Collecting and Analyzing Social Media Data

Montreal Methods Workshop

Alexandra Siegel

March 4, 2021

Assistant Professor, University of Colorado Boulder
Non-Resident Fellow, Brookings
Faculty Affiliate, Stanford Immigration Policy Lab
Faculty Affiliate, NYU Center for Social Media and Politics

Why use social media data to study politics?

Why use social media data to study politics?

- Real-time, scalable, measures of political behavior

- Real-time, scalable, measures of political behavior
- Elites, everyday citizens, extremists, media etc. on same platform

Why use social media data to study politics?

- Real-time, scalable, measures of political behavior
- Elites, everyday citizens, extremists, media etc. on same platform
- Access to politically sensitive content and hard to reach populations


You
Tube

What types of data can we collect?

- Twitter, Facebook, Youtube, Instagram, Reddit, Tiktok

What types of data can we collect?

- Twitter, Facebook, Youtube, Instagram, Reddit, Tiktok
- APIs and Terms of Service

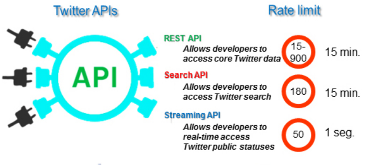
What types of data can we collect?

- Twitter, Facebook, Youtube, Instagram, Reddit, Tiktok
- APIs and Terms of Service
- Available Metadata

What types of data can we collect?

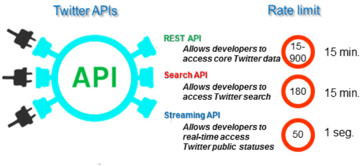
- Twitter, Facebook, Youtube, Instagram, Reddit, Tiktok
- APIs and Terms of Service
- Available Metadata
- Static vs. Ongoing collections

Twitter: Social Scientists' Favorite Platform




```
{ "created_at": "Wed Nov 07 04:16:18 +0000 2012",
"id": 266031293945503744,
"text": "Four more years. http://t.co/bAJE6Vom",
"source": "web",
"user": {
"id": 813286,
"name": "Barack Obama",
"screen_name": "BarackObama",
"location": "Washington, DC",
"description": "This account is run by Organizing
Tweets from the President are signed -bo.",
"url": "http://t.co/8aJ56Jcemr",
"protected": false,
"followers_count": 54873124,
"friends_count": 654580,
"listed_count": 202495,
"created_at": "Mon Mar 05 22:08:25 +0000 2007",
"time_zone": "Eastern Time (US & Canada)",
"statuses_count": 10687,
"lang": "en" },
"coordinates": null,
"retweet_count": 756411,
"favorite_count": 288867,
"lang": "en"
}
```

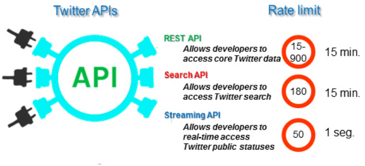
Twitter: Social Scientists' Favorite Platform




```
{ "created_at": "Wed Nov 07 04:16:18 +0000 2012",
"id": 266031293945503744,
"text": "Four more years. http://t.co/bAJE6Vom",
"source": "web",
"user": {
"id": 813286,
"name": "Barack Obama",
"screen_name": "BarackObama",
"location": "Washington, DC",
"description": "This account is run by Organizing
Tweets from the President are signed -bo.",
"url": "http://t.co/8aJ56Jcemr",
"protected": false,
"followers_count": 54873124,
"friends_count": 654580,
"listed_count": 202495,
"created_at": "Mon Mar 05 22:08:25 +0000 2007",
"time_zone": "Eastern Time (US & Canada)",
"statuses_count": 10687,
"lang": "en" },
"coordinates": null,
"retweet_count": 756411,
"favorite_count": 288867,
"lang": "en"
}
```


