Collecting Twitter and reddit Data

Fundamentals of Computing and Data Display

Christoph Kern Ruben Bach[1]

(c.kern, r.bach)@uni-mannheim.de

[1] Thanks to Simon Kuehne (Bielefeld University) for Twitter materials

Outline

Twitter

Twitter

- Micro-blogging
- Online news and social networking service
- Main function: Publicly sharing short texts/photos/links, "Tweets"
- About 326 Million monthly active users and 500 Million Tweets send each day
  - `https://s22.q4cdn.com/826641620/files/doc_financials/2018/q3/TWTR-Q3_18_InvestorFactSheet.pdf`

Twitter Mechanics

- Each user has a profile (page) and can add a photo and information about themselves
- Users can follow each other
- Users can tweet, i.e., publicly sharing a text/photo/link
- Each Tweet is restricted to a maximum of 280 characters
- Users can interact with a Tweet via comments (replys), likes, shares (retweets)
- Users can interact with other users via direct messaging
- Users can create a thread = A series of connected Tweets
- Users use hashtags (#) in order to associate their Tweets with certain topics and to make them easier to find
- Users can search for keywords/hashtags in order to find relevant Tweets and users

Twitter for research

- Analyzing Tweets and social interaction on Twitter can help to answer social science research questions, esp. in communication research and political science
- Contrary to Facebook and Instagram, Twitter data is (easily) **accessible for researchers**
  - Facebook: API shut-down in April 2018
  - Instagram: API shut-down in December 2018

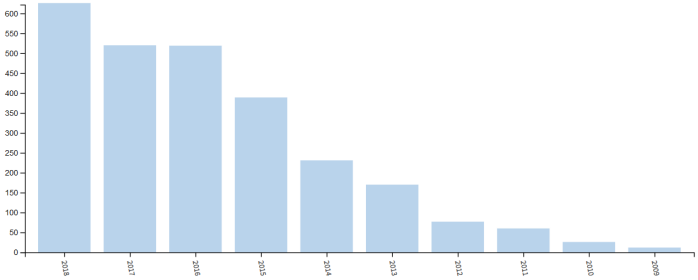


Figure: Number of journal articles with "Twitter" in title (Web of Science database)

Twitter for research

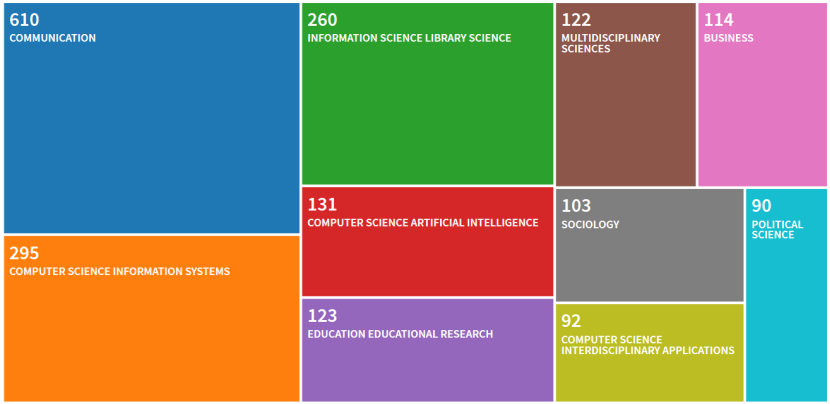


Figure: Journal articles since 2009 with "Twitter" in title (Web of Science database)

Reddit

Reddit

- "the front page of the internet"
- Registered users can post texts, links, images, videos onto thousands of subreddits ("submission")
- Chances are high that any topic you can think of is covered on reddit
- Other individuals may read, vote submissions up or down, or comment on posts
- One of the most visited website in the world (18^{th}) (facebook.com: 5^{th}, twitter.com 51^{st})
  - `https://www.alexa.com/siteinfo/reddit.com` (Aug 2020)

Reddit statistics

Table 1. Reddit Usage Statistics.

