

computational social media

lecture 3: tweeting

part 2

daniel gatica-perez

announcements

project pitch day

5-minute presentation of your projects

structure: title, problem, goals, approach, evaluation

short Q&A after each presentation

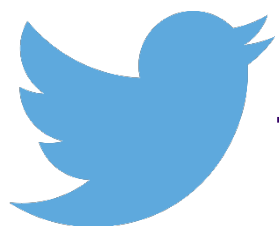
assignment #2 will be announced

this lecture



a human-centric view of twitter

1. introduction
2. twitter users & uses
3. understanding large-scale human behavior
4. inferring real-world events & trends
5. spreading information in the real world



+ ML + NLP to infer:

Box office revenues [1]

Stock market [2]

Election outcomes [3]

Influenza [4]

Cascades [5]

Fake news [6]

Hate speech [7]

Bots [8]

etc. etc.

[1] Asur & Huberman, Predicting the future with social media. In Proc. IEEE Int. Conf. on Web Intelligence, 2010

[2] Bollen, Mao & Zeng, Twitter mood predicts the stock market. Journal of Computational Science, 2011

[3] Gayo-Avello, Metaxas & Mustafarajm Limits of Electoral Predictions Using Twitter. In Proc. AAAI ICWSM, 2011

[4] Broniatowski, Paul & Dredze, National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic, PLOS ONE, 2013

[5] Cheng, Adamic, Dow, Kleinberg & Leskovec. Can cascades be predicted?. In Proc WWW, 2014


[6] Oshikawa, Qian, & Wang, A Survey on Natural Language Processing for Fake News Detection, in Proc. LREC, 2020


[7] Waseem & Hovy, Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter, in Proc. NAACL-HLT 2016

[8] Varol, Ferrara, Davis, Menczer & Flammini Online human-bot interactions: Detection, estimation, and characterization, in Proc. AAAI ICWSM 2017

case study: twitter & the flu


[Redacted]
Some husbands bring their wives flowers. I get NyQuil. And I'm super happy about it. [#flu](#)

 3   9 

[Redacted] 
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](#)




 2   2 

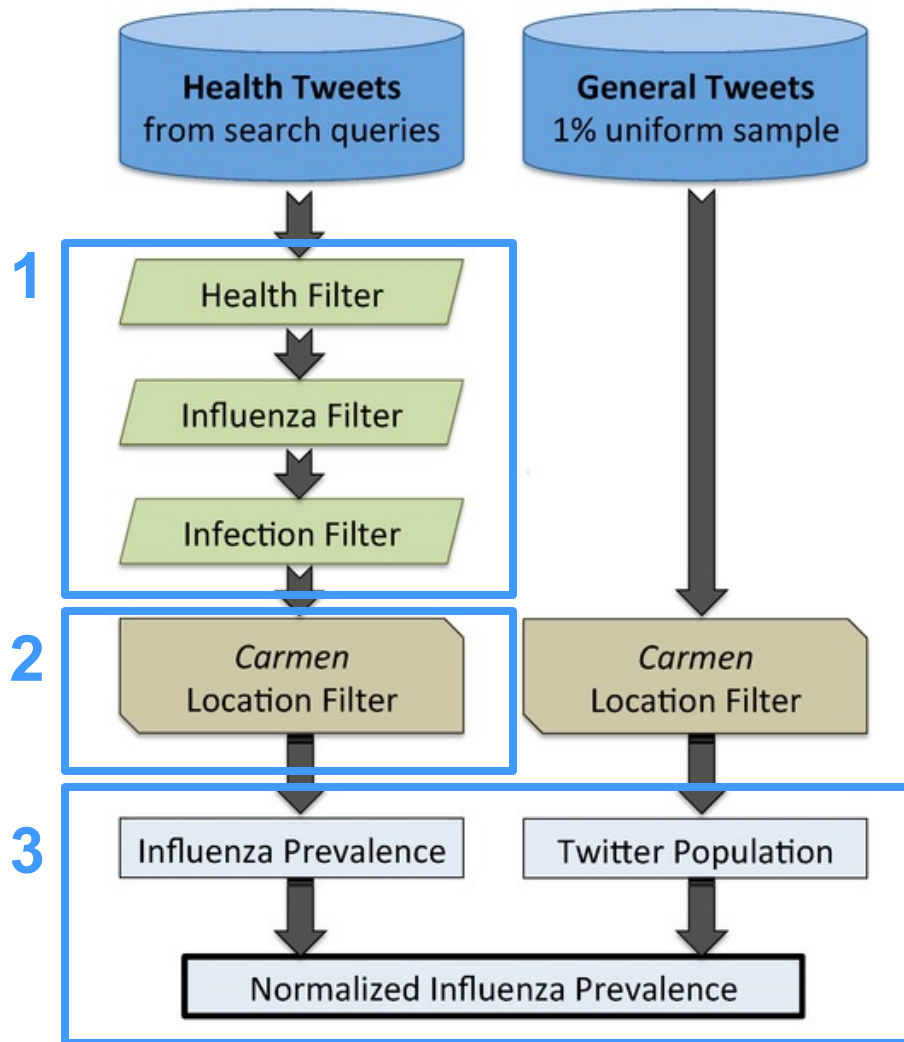
[Redacted] 
Periodic Pennsylvania [#flu](#) update; season has wound down greatly, but is not yet past. (As I've said so many times, this season is depicted by the mountainous red line.)

