Tracking and Predicting the Spread of Disease using Social Media

Henry Kautz
Adam Sadilek
Department of Computer Science
Vincent Silenzio
Department of Psychiatry
JazzStarDiva [0.94634229]
I'm getting sick on this bus
#literally

Micah_88 [1.14140003]
oh ok I was sick n still is
Organic Sensor Networks

- Detailed measurements at a population scale
  - No active user participation
  - Fine granularity
  - Timely

- Inference & Prediction
  - What will happen in the future?
  - What factors (places/events/actions) influence health?
Organic Sensor Networks

- 52% of adults use online social networks
- Smartphone access
  - Real time
  - “In the moment”
  - Location aware
The Data

- New York City
  - 16M tweets / month, 0.6M users
- 6K geo-active users
  - 2.5M tweets by geo-active users
- 103K “follows” relationships
- 32K “friends” relationships
Language Model

- How to automatically classify a tweet as “sick”?
- Previous approaches: keywords, e.g. “sick”
- Problems
  - “Sick of homework”
  - “Under the weather”
- Our approach: statistical word trigram model
Each trigram is a feature (dimension)

Support vector machine: find a hyperplane that separates positive from negative examples
sick

+0.8

+0.8
sick and tired

+0.8

-0.7

+0.6

+0.7
sick and tired of

\[ +0.8 \quad -0.7 \quad +0.6 \quad -0.8 \]

\[ -0.1 \]
sick and tired of flu

How do we get these numbers???
Cascade SVM

Corpus of 5,128 tweets labeled by human workers

Corpus of 1.6 million machine-labeled tweets

Random sample of 200 million tweets

Corpus of "other" tweets

Corpus of "sick" tweets

Final corpus

Training
Fig. 1. SVM Features

<table>
<thead>
<tr>
<th>Positive Features</th>
<th>Weight</th>
<th>Negative Features</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>sick</td>
<td>0.9579</td>
<td>sick of</td>
<td>−0.4005</td>
</tr>
<tr>
<td>headache</td>
<td>0.5249</td>
<td>you</td>
<td>−0.3662</td>
</tr>
<tr>
<td>flu</td>
<td>0.5051</td>
<td>lol</td>
<td>−0.3017</td>
</tr>
<tr>
<td>fever</td>
<td>0.3879</td>
<td>love</td>
<td>−0.1753</td>
</tr>
<tr>
<td>feel</td>
<td>0.3451</td>
<td>i feel your</td>
<td>−0.1416</td>
</tr>
<tr>
<td>coughing</td>
<td>0.2917</td>
<td>so sick of</td>
<td>−0.0887</td>
</tr>
<tr>
<td>being sick</td>
<td>0.1919</td>
<td>bieber fever</td>
<td>−0.1026</td>
</tr>
<tr>
<td>better</td>
<td>0.1988</td>
<td>smoking</td>
<td>−0.0980</td>
</tr>
<tr>
<td>being</td>
<td>0.1943</td>
<td>i’m sick of</td>
<td>−0.0894</td>
</tr>
<tr>
<td>stomach</td>
<td>0.1703</td>
<td>pressure</td>
<td>−0.0837</td>
</tr>
<tr>
<td>and my</td>
<td>0.1687</td>
<td>massage</td>
<td>−0.0726</td>
</tr>
<tr>
<td>infection</td>
<td>0.1686</td>
<td>i love</td>
<td>−0.0719</td>
</tr>
<tr>
<td>morning</td>
<td>0.1647</td>
<td>pregnant</td>
<td>−0.0639</td>
</tr>
</tbody>
</table>
My feet hurt soooooo baddddd I need to be carried home waaaaahhhhhhhhhhhhhhhhhhhhh never wearing heels again.
Twitter Health

- Aggregate accuracy comparable with:
  - Google Flu Trends ($R = 0.73$)
  - CDC statistics

+ we can model fine-grained interactions between specific individuals
Impact of Co-Location

