Day 12: Data from the Web

ME314: Introduction to Data Science and Machine Learning

Jack Blumenau

28th July 2022

Final assessment

- Released noon on Friday
- Due 8pm on Sunday

Day 12 Outline

Social Media Data

Challenges of Social Data

Applications Using Social Data

Accessing Social Media APIs

Web scraping

What next?

Social Media Data

Why social media data?

Volume and coverage

- Twitter: $\approx$ 330 million monthly active users, $\approx$ 550m tweets per day
- Facebook: $\approx$ 1.88 billion daily active users, $\approx$ 2.7 billion monthly active users as of 2017
- Real time — new data is available (somewhat) publicly immediately on current events
- Metadata — geographic location, user device, profile, timestamp and other metadata is accessible.

Why social media data?

Social media and politics

- 99% of Members of the US Congress have an active social media account
- 90% of governments have a presence on Twitter
- “Traditional” media outlets rely on social media to promote their content
- 50% of social media users in U.S. share information about news stories, images or videos about current events
- 46% have discussed a news issue or event on social media

(Sources: Zeitzoff and Barbera, ISQ 2017; Pew Research Center)

Why social media data?

Social media and social science

1. **Unobtrusive** data collection at scale, e.g.in study of networks, censorship, etc
2. **Homogeneity** in data format across actors, countries, and over time, e.g. in the study of political rhetoric
3. **Granularity** of data, both temporal and spatial, e.g. in the study of geographical segregation
4. **Increasing representativeness** of social media users, e.g. in the study of political elites

Social media research

Two different approaches in the growing field of social media research:

1. Social media as a new source of data, e.g.
    - Behavior, opinions, and latent traits
    - Interpersonal networks
    - Elite behavior
    - Affordable field experiments
2. How social media affects social behavior, e.g.
    - Collective action and social movements
    - Political campaigns
    - Social capital and interpersonal communication
    - Political attitudes and behavior

Challenges of Social Data

Challenges

Social media data therefore offers many advantages, and the tools we will focus on today will give you the opportunity to dramatically expand the scope of your data collection efforts.

However, we should also consider some of the challenges of this type of data.

1. Bias
2. Ethics – informed consent
3. Fairness
4. Practical challenges

Big data and bias

- Population bias
  - Sociodemographics are (strongly) correlated with social media use
- Self-selection within samples
  - Partisans are much more likely to post about politics than independents
  - Product lovers/haters are more likely to post about products than those who have weaker opinions
- Priopprietary algorithms and unknown bias
  - e.g. Twitter API returns data which is "is not an accurate representation of the overall platform's data"
- Social media ativity does not generalise easily
  - Twitter is to social scientists what the fruit fly is to biologists – a model organism, but one that generalises poorly

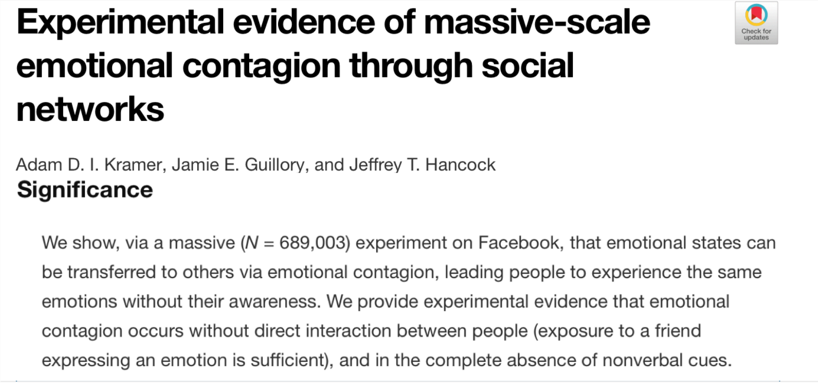
See Ruths and Pfeffer, 2015, "Social media for large studies of behavior"

Ethical concerns: informed consent (I)

> *One of the cornerstones of conducting ethically sound social science research involves the informed consent of participants, obtained through advising them about the study in which they are invited to partake, its possible risks, but also benefits, and the study's projected outcomes. The use of informed consent is important because it allows participants to make a choice and signals their willing participation.*

Gleibs, 2014

Ethical concerns: informed consent (I)



Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock

Significance

We show, via a massive (*N* = 689,003) experiment on Facebook, that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. We provide experimental evidence that emotional contagion occurs without direct interaction between people (exposure to a friend expressing an emotion is sufficient), and in the complete absence of nonverbal cues.

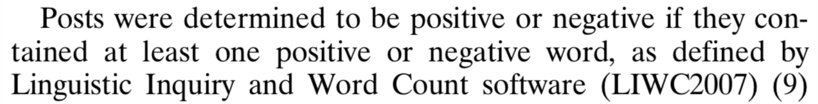
Ethical concerns: informed consent (I)

Research design

1. Measure the "emotional state" of Facebook newsfeed posts
2. Randomly assign Facebook users to three conditions
    - Reduced exposure to "negative" newsfeed items
    - Reduced exposure to "positive" newsfeed items
    - Control
3. Measure the "emotional state" of those users' subsequent newsfeed posts
4. If those in the treatment conditions have different emotional states than those in the control condition $\longrightarrow$ evidence of emotional "contagion"