138,000+ active subreddits
10,700,000 posts and comments per month
330,000,000+ active users per month
14,000,000,000 screen views per month

Amaya et al. (2019)

Reddit Mechanics

- Most reddit content can be read without registering
- To post, comment, up or down vote, users need to register an (anonymous) account (email address sufficient)
- Users can subscribe subreddits
  - e.g., r/worldnews, r/democrats, r/AskStatistics, r/uglycats, r/mommit
- Some subreddits contain explicit content ("nsfw")
- Rules regarding appropriate content and behavior vary by subreddit

Reddit Mechanics

Typical comment structure

- Submission
  - Comment 1
    - Comment 1.1
    - Comment 1.2
  - Comment 2
    - Comment 2.1
  - ...

Reddit for research

- Reddit data is (easily) **accessible for researchers**
- Much less popular for research than other social media data
  - Google Scholar: "Twitter" ∼7,530,000, "Reddit" ∼1,630,000
- Examples
  - Ammari et al. (2018): Parenting roles
  - Choudhury (2014): Mental health discourse
  - En et al. (2013): Construction of sexual identities
  - Sowles et al. (2018): Eating disorders

Outline

Collecting Twitter data

API Authentification

- Twitter uses the OAuth protocol, an "open protocol to allow secure authorization in a simple and standard method from web, mobile and desktop applications."
  - `https://oauth.net/`
- Create a Twitter account
- Login to your Twitter account via `https://developer.twitter.com/`
- Activate Twitter Developer Account
- Create an app
- Create keys, access token & secret
- Use those to authenticate each query to the Twitter API
- Using **rtweet**-command for the first time in a session will open browser
- Login with your username and password

Collecting Twitter data

The Twitter API Platform

- Allows to access (real-time) Twitter data, i.e., Tweets
- Twitter offers a variety of API services, some for free, others not
- `https://developer.twitter.com/en/docs.html`
  - Search for Tweets published in the past
  - Stream Tweets in realtime
  - Manage Twitter accounts and ads
  - etc.
- Realtime Streaming API used in the majority of research projects

Collecting Twitter data

The Streaming API

- `https://developer.twitter.com/en/docs/tweets/filter-realtime/overview`
- *"Establishing a connection to the streaming APIs means making a very long lived HTTP request, and parsing the response incrementally. Conceptually, you can think of it as downloading an infinitely long file over HTTP."*
- Receive up to a maximum of 1% of all Tweets worldwide
- As a query is usually specified by selected **keywords** or **geographic areas**, you will be able to collect (almost) all relevant Tweets
- Three filter parameters
  - 'Follow': Receive Tweets of up to 5,000 users
  - 'Track': Receive Tweets that contain up to 400 keywords
  - 'Location': Receive Tweets from within a set of up to 25 geographic bounding boxes

Collecting Twitter data

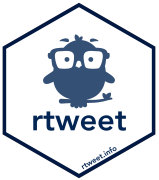
Twitter API response

- The Twitter API sends Tweets and related meta-information.
- Stored in .json format
- Each object ('row') represents a Tweet or ReTweet with:
  - The content of a Tweet + Tweet-URL + Tweet-ID
  - User-name + User-ID
  - Time-stamp
  - Place, country, geocodes (rarely)
  - User self-description, residence, no. of followers, no. of friends
  - URLs to images, videos
  - etc.

rtweet

There are a number of ways to collect Twitter data

- R: Package "**rtweet**"
  - `https://rtweet.info/articles/intro.html`
- Python: Package "tweepy"
- Write your own script
- etc.



Analyzing Twitter data

- Content Analysis
  - What kind of topics are users talking about?
- Sentiment Analysis
  - What kind of opinions/attitudes/emotions towards objects/concepts are users communicating?
- Network Analysis
  - Who is related to whom? Who are important users?
- Geospatial Analysis
  - Where are users/Tweets coming from?

