computational social media

lecture 3: tweeting

part 3

daniel gatica-perez

29.03.2019
announcements

reading #3 will be presented today

S. Vosoughi, D. Roy, S. Aral
The spread of true and false news online, Science, 359, pp. 1146-1151, March 2018

Presenter: A. Pourhabibi
Discussant: J. Wang
Scribe: M. Kramer
this lecture

a human-centric view of twitter

1. introduction
2. twitter users & uses
3. twitter-specific phenomena
4. twitter & large-scale human behavior
5. twitter & real-world events
   twitter & journalism
   twitter & political activism
   inferring real-world events from twitter
twitter & political collective action
“labeling the movements as Twitter revolutions does not accurately capture the medium’s role”

Dhiraj Murthy, Twitter. Social Communication in the Twitter Age, Ch. 6, Polity, 2013
Egypt: Timeline of Communication Shutdown during the Revolution

Population (2011) ~ 80 million

D. Murthy, Twitter, Polity, 2013

- 13.5 million (16.2%) on internet
- 3.5 million (4.21%) on Facebook
- 12,000 (0.014%) on Twitter

About: diagram to illustrate sequence of communications shutdown Egyptians went through from 25 January to 06 February 2011. Times mentioned are according to Egypt local time. Numbers in the diagram are approximate. Diagram gets updated with info when possible.


Reports, Jan 2011: Facebook Users 4,634,600 (7.6% of population). Jan-March 2011: avg. Twitter Users: 131,204 (0.15% of population).

For More Info: ramy.raoof@gmail.com

credit (cc) ramy raoof @ flickr (09.06.2011): https://www.flickr.com/photos/ramyraoof
+ within Egypt, Twitter direct reach was minimal
+ … but government’s perception was one of a threat
+ shutting it down legitimized the service and its role
+ this hugely increased awareness internally (by 03.2011, 131,000 users)
… and created big reaction externally

D. Murthy, Twitter. Social Communication in the Twitter Age, Ch. 6, Polity, 2013
tweet: https://twitter.com/PJCrowley/status/30828460062547968

+ shutting down the internet had a much larger effect, affecting many more groups of society (e.g. business)
“Communication has played a role in spreading revolutionary ideas throughout history” (Motadel, 2011)

“Twitter has potential in organizing activists’ movements… but we should be careful about concluding Twitter causing these movements” (Murthy)

real causes (Warf, 2011; Hasseb, 2011):
repressive state
police brutality
persistent poverty
high unemployment
accumulated political consciousness

D. Murthy, Twitter. Social Communication in the Twitter Age, Ch. 6, Polity, 2013
collective action and pluralistic ignorance

situation: participating in a protest against a regime

+ collective action produces payoffs only if enough people participate
+ choices are visible in threshold model of diffusion of innovations
+ repressive regimes tend to limit communication among citizens

pluralistic ignorance: “erroneous estimates about the prevalence of certain opinions of the population at large”

effect of knowledge on collective action

assume each person has a personal threshold to participate in a protest (at least K people including the person herself)

assume each node only knows the threshold of self and its neighbors

example 1
+w would join if 4 people do; since she only has 2 neighbors with lower thresholds, she does not join
+v knows w’s threshold, so v infers that w will not join. Since v requires 3 people to join, v does not join
+u requires 2 people, but knows the thresholds of w and v, hence infers that neither will join, and does not join
+the protest does not occur

example 2

+ $u$, $v$, and $w$ each know that there are three nodes with threshold 3, but this fact is **common knowledge**: each node knows this fact, and each node knows that each node knows it. As a result, all three nodes join

+ $x$ knows the threshold of $w$, so as $w$ joins, $x$ does too

+ the protest does occur

for collective action, this kind of common knowledge is important

social media platforms can be facilitators of this

---


inferring real-world events from twitter data
+ ML + NLP to infer:

- Box office revenues [1]
- Stock market [2]
- Election outcomes [3]
- Influenza [4]
- Cascades [5]
- etc. etc.

[1] Asur & Huberman, Predicting the future with social media. In Proc. IEEE Int. Conf. on Web Intelligence, 2010
case study: twitter & the flu
Farah Heron @FarahHeron · Apr 17
Some husbands bring their wives flowers. I get NyQuil. And I’m super happy about it. #flu

Becky Fuller @InaraRose42 · Apr 18
Ok, so I’m a week into this #flu and can safely say that at this point I am nothing but Gatorade and NyQuil #dying

Heather Stauffer @HStaufferLNP · 19h
Periodic Pennsylvania #flu update; season has wound down greatly, but is not yet past. (As I’ve said so many times, this season is depicted by the mountainous red line.)

SES LHD Public Health @SESPublicHealth · Apr 22
Parents the FREE #flu vaccine for children aged 6 months to <5 years is now available. Here are top 4 reasons to vaccinate your child.
1. “Tweets matching 100s of health-related keywords passed 3 classification filters to remove irrelevant tweets.

2. Locations are identified with geolocation system and only tweets in the location of interest are saved.

3. The volume of tweets is normalized by the total volume of tweets from a random sample of Twitter to produce a prevalence measure.”