- Collecting Data through traditional APIs ▸ Link

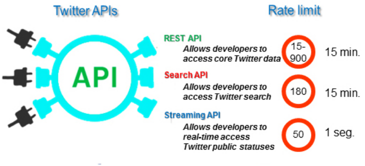
Twitter: Social Scientists' Favorite Platform




```
{ "created_at": "Wed Nov 07 04:16:18 +0000 2012",
"id": 266031293945503744,
"text": "Four more years. http://t.co/bAJE6Vom",
"source": "web",
"user": {
"id": 813286,
"name": "Barack Obama",
"screen_name": "BarackObama",
"location": "Washington, DC",
"description": "This account is run by Organizing
Tweets from the President are signed -bo.",
"url": "http://t.co/8aJ56Jcemr",
"protected": false,
"followers_count": 54873124,
"friends_count": 654580,
"listed_count": 202495,
"created_at": "Mon Mar 05 22:08:25 +0000 2007",
"time_zone": "Eastern Time (US & Canada)",
"statuses_count": 10687,
"lang": "en" },
"coordinates": null,
"retweet_count": 756411,
"favorite_count": 288867,
"lang": "en"
}
```


- Collecting Data through traditional APIs ▸ Link
- Academic Research API ▸ Link

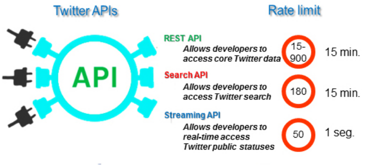
Twitter: Social Scientists' Favorite Platform




```
{ "created_at": "Wed Nov 07 04:16:18 +0000 2012",
"id": 266031293945503744,
"text": "Four more years. http://t.co/bAJE6Vom",
"source": "web",
"user": {
"id": 813286,
"name": "Barack Obama",
"screen_name": "BarackObama",
"location": "Washington, DC",
"description": "This account is run by Organizing
Tweets from the President are signed -bo.",
"url": "http://t.co/8aJ56Jcemr",
"protected": false,
"followers_count": 54873124,
"friends_count": 654580,
"listed_count": 202495,
"created_at": "Mon Mar 05 22:08:25 +0000 2007",
"time_zone": "Eastern Time (US & Canada)",
"statuses_count": 10687,
"lang": "en" },
"coordinates": null,
"retweet_count": 756411,
"favorite_count": 288867,
"lang": "en"
}
```


- Collecting Data through traditional APIs ▸ Link
- Academic Research API ▸ Link
- Accessing Historical Data with Gnip ▸ Link

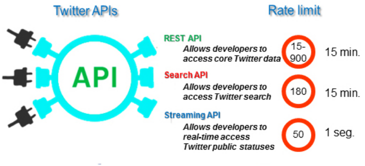
Twitter: Social Scientists' Favorite Platform




```
{ "created_at": "Wed Nov 07 04:16:18 +0000 2012",
"id": 266031293945503744,
"text": "Four more years. http://t.co/bAJE6Vom",
"source": "web",
"user": {
"id": 813286,
"name": "Barack Obama",
"screen_name": "BarackObama",
"location": "Washington, DC",
"description": "This account is run by Organizing
Tweets from the President are signed -bo.",
"url": "http://t.co/8aJ56Jcemr",
"protected": false,
"followers_count": 54873124,
"friends_count": 654580,
"listed_count": 202495,
"created_at": "Mon Mar 05 22:08:25 +0000 2007",
"time_zone": "Eastern Time (US & Canada)",
"statuses_count": 10687,
"lang": "en" },
"coordinates": null,
"retweet_count": 756411,
"favorite_count": 288867,
"lang": "en"
}
```


- Collecting Data through traditional APIs ▸ Link
- Academic Research API ▸ Link
- Accessing Historical Data with Gnip ▸ Link
- Rehydrating Tweets ▸ Link

Twitter: Social Scientists' Favorite Platform




```
{ "created_at": "Wed Nov 07 04:16:18 +0000 2012",
"id": 266031293945503744,
"text": "Four more years. http://t.co/bAJE6Vom",
"source": "web",
"user": {
"id": 813286,
"name": "Barack Obama",
"screen_name": "BarackObama",
"location": "Washington, DC",
"description": "This account is run by Organizing
Tweets from the President are signed -bo.",
"url": "http://t.co/8aJ56Jcemr",
"protected": false,
"followers_count": 54873124,
"friends_count": 654580,
"listed_count": 202495,
"created_at": "Mon Mar 05 22:08:25 +0000 2007",
"time_zone": "Eastern Time (US & Canada)",
"statuses_count": 10687,
"lang": "en" },
"coordinates": null,
"retweet_count": 756411,
"favorite_count": 288867,
"lang": "en"
}
```