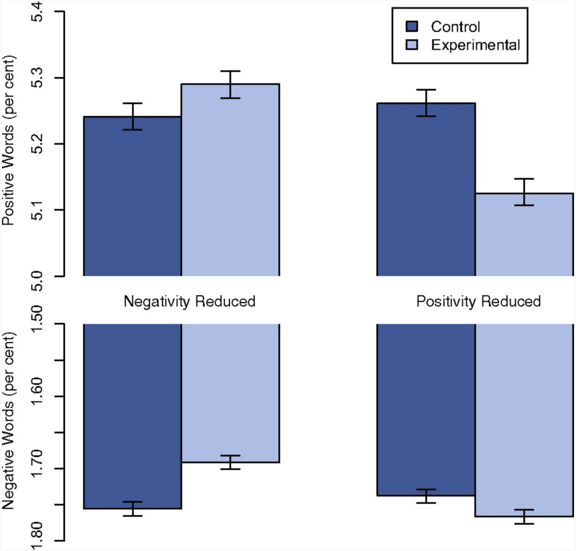
Ethical concerns: informed consent (I)

Measurement: How did the researchers measure positive and negative emotional states?

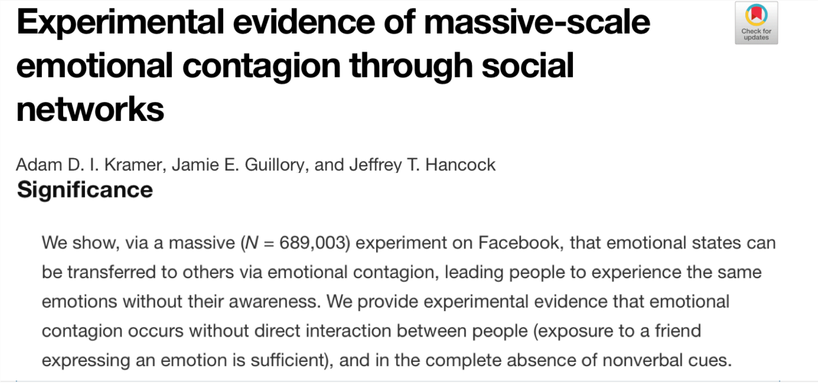


Posts were determined to be positive or negative if they contained at least one positive or negative word, as defined by Linguistic Inquiry and Word Count software (LIWC2007) (9)

Ethical concerns: informed consent (I)



Ethical concerns: informed consent (I)



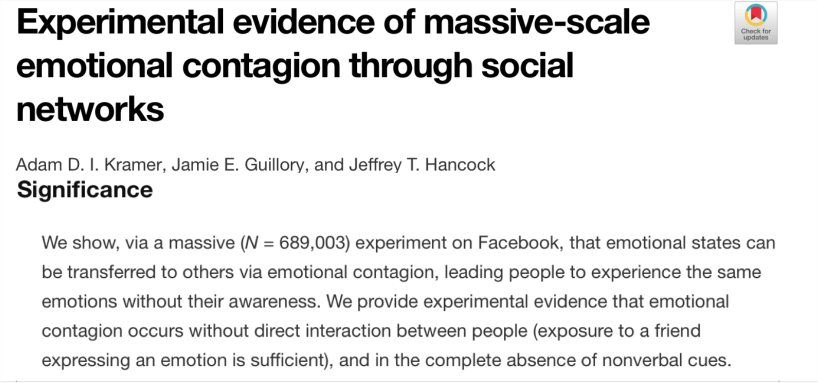
Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock

Significance

We show, via a massive (*N* = 689,003) experiment on Facebook, that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. We provide experimental evidence that emotional contagion occurs without direct interaction between people (exposure to a friend expressing an emotion is sufficient), and in the complete absence of nonverbal cues.

Ethical concerns: informed consent (I)



Experimental evidence of massive-scale emotional contagion through social networks

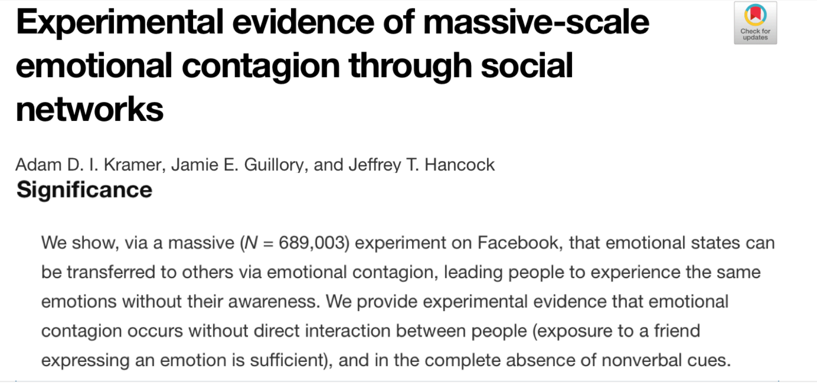
Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock

Significance

We show, via a massive (*N* = 689,003) experiment on Facebook, that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. We provide experimental evidence that emotional contagion occurs without direct interaction between people (exposure to a friend expressing an emotion is sufficient), and in the complete absence of nonverbal cues.

