RESEARCH QUESTION

- Is there a public health signal that can be detected within the chatter of Twitter?
- If so, what can we do with that signal?
MINING HEALTH TRENDS

- Google search query: flu medicine
  - It’s likely this person has the flu

- aggregate and correlate queries to predict flu activity

- Similar results can be replicated with Twitter messages in place of search queries
  - Tweets also contain more info than search queries


More Twitter papers in 2011:

Achrekar, Harshavardhan; Gandhe, Avinash; Lazarus, Ross; Ssu-Hsin Yu; Liu, Benyuan. Predicting Flu Trends using Twitter. 2011.

Prier, Smith, Giraud-Carrier, Hanson. Identifying health related topics on twitter: an exploration of tobacco related tweets as a test topic. 2011.

A GENERAL APPROACH

- Most previous studies were very focused
  - One disease of interest
  - Supervised approaches with training data
- Our assumption: don’t know a priori what to look for
  - General approach to look for many diseases
- Use unsupervised or semi-supervised models
PART 1: MODELING HEALTH TWEETS

- Not all Tweets actually talk about health
  - I FEEL LIKE I'M GOING TO DIE OF BIEBER FEVER. NO JOKE.
  - Web design class gives me a huge headache everytime.

- Step 1: find health related tweets
  - Method: supervised machine learning

- Step 2: group tweets by disease / ailment
  - Method: unsupervised topic models
Trained SVM classifier on 5,128 hand labeled tweets
- Cross-validation precision: 90%

Corpus: 2B tweets from May 2009 to October 2010

Keyword filter

Classifier

2 billion

11.7 million

1.63 million Health Tweets
CATEGORIZING TWEETS

- Now we have a set of tweets we know are about health
- Can we group them by ailment?
  - Solution: structured topic models
  - Use symptom/treatment structure to separate illness text from other text
UNSUPERVISED TOPIC MODELS

- Topic Models: a popular tool for modeling corpora
  - Bayesian probabilistic model for generating text

- Basic idea:
  - Each document is a distribution over topics
  - Each topic is a distribution over words
  - Infer these distributions automatically through posterior inference methods -> unsupervised

A MODEL FOR HEALTH IN TWITTER

- Each tweet is about an “ailment” (medical condition)
- Each word in a tweet comes from one of two sources:
  - General topics or background noise (not about health)
  - Ailment words: broken down into three facets (“aspects”)
    - General words, symptoms, treatments
    - Symptoms and treatments identified based on scraped list

Flu: runny nose, headache, advil!!!!! home sick watching TV
ATAM

- **Ailment Topic Aspect Model (ATAM)**
- For each tweet (to $D$):
  - Select an ailment $a$ from a distribution $\eta$
  - Select a topic distribution from $\theta$
  - Select a switching distribution $\pi$
- For each word (to $N$):
  - Select switching variable $x$ from $\pi$
  - If $x == \text{topic}$:
    - Generate topic $z$ from $\theta$ and then word $w$ from $\phi_z$
  - If $x == \text{ailment}$:
    - Observe $y$ and generate word $w$ from $\phi_{a,y}$
LABELING AILMENTS

- Inference: Gibbs sampling (see paper)
- Two annotators labeled model output (based on top 20 ailment words) with ailment name or as “incoherent”
  - Each ailment has top 20 general words, symptom words, treatment words
- Agreed on labels for 15/20 ailments
- Focused on these 15 in further
# AILMENTS: EXAMPLE OUTPUT

<table>
<thead>
<tr>
<th>Ailment</th>
<th>Allergies</th>
<th>Aches/Pains</th>
<th>Dental</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Words</td>
<td>allergies stop</td>
<td>body head</td>
<td>meds killers</td>
</tr>
<tr>
<td></td>
<td>eyes allergic</td>
<td>need hurts</td>
<td>dentist teeth</td>
</tr>
<tr>
<td>Symptoms</td>
<td>sneezing cold</td>
<td>pain aches</td>
<td>pain toothache</td>
</tr>
<tr>
<td></td>
<td>coughing</td>
<td>stomach</td>
<td>sore</td>
</tr>
<tr>
<td>Treatments</td>
<td>medicine benadryl</td>
<td>massage “hot bath”</td>
<td>braces surgery</td>
</tr>
<tr>
<td></td>
<td>claritin</td>
<td>ibuprofen</td>
<td>antibiotics</td>
</tr>
</tbody>
</table>
PART 2: ANALYZING AILMENTS

- We now have groups of tweets categorized by ailment
- We can analyze each ailment
  - Trends over time
  - Trends across geography
- Deeper symptom and treatment analysis
FLU TRENDS REDUX

- Correlation coefficient: 0.958
Previous work focused on influenza surveillance

We have a richer model!