- Collecting Data through traditional APIs ▸ Link
- Academic Research API ▸ Link
- Accessing Historical Data with Gnip ▸ Link
- Rehydrating Tweets ▸ Link
- Scraping Tweets ▸ Link

Facebook: Challenges and Opportunities for Research

Facebook: Challenges and Opportunities for Research

- Collecting Public Page Data with Crowdtangle Link

Facebook: Challenges and Opportunities for Research

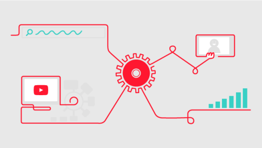
- Collecting Public Page Data with Crowdtangle ▸ Link
- Accessing Data through Social Science One ▸ Link

Facebook: Challenges and Opportunities for Research

- Collecting Public Page Data with Crowdtangle Link
- Accessing Data through Social Science One Link
- Using Facebook Ads for Academic Research Link

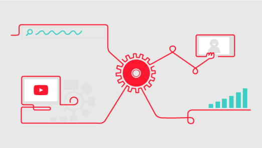
```
type     by       post_id post_link       post_message    picture full_picture    link
link_domain       post_published  post_published_unix     post_published_sql
likes_count_fb    comments_count_fb       reactions_count_fb      shares_count_fb
engagement_fb     comments_retrieved      comments_base   comments_replies
comment_likes_count       rea_LOVE        rea_WOW rea_HAHA        rea_SAD rea_ANGRY
rea_THANKFUL
```

Youtube: An Underutilized Resource




```
{
  "kind": "youtube#caption",
  "etag": etag,
  "id": string,
  "snippet": {
    "videoId": string,
    "lastUpdated": datetime,
    "trackKind": string,
    "language": string,
    "name": string,
    "audioTrackType": string,
    "isCC": boolean,
    "isLarge": boolean,
    "isEasyReader": boolean,
    "isDraft": boolean,
    "isAutoSynced": boolean,
    "status": string,
    "failureReason": string
  }
}
```

Youtube: An Underutilized Resource

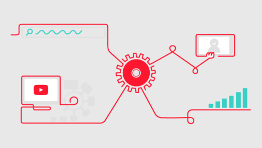



```
{
  "kind": "youtube#caption",
  "etag": etag,
  "id": string,
  "snippet": {
    "videoId": string,
    "lastUpdated": datetime,
    "trackKind": string,
    "language": string,
    "name": string,
    "audioTrackType": string,
    "isCC": boolean,
    "isLarge": boolean,
    "isEasyReader": boolean,
    "isDraft": boolean,
    "isAutoSynced": boolean,
    "status": string,
    "failureReason": string
  }
}
```


- Incredibly generous API

  Link

Youtube: An Underutilized Resource



```
{
  "kind": "youtube#caption",
  "etag": etag,
  "id": string,
  "snippet": {
    "videoId": string,
    "lastUpdated": datetime,
    "trackKind": string,
    "language": string,
    "name": string,
    "audioTrackType": string,
    "isCC": boolean,
    "isLarge": boolean,
    "isEasyReader": boolean,
    "isDraft": boolean,
    "isAutoSynced": boolean,
    "status": string,
    "failureReason": string
  }
}
```

- Incredibly generous API
  ▸ Link
- Channel, Video, Metadata (including comments) & a computer generated TRANSCRIPT (!) in any language

Instagram: There's politics here too!

```
Default: {urlname}
Options:
{username}: Scraped user
{shortcode}: Post shortcode (profile_pic and story are empty)
{urlname}: Original file name from url.
{mediatype}: The type of media being downloaded.
{datetime}: Date and time of upload. (Format: 20180101 01h01m01s)
{date}: Date of upload. (Format: 20180101)
{year}: Year of upload. (Format: 2018)
{month}: Month of upload. (Format: 01-12)
{day}: Day of upload. (Format: 01-31)
{h}: Hour of upload. (Format: 00-23h)
{m}: Minute of upload. (Format: 00-59m)
{s}: Second of upload. (Format: 00-59s)
```