 2   1 

[Show this thread](#)

[Redacted] 
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

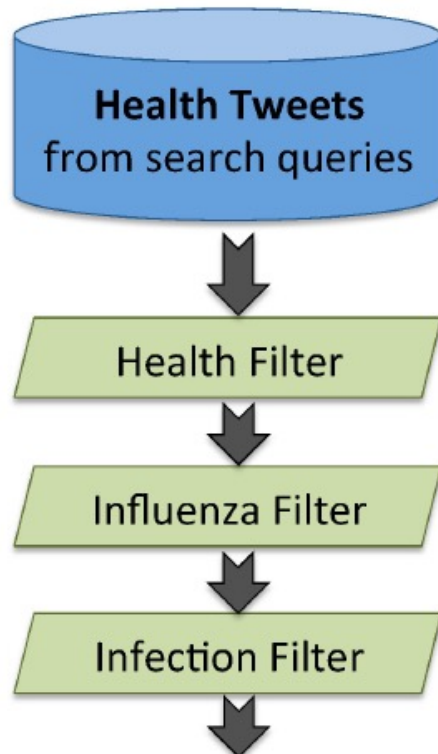


1. "Tweets matching hundreds of health-related keywords passed 3 classification **filters** to remove irrelevant tweets.
2. Locations are identified with **geolocation** system and only tweets in the location of interest are saved.
3. The volume of tweets is **normalized** by the total volume of tweets from a random Twitter sample to produce a prevalence measure."

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



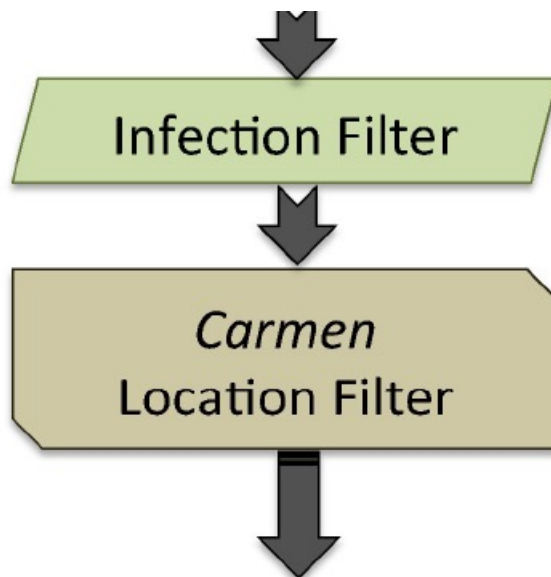
570,000 influenza infection tweets during 8 months

Filter 1 (health-relevant vs. irrelevant):
“combination of keyword filtering and 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; 74% precision, 87% recall”

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.

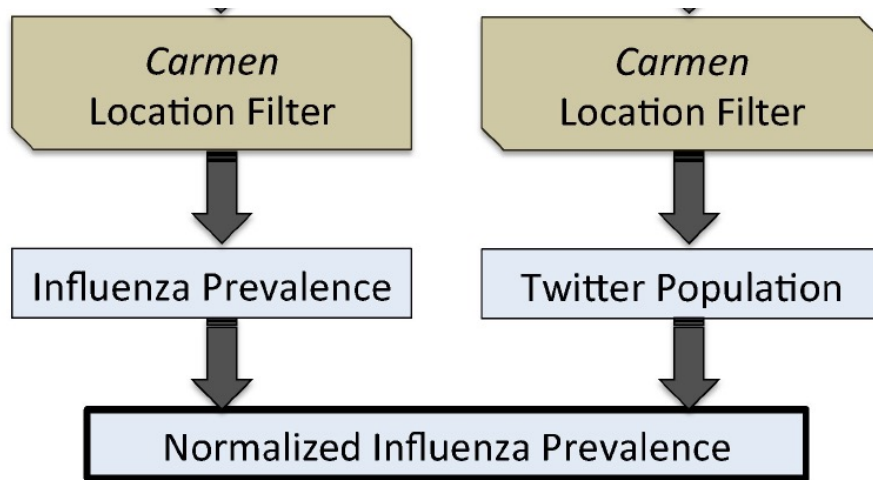
In evaluation set: 56,000 tweets, two locations:

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:

“Weekly number of infection tweets divided 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

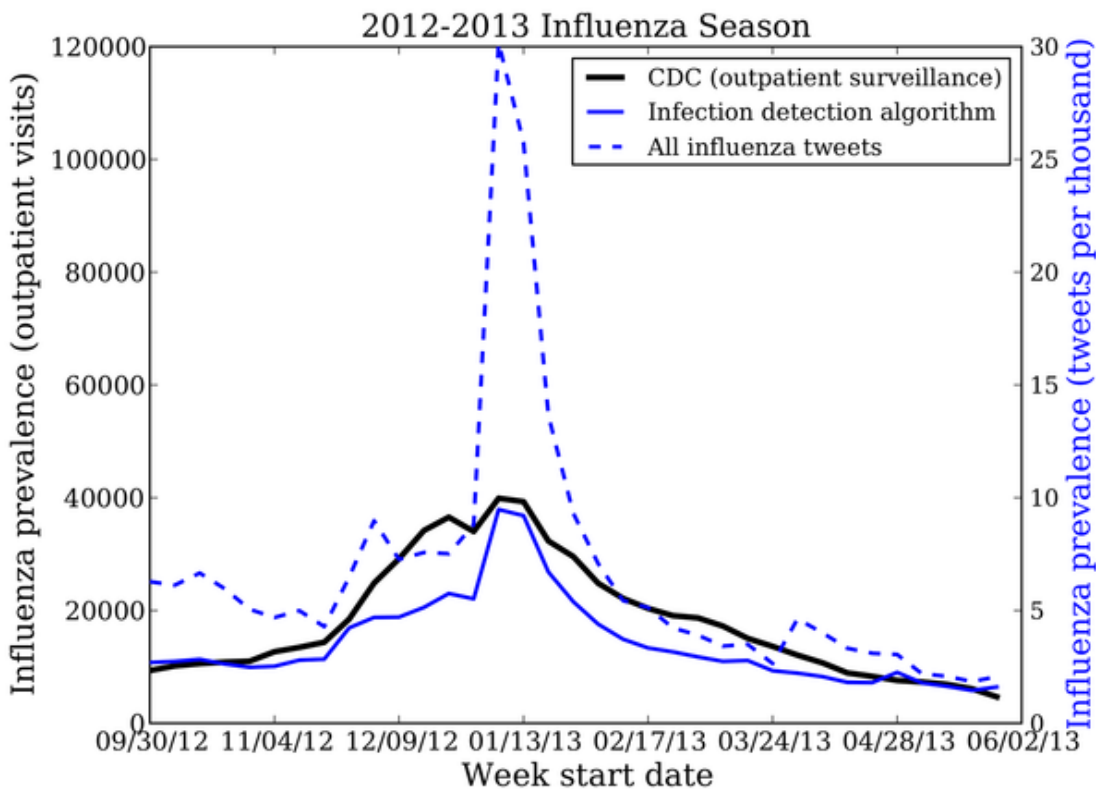
National Level (104,200 influenza infection tweets from US)

- * 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 ILI rates, yet they are available up to two weeks earlier than CDC data.”

Twitter Improves Influenza Forecasting

OCTOBER 28, 2014 · RESEARCH ARTICLE

 [Print or Save PDF](#)

 [Citation](#)

 [XML](#)

 [Email](#)

■ 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 Parable of Google Flu: Traps in Big Data Analysis

David Lazer,^{1,2*} Ryan Kennedy,^{1,3,4} Gary King,³ Alessandro Vespignani^{5,6,3}

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

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

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



discontinued in 2015
public data available on the webpage

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/>
<https://ai.googleblog.com/2015/08/the-next-chapter-for-flu-trends.html>

fast forward 2020



Cornell University

arXiv.org > cs > arXiv:2003.07372

Computer Science > Social and Information Networks

COVID-19: The First Public Coronavirus Twitter Dataset

Emily Chen, Kristina Lerman, Emilio Ferrara

(Submitted on 16 Mar 2020)

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](https://github.com/echen102/COVID-19-TweetIDs)). 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.