“The study was consistent with Facebook’s Data Use Policy, to which all users agree prior to creating an account on Facebook, constituting informed consent for this research.” Kramer et al, PNAS, 2014

Ethical concerns: informed consent (I)



Experimental evidence of massive-scale emotional contagion through social networks

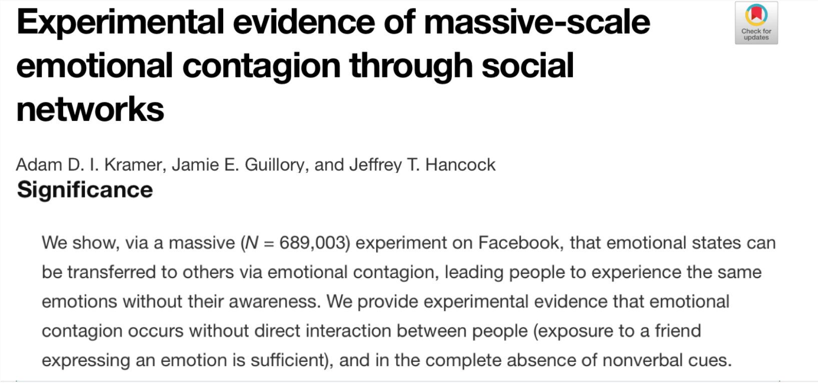
Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock

Significance

We show, via a massive (*N* = 689,003) experiment on Facebook, that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. We provide experimental evidence that emotional contagion occurs without direct interaction between people (exposure to a friend expressing an emotion is sufficient), and in the complete absence of nonverbal cues.

> *"The collection of the data by Facebook may have involved practices that were not fully consistent with the principles of obtaining informed consent and allowing participants to opt out." Editor-in-Chief, PNAS, 2015*

Ethical concerns: informed consent (I)



Experimental evidence of massive-scale emotional contagion through social networks

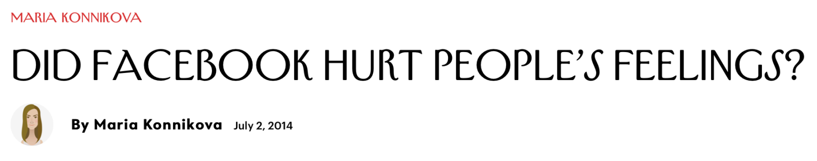
Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock

Significance

We show, via a massive (*N* = 689,003) experiment on Facebook, that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. We provide experimental evidence that emotional contagion occurs without direct interaction between people (exposure to a friend expressing an emotion is sufficient), and in the complete absence of nonverbal cues.

Are EULAs (End-User License Agreement) too complex to allow ‘informed consent’?

Ethical concerns: informed consent (I)



MARIA KONNIKOVA

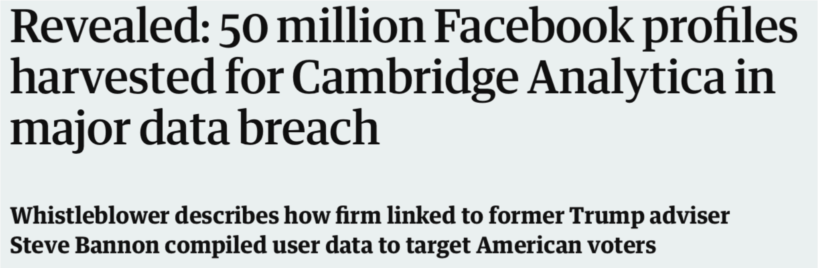
DID FACEBOOK HURT PEOPLE'S FEELINGS?

By Maria Konnikova July 2, 2014

No, they used a poorly designed text measure to detect tiny differences in word use.

But the ethical point stands!

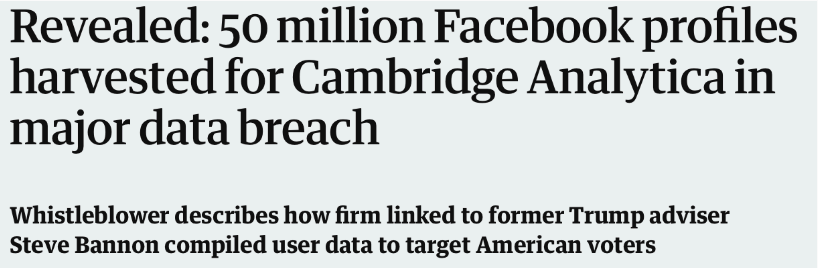
Ethical concerns: informed consent (II)



Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach

Whistleblower describes how firm linked to former Trump adviser Steve Bannon compiled user data to target American voters

Ethical concerns: informed consent (II)

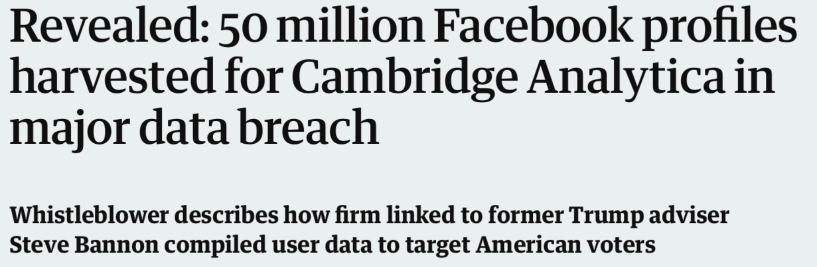


Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach

Whistleblower describes how firm linked to former Trump adviser Steve Bannon compiled user data to target American voters