Number of estimated encounters with sick individuals at time $t$

Conditional probability of getting sick at $t+1$

1 hour time window ($T=1h$)
4 hour time window ($T=4h$)
12 hour time window ($T=12h$)

Prior probability of being sick

$f(x) = 0.002e^{(0.054x)}$

$f(x) = 0.013e^{(0.055x)}$

$f(x) = 0.001e^{(0.055x)}$
Impact of Friendships

Conditional probability of getting sick

- Prob. of getting sick at $t+1$ given $n$ friends are sick at $t$
- Prob. of getting sick given having $n$ friends (any)
- Prob. of getting sick at $t+1$ given $n$ unencountred friends are sick at $t$
- Prior probability of being sick

$$f(x) = 0.003 e^{(0.413x)}$$

(Sadilek et al AAAI 2012)
Health Prediction

\[ h_{t-1} \quad h_t \quad h_{t+1} \quad \ldots \]

```
O_t +1
```

```
O_t
```

```
O_t -1
```

---

```
1
```

```
0.8
```

```
0.6
```

```
0.4
```

```
0.2
```

```
0
```

---

```
Precision Viterbi
```

```
Precision smoothing
```

```
Recall Viterbi
```

```
Recall smoothing
```

```
Recall smoothing
```

```
Precision baseline
```

```
Recall baseline
```

---

```
Number of days into the future (x)
```

```
0
```

```
2
```

```
4
```

```
6
```

```
8
```

---

```
0
```

```
2
```

```
4
```

```
6
```

```
8
```
Social Network Centrality Correlates with Health

- PageRank ($R=0.26$)
- Degree Centrality ($R=0.28$)
- Betweenness Centrality ($R=0.28$)
- Flow Centrality ($R=0.22$)
- Load Centrality ($R=0.28$)
- Reciprocity ($R=0.26$)
CON EDISON - HUDSON AVE STATION:
1 HUDSON AVE BROOKLYN NY 11201
NAICS: 22133
Steam and Air-Conditioning Supply

Annual Air Emissions

<table>
<thead>
<tr>
<th></th>
<th>2008</th>
<th>Emissions Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbon Monoxide</td>
<td></td>
<td>47</td>
</tr>
<tr>
<td>Lead</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Nitrogen Oxides</td>
<td></td>
<td>318</td>
</tr>
<tr>
<td>Particulate Matter 10</td>
<td></td>
<td>27</td>
</tr>
<tr>
<td>Particulate Matter 2.5</td>
<td></td>
<td>27</td>
</tr>
<tr>
<td>Sulfur Dioxide</td>
<td></td>
<td>416</td>
</tr>
<tr>
<td>Volatile Organic Compounds</td>
<td>2008</td>
<td>7</td>
</tr>
</tbody>
</table>
Factors Influencing Health

(Sadilek & Kautz WSDM 2013)
Disease Hubs & Vectors

(Brennan et al IJCAI 2013)
Graphical Summary
The Data

<table>
<thead>
<tr>
<th>Dataset Statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of days</td>
<td>95</td>
</tr>
<tr>
<td>Number of airports tracked</td>
<td>100</td>
</tr>
<tr>
<td>Number of metropolitan areas</td>
<td>75</td>
</tr>
<tr>
<td>Users total</td>
<td>632,611</td>
</tr>
<tr>
<td>Target users</td>
<td>51,137</td>
</tr>
<tr>
<td>Tweets by target users</td>
<td>6,287,446</td>
</tr>
<tr>
<td>User flights inferred through tweets</td>
<td>73,460</td>
</tr>
<tr>
<td>Number of meetings (target users only)</td>
<td>445,812</td>
</tr>
</tbody>
</table>

*target users: tweeted from more than one airport*
Validating the Twitter Health Signal Feature $T_f$

