (Upstream-ada-words)

- Combined upstream model

\[ \alpha_m \sim \mathcal{N}(y_{cm}, \sigma_\alpha) \]

\[ \tilde{\theta}_{mk} = \exp(b_k + \alpha_m \eta_k) \]

\[ \tilde{\phi}_{kv} = \exp(b_v + \omega_v \eta_k) \]
Behavioral Risk Factor Surveillance System: annual survey by US federal government to learn about health/behavior of population.

We selected three questions from BRFSS phone surveys:

- **Guns**: Do you have a firearm in your house? (2001)
- **Vaccines**: Have you had a flu shot in the past year? (2013)
- **Smoking**: Are you a current smoker? (2013)

Survey responses are aggregated at the level of US state.
Twitter Data

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Vocab</th>
<th>BRFSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guns</td>
<td>12,358</td>
<td>Owns firearm</td>
</tr>
<tr>
<td>Vaccines</td>
<td>13,451</td>
<td>Had flu shot</td>
</tr>
<tr>
<td>Smoking</td>
<td>13,394</td>
<td>Current smoker</td>
</tr>
</tbody>
</table>

- 100,000 tweets per dataset (filtered by relevant keywords)
  - collected between Dec. 2012 - Jan. 2015
- Identified as English using langid
  - https://github.com/saffsd/langid.py
- Stopwords removed and low-frequency tokens excluded
- Location inferred using Carmen
  - https://github.com/mdredze/carmen-python
For each dataset:

- Each collection is defined as the set of tweets per US state
  - 50 collections
- Each collection’s $y_c$ value is the proportion respondents answering “Yes” to the BRFSS question

Predicting survey values:

- L2-regularized linear regression model
- Features: mean topic distributions $\theta$ per collection
Lots of hyperparameters – selected hyperparameters that maximized perplexity on heldout sample

Optimized each model using Spearmint: 
https://github.com/JasperSnoek/spearmint

Fit models using Gibbs sampling with AdaGrad for parameter ($\eta$) optimization

Prediction task tuned with 5-fold cross validation: 80% train, 10% dev, 10% test.
## Results

<table>
<thead>
<tr>
<th>Features</th>
<th>Model</th>
<th>Guns</th>
<th>Vaccines</th>
<th>Smoking</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>RMSE</td>
<td>Perplexity</td>
<td>RMSE</td>
</tr>
<tr>
<td>None</td>
<td>LDA</td>
<td>17.44</td>
<td>2313 (±52)</td>
<td>8.67</td>
</tr>
<tr>
<td>Survey</td>
<td>Upstream</td>
<td>15.37</td>
<td>1529 (±12)</td>
<td>6.54</td>
</tr>
<tr>
<td></td>
<td>Upstream-words</td>
<td>11.50</td>
<td>1429 (±22)</td>
<td>6.37</td>
</tr>
<tr>
<td></td>
<td>Upstream-ada</td>
<td>11.48</td>
<td>1506 (±67)</td>
<td>5.82</td>
</tr>
<tr>
<td></td>
<td>Upstream-ada-words</td>
<td>11.47</td>
<td>1535 (±28)</td>
<td>7.20</td>
</tr>
<tr>
<td></td>
<td>Downstream-sLDA</td>
<td>11.52</td>
<td>1561 (±22)</td>
<td>11.22</td>
</tr>
<tr>
<td></td>
<td>Downstream-collective</td>
<td>12.81</td>
<td>1573 (±20)</td>
<td>9.17</td>
</tr>
</tbody>
</table>

Michael J. Paul (University of Colorado, Boulder)

Collective Supervision of Topic Models for Predicting Surveys with Social Media

February 14, 2016
UBCs were a big US political issue in 2013, when national gun control legislation was floated.

We collected surveys on support for UBCs for 22 states from various polls (mostly Public Policy Polling).

Baseline: use older 2001 survey of proportion households containing a firearm.
Use Case – Support for Universal Background Checks

<table>
<thead>
<tr>
<th>Features</th>
<th>Model</th>
<th>RMSE (2001 Y included)</th>
<th>RMSE (2001 Y omitted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>No model</td>
<td>7.26</td>
<td>7.59</td>
</tr>
<tr>
<td></td>
<td>Bag of words</td>
<td>5.16</td>
<td>7.31</td>
</tr>
<tr>
<td></td>
<td>LDA</td>
<td>6.40</td>
<td>7.59</td>
</tr>
<tr>
<td>Survey</td>
<td>Upstream-ada-words</td>
<td>5.11</td>
<td>5.48</td>
</tr>
</tbody>
</table>

Michael J. Paul (University of Colorado, Boulder)
Collective Supervision of Topic Models for Predicting Surveys with Social Media
February 14, 2016 18 / 22
Use Case – Support for Universal Background Checks
Thank You

http://socialmediahealthresearch.org

mdredze@cs.jhu.edu

www.dredze.com

John Ayers
David Broniatowski
Michael Paul
Adrian Benton
Michael Smith
Glen Coppersmith
Craig Harman

Ben Althouse
Morgan Johnson
Jon-Patrick Allem
Matt Childers
Karen Hilyard
Angie Chen