MAKING SENSE OF THE WEB FOR PUBLIC HEALTH USING NLP

MICHAEL J. PAUL
JOHNS HOPKINS UNIVERSITY
TALK OVERVIEW

• Introductory examples
• Brief background on NLP
• Two applications of NLP:
  • Influenza surveillance
  • Cancer search behavior
Google Flu Trends:
HealthMap:
COMPUTATIONAL EPIDEMIOLOGY

Data act as a “sensor” of population-wide behavior.
Air pollution in Chinese social media:

Relationship between pollution levels and weibos
Understanding healthcare quality through online doctor reviews:

Text from reviews is significantly predictive of external measures of healthcare quality

RateMDs.com: Find a Doctor Bro.
Monitoring allergy prevalence in Twitter:
COMPUTATIONAL EPIDEMIOLOGY

Data act as a “sensor” of population-wide behavior.

How do we convert this text into useable data?
Most common approach: regression

**Detecting influenza epidemics using search engine query data**

Jeremy Ginsberg\(^1\), Matthew H. Mohebbi\(^1\), Rajan S. Patel\(^1\), Lynnette Brammer\(^2\), Mark S. Smolinski\(^1\) & Larry Brilliant\(^1\)

\(^1\)Google Inc. \(^2\)Centers for Disease Control and Prevention

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

where \( P \) is the percentage of ILI physician visits, \( Q \) is the ILI-related query fraction, \( \beta_0 \) is the intercept, \( \beta_1 \) is the multiplicative coefficient, and \( \varepsilon \) is the error term. \( \text{logit}(P) \) is the natural log of \( P/(1-P) \).
• Most common approach: regression

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

where \( P \) is the percentage of ILI physician visits, \( Q \) is the ILI-related query, \( \beta_1 \) is the multiplicative term, \( \text{logit}(P) \) is the natural log of \( P/(1-P) \).

This is a univariate model. Not very rich...
FROM TEXT TO DATA

- Most 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 \( i \)

<table>
<thead>
<tr>
<th>2009-2010</th>
<th>2012-2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>flu</td>
<td>christmas</td>
</tr>
<tr>
<td>sick</td>
<td>sick</td>
</tr>
<tr>
<td>swine</td>
<td>flu</td>
</tr>
<tr>
<td>shot</td>
<td>strong</td>
</tr>
<tr>
<td>cancer</td>
<td>processing</td>
</tr>
<tr>
<td>fever</td>
<td>snow</td>
</tr>
<tr>
<td>h1n1</td>
<td>new</td>
</tr>
<tr>
<td>#beatcancer</td>
<td>want</td>
</tr>
<tr>
<td>better</td>
<td>hard</td>
</tr>
<tr>
<td>getting</td>
<td>better</td>
</tr>
<tr>
<td>home</td>
<td>body</td>
</tr>
<tr>
<td>halloween</td>
<td>best</td>
</tr>
<tr>
<td>breast</td>
<td>coughing</td>
</tr>
<tr>
<td>cough</td>
<td>festivities</td>
</tr>
<tr>
<td>throat</td>
<td>eve</td>
</tr>
</tbody>
</table>

words with highest \( \beta \) values
• Most 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 \(i\)

2009-2010  
- flu  
- sick  
- swine  
- shot  
- cancer  
- fever  
- h1n1  
- #beatcancer  
- better  
- getting  
- home  
- halloween  
- breast  
- cough  
- throat  

2012-2013  
- christmas  
- sick  
- flu  
- strong  
- processing  
- snow  
- new  
- want  
- hard  
- better  
- body  
- best  
- coughing  
- festivities  
- eve

Conclusion: our tools need better language understanding
TALK OVERVIEW

• Introductory examples

• **Brief background on NLP**

• Two applications of NLP:
  • Influenza surveillance
  • Cancer search behavior
NATURAL LANGUAGE PROCESSING

Teaching computers to understand human language

For example...