Instagram: There's politics here too!

```
Default: {urlname}
Options:
{username}: Scraped user
{shortcode}: Post shortcode (profile_pic and story are empty)
{urlname}: Original file name from url.
{mediatype}: The type of media being downloaded.
{datetime}: Date and time of upload. (Format: 20180101 01h01m01s)
{date}: Date of upload. (Format: 20180101)
{year}: Year of upload. (Format: 2018)
{month}: Month of upload. (Format: 01-12)
{day}: Day of upload. (Format: 01-31)
{h}: Hour of upload. (Format: 00-23h)
{m}: Minute of upload. (Format: 00-59m)
{s}: Second of upload. (Format: 00-59s)
```

- API is increasingly restricted ▸ Link

Instagram: There's politics here too!

```
Default: {urlname}
Options:
{username}: Scraped user
{shortcode}: Post shortcode (profile_pic and story are empty)
{urlname}: Original file name from url.
{mediatype}: The type of media being downloaded.
{datetime}: Date and time of upload. (Format: 20180101 01h01m01s)
{date}: Date of upload. (Format: 20180101)
{year}: Year of upload. (Format: 2018)
{month}: Month of upload. (Format: 01-12)
{day}: Day of upload. (Format: 01-31)
{h}: Hour of upload. (Format: 00-23h)
{m}: Minute of upload. (Format: 00-59m)
{s}: Second of upload. (Format: 00-59s)
```

- API is increasingly restricted ▸ Link
- BUT we can data from public accounts ▸ Link

Reddit: Naturally Annotated Political Texts

Reddit: Naturally Annotated Political Texts

- Easiest to collect with Google Big Query ▸ Link

Reddit: Naturally Annotated Political Texts

- Easiest to collect with Google Big Query ▸ Link
- Can query by subreddit, time, keywords etc.

TikTok:

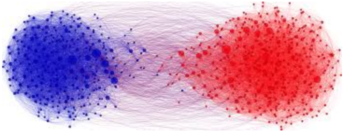
TikTok:

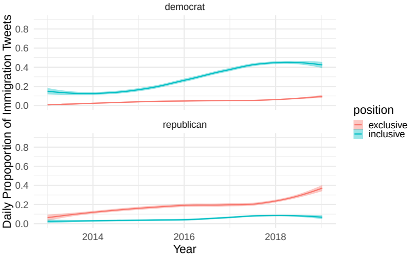
- TikTok API ▸ Link

TikTok:

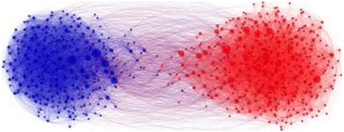
- TikTok API ▸ Link
- Query by user, hashtags, trending etc.

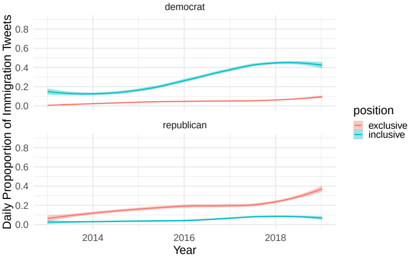
What can we do with social media data?





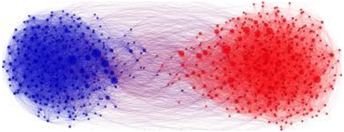
What can we do with social media data?

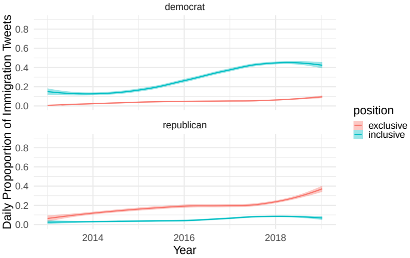




- Text/Image/Video as data

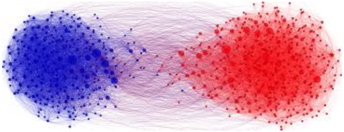
What can we do with social media data?

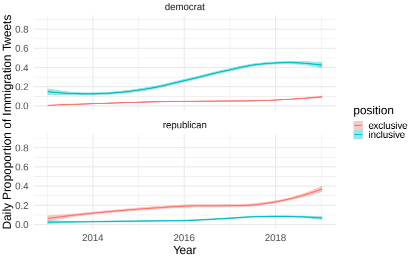




- Text/Image/Video as data
- Network analysis

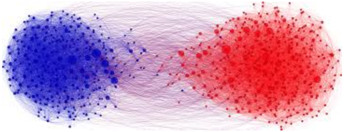
What can we do with social media data?

