MAKING SENSE OF THE WEB FOR PUBLIC HEALTH USING NLP

MICHAEL J. PAUL
JOHNS HOPKINS UNIVERSITY

Google Flu Trends:

Explore flu trends - United States

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

National

- Intense
- High
- Moderate
- Low
- Minimal

2014-2015 - Past years
PUBLIC HEALTH + WEB

HealthMap:

![HealthMap World Map](image)
EPIDEMIOLOGY

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

How do we convert this text into useable data?

Data act as a “sensor” of population-wide behavior
• Simplest approach: keyword counting
FROM TEXT TO DATA

- Simplest approach: keyword counting
FROM TEXT TO DATA

• Simplest approach: keyword counting

Do Twitter users really describe colds this way?
Do Twitter users really describe colds this way?
• Most 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

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

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 \( \epsilon \) is the error term. \( \text{logit}(P) \) is the natural log of \( P/(1-P) \).
FROM TEXT TO DATA

• Most 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

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

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 scalar. Seems crazy to an NLPPer!
• 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
WE NEED LANGUAGE UNDERSTANDING!

(This is the point of my talk)
TALK OVERVIEW

• Three applications for NLP:
  • Influenza surveillance
  • Air pollution monitoring
  • Medical search behavior

• What’s next?
TALK OVERVIEW

• Three applications for NLP:
  • Influenza surveillance
  • Air pollution monitoring
  • Medical search behavior

• What’s next?
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”
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 labeled data: Infection vs Awareness

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

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

\[
\frac{\text{# tweets about flu infection that week}}{\text{# of all tweets that week}}
\]
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
TWITTER FLU PREDICTION

F1 = 0.7891

F1 = 0.7985

Recall vs. Precision graph with two curves:
- Red: N-Grams
- Blue: All Features

Precision range: 40 to 100
Recall range: 0 to 100
TWITTER FLU PREDICTION

Correlation with classifier: 0.990
Correlation with keywords: 0.977
Correlation with classifier: **0.93**
Correlation with keywords: **0.75**
TALK OVERVIEW

• Three applications for NLP:
  • Influenza surveillance
  • **Air pollution monitoring**
  • Medical search behavior

• What’s next?
AIR POLLUTION IN CHINA

What do people have to say about air quality on Sina Weibo?

• Can social media detect pollution levels?
• Can we learn about health effects and behavioral response?
AIR POLLUTION IN CHINA

What do people have to say about air quality on Sina Weibo?

• Can social media detect pollution levels?
• Can we learn about health effects and behavioral response?
AIR POLLUTION IN CHINA

Data pipeline:

1. Started with 93 million crawled Weibo messages
2. Filtered to 1 million messages with health keywords
3. Ran LDA with 100 topics
4. Analyzed messages with the air pollution topic
Validation: We compared the volume of messages with this topic to government-provided pollution rates.

Correlation across 74 cities: 583
We then annotated a small sample of topical messages with detailed codes.
AIR POLLUTION IN CHINA

We then annotated a small sample of topical messages with detailed codes.

As with flu, we trained a cascade of classifiers:

1. **Relevant** to air pollution?
2. **Experiential**?

Features: 1, 2, 3-grams
Validation: We compared the volume of messages with this topic to government-provided pollution rates.

Correlation across 74 cities: .583

with experiential classifiers: .718
TALK OVERVIEW

• Three applications for NLP:
  • Influenza surveillance
  • Air pollution monitoring
  • Medical search behavior

• What’s next?
MEDICAL SEARCH

Scientific questions:

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

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

Engineering goals:

• How can we make **search engines better** to support these goals?
What do people search when confronted with a major illness?

Our project focused on breast and prostate cancer.
- I’ll just talk about the first today.

Approach: large scale analysis of anonymized logs.
- Step 1: retrieve search histories about breast cancer.
SEARCH AND BREAST CANCER

**Starting point:** filter for search histories containing “breast cancer” ≥ 3 times
Starting point: filter for search histories containing "breast cancer" ≥ 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

**Classifier:**

- Annotated 480 partial histories
  - filtered for relevant queries
- Trained with boosted decision trees
### EXPERIENTIAL SEARCH PREDICTION