- Automatic translation

- Text message autocorrection

- Question answering (e.g., Siri)
NATURAL LANGUAGE PROCESSING

- Standard task: automatic text categorization
NATURAL LANGUAGE PROCESSING

• Standard task: automatic text categorization

• Classifier:
  • takes a document as input
  • outputs a guess for the correct category
**Machine learning:**

- *train* a classifier by giving it labeled examples
- the classifier can then give labels to new documents

**Training:**
- Knicks End Losing Streak
- Apple may launch interactive iPad stylus soon
- Seahawks vs. Packers rewind: What we learned
- Meet the $1 million robot you can buy on Amazon

**Classification:**
- No shortage of Seahawks-Patriots storylines in Super Bowl
- ????
Common classification model: **logistic regression**

- Takes a set of predictors about a document
  - typically counts of words or phrases
- Outputs a value close to 0 or 1
  - binary classification
Common classification model: **logistic regression**

- You’ve probably used LR, but maybe in different ways…

<table>
<thead>
<tr>
<th>Epidemiology</th>
<th>NLP / Machine Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model is applied to a fixed set of data (e.g. about a population)</td>
<td>Model is intended to be applied to unseen data in the future</td>
</tr>
<tr>
<td>Goal is usually to analyze effect of predictors</td>
<td>Don’t care what the predictors are; just want good predictions</td>
</tr>
<tr>
<td>Typically on the order of 10 or 100 predictors</td>
<td>Potentially on the order of 100K or 1M predictors (aka features)</td>
</tr>
<tr>
<td>Evaluation: goodness of fit ($R^2$, AIC, etc.)</td>
<td>Evaluation: classification accuracy, ROC curves, etc.</td>
</tr>
</tbody>
</table>
TALK OVERVIEW

• Introductory examples
• Brief background on NLP
• Two applications of NLP:
  • Influenza surveillance
  • Cancer search behavior
INFLUENZA SURVEILLANCE

• Government flu monitoring is the gold standard
  • But reports have a delay of ~2 weeks (or longer, if the government shuts down 😊)

• Text-driven systems can produce estimates immediately
  • This talk: let’s use tweets
    • advantage: huge, public, free
We only want to count tweets about the flu
- Not about Christmas or breast cancer

We want to include only tweets that are experiential

“think I’m coming down with the flu” vs “tired of hearing about the flu”
TWITTER FLU PREDICTION

We only want to count tweets about the flu
- Not about Christmas or breast cancer

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 classifier: **Infection** vs **Awareness**

- What we’re trying to measure
- Affected by panic, undue media attention
The \textbf{infection vs awareness} distinction matters!

Google concluded that media attention was a primary cause of their huge overestimate in 2012-2013.

"flu symptoms" – not an experiential query
Our current system uses 3 logistic regression 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

Correlation with classifier: 0.990
Correlation with keywords: 0.977
TWITTER FLU PREDICTION

Correlation with classifier: 0.93
Correlation with keywords: 0.75
TWITTER FLU AWARENESS

Our classifier: Infection vs Awareness

What most researchers have looked at

Another interesting trend we can examine!

We can ask things like...

• How high is public awareness?

• Does public awareness track infection rates?
  • Or does it more closely follow news media attention?
TWITTER FLU AWARENESS

Infection:

Awareness:
TALK OVERVIEW

• Introductory examples
• Brief background on NLP
• Two applications of NLP:
  • Influenza surveillance
  • Cancer search behavior
MEDICAL SEARCH

• **What information** do patients need?
  • and **when**?

• How do people use the web to make **decisions**?
  • e.g. choice of treatment, choice of doctor

Our projects focused on **breast and prostate cancer**
• Approach: large scale analysis of anonymized logs
Starting point: filter for search histories containing "breast cancer" ≥ 3 times
Starting point: filter for search histories containing "breast cancer" $\geq 3$ times

• But people search for lots of reasons...

Internet Opens Up Whole New World Of Illness For Local Hypochondriac

MERIDEN, CT—All her life, Janet Hartley has suffered from a host of ill-defined viruses and inexplicable aches and pains, diagnosing herself with everything from diabetes to cancer. But ever since discovering such online medical resources as WebMD, drkoop.com, and Yahoo! Health, the 41-year-old hypochondriac has had a whole new world of imaginary illnesses opened up to her.