<https://arxiv.org/abs/2003.07372v1>

<https://github.com/echen102/COVID-19-TweetIDs>

BERT-based analysis of covid-19 tweets

BERT (Bidirectional Encoder Representations from Transformers): deep learning technique for NLP pretraining (Devlin, 2018)

CT-BERT (COVID Twitter BERT):

- trained on a corpus of 160M tweets Jan-Apr 2020 (22.5M after cleaning)
- used for downstream classification tasks

Dataset	Classes	Train	Dev	Labels	mean F1 score	
					BERT-LARGE	CT-BERT
COVID-19 Category (CC)	2	3094	1031	Personal News	0.931	0.949
Vaccine Sentiment (VC)	3	5000	3000	N Neutral Positive	0.824	0.869
Maternal Vaccine Stance (MVS)	4	1361	817	Disc A N Promotional	0.696	0.748
Stanford Sentiment Treebank 2 (SST-2)	2	67 349	872	Negative Positive	0.937	0.944
Twitter Sentiment SemEval (SE)	3	6000	817	Neg Neutral Positive	0.620	0.654
average					0.802	0.833

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv:1810.04805, 2018.

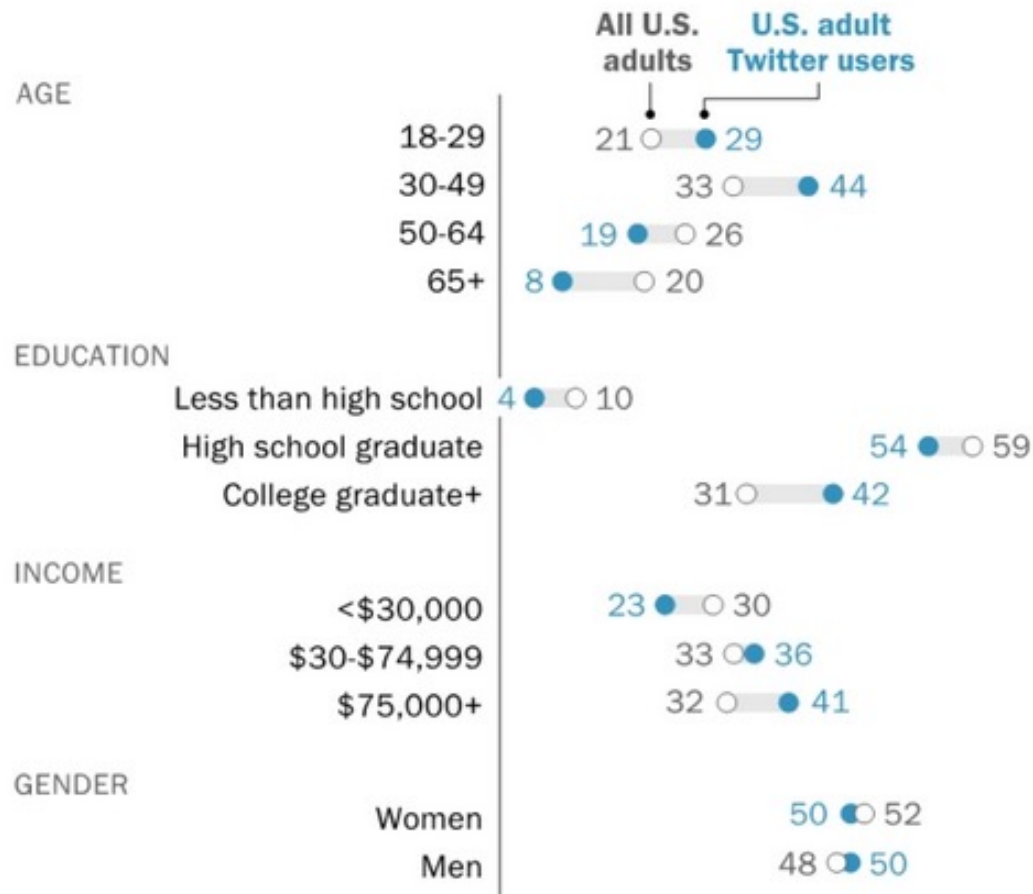
M. Müller, M. Salathé and P. E. Kummervold, COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter, arXiv 2005.07503, 2020

caution: Twitter is not everybody

representative survey of US adult Twitter users (N=2791)

Twitter users are younger, more highly educated and wealthier than general public

% of _____ who are ...



what to remember

inferring real-world trends from twitter

influenza trend detection as a case study

high-quality data annotation & pre-processing are essential

overoptimistic expectations of full automation in the “early days”

rather, a tool that could complement more established methods

beware of methodological limitations

sampling biases: not everybody is on Twitter

data pre-processing choices need to be made visible

overemphasis on single platform

questions?

daniel.gatica-perez@epfl.ch