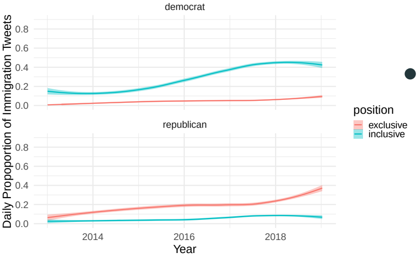




- Text/Image/Video as data
- Network analysis
- Spatial analysis

What can we do with social media data?





- Text/Image/Video as data
- Network analysis
- Spatial analysis
- Time series analysis

Supervised Approaches

Popular Approaches to Text Analysis for Social Media Data

Supervised Approaches

- Dictionary-based methods

Popular Approaches to Text Analysis for Social Media Data

Supervised Approaches

- Dictionary-based methods
- Training classifiers on human coded or naturally annotated data

Popular Approaches to Text Analysis for Social Media Data

Supervised Approaches

- Dictionary-based methods
- Training classifiers on human coded or naturally annotated data
- Semantic similarity measures (eg. using fasttext, but also cosine similarity etc.)

Unsupervised Approaches

Popular Approaches to Text Analysis for Social Media Data

Supervised Approaches

- Dictionary-based methods
- Training classifiers on human coded or naturally annotated data
- Semantic similarity measures (eg. using fasttext, but also cosine similarity etc.)

Unsupervised Approaches

- LDA & Structural topic models (but watch out for short texts!)

Popular Approaches to Text Analysis for Social Media Data

Supervised Approaches

- Dictionary-based methods
- Training classifiers on human coded or naturally annotated data
- Semantic similarity measures (eg. using fasttext, but also cosine similarity etc.)

Unsupervised Approaches

- LDA & Structural topic models (but watch out for short texts!)
- Neural networks (including word2vec)

Illustration: Was there a “Trump effect” on Twitter?

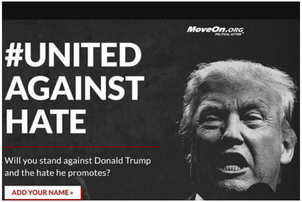
POLITICS SPECIAL REPORTS | Mon Nov 7, 2016 | 10:46pm EST

Hate speech seeps into U.S. mainstream amid bitter campaign

NEWS DESK

HATE ON THE RISE AFTER TRUMP’S ELECTION

By Alexis Okeowo November 17, 2016



DEMOCRACY & GOVERNMENT

Donald Trump and the Escalation of Hate

A number of civil-rights organizations have spoken out about the rise of hate speech and violent threats by groups and individuals who support the presumptive Republican presidential nominee.

BY KARIN KAMP | JUNE 15, 2016

'Massive rise' in hate speech on Twitter during presidential election

Jessica Guynn , USA TODAY Published 5:00 p.m. ET Oct. 21, 2016 | Updated 7:00 p.m. ET Oct. 23, 2016

How do we measure online hate speech?

HATEBASE

reddit

How do we measure online hate speech?

- On Twitter

How do we measure online hate speech?

- On Twitter
- Machine-Learning-Augmented-Dictionary Method

How do we measure online hate speech?

- On Twitter
- Machine-Learning-Augmented-Dictionary Method
- Leveraging Data from Hateful Sub-Reddits

How do we measure online hate speech?

- On Twitter
- Machine-Learning-Augmented-Dictionary Method
- Leveraging Data from Hateful Sub-Reddits
- Political Datasets & Random Sample of American Twitter Users (June 2015 - June 2017)

Dictionary-based Hate Speech Detection on Twitter

1. Create dictionaries of slurs and terms from existing dictionaries of hate speech and white nationalist rhetoric (Hatebase, Racial Slur Database, ADL) $\rightarrow$ (4,477 terms, including variations)

Dictionary-based Hate Speech Detection on Twitter

1. Create dictionaries of slurs and terms from existing dictionaries of hate speech and white nationalist rhetoric (Hatebase, Racial Slur Database, ADL) $\rightarrow$ (4,477 terms, including variations)
2. Remove terms that are primarily not used as hate speech in a random sample of our Political Twitter dataset. $\rightarrow$ (e.g. pizza, newspaper, soak, taco)$\rightarrow$ (538 terms)