- CA gained access to Facebook data through a partnership with Aleksandre Kogan, a UK academic
- Kogan presented his data gathering as academic, but agreed to share information with CA
- Facebook claims that users gave consent to share data with Kogan, but not to the secondary sharing with CA
- Kogan's app collected profile data from participants' networks using the social graph API

Ethical concerns: informed consent (II)



Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach

Whistleblower describes how firm linked to former Trump adviser Steve Bannon compiled user data to target American voters

Two privacy issues:

1. Lack of consent for data exposure to a third party
2. Lack of consent for harvesting of social graph data

Big data and fairness

"There is nothing about doing data analysis that is neutral. What and how data is collected, how the data is cleaned and stored, what models are constructed, and what questions are asked – all of this is political."
Danah Boyd, NYU

1. Computers can learn to acquire existing human baises
    - Particularly problematic for decision-making (hiring, policing, etc)
2. Even in very large datasets, there is always proportionally less data available about minorities
    - If the training data reflect existing social biases against a minority, the algorithm is likely to incorporate these biases
    - Statistical patterns that apply to the majority might be invalid within a minority group

Big data and fairness: Example

Word-embeddings are an unsupervised learning method for discovering the “meaning” of words inductively from a corpus of texts.

Intuition:

> *“You shall know a word by the company it keeps.’’ J.R. Firth 1957*

The basic idea behind word-embedding models is to use the co-occurance of terms within a corpus to create vectors that encode the meaning of each term.

There has been a recent surge of work that uses word-embeddings for many downstream predictive tasks, including in decision-making systems.

One way of understanding the resulting embeddings is to see which words are “close” to one another in the embedding space.

Big data and fairness: Example

Let’s use a matrix of word-embeddings that I trained on the corpus of parliamentary speeches we have been using:

```
word_vectors[1:5,1:5]
```

```
##                [,1]        [,2]        [,3]        [,4]       [,5]
## house    0.09834925  0.34462858  0.43410388 -0.01537683  0.3328848
## proceeds 0.10935879 -0.69976782 -0.11314722  0.40691536 -0.6123208
## choice   0.21215889  0.54387728 -0.51125106 -0.56793830  0.8246459
## speaker  0.15791494 -0.05892315  0.21089931  0.05878700  0.3328526
## may      0.13679385  0.59354320  0.08695598  0.07544566  0.3619411
```

This shows us the first 5 embedding-dimensions (150 total) of the first 5 words in our corpus.

Big data and fairness: Example

We can calculate the cosine-similarity between word embeddings by using the `sim2` function from the `text2vec` package:

```r
library(text2vec)
```

```r
# Calculate the similarity between the word "nhs" and all other words
cosine_sim <- sim2(word_vectors, word_vectors["nhs",, drop = F])

# Show the top 5 most similar words to the word "nhs"
rownames(word_vectors)[order(cosine_sim, decreasing = T)][1:5]
```

```
## [1] "nhs"       "trusts"    "hospitals" "patients"  "trust"
```

```r
# Calculate the similarity between the word "brexit" and all other words
cosine_sim <- sim2(word_vectors, word_vectors["brexit",, drop = F])

# Show the top 5 most similar words to the word "brexit"
rownames(word_vectors)[order(cosine_sim, decreasing = T)][1:5]
```

```
## [1] "brexit"       "no-deal"      "negotiations" "referendum"   "scenario"
```

Big data and fairness: Example

We can use these similarity measures to test whether embeddings trained on this corpus embed gender bias.

1. List a set of job titles

```
career_words <- c("policeman", "cleaner", "surgeon", "politician",
                  "author", "librarian", "cashier", "waiter",
                  "waitress", "banker", "doctor", "academic","nurse")
```

2. Calculate the similarity between each of these jobs and the words “man” and “woman”

```
female_sim <- sim2(word_vectors["woman",,drop = F], word_vectors[career_words,,drop = F])
male_sim <- sim2(word_vectors["man",,drop = F], word_vectors[career_words,,drop = F])
```

3. See whether each careers is considered more male or female according to the embeddings

```
more_female_careers <- career_words[female_sim > male_sim]
more_male_careers <- career_words[female_sim < male_sim]
```

Big data and fairness: Example

```r
print(more_male_careers)
```

```
## [1] "policeman"  "surgeon"    "politician" "waiter"     "banker"
## [6] "doctor"     "academic"
```

```r
print(more_female_careers)
```

```
## [1] "cleaner"   "author"    "librarian" "cashier"   "waitress"  "nurse"
```

More practical concerns

- Legal issues need to catch up with the technology
- Large amounts of data
  - storage problems
  - analysis problems
- Language is informal and often non-textual (emoticons, links, images) - and slang, txtspk, emoticons :-(
- Lots of fake users
- Commercial interfaces are brittle and opaque

Applications Using Social Data

Example applications

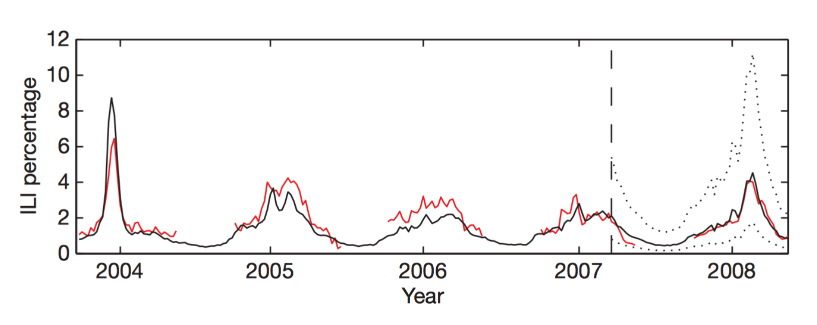
Nevertheless, social media data can provide interesting (and important) insights into human behaviour.

