Large Scale Analysis of Health Communications on the Social Web

Michael J. Paul
Johns Hopkins University

August 24, 2012
Who am I?

• 3rd-year PhD student
• Background: **computer science**
  • (not an expert in health, medicine, or sociology 😊)
  • machine learning, natural language processing
• Affiliations:
  • Center for Language and Speech Processing
  • Human Language Technology Center of Excellence
• Collaborators (re: health + social media):
  • **Mark Dredze** (my co-advisor)
  • Hieu Tran, Alex Lamb, more students on the way...

August 24, 2012
This Talk

- Methodology: simple **statistical** tools
  - Topic Models

- Health analytics through social media
  - Disease surveillance on Twitter
  - Analysis of specialized discussion forums

- Communication on the Web
  - Simple, scalable models of discourse
Probabilistic Topic Models

- “Unsupervised” machine learning
  - Doesn’t require human input

- Idea: associate words and documents with different categories called **topics**
  - But we don’t know what the topics are!
  - Learn automatically through pattern recognition
    - called **clustering**

- Linguistically bare, but useful for shallow analysis
  - Helpful way to explore massive text collections
  - Scalable and robust
  - Really really popular – LDA (2003) has 4,600 citations
Probabilistic Topic Models

- Statistical model of text generation
- Two types of parameters:
  - $p(\text{topic} \mid \text{document})$ for each document
  - $p(\text{word} \mid \text{topic})$ for each topic
- Optimize parameters to fit model to data (i.e. a collection of documents)
  - maximum likelihood estimation, Monte Carlo, etc.
Probabilistic Topic Models

Probabilistic Topic Models

- Why does this work?
- Intuition: words which co-occur in non-random ways are likely to be generated from the same topic

**Distributional Hypothesis** (Harris 1954)
- words in similar contexts have similar meaning
- part of distributional semantics
This Talk

- Method: simple statistical tools
  - Topic Models

- Health analytics through social media
  - Disease surveillance on Twitter
  - Analysis of specialized discussion forums

- Communication on the Web
  - Simple, scalable models of discourse
Twitter for Public Health

- Huge amount of data
  - 200+ million tweets every day
  - (but only small sample is free)

- Decent proxy for what’s going on in the world
  - including population health

- Tweets are usually publicly available through API
  - vs Facebook (more private accounts)
  - Usually qualifies for IRB waiver
Twitter Data Collection

- **1.6 million** health tweets from 2009-2010
  - distilled from a stream of **2 billion** total tweets
  - **machine learning** to distinguish pertinent tweets ("stuck in bed sick") vs false positives ("so sick of school!")

- Today (since Fall 2011):
  - downloading **~2 million** tweets a day using 400 targeted health search queries
    - estimated 10% are pertinent

- In progress:
  - Public web interface to analyze data
  - Languages beyond English (next up: **Spanish**)

August 24, 2012
This is a lot of text!

We don’t (or at least, we didn’t) know what health issues people actually tweet about.

How to make sense of this much data?
  - Topic models
Ailment Topic Aspect Model

- Topic model with special structure designed to find diseases and ailments
- Breaks down words by aspect: general words, symptom words, or treatment words
- We used the model output to analyze disease over time and location

