DATA MINING SOCIAL WEBSITES FOR PUBLIC HEALTH

Michael J. Paul
Department of Computer Science
Johns Hopkins University

Association of Public Data Users
September 16, 2013

Social Web as a Data Source

- Millions of people share on the web what they are **doing** and **thinking** every day

- Can analyze social websites to infer:
  - what is **happening** in a population
  - the **attitudes/thoughts** of a population

- Faster, cheaper than traditional data collection

- Noisier, less formal data than traditional sources
  - Need to convert web content into usable statistics

Social Web as a Data Source

- Millions of people share on the web what they are **doing** and **thinking** every day

- Can analyze social websites to infer:
  - what is **happening** in a population
  - the **attitudes/thoughts** of a population

- This talk:
  - what can we learn about the **health** of a population?

Social Websites and Health

- **Twitter**
  - Millions of messages every hour
  - Large-scale influenza surveillance
- **RateMDs.com**
  - Reviews of doctors by patients
  - Insights into patient perception of provider quality
- **Drugs-Forum.com**
  - Discussion forums about illicit drug activity
  - Insights into drug trends, including novel/emerging drugs

Social Websites and Health

- **Twitter**
  - Millions of messages every hour
  - Large-scale influenza surveillance
- **RateMDs.com**
  - Reviews of doctors by patients
  - Insights into patient perception of provider quality
- **Drugs-Forum.com**
  - Discussion forums about illicit drug activity
  - Insights into drug trends, including novel/emerging drugs

Twitter: Data

- Free streams of data provide 1% random sample of public status messages (tweets)

- Search streams provide tweets that match certain keywords
  - Still capped at 1%, but more targeted

- Geolocation: *Carmen*
  - Identifies where a tweet is from (e.g. New York City, US)
  - `https://github.com/mdredze/carmen`

Twitter: Flu Surveillance

- New system automatically identifies tweets that indicate influenza infection

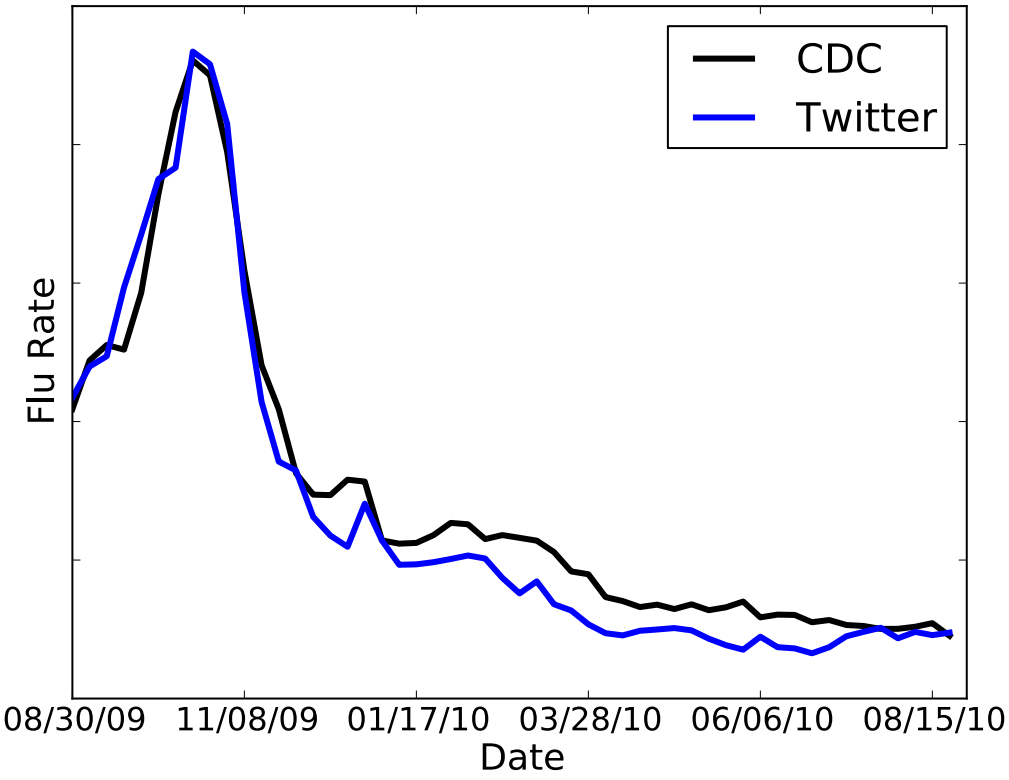
12/30/12	1630
12/31/12	1826
1/1/13	1578
1/2/13	1924
1/3/13	2167
1/4/13	2316
1/5/13	2133
1/6/13	2076
1/7/13	2441
1/8/13	2588
1/9/13	3536
1/10/13	4123
1/11/13	4627
1/12/13	3716
1/13/13	3263
1/14/13	3811

- Goal: estimate influenza prevalence quickly
  - Can complement existing surveillance systems

