computational social media

lecture 3: tweeting
part 2

daniel gatica-perez

27.03.2020
announcements

discussion about assignment #2

reading #3 will be presented today:
this lecture

a human-centric view of twitter

1. introduction
2. twitter users & uses
3. large-scale human behavior & real-world events
inferring real-world trends 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
Some husbands bring their wives flowers. I get NyQuil. And I’m super happy about it. #flu

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

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.)

Show this thread

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.
estimating influenza prevalence from Twitter

1. “Tweets matching hundreds 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 Twitter sample 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
1. 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
2. extracting location of flu-infection tweets

Challenges
* GPS existing for only small fraction of tweets
* Self-reported location from users’ public profile: “New York, NY”, “NYC”, “Candy Land”

Location filter output
(country, state, county, city)

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

104,200 US influenza infection tweets
3. Extracting normalized influenza prevalence

**Normalized Influenza Prevalence Measure:**

“Normalized weekly number of infection tweets by total number of tweets in the general stream for same week and location”

Gold standard: “US CDC Outpatient Influenza-Like Illness Surveillance Network: 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 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 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: national influenza weekly rates (Twitter vs. 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
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.
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 x 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 auto-correlation), 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
Thank you for stopping by.

Google Flu Trends and Google Dengue Trends are no longer publishing current estimates of Flu and Dengue fever based on search patterns. The historic estimates produced by Google Flu Trends and Google Dengue Trends are available below. It is still early days for nowcasting and similar tools for understanding the spread of diseases like flu and dengue – we’re excited to see what comes next. Academic research groups interested in working with us should fill out this form.

Sincerely,

The Google Flu and Dengue Trends Team.

**Google Flu Trends Data:**

You can also see this data in Public Data Explorer

- World
- Argentina
- Australia
- Austria
- Belgium
- Bolivia

https://www.google.org/flutrends/about/
At the time of this writing, the novel coronavirus (COVID-19) pandemic outbreak has already put tremendous strain on many countries' citizens, resources and economies around the world. Social distancing measures, travel bans, self-quarantines, and business closures are changing the very fabric of societies worldwide. With people forced out of public spaces, much conversation about these phenomena now occurs online, e.g., on social media platforms like Twitter. In this paper, we describe a multilingual coronavirus (COVID-19) Twitter dataset that we have been continuously collecting since January 22, 2020. We are making our dataset available to the research community (this https URL). It is our hope that our contribution will enable the study of online conversation dynamics in the context of a planetary-scale epidemic outbreak of unprecedented proportions and implications. This dataset could also help track scientific coronavirus misinformation and unverified rumors, or enable the understanding of fear and panic --- and undoubtedly more. Ultimately, this dataset may contribute towards enabling informed solutions and prescribing targeted policy interventions to fight this global crisis.
spreading information in the real world

a. who says what to whom on Twitter
b. cascading behavior in networks
c. structural virality of online diffusion
who says what to whom on Twitter

the goal of media communication research

Harold Lasswell (1948):
“who says what to whom in what channel with what effect”

“difficult to examine information flow in large populations“

"communication channels may have different effects"


photo credit: United Workers (cc) https://www.flickr.com/photos/unitedworkers/14138566864
three models of communication

mass communication:
“one-way message transmission from one source to a large, relatively undifferentiated and anonymous audience”

interpersonal communication:
“two-way message exchange between two or more individuals”

two-step flow of communication:
“mass media influence the public only indirectly”
“the critical intermediate layer are media-savvy individuals – the opinion leaders”

### who is on twitter?

<table>
<thead>
<tr>
<th>Communication Type</th>
<th>User Category Examples</th>
<th>User Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mass-personal</td>
<td>Celebrities, Bloggers</td>
<td>Barack Obama, Taylor Swift</td>
</tr>
<tr>
<td>Personal</td>
<td>Others (the rest of us)</td>
<td>Sheryl Sandberg, Steve Jobs</td>
</tr>
</tbody>
</table>
goals of the study

who?
user categories

who listens to whom?
information flow & consumption

who says what?
information production
quick detour: what is the "full data"?

Q1. all people currently living in a country?
Q2. all Twitter user accounts?

A1: exact number unknown
A2: exact number known only to Twitter

estimates for each case might exist (with varying levels of uncertainty)

more often than not, we work with partial data, with samples
sampling

assume that $X$ is a random variable with distribution $p(X)$

Monte Carlo: sampling $p(X)$ generates a finite number of samples that can be used to approximate functions of $X$ (e.g. expected value)

![Diagram showing distribution and approximation](https://www.slideshare.net/kohta/particle-filter-tracking-in-python)

a random sample of $X$: $(X_1,\ldots,X_N)$ is representative in this sense
sampling in the social sciences

access to full populations is impossible or impractical

X is a vector of individual attributes: sex, age, zip code, etc.

how to obtain representative population samples has been studied in depth in the social sciences

non-probabilistic sampling techniques exist, e.g. convenience sampling, known to be non-representative of the population

bias: systematic error arising from many factors, including but not limited to the lack of representativeness of the sample
Twitter data samples

fully random sampling is impossible unless someone partners with Twitter or pays for data: Twitter API, Enterprise category

convenience sample: Twitter API, Standard category

https://developer.twitter.com/en/pricing
the real-name web, revisited on twitter

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

M. Ingram, Why Twitter doesn’t care what your real name is. GigaOM, Sep 2011
back to the main topic: datasets

1. follower graph [Kwak et al, WWW 2010]
   collected July 2009, 42M users, 1.5B edges
   median number of followers < 100
   few users have millions of followers