- Offers measures of actual behaviors, as compared with self-reports of behaviors
  - Particularly important for studying phenomena where people either deliberately or accidently misreport their behaviour
  - i.e. reporting of social ties, anti-social behaviour, etc
- Offers instantaneous information about potentially important trends
  - i.e. employment (Toole et al, 2015); public opinion (Beauchamp, 2016); health (Ginsberg et al, 2009)
- Offers the opportunity to study the mechanics of social systems
  - How do individuals interact? How do they form social ties? How does segregation occur?

Example application (I)

- Tracking disease through google search terms (Ginsberg et al 2009)
  - Aggregate logs of online search queries between 2003 and 2008
  - Compute a time series of weekly counts for common search terms, decomposed by region
  - Collect data on the number of visits to doctors with influenza-like illness (ILI) symptoms
  - Use a series of logistic regressions to learn which search terms are associated with flu outbreaks
  - Select the 50 most predictive terms
  - Use this association to observe current outbreaks

Example application (I)



"Harnessing the collective intelligence of millions of users, Google web search logs can provide one of the most timely, broad-reaching influenza monitoring systems available today." (Ginsberg et al 2009)

Example application (I)

However...

- Although there was a lot of search data to use for training, there were relatively few outcome points (flu outbreaks)
- This meant that the researchers ended up overfitting the data
- This had considerable consequences for predicting in other years:
    - Missed the non-seasonal outbreak in 2009
    - Hugely overestimates flu outbreaks in 2011-2013
    - Estimates were almost double official statistics in 2012

"Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data." (Lazer et al, 2014)

“Fixing” the biased twitter sample

Barbera, 2018, Working Paper

- Using twitter in studies of social behaviour is difficult because we lack information about the sociodemographic characteristics of twitter users
- Researchers cannot estimate “survey” weights to recover representativeness of their samples as we do with traditional surveys
- (Additional problem: many interesting questions require demographic information!)
- Solution: match a large (250,000) sample of twitter users to voter registration records, which provide information on age, gender, race, party identification, and - indirectly - house value
- Train a classifier to learn the text features most associated with these demographics
- Predict demographics for many other users

“Fixing” the biased twitter sample

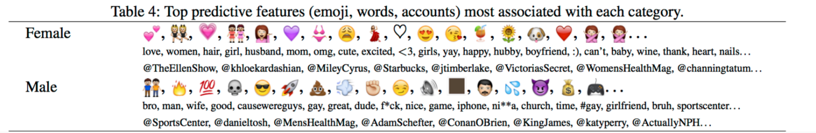
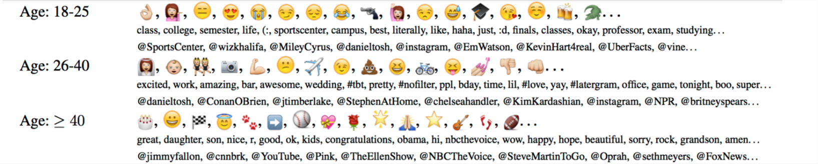


Table 4: Top predictive features (emoji, words, accounts) most associated with each category.

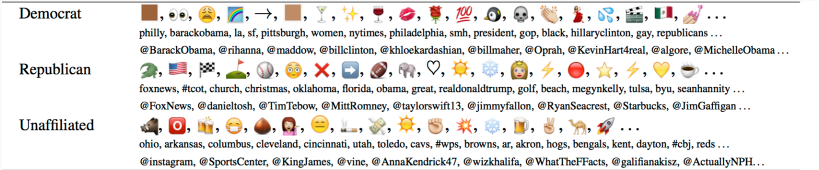
Female	💕, 👯, 💗, 👭, 💁, 💜, 👙, 😩, 💃, ♡, 😍, 😘, 🍹, 🌻, 🐶, ❤️, 🙅, 🙅. . . love, women, hair, girl, husband, mom, omg, cute, excited, <3, girls, yay, happy, hubby, boyfriend, :), can’t, baby, wine, thank, heart, nails. . . @TheEllenShow, @khloekardashian, @MileyCyrus, @Starbucks, @jtimberlake, @VictoriasSecret, @WomensHealthMag, @channingtatum. . .
Male	👬, 🔥, 💯, 💀, 😎, 🚀, 💩, 💨, ✊, 😏, 🔊, 🏾, 👨, 💦, 😈, 💰, 🎮. . . bro, man, wife, good, causewereguys, gay, great, dude, f*ck, nice, game, iphone, ni**a, church, time, #gay, girlfriend, bruh, sportscenter. . . @SportsCenter, @danieltosh, @MensHealthMag, @AdamSchefter, @ConanOBrien, @KingJames, @katyperry, @ActuallyNPH. . .