- Computed weekly:
  - Center for Disease Control flu measure $C_f$
  - Google Flu measure $G_f$
- Computed daily: Twitter Health measure $T_f$

$$\sum_{u \in \text{UsersTweetingAt}(d, r)} \Pr(u \text{ indicates sickness on day } d)$$
Validating $T_f$

- NYC, Boston, Los Angeles, Seattle, San Francisco
- $T_f$ correlated with $C_f$ ($R=0.80$, $p=0.002$)
- $T_f$ correlated with $G_f$ ($R=0.87$, $p=0.0002$)
Volume and Sick Traveller Features

- $f(t, x \rightarrow y) = \#$ Twitter users who flew from airport $x$ to airport $y$
  - User tweeted from $x$ on day $t$
  - User tweeted from $y$ earlier on day $t$ or on day $t-1$
- $V(t,x) = \#$ Twitter users who flew into $x$ on day $t$
- $f^s(t, x \rightarrow y) = \#$ sick Twitter users who flew from airport $x$ to airport $y$
  - User made “sick” tweet on day $t$ or $t-1$
- $S(t,x) = \#$ sick Twitter users who flew into $x$ on day $t$
Meeting Feature

• Two users assumed to meet if they appear within 100 meters of each other within one hour

• $M(t,x) = \#$ meetings that users traveling to airport $x$ on day $t$ had with sick users on days $t$ or $t-1$

• Captures number of exposed individuals traveling to $x$
Measuring Explanatory Power of Features

- Goal: explain weekly change in Google Flu measure, $\Delta G_f$, in each city $x$
- Linear regression over features from prior 7 days

<table>
<thead>
<tr>
<th>features</th>
<th>explains % of $\Delta G_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V(t, x)$</td>
<td>56%</td>
</tr>
<tr>
<td>$V(t, x), S(t,x)$</td>
<td>73%</td>
</tr>
<tr>
<td>$V(t, x), S(t,x), M(t,x)$</td>
<td>78%</td>
</tr>
</tbody>
</table>
Prediction

- Goal: predict $T_f$ for city $x$ on a given day using $V(x,t)$, $S(x,t)$, $M(x,t)$ for 3 previous days
- Single linear regression model for all cities
- Our prediction of a city's flu index next week is within 7% of the true value 95% of the time
There’s a Fly in My Tweets

By HENRY KAUTZ
Published: June 21, 2013

ROCHESTER — MANY important public health questions are difficult and costly to answer. What kind of sources of pollution, like dry cleaners that use perchloroethylene, affect the health of nearby residents? Are people who live near these sources more or less healthy, or do those friendships increase the risk of disease? Do frequent visits to public spaces like parks or restaurants affect a person’s health?

Researchers are using health information from social media to build new surveillance systems using health information from social media to build new surveillance systems. The millions of people who use Facebook are providing a real-time “signal” of the “biological and cultural” factors that affect our health.

For example, Rochester has a project using cellphone use...
Learning a Food Poisoning Language Model