# Ailment Topic Aspect Model

<table>
<thead>
<tr>
<th>General</th>
<th>Allergies</th>
<th>Insomnia</th>
<th>Obesity</th>
<th>Injuries</th>
<th>Respiratory</th>
<th>Dental</th>
<th>Aches/Pains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allergies</td>
<td>allergies</td>
<td>sleep</td>
<td>blood</td>
<td>knee</td>
<td>throat</td>
<td>ow</td>
<td>body</td>
</tr>
<tr>
<td></td>
<td>nose</td>
<td>asleep</td>
<td>weight</td>
<td>leg</td>
<td>stop</td>
<td>teeth</td>
<td>need</td>
</tr>
<tr>
<td></td>
<td>eyes</td>
<td>fell</td>
<td>eat</td>
<td>right</td>
<td>better</td>
<td>tooth</td>
<td>neck</td>
</tr>
<tr>
<td></td>
<td>allergy</td>
<td>awake</td>
<td>healthy</td>
<td>ankle</td>
<td>voice</td>
<td>wisdom</td>
<td>hurts</td>
</tr>
<tr>
<td></td>
<td>allergic</td>
<td>hours</td>
<td>fat</td>
<td>shoulder</td>
<td>hurts</td>
<td>dentist</td>
<td>head</td>
</tr>
<tr>
<td>Symptoms</td>
<td>sneezing</td>
<td>insomnia</td>
<td>pressure</td>
<td>cough</td>
<td>pain</td>
<td>pain</td>
<td>aches</td>
</tr>
<tr>
<td></td>
<td>coughing</td>
<td>fall</td>
<td>weight</td>
<td>coughing</td>
<td>sore</td>
<td>toothache</td>
<td>pain</td>
</tr>
<tr>
<td></td>
<td>cold</td>
<td>burning</td>
<td>loss</td>
<td>cold</td>
<td>arthritis</td>
<td>sore</td>
<td>sore</td>
</tr>
<tr>
<td></td>
<td>nose</td>
<td>pain</td>
<td>blood</td>
<td>sneezing</td>
<td>limping</td>
<td>infection</td>
<td>muscle</td>
</tr>
<tr>
<td></td>
<td>runny</td>
<td>falling</td>
<td>high</td>
<td>sneeze</td>
<td>neck</td>
<td>tooth</td>
<td>aching</td>
</tr>
</tbody>
</table>

| Treatments       | medicine            | sleeping         | diet          | surgery     | medicine    | braces   | massage    |
|                  | benadryl            | pills            | exercise      | brace       | antibiotics | pain     | exercise    |
|                  | claritin            | caffeine         | dieting       | crutches    | codeine     | relief   | massages   |
|                  | zyrtec              | tylenol          | insulin       | physical    | vitamin     | muscle   | bath        |
|                  | drops               | pill             | exercising    | therapy     | tylenol     | surgery  | hot         |
Twitter: Influenza Rates

![Graph showing influenza rates comparison between Twitter and CDC data over weeks from August 2009 to September 2010. The graph includes a line for Twitter data and a dashed line for CDC data, with X-axis representing weeks and Y-axis representing the percentage of specimens positive for influenza.]
Twitter: Allergy Prevalence

(a) February

(b) April

(c) June

(d) August
Beyond Twitter

- Twitter is good for general population surveys
- More specialized applications would benefit from focused data sources, e.g.
  - medical forums and question answering services
  - online support groups
  - etc.
- We are currently investigating forums for drug users
Drug Abuse on the Web

- Data: Drugs-Forum.com
  - Discussion forums about recreational drug use
  - Our working data set: 100K messages

- Why?
  - Offers candid survey of drug use
  - User base much larger than could be reached through surveys

- End goal: automatically extract and summarize information about various drugs, especially novel and emerging drugs

August 24, 2012
Drug Abuse on the Web

- How to explore giant discussion forum?
  - Probabilistic topic models!

- Our previous topic model included structure about **symptoms** and **treatments** of ailments

- Here, we use a factored topic model to include **delivery** method (e.g. smoking, oral) and other **aspects** (e.g. effects, usage, chemistry).

Drug Abuse on the Web

- Can model different delivery methods of drugs:

  - **cocaine, smoking**
    - crack
    - smoke
    - smoking
    - pipe
    - hit
    - rock
    - powder
    - smoked
    - rocks
    - ash
    - cigarette

  - **cocaine, snorting**
    - nose
    - water
    - nasal
    - spray
    - mouth
    - nostril
    - snorting
    - sinuses
    - snort
    - coke
    - blow
Drug Abuse on the Web