"The Internet has really revolutionized my ability to keep on top of my medical problems," said Hartley, speaking from her bed. "For instance, I used to think my headaches were just really bad migraines. But then last week, while searching Mt. Sinai Hospital's online medical database, I learned about something much more serious called cranial AVM, or arteriovenous malformation, which, along with headache pain, may also result in dizziness, loss of concentration, and..."
As before, we need to identify experiential search

- Built **NLP classifiers** by labeling 480 anonymized search histories

We looked for search queries that were consistent with a cancer diagnosis:

<table>
<thead>
<tr>
<th>Time</th>
<th>Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nov 13 2013 7:40pm</td>
<td>feels like lump in breast</td>
</tr>
<tr>
<td>Dec 1 2013 11:21am</td>
<td>pain after biopsy</td>
</tr>
<tr>
<td>Dec 1 2013 11:31am</td>
<td>what happens after breast biopsy</td>
</tr>
<tr>
<td>Dec 9 2013 6:33pm</td>
<td>how often are breast lumps cancer</td>
</tr>
<tr>
<td>Dec 9 2013 6:45pm</td>
<td>does cancer make you thirsty</td>
</tr>
<tr>
<td>Dec 9 2013 6:49pm</td>
<td>how long does it take for biopsy results</td>
</tr>
<tr>
<td>Dec 12 2013 12:08pm</td>
<td>stage 2a breast cancer</td>
</tr>
<tr>
<td>Dec 12 2013 12:15pm</td>
<td>invasive ductal carcinoma</td>
</tr>
<tr>
<td>Dec 12 2013 12:17pm</td>
<td>poorly differentiated idc breast cancer</td>
</tr>
<tr>
<td>Dec 12 2013 12:29pm</td>
<td>breast cancer survival rate</td>
</tr>
<tr>
<td>Dec 12 2013 12:32pm</td>
<td>stage 2 breast cancer survival rate</td>
</tr>
<tr>
<td>Dec 12 2013 7:44pm</td>
<td>breast reconstruction surgery</td>
</tr>
<tr>
<td>Dec 12 2013 7:46pm</td>
<td>breast reconstruction after cancer</td>
</tr>
<tr>
<td>Dec 13 2013 8:05am</td>
<td>breast cancer treatment</td>
</tr>
<tr>
<td>Dec 13 2013 8:16am</td>
<td>recovering from breast cancer</td>
</tr>
<tr>
<td>Dec 15 2013 09:20am</td>
<td>breast cancer surgeon</td>
</tr>
<tr>
<td>Dec 15 2013 10:22am</td>
<td>full mastectomy</td>
</tr>
<tr>
<td>Dec 15 2013 10:23am</td>
<td>mastectomy pros and cons</td>
</tr>
<tr>
<td>Dec 15 2013 10:29am</td>
<td>do you need chemo after mastectomy</td>
</tr>
</tbody>
</table>

(Separately, we are working on obtaining ground truth from patients)
SEARCH AND BREAST CANCER

Chemotherapy searches

Average number of queries

Days since DDX

- Stage 0
- Stage I
- Stage II
- Stage III
- Stage IV
External validation

• **Geography**: correlation with US state incidence
  - Keyword filter: .036 (i.e. “breast cancer” 3 times)
  - With classifier: .348 (a tenfold increase!)

• **Gender** (100 times more common in women):
  - Keyword filter: 70.0% women
  - With classifier: 88.9% women

• **Age** (6 times more common in elderly):
  - Keyword filter: 5.4% aged 65+
  - With classifier: 22.2% aged 65+
CONCLUSION

It’s clear that experiential classification is important. This requires NLP. But there’s much more to do!

Interesting problems for language understanding:

• mining *attitudes, perceptions, and behaviors*
THANKS TO MANY COLLABORATORS:

Mark Dredze (advisor)

Flu:
• David Broniatowski
• Alex Lamb
• Michael Smith

Cancer search:
• Eric Horvitz
• Ryen White
• Sara Javid
• Janice Tsai

Air pollution:
• Shiliang Wang
• Angie Chen
• Brian Schwartz

Doctor reviews:
• Byron Wallace
• Urmimala Sarkar
• Thomas Trikalinos
• Andrea Lopez