How to learn accurate model for very rare messages?
Leverage Human-Machine Computation
Iterative Improvement of Language Model
<table>
<thead>
<tr>
<th>Positive Features</th>
<th>Weight</th>
<th>Negative Features</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>stomach</td>
<td>1.7633</td>
<td>think i’m sick</td>
<td>-0.8411</td>
</tr>
<tr>
<td>stomachache</td>
<td>1.2447</td>
<td>i feel sooo</td>
<td>-0.7156</td>
</tr>
<tr>
<td>nausea</td>
<td>1.0935</td>
<td>fuck i’m</td>
<td>-0.6393</td>
</tr>
<tr>
<td>tummy</td>
<td>1.0718</td>
<td>@MENTION sick to</td>
<td>-0.6212</td>
</tr>
<tr>
<td>#upsetstomach</td>
<td>0.9423</td>
<td>sick of being</td>
<td>-0.6022</td>
</tr>
<tr>
<td>nauseated</td>
<td>0.8702</td>
<td>ughhh cramps</td>
<td>-0.5909</td>
</tr>
<tr>
<td>upset</td>
<td>0.8213</td>
<td>cramp</td>
<td>-0.5867</td>
</tr>
<tr>
<td>nautious</td>
<td>0.7024</td>
<td>so sick omg</td>
<td>-0.5749</td>
</tr>
<tr>
<td>ache</td>
<td>0.7006</td>
<td>tired of</td>
<td>-0.5410</td>
</tr>
<tr>
<td>being sick man</td>
<td>0.6859</td>
<td>cold</td>
<td>-0.5122</td>
</tr>
<tr>
<td>diarrhea</td>
<td>0.6789</td>
<td>burn sucks</td>
<td>-0.5085</td>
</tr>
<tr>
<td>vomit</td>
<td>0.6719</td>
<td>course i’m sick</td>
<td>-0.5014</td>
</tr>
<tr>
<td>@MENTION i’m getting</td>
<td>0.6424</td>
<td>if i’m</td>
<td>-0.4988</td>
</tr>
<tr>
<td>#tummyache</td>
<td>0.6422</td>
<td>is sick</td>
<td>-0.4934</td>
</tr>
<tr>
<td>#stomachache</td>
<td>0.6408</td>
<td>so sick and</td>
<td>-0.4904</td>
</tr>
<tr>
<td>i’ve never been</td>
<td>0.6353</td>
<td>omg i am</td>
<td>-0.4862</td>
</tr>
<tr>
<td>threw up</td>
<td>0.6291</td>
<td>@LINK</td>
<td>-0.4744</td>
</tr>
<tr>
<td>i’m sick great</td>
<td>0.6204</td>
<td>@MENTION sick</td>
<td>-0.4704</td>
</tr>
<tr>
<td>poisoning</td>
<td>0.5879</td>
<td>if</td>
<td>-0.4695</td>
</tr>
<tr>
<td>feel better tomorrow</td>
<td>0.5643</td>
<td>i feel better</td>
<td>-0.4670</td>
</tr>
</tbody>
</table>
Figure 5: Probability distributions over violation scores (higher is worse) for restaurants, where we have not observed evidence of illness ($Pr(s \mid e = 0)$; blue), and restaurants in which we observed at least one individual who subsequently became ill ($Pr(s \mid e = 1)$; orange). Nonparametric Kolmogorov-Smirnov test shows that the two distributions are significantly different (p-value of $1.5 \times 10^{-11}$).
<table>
<thead>
<tr>
<th>Feature</th>
<th>Regression Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant term $c$</td>
<td>+16.1585 ***</td>
</tr>
<tr>
<td>Number of visits</td>
<td>-0.0015 ***</td>
</tr>
<tr>
<td>Number of distinct visitors</td>
<td>-0.0014 ***</td>
</tr>
<tr>
<td>Number of sick visitors ($f^T$)</td>
<td>+3.1591 ***</td>
</tr>
<tr>
<td>Proportion of sick visitors ($f$)</td>
<td>+19.3370 ***</td>
</tr>
<tr>
<td>Number of sick days of visitors</td>
<td>0 ***</td>
</tr>
</tbody>
</table>

Table 3: Regression coefficients for predicting $s$, the DOHMH violation score, from Twitter data. *** denotes statistical significance with p-value less than 0.001.
Figure 4: We obtain increasingly stronger signal as we concentrate on restaurants with larger amounts of associated Twitter data. Pearson correlation coefficient increases linearly as we consider venues with at least $n$ visits recorded in the data (horizontal axis). At the same time, the correlation is increasingly significant in terms of p-value as we observe more data. Note that even sparsely represented restaurants (e.g., with one recorded visit) exhibit weak, but significant correlation.
Summary

• Social media can be mined to estimate the impact of physical encounters, social ties, and behavioral and environmental factors on public health

• Cheap and fast (“data exhaust”)

• Unique new tool for identifying movement of disease and individual disease carrying individuals on a global scale

• Could support novel interventions