D. Broniatowski, M. Paul & M. Dredze, National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic, PLOS ONE, 2013
data & filters to extract flu-infection tweets

Start: 30.09.2012 (first week of 2012-2013 influenza season defined by US CDC: Centers for Disease Control and Prevention)
End: 31.05.2013

Filter 1 (health-relevant vs. irrelevant):
“combination of keyword filtering and support vector machine (SVM) trained on 5,128 annotated tweets; 90% precision, 32% recall.”

Filter 2 (discussed influenza vs. not):
“logistic regression trained on 11,990 tweets. Features: unigrams, bigrams, trigrams & linguistic information about semantics, syntax, and writing style; 67% precision, 87% recall”

Filter 3 (indicated infection vs. just awareness): “logistic regression trained on same 11,990 labeled tweets, and same features: have 74% precision, 87% recall”

570,000 influenza infection tweets during 8 months
extracting location of flu-infection tweets

**Geolocation system:**
* GPS associated to small percentage of tweets
* Self-reported location from users’ public profile: “New York, NY”, “NYC,” “Candy Land”

Output: country, state, county, city

Result: identified location of 22% of tweets. In evaluation set of 56,000 tweets, for two locations (USA, NYC) accuracy was 92%, 61% (within 50 miles of NYC).

104,200 US influenza infection tweets
extracting normalized influenza prevalence

Normalized Influenza Prevalence:
“Normalized weekly number of infection tweets by total number of tweets in the general stream for same week and location to produce a Twitter-based influenza prevalence measure”

“Compare to the CDC’s US Outpatient Influenza-Like Illness Surveillance Network, which includes number of visits for influenza-like illness (ILI)”
results: correlation for national influenza rates between Twitter and CDC

**National Level** (104,200 influenza infection tweets from USA):
* Weekly # tweets indicating *influenza infection* is strongly correlated with weekly CDC ILI outpatient counts ($r = 0.93; p < 0.001$).
* Weekly # tweets containing *influenza keywords* provided by US Dept. of Health and Human Services is less strongly correlated ($r = 0.75; p < 0.001$).
* 45% reduction in mean absolute error over the keyword filter.

**Municipal Level** (4,800 influenza infection tweets from NYC):
* Weekly # tweets indicating *influenza infection* is strongly correlated with NYC city’s weekly emergency department visits for ILI ($r = 0.88; p < 0.001$).
* Weekly # tweets containing *influenza keywords* is less strongly correlated ($r = 0.72; p < 0.001$);
results (2): national influenza rates (Twitter and CDC)

“Dashed blue line: measure estimated by simple model (keyword matching)

Solid blue line: measure estimated by infection detection model

Black line: CDC data

Twitter estimates neither lead nor lag the CDC ILI rates, yet Twitter data are available up to two weeks earlier than CDC data.”

D. Broniatowski, M. Paul & M. Dredze, National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic, PLOS ONE, 2013
In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (1, 2). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (3, 4), what lessons can we draw from this error?

The problems we identify are not limited to GFT. Research on whether search or social media can predict Covid has become commonplace (5–7) and is often put in sharp contrast with traditional methods and hypotheses. Although these studies have shown the ability and dependencies among data (12), the core challenge is that most big data that have received popular attention are not the run ever since, with a few changes announced in October 2013 (10, 15).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week’s errors predict this week’s errors (temporal autocorrelation), and the direction and magnitude of error varies with the time of year (seasonality). These patterns mean that GFT overlooks considerable information that could be extracted by traditional statistical methods.

Even after GFT was updated in 2009, the comparative value of the algorithm as a stand-alone flu monitor is questionable. A study in 2010 demonstrated that
Twitter Improves Influenza Forecasting

OCTOBER 28, 2014 · RESEARCH ARTICLE

Authors

Michael J. Paul  Mark Dredze  David Broniatowski

Abstract

Accurate disease forecasts are imperative when preparing for influenza epidemic outbreaks; nevertheless, these forecasts are often limited by the time required to collect new, accurate data. In this paper, we show that data from the microblogging community Twitter significantly improves influenza forecasting. Most prior influenza forecast models are tested against historical influenza-like illness (ILI) data from the U.S. Centers for Disease Control and Prevention (CDC). These data are released with a one-week lag and are often initially inaccurate until the CDC revises them weeks later. Since previous studies utilize the final, revised data in evaluation, their evaluations do not properly determine the effectiveness of forecasting. Our experiments using ILI data available at the time of the forecast show that models incorporating data derived from Twitter can reduce forecasting error by 17-30% over a baseline that only uses historical data. For a given level of accuracy, using Twitter data produces forecasts that are two to four weeks ahead of baseline models. Additionally, we find that models using Twitter data are, on average, better predictors of influenza prevalence than are models using data from Google Flu Trends, the leading web data source.
the real-name web, revisited
is twitter interested in real names?

Twitter does not require real names. Pseudonyms are valuable in information networks: “not real names but persistent identity with reputation attached.” It works as identity service for individuals & entities whose long-time presence depends on being identified.

Downside: fake accounts.
what to remember

society shapes twitter; twitter influences society
  case study 1: journalistic practices & the news
  case study 2: political activism

inferring real-world events from twitter
  inferring everything from twitter, really?
  case study: flu patterns
  beware of methodological limitations:
    representativeness & sampling biases
    overlook of cultural issues
    overemphasis on single platform
questions?

gatica@idiap.ch
daniel.gatica-perez@epfl.ch