Social media data - tips and tricks

A presentation for Hacky Hour

Noel Zeng, Centre for eResearch
noel.zeng@auckland.ac.nz

Links from the 30 July presentation

- “Where to get Twitter data for academic research” https://gwu-libraries.github.io/sfm-ui/posts/2017-09-14-twitter-data
- Documenting the Now Project - a community of US academics using social media data (with a Slack), software tools and a catalogue of existing Tweet datasets. https://www.docnow.io/
- Tools for scraping and visualisation https://github.com/twintproject/twint, https://github.com/drawrowfly/tiktok-scraper and https://github.com/arc298/instagram-scraper
- Language detection: https://github.com/wooorm/franc, https://github.com/Mimino666/langdetect
- Sarcasm detection: http://eprints.whiterose.ac.uk/130763/
- Ethics - Netflix documentary about the Cambridge Analytica scandal ($): https://www.netflix.com/nz/Title/80117542

What we're doing today

- An overview of using social media data for research with focus on Twitter
- Example - collecting and analysing Tweets about climate change
- Questions and answers, community sharing

An overview of social media research

What kind of data can you collect?

- From a variety of sites: social networks, comments sections, forums...
- The post itself - text, images, videos, who it mentions, popularity, replies, geolocation.
- Information about the author - location, followers & followed, biography
- You can collect posts that match a search criteria, hashtag, trends.

Where to get data?

- **Manual retrieval**

- **Existing datasets** - other researchers or institutions have started releasing datasets.
  E.g. a #BlackLivesMatter Tweets dataset from 2017, NLNZ Tweet collections (e.g. Election 2017 candidate Tweets), DocNow Catalog

- **Official API** - sanctioned by the platform, used as a source in many journal articles.
  BUT expensive, limits on what and how much data can be retrieved.
  Tools available to talk with them - e.g. SFM https://sfm.readthedocs.io/en/latest/

- **Scraping** - may be the only option some times, can cheaply gather large amounts of data.
  BUT technically violating Terms of Service, can stop working at any time, may not return all data.
  Also tools available - e.g. https://github.com/twintproject/twint,

What can you do with the data?

- **Manual analysis** - can read through the posts manually.
    - Works for small number of posts. This could be feasible for hundreds or thousands of Tweets, but usually, the volume is much higher.

All Lives Matter, but so Does Race: Black Lives Matter and the Evolving Role of Social Media

Nikita Carney

First Published April 13, 2016 | Research Article | Check for updates

https://doi.org/10.1177/0160597616643868

Article information

Abstract

This article demonstrates the ways in which youth of color played an active role in debates that erupted on Twitter following the tragic deaths of Michael Brown and Eric Garner in 2014. These debates on social media represent a larger struggle over discourse on race and racism across the nation. Drawing from critical theory and race theory, and engaging in the relatively new practice of using Twitter as a source of data for sociological analysis, this article examines Twitter as an emerging public sphere and studies the hashtags "#AllLivesMatter" and "#BlackLivesMatter" as contested signs that represent dominant ideologies. This article consists of a qualitative textual analysis of a selection of Twitter posts from December 3 to 7, 2014, following the nonindictments of officers in the murders of Michael Brown and Eric Garner. The debates on Twitter reveal various strategies that youth of color employed to shape the national discourse about race in the wake of these high-profile tragedies.

Carney 2016 (Original paper)

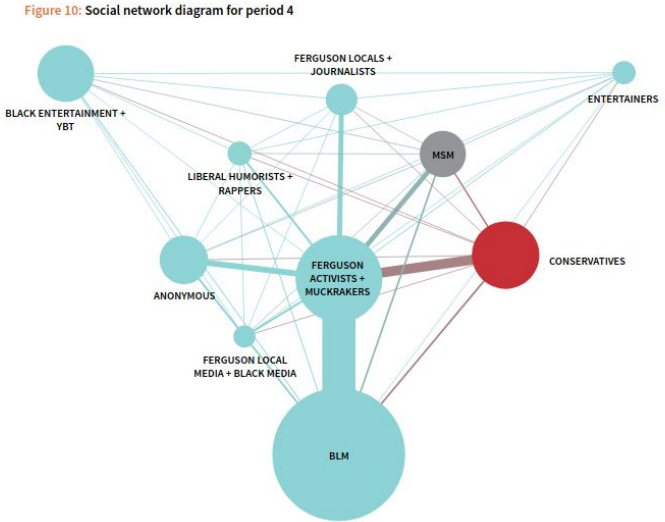
What can you do with the data?