What other public health information can we learn from Twitter?
GEOGRAPHIC SURVEILLANCE

- Track ailments by time and location
  - Compute ailment per capita in each state for each month
  - Determine state for 200,000 tweets with simple keyword filtering
- Seasonal allergies
  - Allergy season starts in different months in different regions
ALLERGIES

February
ALLERGIES
ALLERGIES

August
SELF-REPORTED MEDICATION USAGE

- We have questions about how populations are medicating
- Since many patients self-medicate, how to track?

Whhhhhhat?!?!?! I don't always **sleep**! But I did have a drug-induced slumber last night. I told you Benadryl is my friend

Didn't take a benadryl last night so therefore my allergies f****** up my sleep. I was coughing and blowing my nose all night :-(

- What ailments are most associated with treatments?
# PAIN RELIEF MEDS

<table>
<thead>
<tr>
<th>Medicine</th>
<th>Entropy</th>
<th>Most Common Ailments</th>
</tr>
</thead>
<tbody>
<tr>
<td>tylenol</td>
<td>1.57</td>
<td>Headache (39%), Insomnia (30%), Cold (9%)</td>
</tr>
<tr>
<td>ibuprofen</td>
<td>1.54</td>
<td>Headache (37%), Dental (21%), Aches (17%)</td>
</tr>
<tr>
<td>advil</td>
<td>1.08</td>
<td>Headache (61%), Cold (6%), Dental (5%)</td>
</tr>
<tr>
<td>aspirin</td>
<td>1.04</td>
<td>Headache (69%), Insomnia (10%), Aches (10%)</td>
</tr>
<tr>
<td>vicodin</td>
<td>1.33</td>
<td>Dental (61%), Injuries (11%), Headache (10%)</td>
</tr>
<tr>
<td>codeine</td>
<td>1.94</td>
<td>Cold (25%), Dental (19%), Headache (17%)</td>
</tr>
<tr>
<td>morphine</td>
<td>1.17</td>
<td>Dental (59%), Infection (22%), Aches (9%)</td>
</tr>
</tbody>
</table>
## ALLERGY MEDS

<table>
<thead>
<tr>
<th>Medicine</th>
<th>Entropy</th>
<th>Most Common Ailments</th>
</tr>
</thead>
<tbody>
<tr>
<td>benadryl</td>
<td>1.24</td>
<td>Allergies (64%), Skin (13%), Insomnia (12%)</td>
</tr>
<tr>
<td>claritin</td>
<td>0.54</td>
<td>Allergies (88%), Headache (5%)</td>
</tr>
<tr>
<td>zyrtec</td>
<td>0.49</td>
<td>Allergies (90%)</td>
</tr>
<tr>
<td>sudafed</td>
<td>1.61</td>
<td>Allergies (39%), Cold (21%), Headache (20%)</td>
</tr>
</tbody>
</table>
OTHER ANALYSES

- More experiments in the paper
- We look at symptoms in addition to treatments
- We find correlations between ailments and other known factors
  - Spoiler alert: cancer is correlated with tobacco rates
LOOKING FORWARD

- User population
  - Most users are in the US
  - Young population
    - Many under 35
    - Less than 2% are older than 65
- Privacy
  - Limits to what people will share
    - Ex. STD
@maikel3000
Maikel O’Hanlon

U can find healthiness in Twitter. gigaom.com/2011/07/07/can... Just not while having ur nose in ur mobile phone while driving or crossing street

11 Jul via Tweet Button

@ej_butler
Ed Butler

Some resourceful researchers are figuring out ways to mine Twitter data to find health trends http://n.pr/rlfRDy /via @NPRHealth #hcmsanz

@afsh_ahmed
Afshan

Twitter: essential research tool or means 2 make researchers jus plain lazy? Researchers take US temperature via Twitter bbc.co.uk/news/technolog...

@SSinSF
Scott Shadiow

You never cease to amaze me @Twitter... RT @NPRHealth Twitter Provides A Trove Of Health Trends n.pr/oI7WXr

@ashdonaldson
Ash Donaldson

Perfect storm for misinfo: HuffPo article about comp scientists studying health, reported by Fox & Daily Mail huff.to/pS1ylB

8 Jul via Echofon

@heekeuri
Heekmah(dictatedby希)

"Mark Dredze and Michael J. Paul fed 2 billion public tweets posted between May 2009 and October 2010 into computers" WE ARE BEING WATCHED