2. twitter firehose (full stream)
   223 days (Jul 2009 – Mar 2010)
   5B tweets
   260M tweets with bit.ly URL links

restriction to URLs motivated by
easier to track content
give access to richer content
lists: feature that groups users into sets

useful to organize users into sets

list names are meaningful labels to describe listed users → user categorization
user crawling: snowball sample of lists of popular users

Manual seed users (4 categories)

Check lists seed users belong to & manually select keywords

Crawl all lists seed users appear in

Prune lists to keep those that contain keywords

Crawl all users in pruned lists

Media (news, news-media), Celebrities (stars, celebs), Organizations (company, ngo, brand), Blogs (blog, blogger)
the concept of elite users: top 5000 users (ranked by how frequently they are listed in each category)

<table>
<thead>
<tr>
<th>category</th>
<th>Snowball Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># of users</td>
</tr>
<tr>
<td>celeb</td>
<td>82,770</td>
</tr>
<tr>
<td>media</td>
<td>216,010</td>
</tr>
<tr>
<td>org</td>
<td>97,853</td>
</tr>
<tr>
<td>blog</td>
<td>127,483</td>
</tr>
<tr>
<td>total</td>
<td>524,116</td>
</tr>
</tbody>
</table>

statistics of the snowball sample

top 5 users per category (ranked by #lists in that category)

<table>
<thead>
<tr>
<th>Celebrity</th>
<th>Media</th>
<th>Org</th>
<th>Blog</th>
</tr>
</thead>
<tbody>
<tr>
<td>aplusk</td>
<td>cnnbrk</td>
<td>google</td>
<td>mashable</td>
</tr>
<tr>
<td>ladygaga</td>
<td>nytimes</td>
<td>Starbucks</td>
<td>problogger</td>
</tr>
<tr>
<td>TheEllenShow</td>
<td>asahi</td>
<td>twitter</td>
<td>kibeloco</td>
</tr>
<tr>
<td>taylorswift13</td>
<td>BreakingNews</td>
<td>joinred</td>
<td>naosalvo</td>
</tr>
<tr>
<td>Oprah</td>
<td>TIME</td>
<td>ollehkt</td>
<td>dooce</td>
</tr>
</tbody>
</table>

counts of URLs initiated by each category composed of 5000 elite users

<table>
<thead>
<tr>
<th>category</th>
<th># of URLs</th>
<th># of URLs per-capita</th>
</tr>
</thead>
<tbody>
<tr>
<td>celeb</td>
<td>139,058</td>
<td>27.81</td>
</tr>
<tr>
<td>media</td>
<td>5,119,739</td>
<td>1023.94</td>
</tr>
<tr>
<td>org</td>
<td>523,698</td>
<td>104.74</td>
</tr>
<tr>
<td>blog</td>
<td>1,360,131</td>
<td>272.03</td>
</tr>
<tr>
<td>ordinary</td>
<td>244,228,364</td>
<td>6.10</td>
</tr>
</tbody>
</table>
elite users: how do they relate to ordinary users?

start with 100K ordinary (non-elite) users

tweets produced by the accounts a user follows

measure the proportion of elite accounts that ordinary users receive tweets from for 4 categories of interest

celebrities dominate: users get 25% of their tweets from the top 1000 celebrities

average fraction of #tweets for a random user that are accounted for by top K elite users.
who listens to whom?

"Ordinary users receive their information from thousands of distinct sources, many of which are not the media"

“Audiences have become increasingly fragmented."

“Only ~15% of tweets received by ordinary users are received directly from the media"

"20K elite users attract ~50% of all attention"

add values for k=5000 for 4 categories

credit (cc): institute of network cultures
who listens to whom among the 4 categories?

% of tweets received from

<table>
<thead>
<tr>
<th></th>
<th>Celeb</th>
<th>Media</th>
<th>Org</th>
<th>Blog</th>
</tr>
</thead>
<tbody>
<tr>
<td>Celeb</td>
<td>38.27</td>
<td>6.23</td>
<td>1.55</td>
<td>3.98</td>
</tr>
<tr>
<td>Media</td>
<td>3.91</td>
<td>26.22</td>
<td>1.66</td>
<td>5.69</td>
</tr>
<tr>
<td>Org</td>
<td>4.64</td>
<td>6.41</td>
<td>8.05</td>
<td>8.70</td>
</tr>
<tr>
<td>Blog</td>
<td>4.94</td>
<td>3.89</td>
<td>1.58</td>
<td>22.55</td>
</tr>
</tbody>
</table>
two-step flow of information

media has an indirect influence over the public via an **intermediate** layer of opinion leaders (Katz 1955)

information passes through intermediaries via
(1) retweets
(2) tweets of URLs

for 1M random **ordinary** users, **46%** of **received URLs** generated by **top 5000 media** users were received via **intermediaries**
two-step flow of information (2)

**ordinary users**: two patterns
* users receiving up to 100 media URLs, receive them essentially all through intermediaries
* others receive them most of them via the media

**intermediaries**: pass along content to at least one other user
* 99% are ordinary users, not elite
* exposed to more media than random ordinary users (9165 vs. 1377 URLs)
* more active (543 vs. 34 followers; 180 vs. 7 tweets)
what to remember

inferring real-world trends from twitter
  influenza trend detection as a case study

who talks to whom about what on Twitter
  fragmented audiences: no longer ruled by classical media
  concentrated attention: 20K elite users get half the attention
  homophily: celebrities follow celebrities; media follows media
  two-step info flow: half of media URLs pass via intermediaries

beware of methodological limitations
  sampling biases
  data preprocessing choices
  overemphasis on single platform
questions?

Daniel Gatica-Perez
daniel.gatica-perez@epfl.ch