- **Natural language processing (NLP)**
  - Preprocessing: stemming, removing articles.
  - Use algorithms to analyse the text of the posts:
    - What's being talked about: Named-Entity Recognition
    - Classifying what themes are being talked about: topic modelling.
    - What kind of language is being used: sentiment analysis
    - Relative word frequencies

Anderson et al 2018 (original paper)

What can you do with the data?

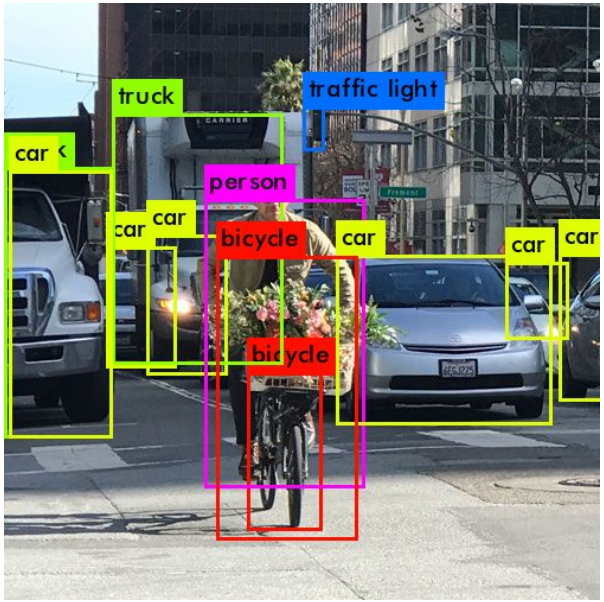
- **Social network analysis**
  - Analyse the social structure underlying a particular issue based on who retweets whom and who follows whom.
  - These networks can be seen as graphs, which can be visualised and analysed.
  - NetworkX in Python, dot files.



Freelon et al. 2016 (Original paper)

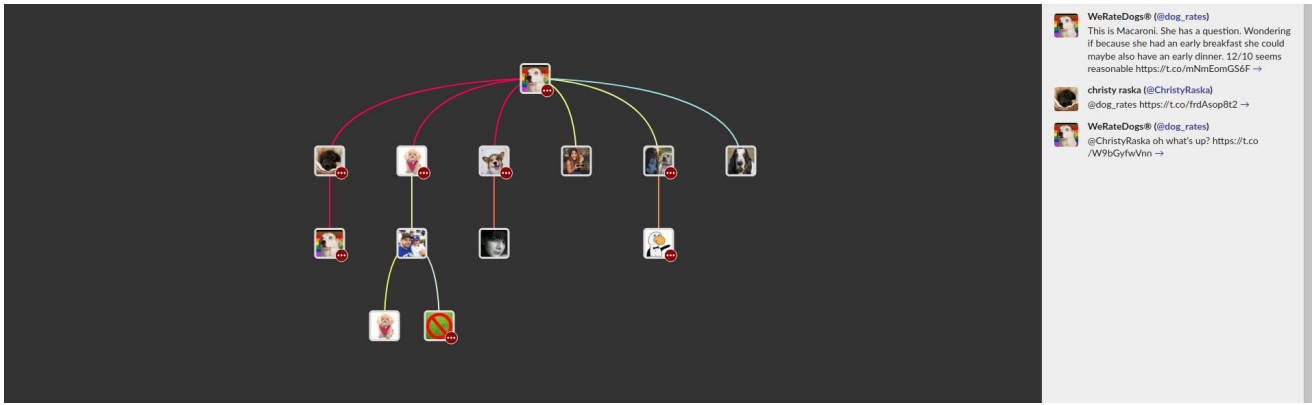
What can you do with the data?

- **Image Processing on media posts**
  - Finding objects and figure out what they are: Object detection and object recognition
  - More Machine Learning heavy. There are pretrained models which recognise some objects.



What can you do with the data?

- Visualisation
- Mapping



Treeverse - scrape and visualise a Twitter post and its replies.

Ethics

- Social media research is a new field, rules around what's ethical are still being developed. E.g. Cambridge Analytica

- Check with your supervisor and Human Ethics on what is acceptable.
  - Ethics team
  - Faculty/LSRI based ethics advisors

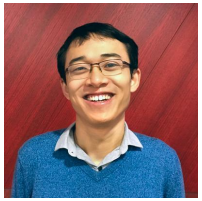
- Preserve choice for Tweeter by only publishing Tweet IDs. (Required by Twitter)