- Can model different aspects of drugs:

  ecstasy, effects
  - experience
  - time
  - people
  - experiences
  - feeling
  - drug
  - feel
  - magic
  - first
  - feelings
  - euphoria
  - mind

  ecstasy, usage
  - pills
  - mdma
  - pill
  - test
  - ecstasy
  - pure
  - quality
  - kit
  - contain
  - testing
  - people
  - know

  ecstasy, culture
  - music
  - rolling
  - rave
  - people
  - great
  - mp3
  - best
  - dance
  - friends
  - club
  - love
  - fun
Drugs Abuse on the Web

- Can drill down to **triples** of (drug, delivery, aspect):

<table>
<thead>
<tr>
<th>cocaine, health</th>
<th>cocaine, snorting, health</th>
</tr>
</thead>
<tbody>
<tr>
<td>cocaine</td>
<td>nose</td>
</tr>
<tr>
<td>addiction</td>
<td>pain</td>
</tr>
<tr>
<td>drug</td>
<td>damage</td>
</tr>
<tr>
<td>dopamine</td>
<td>blood</td>
</tr>
<tr>
<td>people</td>
<td>cocaine</td>
</tr>
<tr>
<td>brain</td>
<td>problem</td>
</tr>
<tr>
<td>drugs</td>
<td>cause</td>
</tr>
<tr>
<td>substance</td>
<td>cola</td>
</tr>
<tr>
<td>addictive</td>
<td>using</td>
</tr>
<tr>
<td>using</td>
<td>doctor</td>
</tr>
<tr>
<td>body</td>
<td>problems</td>
</tr>
<tr>
<td>physical</td>
<td>sinus</td>
</tr>
</tbody>
</table>
This Talk

- Method: simple statistical tools
  - Topic Models

- Health analytics through social media
  - Disease surveillance on Twitter
  - Analysis of specialized discussion forums

- Communication on the Web
  - Simple, scalable models of discourse
Mixed membership Markov models
- Like topic models but aware of thread structure

“Topics” in document depend on previous document
- i.e. speech acts in message depend on speech acts of parent message in thread
- Your response to me depends on what I wrote

Topic Models for Dialog

likely to begin threads

likely to follow questions

not likely to follow itself

“question”

“I
my
have
computer
tried
help"

“answer”

“you
your
http
com
windows”
In the future, I plan to add additional features to the model to potentially consider...

- participant attributes (e.g. gender)
- participant role (e.g. doctor vs patient)
- relationship between participants
Lastly…

- How to fit conversation modeling with health applications?
Potential Avenues for Future Work

- Study of intra-clinic communications
  - e.g. internal SMS system
  - (if such data can be obtained)

- Spread of (mis)information
  - models of rumor propagation
  - models of influence, ideology

- Modeling doctor-patient interactions
  - more generally, interactions between experts and non-experts on the web

August 24, 2012
Other Twitter Studies

- Learning to distinguish concern and awareness of the flu vs infection with the flu
  - “home with the flu” vs
  - “roommates all have the flu, worried i’m next”

- Application area: behavioral epidemiology
  - use tweets as a data source to validate behavioral models

Other Twitter Studies

- Analyze mentions of patient safety events (e.g. medical mistakes)

<table>
<thead>
<tr>
<th>Error Source</th>
<th>Percentage</th>
<th>Error Type</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surgeon</td>
<td>18.1%</td>
<td>Surgical</td>
<td>7.2%</td>
</tr>
<tr>
<td>Doctor</td>
<td>51.5%</td>
<td>Diagnostic</td>
<td>15.0%</td>
</tr>
<tr>
<td>Nurse</td>
<td>18.7%</td>
<td>Medication</td>
<td>22.2%</td>
</tr>
<tr>
<td>Pharmacist</td>
<td>0.0%</td>
<td>Procedure</td>
<td>36.5%</td>
</tr>
<tr>
<td>Other Medical</td>
<td>4.7%</td>
<td>Infection</td>
<td>0.0%</td>
</tr>
<tr>
<td>Dentist</td>
<td>0.0%</td>
<td>Communication</td>
<td>9.6%</td>
</tr>
<tr>
<td>Unknown</td>
<td>7.0%</td>
<td>Unknown</td>
<td>9.6%</td>
</tr>
</tbody>
</table>