Twitter: Flu Surveillance (2009-10)

- Inferred from **12 million** tweets with health keywords

- Correlation:
  - **99%**

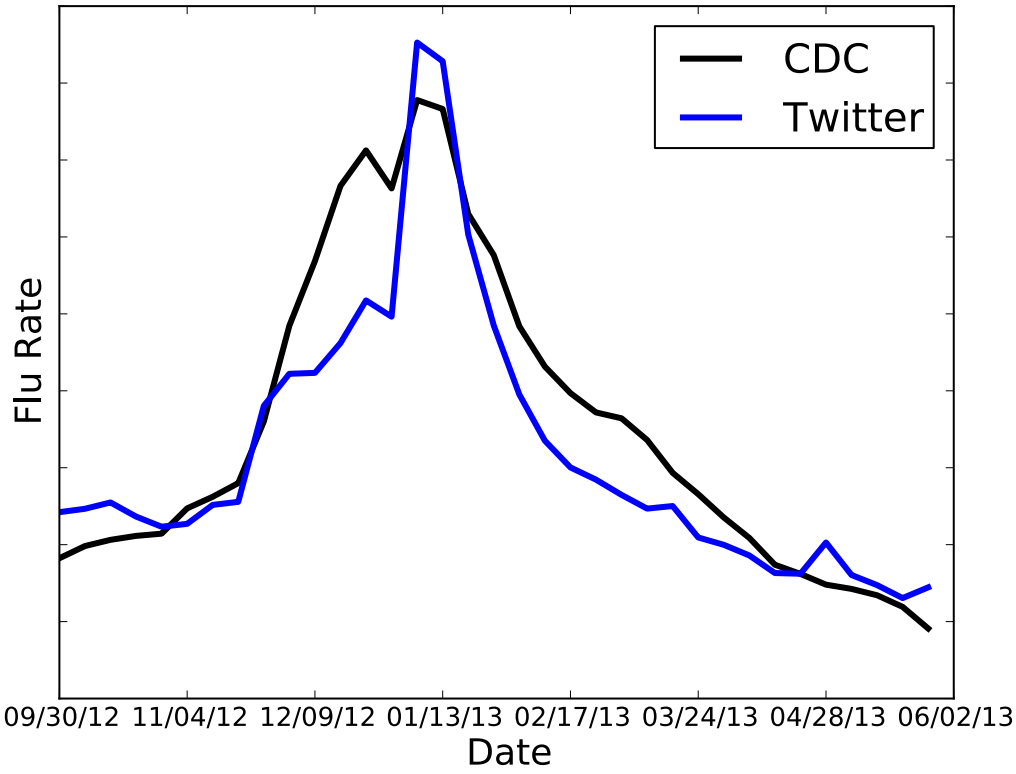


A Lamb, MJ Paul, M Dredze. 2013. Separating fact from fear: Tracking flu infections on Twitter. North American Chapter of the Association for Computational Linguistics.

Twitter: Flu Surveillance (2012-13)

- Inferred from **300 million** tweets with health keywords

- Correlation:
  - **93%**



DA Broniatowski, MJ Paul, M Dredze. Under review. National and local influenza surveillance through Twitter: An analysis of the 2012-2013 influenza epidemic.

Social Websites and Health

- **Twitter**
  - Millions of messages every hour
  - Large-scale influenza surveillance
- **RateMDs.com**
  - Reviews of doctors by patients
  - Insights into patient perception of provider quality
- **Drugs-Forum.com**
  - Discussion forums about illicit drug activity
  - Insights into drug trends, including novel/emerging drugs

Rate MDs: Data

- Online site for doctor reviews
- We analyzed **52,226** reviews from the United States
  - Reviews include 1–5 ratings plus free text