Dictionary-based Hate Speech Detection on Twitter

1. Create dictionaries of slurs and terms from existing dictionaries of hate speech and white nationalist rhetoric (Hatebase, Racial Slur Database, ADL) $\rightarrow$ (4,477 terms, including variations)
2. Remove terms that are primarily not used as hate speech in a random sample of our Political Twitter dataset. $\rightarrow$ (e.g. pizza, newspaper, soak, taco)$\rightarrow$ (538 terms)
3. Add common Twitter specific terms using word2vec dictionary $\rightarrow$ (+ 500 terms)

Dictionary-based Hate Speech Detection on Twitter

1. Create dictionaries of slurs and terms from existing dictionaries of hate speech and white nationalist rhetoric (Hatebase, Racial Slur Database, ADL) → (4,477 terms, including variations)
2. Remove terms that are primarily not used as hate speech in a random sample of our Political Twitter dataset. → (e.g. pizza, newspaper, soak, taco)→ (538 terms)
3. Add common Twitter specific terms using word2vec dictionary → (+ 500 terms)

Problems with Dictionary Methods:

Dictionary-based Hate Speech Detection on Twitter

1. Create dictionaries of slurs and terms from existing dictionaries of hate speech and white nationalist rhetoric (Hatebase, Racial Slur Database, ADL) → (4,477 terms, including variations)
2. Remove terms that are primarily not used as hate speech in a random sample of our Political Twitter dataset. → (e.g. pizza, newspaper, soak, taco)→ (538 terms)
3. Add common Twitter specific terms using word2vec dictionary → (+ 500 terms)

Problems with Dictionary Methods:

- Term can be part of a Twitter handle: @angrybitch

Dictionary-based Hate Speech Detection on Twitter

1. Create dictionaries of slurs and terms from existing dictionaries of hate speech and white nationalist rhetoric (Hatebase, Racial Slur Database, ADL) $\rightarrow$ (4,477 terms, including variations)
2. Remove terms that are primarily not used as hate speech in a random sample of our Political Twitter dataset. $\rightarrow$ (e.g. pizza, newspaper, soak, taco)$\rightarrow$ (538 terms)
3. Add common Twitter specific terms using word2vec dictionary $\rightarrow$ (+ 500 terms)

   Problems with Dictionary Methods:
   - Term can be part of a Twitter handle: @angrybitch
   - Dictionary terms can be parts of other words: spicy

Dictionary-based Hate Speech Detection on Twitter

1. Create dictionaries of slurs and terms from existing dictionaries of hate speech and white nationalist rhetoric (Hatebase, Racial Slur Database, ADL) → (4,477 terms, including variations)
2. Remove terms that are primarily not used as hate speech in a random sample of our Political Twitter dataset. → (e.g. pizza, newspaper, soak, taco)→ (538 terms)
3. Add common Twitter specific terms using word2vec dictionary → (+ 500 terms)

Problems with Dictionary Methods:

- Term can be part of a Twitter handle: @angrybitch
- Dictionary terms can be parts of other words: spicy
- Dictionary terms can be homonyms: "a chink in his armor"

Dictionary-based Hate Speech Detection on Twitter

1. Create dictionaries of slurs and terms from existing dictionaries of hate speech and white nationalist rhetoric (Hatebase, Racial Slur Database, ADL) → (4,477 terms, including variations)
2. Remove terms that are primarily not used as hate speech in a random sample of our Political Twitter dataset. → (e.g. pizza, newspaper, soak, taco)→ (538 terms)
3. Add common Twitter specific terms using word2vec dictionary → (+ 500 terms)

Problems with Dictionary Methods:

- Term can be part of a Twitter handle: @angrybitch
- Dictionary terms can be parts of other words: spicy
- Dictionary terms can be homonyms: "a chink in his armor"

- Examples of Anti-Hate Speech that include dictionary terms:

Dictionary-based Hate Speech Detection on Twitter

1. Create dictionaries of slurs and terms from existing dictionaries of hate speech and white nationalist rhetoric (Hatebase, Racial Slur Database, ADL) → (4,477 terms, including variations)
2. Remove terms that are primarily not used as hate speech in a random sample of our Political Twitter dataset. → (e.g. pizza, newspaper, soak, taco)→ (538 terms)
3. Add common Twitter specific terms using word2vec dictionary → (+ 500 terms)