**Features:**
- Ontology of terms
- each category is a feature

<table>
<thead>
<tr>
<th>Category</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cosmetic</td>
<td>Post-Surgery</td>
<td>Post-Surgery</td>
<td>{cosmetic, plastic} {surgery, surgeon}, prosthesis, prosthetic(s),</td>
</tr>
<tr>
<td>Cosmetic</td>
<td>Hair Loss</td>
<td>Hair Loss</td>
<td>implant(s), reconstruction (\text{wig(s)}, \text{head {scarf, scarves, covering(s)}}, \text{hair (re)grow(\text{th})})</td>
</tr>
<tr>
<td>Description</td>
<td>Type</td>
<td>Cancer Type</td>
<td>DCIS, LCIS, IDC, ILC, lobular, ductal, in situ, metastatic, mucinous,</td>
</tr>
<tr>
<td>Description</td>
<td>Stage</td>
<td>Grade</td>
<td>inflammatory (\text{what stage, stages, staging, what grade, grades, grading, differentiated})</td>
</tr>
<tr>
<td>Description</td>
<td>Grade</td>
<td>Grade</td>
<td>pre(\text{cancer}, early stage, stage {0–4}, zero–four, [I–IV]}{a,b,c})</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Diagnosis</td>
<td>Diagnosis</td>
<td>\text{grade {I–III}, {low, moderate, intermediate, high} grade}</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Diagnostics</td>
<td>Biopsy</td>
<td>\text{diagnosis, diagnosed}</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Screening</td>
<td>Mammography</td>
<td>biopsy, biopsies</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Screening</td>
<td>Ultrasound</td>
<td>\text{mammogram(s), mammography}</td>
</tr>
<tr>
<td>Lifestyle</td>
<td>Diet</td>
<td>Diet</td>
<td>ultrasound(s)</td>
</tr>
<tr>
<td>Lifestyle</td>
<td>Fitness</td>
<td>Diet</td>
<td>diet(s), eat(ing), food(s), vitamin(s), supplements, nutrition, protein,</td>
</tr>
<tr>
<td>Professional</td>
<td>Healthcare</td>
<td>Provider</td>
<td>\text{recipe(s), cookbook}</td>
</tr>
<tr>
<td>Professional</td>
<td>Healthcare</td>
<td>Doctor</td>
<td>\text{fitness, exercise(s), yoga}</td>
</tr>
<tr>
<td>Professional</td>
<td>Healthcare</td>
<td>Oncologist</td>
<td>\text{clinic(s), hospital(s), cancer center(s)}</td>
</tr>
<tr>
<td>Treatment</td>
<td>Treatment</td>
<td>Treatment</td>
<td>\text{doctor(s), physician(s)}</td>
</tr>
<tr>
<td>Treatment</td>
<td>Treatment</td>
<td>Side Effects</td>
<td>oncologist(s)</td>
</tr>
<tr>
<td>Treatment</td>
<td>Chemotherapy</td>
<td>Chemotherapy</td>
<td>treatment(s), medication(s), med(s)</td>
</tr>
<tr>
<td>Treatment</td>
<td>Chemotherapy</td>
<td>Side Effects</td>
<td>side effect(s)</td>
</tr>
<tr>
<td>Treatment</td>
<td>Chemotherapy</td>
<td>Side Effects</td>
<td>chemotherapy, chemo, cemo, kemo</td>
</tr>
<tr>
<td>Treatment</td>
<td>Chemotherapy</td>
<td>Side Effects</td>
<td>hair loss, hair fall(ing), {lose, losing} {my, your} hair</td>
</tr>
</tbody>
</table>
EXPERIENTIAL SEARCH PREDICTION

Features:

- Language features:
  - First/second person pronouns (including possessives)
  - Experiential phrases (e.g. “i have”, “i was diagnosed”)
  - Starts with a question word

- Volume and temporal patterns:
  - % of queries/sessions containing ontology terms
  - Length of cancer-related sessions
  - Time between cancer-related sessions
  - Ordering of categories searched
  - ... and a lot more
External validation

• **Geography**: correlation with state rates
  - 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+
We also built classifiers to identify the inferred **day of diagnosis (DDX)**

• I’ll skip the details today

This gives a common point for **aligning** the 1700 histories tagged by the classifier
BREAKDOWN BY STAGE

Days since DDX

Average number of queries

Screening/workup searches

- Stage 0
- Stage I
- Stage II
- Stage III
- Stage IV
BREAKDOWN BY STAGE

Average number of queries vs. Days since DDX, showing surgery searches for different stages.
WHAT’S NEXT?

(And what role will NLP play?)
Another application: medical mistakes in Twitter
- important public health issue
- not well understood

Our qualitative study:
- just need to find some relevant tweets to examine
- so we came up with reasonable search terms...
Surprisingly many false positives...

• I hope the **doctor was wrong** and a miracle happens
• The antibiotics were just to prevent **surgery infection**.
• I think the **hospital gave me the wrong kid lol**
• I hate going to the **stupid doctor**
• on my way to the **hospital fucked up my knee**
• I'm just drowsy... I bought the **wrong meds**.
• You must be on some **wrong pills bro**
Why Big Data Missed the Early Warning Signs of Ebola

Hint: Ils ne parlent pas le français.

By Kalev Leetaru | September 26, 2014
We still need better NLP

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 PEOPLE

- Mark Dredze (advisor)
- Microsoft Research (funding)

Flu:
- David Broniatowski
- Alex Lamb
- Nicholas Generous

Pollution:
- Shiliang Wang
- Angie Chen
- Brian Schwartz

Medical search:
- Eric Horvitz
- Ryen White
- Sara Javid
- Janice Tsai

Patient safety:
- Sarah Bell
- Atul Nakhasi
- Ralph Passarellá
- Peter Pronovost