Analyzing Twitter data

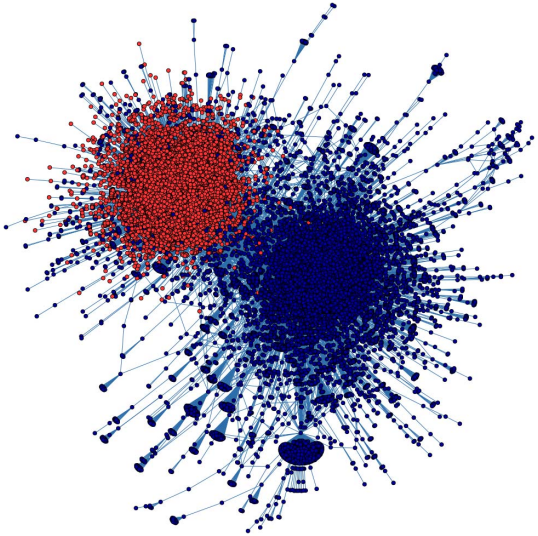


Figure: A political retweet network (Conover et al. 2011)

Analyzing Twitter data

Twitter text analysis in R

- Text mining of tweets
  - `https://www.earthdatascience.org/courses/earth-analytics/get-data-using-apis/text-mining-twitter-data-intro-r/`
- Twitter bot detection
  - `https://mikewk.shinyapps.io/botornot/`
  - `https://tweetbotornot.mikewk.com/`
- Sentiment analysis
  - `https://quanteda.io/`
  - `https://cran.r-project.org/web/packages/SentimentAnalysis/vignettes/SentimentAnalysis.html`

Bias and ethics

Bias, 'representativeness' and replication

- Twitter is far from representing a random sample of a given population
- Population → Internet Users → Twitter Users → Active Twitter Users → (Users sharing geo-information)
  - Barbera, P. and Rivero, G. (2015). Understanding the Political Representativeness of Twitter Users. *Social Science Computer Review 33*(6), 712–729.
  - Mellon, J. and Prosser, C. (2017). Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research and Politics 2017*, 1–9.
- Real-time Twitter data collection is not reproducible and for a given query you can only hope for Twitter providing you with a true random sample of Tweets
- Further issues: Social bots, company & institutional accounts

Bias and ethics

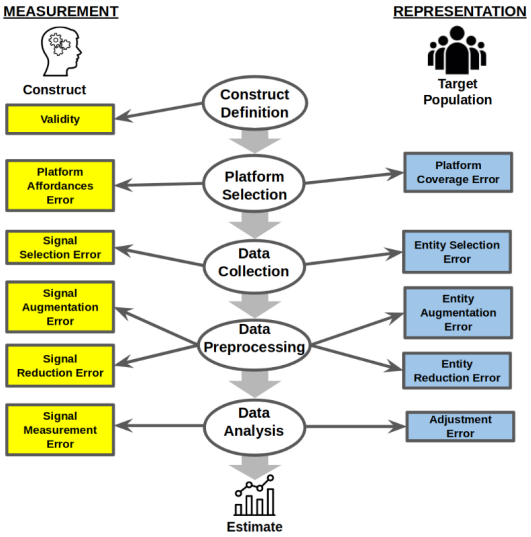


Figure: Potential measurement and representation errors in digital trace data (Sen et. al 2019)

Bias and ethics

Research ethics and privacy

- Further anonymize data? Separate user-ID's from Tweet content and meta-information? Handling Twitter data how we handle survey data?
- Open Science? Share data? Publish user-names? What about celebrities and politicians?

Bias and ethics

Data access uncertainty

- Always have in mind that data access is dependent upon Twitter's willingness to share their data
- Potential legal issues and restrictions
- Facebook's and Instagram's data access was basically shut-down completely within a couple of weeks/months
- Given that, research projects relying on Twitter data are risky

Collecting reddit data

Collecting reddit data

There are several ways to collect reddit data

- R: "**RedditExtractoR**"
- Python: "praw" / "psaw"[2]
- DIY using OAuth protocol + reddit API
  1. Create a reddit account
  2. Login to your reddit account and create an app via `https://reddit.com/prefs/apps`
     - Select "script"
     - Name: anynameyouwant
     - Redirect url: "http://localhost:8080"
  3. Make note of client id and client secret
  4. Access API with your app credentials (may need to request access token)

[2] `https://www.reddit.com/r/pushshift/comments/bcxguf/new_to_pushshift_read_this_faq/`