Collecting and analysing Tweets about climate change

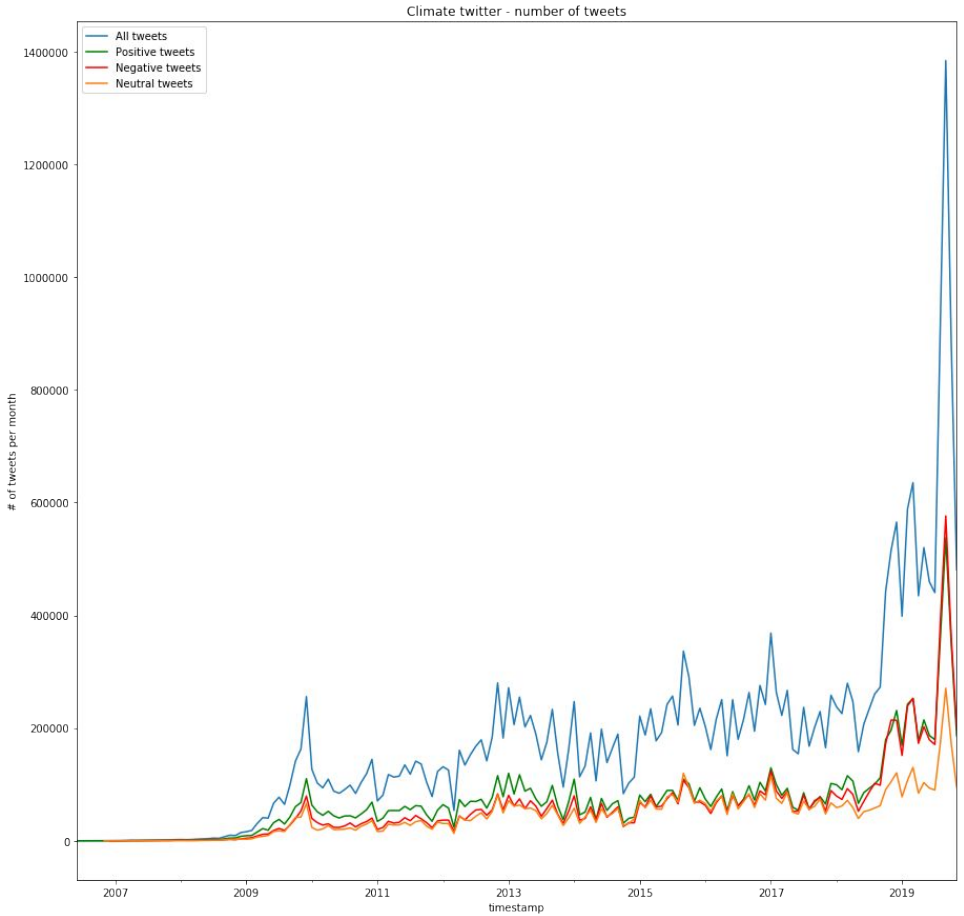
Summary

- Exploratory data gathering about how climate change is being talked about in social media.
- Collected 28.5 million Tweets from 2006 - 2019 that mentioned “climate change” and variants of the word.
- Scripted a tool called twitterscraper to scrape Tweets.
- Analysis through VADER, pandas on University Nectar Cloud, and visualised with Matplotlib.

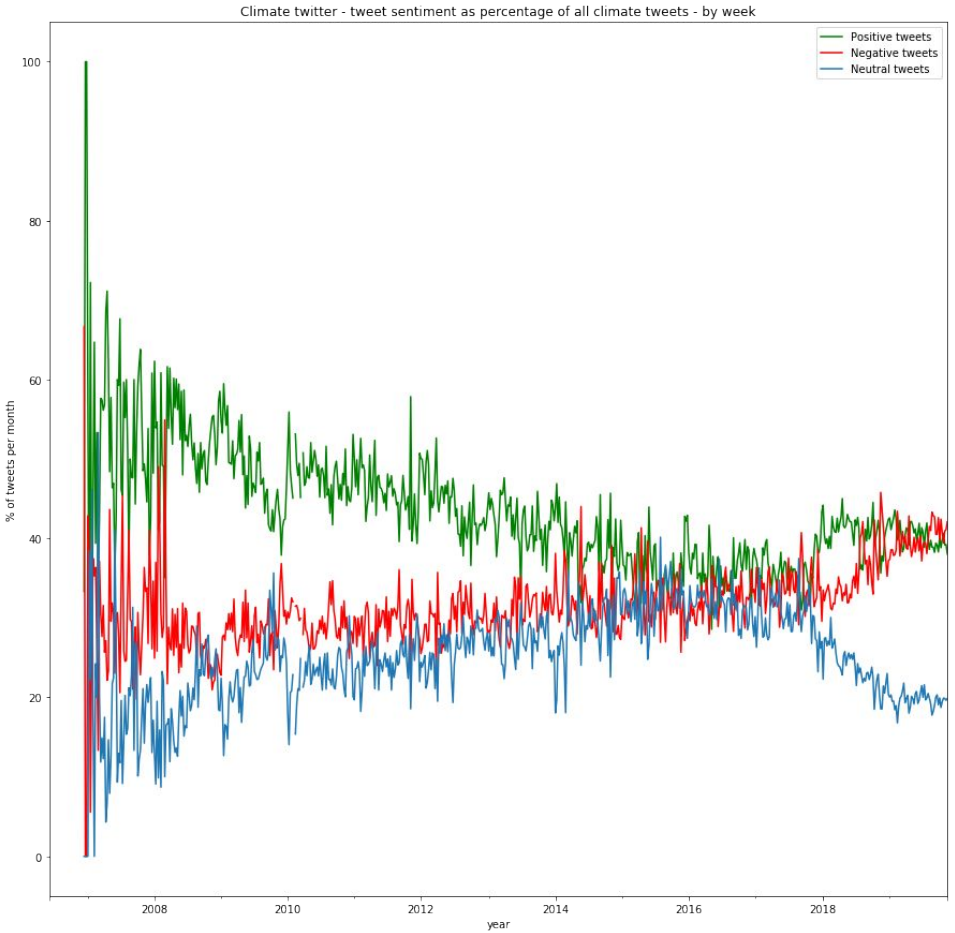
Code and analysis results are available on https://github.com/UoA-eResearch/twitter_analysis