“Fixing” the biased twitter sample



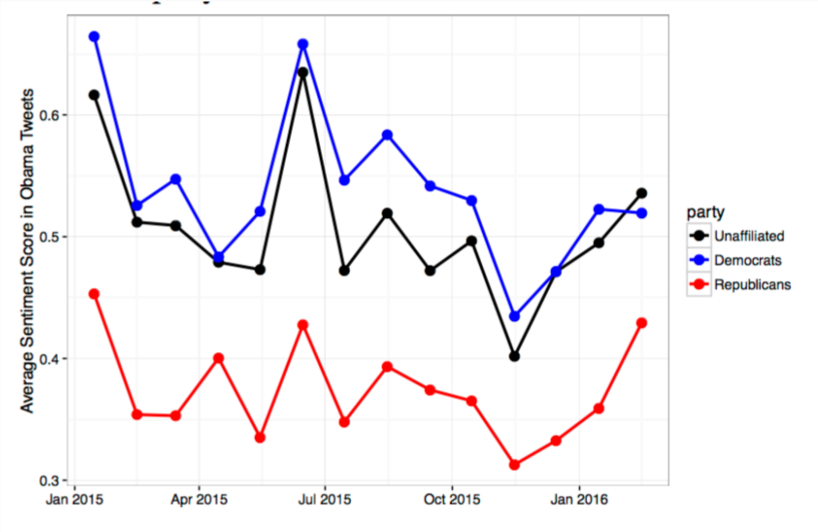
Age: 18-25	👌, 💁, 😑, 😍, 😭, 😏, 😔, 😂, 🔫, 🙋, 😒, 😅, 🎓, 😘, ☺️, 🍻, 🐊. . . class, college, semester, life, (:, sportscenter, campus, best, literally, like, haha, just, :d, finals, classes, okay, professor, exam, studying. . . @SportsCenter, @wizkhalifa, @MileyCyrus, @danieltosh, @instagram, @EmWatson, @KevinHart4real, @UberFacts, @vine. . .
Age: 26-40	👰, 👶, 👯, 📷, 💪, 😕, ✈️, 😉, 💩, 😆, 🚲, 😝, 💅, 👎, 👊. . . excited, work, amazing, bar, awesome, wedding, #tbt, pretty, #nofilter, ppl, bday, time, lil, #love, yay, #latergram, office, game, tonight, boo, super. . . @danieltosh, @ConanOBrien, @jtimberlake, @StephenAtHome, @chelseahandler, @KimKardashian, @instagram, @NPR, @britneyspears. . .
Age: ≥ 40	🎂, 😀, 🏁, 😇, 🐾, ➡️, ⚾, 💝, 🌹, 🌟, 🙏, ⭐, 🎸, 👣, 🏈. . . great, daughter, son, nice, r, good, ok, kids, congratulations, obama, hi, nbcthevoice, wow, happy, hope, beautiful, sorry, rock, grandson, amen. . . @jimmyfallon, @cnnbrk, @YouTube, @Pink, @TheEllenShow, @NBCTheVoice, @SteveMartinToGo, @Oprah, @sethmeyers, @FoxNews. . .

“Fixing” the biased twitter sample



Democrat	🟫, 👀, 😩, 🌈, →, 🟫, 🍸, ✨, 🍷, 💋, 🌹, 💯, 🐧, 💀, 👏, 💃, 💦, 🎬, 🇲🇽, 💅 . . . philly, barackobama, la, sf, pittsburgh, women, nytimes, philadelphia, smh, president, gop, black, hillaryclinton, gay, republicans . . . @BarackObama, @rihanna, @maddow, @billclinton, @khloekardashian, @billmaher, @Oprah, @KevinHart4real, @algore, @MichelleObama . . .
Republican	🐊, 🇺🇸, 🏁, ⛳, ⚾, 😳, ❌, ➡️, 🏈, 🐘, ♡, ☀️, ❄️, 👸, ⚡, 🔴, ⭐, ⚡, 💛, ☕ . . . foxnews, #tcot, church, christmas, oklahoma, florida, obama, great, realdonaldtrump, golf, beach, megynkelly, tulsa, byu, seanhannity . . . @FoxNews, @danieltosh, @TimTebow, @MittRomney, @taylorswift13, @jimmyfallon, @RyanSeacrest, @Starbucks, @JimGaffigan . . .
Unaffiliated	🐗, 🅾️, 🍻, 😷, 🌰, 💁, 😑, 🚬, 💸, ☀️, ✊, 💥, ❄️, 🍺, ✌️, 🐪, 🚀 . . . ohio, arkansas, columbus, cleveland, cincinnati, utah, toledo, cavs, #wps, browns, ar, akron, hogs, bengals, kent, dayton, #cbj, reds . . . @instagram, @SportsCenter, @KingJames, @vine, @AnnaKendrick47, @wizkhalifa, @WhatTheFFacts, @galifianakisz, @ActuallyNPH. . .

Validation


Average Sentiment Score in Obama Tweets
0.3
0.4
0.5
0.6
Jan 2015
Apr 2015
Jul 2015
Oct 2015
Jan 2016
party
Unaffiliated
Democrats
Republicans

Accessing Social Media APIs

How can we access this data?