- Our data is available:
  - `http://www.cebm.brown.edu/static/dr-sentiment.zip`

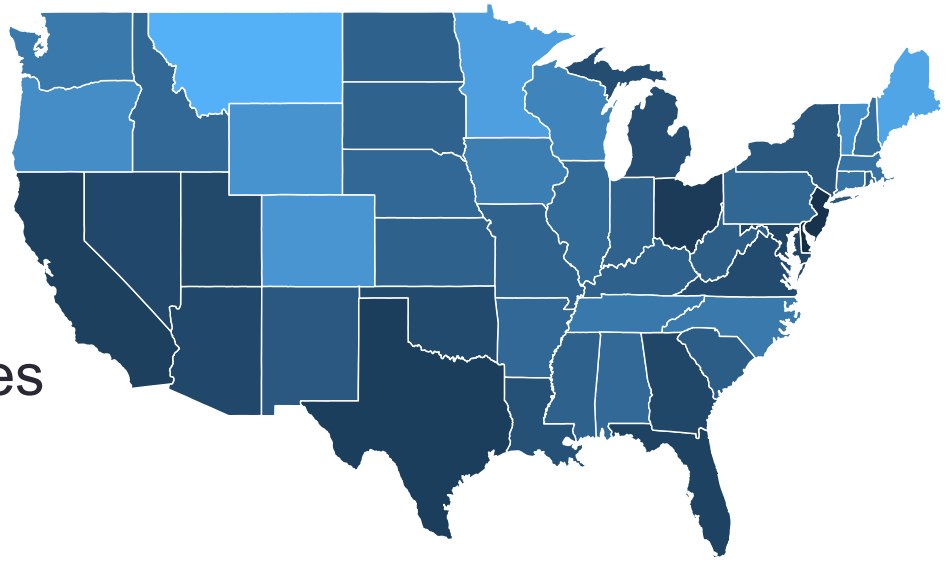
Rate MDs: Analysis

- Automatically grouped words into 3 primary themes
  - Technical competence
  - Interpersonal manner
  - System issues (staff, wait time, etc)
- Automatic sentiment analysis within each theme
- Goal: large-scale understanding of the issues discussed by patients

MJ Paul, B Wallace, M Dredze. 2013. What Affects Patient (Dis)satisfaction? Analyzing Online Doctor Ratings with a Joint Topic-Sentiment Model. AAAI-HIAI Workshop.

Rate MDs: Analysis

- Text is about twice as likely to be **positive** than **negative**
- Text is about twice as likely to describe **interpersonal manner** or **system issues** than **technical competence**
- Can look at geographic variation in prominence of each issue
  - Right: Darker states more likely to discuss system issues

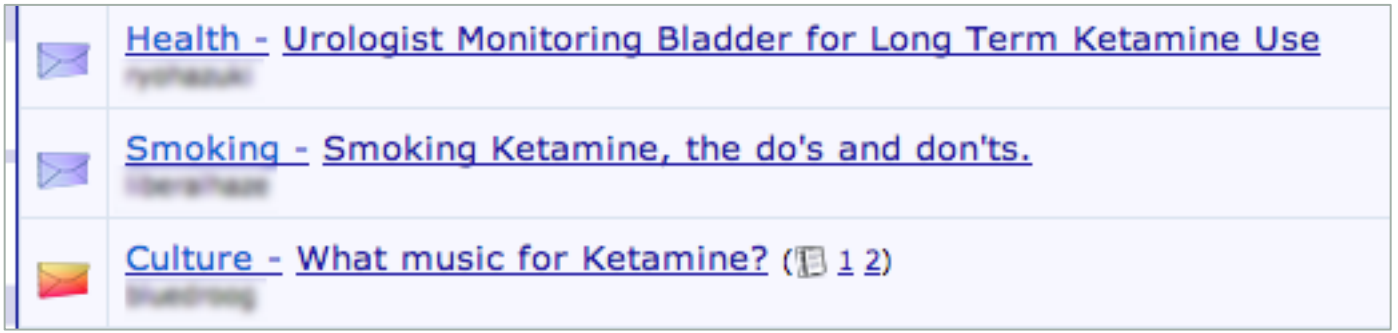


Social Websites and Health

- **Twitter**
  - Millions of messages every hour
  - Large-scale influenza surveillance
- **RateMDs.com**
  - Reviews of doctors by patients
  - Insights into patient perception of provider quality
- **Drugs-Forum.com**
  - Discussion forums about illicit drug activity
  - Insights into drug trends, including novel/emerging drugs

Drugs Forum: Data

- Venue for anonymous users to openly discuss recreational drug usage



- We analyzed **410,000** public messages from **20,000** users

- Some users provide optional demographic information
  - Age, Gender, Country

Drugs Forum: Analysis

- Goal: understand patterns of drug use
  - Which drugs are becoming **more or less popular**?
  - Which **demographic groups** are using which drugs?
  - What are **drug users saying** about various drugs?

- Motivation: record numbers of new drugs created in recent years; hard for researchers to keep up

MJ Paul, M Dredze. 2013. Summarizing Drug Experiences with Multi-Dimensional Topic Models. North American Chapter of the Association for Computational Linguistics.

Drugs Forum: Analysis

- Compared demographics of forum users to demographics in US government survey data (NSDUH)

- Forum users are much more **male** and slightly **younger** than true population of drug users

- But the demographic **variation** across drugs matches the survey data in almost all cases
  - Prescription drug abuse associated with **women** and **older** users
  - Marijuana and hallucinogens associated with **men** and **younger** users

Social Websites and Health

- **Twitter**
  - Millions of messages every hour
  - Large-scale influenza surveillance
- **RateMDs.com**
  - Reviews of doctors by patients
  - Insights into patient perception of provider quality
- **Drugs-Forum.com**
  - Discussion forums about illicit drug activity
  - Insights into drug trends, including novel/emerging drugs

Conclusion

- Social web data can help our understanding of a variety of health questions

- Quality/quantity tradeoff
  - Social websites contain large amount of informal data
  - Quality can be evaluated by comparing to standard data sources

- Demographic biases exist but can be measured (sometimes)
  - We are working to understand and correct for these biases