Collecting reddit data

Collecting reddit data

- R: "**RedditExtractoR**": limited functionality
  - Search for submissions using keywords (in specific subreddits)
  - Extract comments for those submissions
  - Ex post filtering (timestamps)
- Python: "**psaw**": Probably the best solution for most social science research projects
  - Requires only a few lines of code to extract submissions and comments from a pre-scraped database (see `https://pushshift.io/` and `https://files.pushshift.io/`)
  - Search using keywords, subreddits, users, timestamps, ...
  - Almost all content can be accessed
  - Load resulting data in R for cleaning, processing, analyzing
- Python: "**praw**"
  - Python wrapper to access reddit API
  - Limited to 60 items (submission/comment) per minute

Collecting reddit data

Collecting reddit data Reddit API response

- Reddit API sends submissions, comments, and tons of other information in .json format (R and python wrappers parse them automatically)
- Can link comments to submissions
- Submissions may be text, pictures, links, videos, ...
- Each row is a submission or comment

Analyzing reddit data

Similar to Twitter data

- Content Analysis
  - What kind of topics are users talking about?
- Sentiment Analysis
  - What kind of opinions/attitudes/emotions towards objects/concepts are users communicating?
- Compare discussions between groups (subreddits)
  - E.g., do democrats and republicans differ in the ways they discuss Kavanaugh (student project)
  - E.g., differences in moms and dads parenting roles (Ammari et al. 2018)
- Study hard-to-reach populations
  - e.g., LGBTQ, victims of hate crimes, supporters of XYZ

Analyzing reddit data

Downsides

- Users are completely anonymous
- Multiple accounts and bots
- Deleted content cannot be recovered
- Absolutely no background information available (but see `https://snoopsnoo.com/`)

Summary

- Collecting Twitter and reddit data is comparatively easy and cheap
- However, severe representation and measurement issues arise
- Research potential for social science is limited when we're interested in questions that address 'outside-social-media' phenomena
- Twitter data may be (best) treated as auxiliary/cheap proxy information and combined with other data sources
- Reddit data useful for exploratory research, hard-to-reach populations, rare phenomena, comparative research

Resources

- rtweet
  - Documentation: https://rtweet.info/index.html
  - Workshop: https://mkearney.github.io/nicar_tworkshop/#1
- streamR
  - https://github.com/pablobarbera/streamR
- https://www.tidytextmining.com/
- RedditExtractoR
  - https://github.com/ivan-rivera/RedditExtractoR
- PRAW
  - https://praw.readthedocs.io/en/latest/index.html
- PSAW / Pushshift.io
  - https://reddit-api.readthedocs.io/en/latest/#
  - https://pushshift.io/

References

Amaya, A., Bach, R. L., Keusch, F., and Kreuter, F. (2019). New Data Sources in Social Science Research: Things to Know Before Working With Reddit Data. *Social Science Computer Review*, online first. https://doi.org/10.1177/0894439319893305

Ammari, T., Schoenebeck, S., and Romero, D. (2018). Pseudonymous parents: Comparing parenting roles and identities on the Mommit and Daddit subreddits. *Proceedings of ACM 2018 CHI conference on human factors in computing systems*. , April 21–26, ACM.

Choudhury, M., and De, S. (2014). Mental health discourse on reddit: Self-disclosure, social support, and anonymity. *Proceedings of the eighth international AAAI conference on weblogs and social media*.

Conover, M. D., Goncalves, B., Ratkiewicz, J., Flammini A., and Menczer, F. (2011). Predicting the Political Alignment of Twitter Users. *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing* , Boston, MA, 192–199.

En, B., En, M., and Griffiths, D. (2013). Gay stuff and guy stuff: The construction of sexual identities in sidebars on reddit. *Networking Knowledge: Journal of the MeCCSA Postgraduate Network*, 6.

Sen, I., Floeck, F., Weller, K., Weiss, B., and Wagner, C. (2019). A Total Error Framework for Digital Traces of Humans. https://arxiv.org/abs/1907.08228

Sowles, S., McLeary, M., Optican, A., Cahn, E., Krauss, M., Fitzsimmons-Craft, E., ... Cavazos-Rehg, P. (2018). A content analysis of an online pro-eating disorder community on reddit. *Body Image*, 24, 137–144.