- API: Application Programming Interface — a way for two pieces of software to talk to each other
- Twitter, facebook, google — all expose public web services
- Your software can receive (and also send) data automatically through these services
- Data is sent by `http` — the same way your browser does it
- Most services have helping code (known as a wrapper) to construct http requests
- both the wrapper and the service itself are called APIs
- http service also sometimes known as REST (REpresentational State Transfer)

Anatomy of a http request

It is helpful to start paying attention to the structure of basic http requests.

For instance, let's say we want to get some data from the TheyWorkForYou api.

A test request:

```
https://www.theyworkforyou.com/api/getDebates&
output=xml&search=brexit&num=1000&key=XXXXX
```

- Parameters to the API are encoded in the URL
  - `output` = Which format do you want returned?
  - `search` = Return speeches with which words?
  - `num` = number requested
  - `key` = access key

API packages

- It’s not usually necessary to construct these kind of requests yourself
- R, Python, and other programming languages have libraries to make it easier – but you have to find them!
- The documentation for the API will describe the parameters that are available. Though normally in a way that is intensely frustrating.

Available social media APIs

- Wikipedia: mediawiki
- Google (various)
- reddit
- foursquare
- facebook
- twitter: REST, Streaming, firehose

Note: both Twitter and Facebook have increased the registration hurdles required for accessing their APIs recently.

The twitter APIs: REST

- Queries for specific information about users and tweets
- Returns a sample of historical data from the last 7 days.
- Examples: user profile, list of followers and friends, tweets generated by a given user (“timeline”), user lists, etc
- The manual: https://developer.twitter.com/en/docs/tweets/search/overview
- R package : `twitteR` or `rtweet`

The twitter APIs: Streaming

- Connect to the twitter server and collect tweets as they fly by
- Three streaming APIs
  - Filter stream: tweets filtered by keywords
  - Geo stream: tweets filtered by location
  - Sample stream: 1% random sample of tweets
- The manual: https://developer.twitter.com/en/docs/tweets/filter-realtime/overview
- R package: `streamR` or `rtweet`

Authentication

- Username and Password and additional security keys
- Oauth (ROauth): share a key without sharing a username and password
- IP address limitations
- Rate limitations
- Per-user and per-application

The Output: JSON and XML

- XML: eXtensible Markup Language: encodes documents in a form that is both human-readable and machine readable
- JSON : JavaScript Object Notation
- If you have a choice, you probably want JSON
- JSON uses key:value pairs, XML uses trees
- JSON is easily read into a programming language
- `json_lite` and `xml2` are the relevant R packages

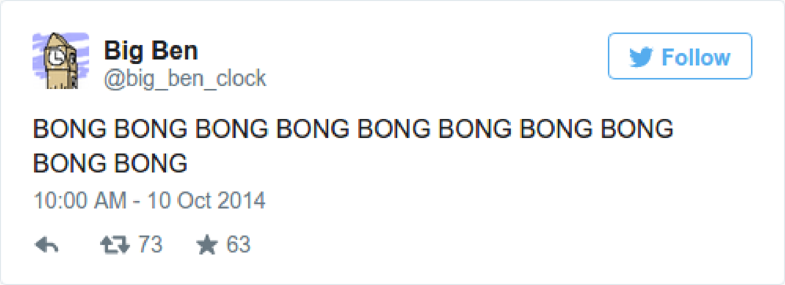
And finally… the data

- Full of spam, bots, unicode, and gibberish
- Lots of retweets (approximately one-third retweets, replies, tweets)
- Only 1% show location — some methods exist to infer location
- All aspects of metadata and reply/retweet structure are available
- All aspects of network structure: followers and ‘friends’, profile information

Twitterbots

- API also allows actions such as posting tweets (POST)
- Examples:
  - @earthquakesLA posts earthquake warnings for LA
  - @year_progress posts a progress bar for the year every day

Twitterbots



Big Ben
@big_ben_clock

BONG BONG BONG BONG BONG BONG BONG BONG BONG BONG

10:00 AM - 10 Oct 2014

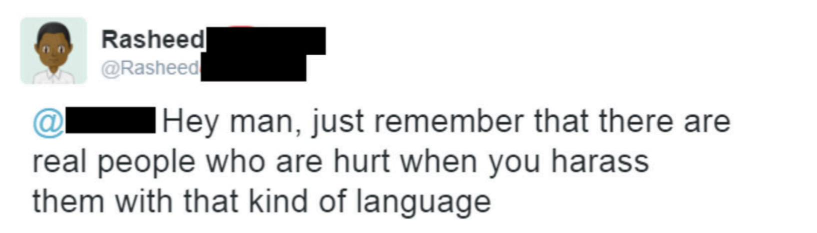
73 63

Twitterbots in research

Munger, 2017, Political Behaviour

- Does social sanctioning reduce racist online harassment?
- Design:
    - Randomly assign a sample of racist Twitter users to a treatment and control group
- Treatment group: direct ‘bots’ to sanction users for their use of racist terms
    - Vary whether the bot is in-group (white) or out-group (black)
    - Vary the number of followers the bot has
- Control group: leave users alone
- Measure whether treatment group reduce their use of racist language in subsequent weeks