Results

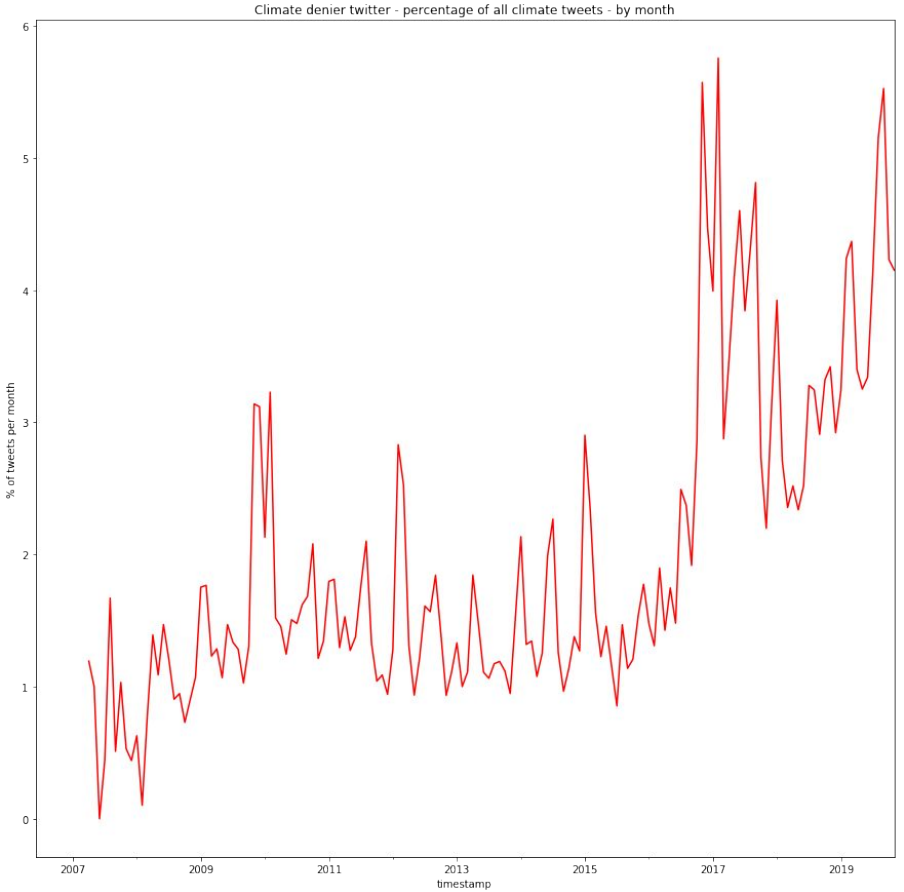

Climate twitter - number of tweets
All tweets
Positive tweets
Negative tweets
Neutral tweets
# of tweets per month
1400000
1200000
1000000
800000
600000
400000
200000
0
2007
2009
2011
2013
2015
2017
2019
timestamp

Trend of sentiment



Climate denial trends

- Within our dataset, we searched for Tweets containing “hoax”, “isn’t real”.



Questions and discussions

Some instigating questions...

- Are you working on any projects that involve social media data?
- Which social network sites are you looking into? Which are you interested in?
- What tools have you come across that's been helpful?
- What's the one thing you wish you knew when you started with social media data research?