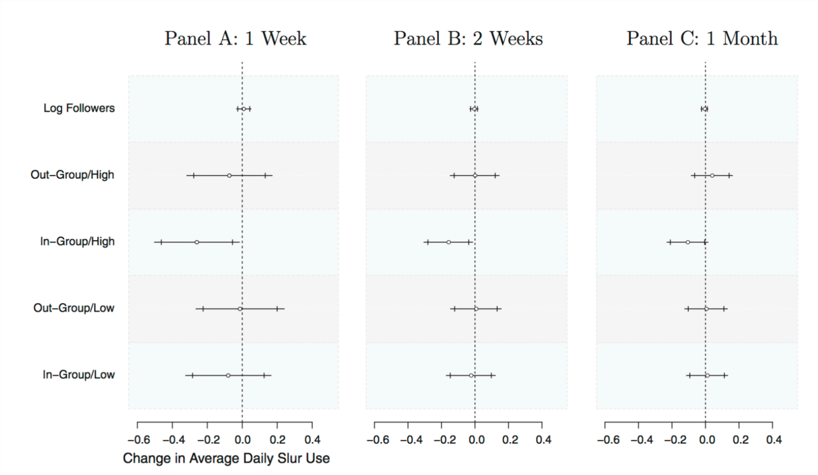
Twitterbots in research



Rasheed [redacted]
@Rasheed [redacted]

@ [redacted] Hey man, just remember that there are real people who are hurt when you harass them with that kind of language

Twitterbots in research



Twitterbots in research

But!

> *“After it was published, a pair of grad students wanted to use the ‘Twitter bot RCT’ method to test different hypotheses but they had a problem: their bots were getting shut down by Twitter after only a few uses. My experiment could not be replicated. This is not in the sense of the ‘replication crisis’, where published effects are found to be null when the experiment is repeated; rather, the rules of the universe (here, Twitter) had changed so that even conducting the experiment is impossible.” (Munger, 2019)*

Twitter uses: Exploiting the meta-data (non-textual)

- location
- time
- username
- user descriptions
- networks of followers
- retweets of followers and texts

Connecting through R

R packages

- Twitter: `twitteR` or `rtweet` for REST, `streamR` or `rtweet` for Streaming
- Facebook: `Rfacebook`
- Integration with `quanteda` is fairly straightforward

Python

- `tweepy`
- `facebook-sdk`

Other open-source tools exist

Break

Demonstration One

Web scraping

Web scraping overview

Overview of the key steps in any web-scraping project

1. Work out how the website is structured
2. Work out how links connect different pages
3. Isolate the information you care about on each page
4. Write a loop which connects 3 to 2, and saves the information you want from each page
5. Put it all into a nice and tidy `data.frame`
6. Feel like a superhero

(This is missing the steps in which you scream at your computer because you can’t figure out how to do steps 1-5.)

Demonstration Two

Superhero

WHENEVER I LEARN A NEW SKILL I CONCOCT ELABORATE FANTASY SCENARIOS WHERE IT LETS ME SAVE THE DAY.
OH NO! THE KILLER MUST HAVE FOLLOWED HER ON VACATION!
BUT TO FIND THEM WE'D HAVE TO SEARCH THROUGH 200 MB OF EMAILS LOOKING FOR SOMETHING FORMATTED LIKE AN ADDRESS!
IT'S HOPELESS!
EVERYBODY STAND BACK.
I KNOW REGULAR EXPRESSIONS.
PERL!
TAP TAP

What next?

Update your CV

You could all now legitimately add something like this to your CV:

> *Training in data science and machine learning, including experience with: data manipulation and visualisation; supervised and unsupervised learning methods; linear and logistic regression; classification methods; non-linear methods (local regression, splines, GAMs); tree-based methods (bagging, RandomForests); unsupervised learning methods (k-means; principal components analysis); quantitative text analysis (dictionaries, supervised learning for text, topic models); web-scraping.*

Further study

1. Machine Learning and Causality
    - How can we use these tools to (help) make causal statements?
2. Machine Learning Theory
    - What is *really* going on here?
3. Measurement
    - What does it mean to have a good measure of X or Y? Which tools are available to us for measuring our concepts of interest?
4. Advanced Text Analysis
    - Beyond topic models!

Further study

There are some very good MSc courses near here that offer a lot of this material!

- LSE option: MSc Applied Social Data Science
- UCL option: MSc Data Science and Public Policy

Please feel free to email me if you are interested in the UCL course.

Thank you!

Commercial interfaces are brittle and opaque

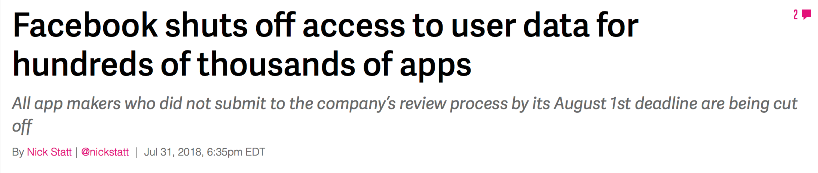
This was an example in slides for this course for a previous year:

```
## Code for API examples from Social Media class, 2017
library(Rfacebook)

## Scraping most recent 200 posts from Trump FB page
trump <- getPage("DonaldTrump", token = token, n = 100)

> Error in callAPI(url = url, token = token, api = api) :
```

Why doesn't it work?



Facebook shuts off access to user data for hundreds of thousands of apps

All app makers who did not submit to the company's review process by its August 1st deadline are being cut off

By Nick Statt | @nickstatt | Jul 31, 2018, 6:35pm EDT

Back