Problems with Dictionary Methods:

- Term can be part of a Twitter handle: @angrybitch
- Dictionary terms can be parts of other words: spicy
- Dictionary terms can be homonyms: "a chink in his armor"
- Examples of Anti-Hate Speech that include dictionary terms:
  - Already been flicked off and called a wetback and it's only been 3 days... thanks Donald trump

Dictionary-based Hate Speech Detection on Twitter

1. Create dictionaries of slurs and terms from existing dictionaries of hate speech and white nationalist rhetoric (Hatebase, Racial Slur Database, ADL) → (4,477 terms, including variations)
2. Remove terms that are primarily not used as hate speech in a random sample of our Political Twitter dataset. → (e.g. pizza, newspaper, soak, taco)→ (538 terms)
3. Add common Twitter specific terms using word2vec dictionary → (+ 500 terms)

Problems with Dictionary Methods:

- Term can be part of a Twitter handle: @angrybitch
- Dictionary terms can be parts of other words: spicy
- Dictionary terms can be homonyms: "a chink in his armor"
- Examples of Anti-Hate Speech that include dictionary terms:
  - Already been flicked off and called a wetback and it's only been 3 days... thanks Donald trump
  - RT @ShaunKing: This just happened in Indiana. "F*** you n**** bitch. Trump is going to deport you back to Africa." Day 1 of Donald

Supervised Classification (Dictionary-based Method):

- Trained undergraduates and crowd-sourced coders on Crowdflower coded a random sample of 25,000 tweets (each tweet coded by 3 people) containing hate speech OR white nationalist rhetoric terms identified using our dictionary method.

Supervised Classification (Dictionary-based Method):

- Trained undergraduates and crowd-sourced coders on Crowdflower coded a random sample of 25,000 tweets (each tweet coded by 3 people) containing hate speech OR white nationalist rhetoric terms identified using our dictionary method.
    - Does this tweet contain hate speech? (yes or no)
    - Does this tweet contain white nationalist rhetoric? (yes or no)
    - Instructions contained detailed definitions and examples.
    - Test questions were used to weed out ineffective coders.

Supervised Classification (Dictionary-based Method):

- According to human coders, fewer than half of the tweets identified *by the dictionary method* in our random sample contained hate speech or white nationalist language.

Supervised Classification (Dictionary-based Method):

- According to human coders, fewer than half of the tweets identified *by the dictionary method* in our random sample contained hate speech or white nationalist language.
- Trained two Naive Bayes classifiers (a hate speech and a white nationalist rhetoric classifier).

Supervised Classification (Dictionary-based Method):

- According to human coders, fewer than half of the tweets identified *by the dictionary method* in our random sample contained hate speech or white nationalist language.
- Trained two Naive Bayes classifiers (a hate speech and a white nationalist rhetoric classifier).
- We measure the popularity of hate speech and white nationalist rhetoric (WNR) as:
  - The daily proportion of tweets containing hate speech or WNR in each of our datasets.
  - The daily proportion of unique users tweeting hate speech or WNR in each of our datasets.

Method II: Bag of Communities Approach

- Concern: are we missing other kinds of hate speech?

Method II: Bag of Communities Approach

- Concern: are we missing other kinds of hate speech?
- Idea: Find a place with known hate speech, then compare daily tweets with that speech

Method II: Bag of Communities Approach

- Concern: are we missing other kinds of hate speech?
- Idea: Find a place with known hate speech, then compare daily tweets with that speech
- Concept: Measure the average predicted probability that tweets are classified as belonging to a corpus of real-world hate speech.

Method II: Bag of Communities Approach

- 1) Download Reddit comments, remove comments with negative scores, and preprocess data.

Method II: Bag of Communities Approach

- 1) Download Reddit comments, remove comments with negative scores, and preprocess data.
- 2) Train supervised classifier (Fasttext) to predict which subreddit each comment was posted in.

Method II: Bag of Communities Approach

- 1) Download Reddit comments, remove comments with negative scores, and preprocess data.
- 2) Train supervised classifier (Fasttext) to predict which subreddit each comment was posted in.
- 3) Use subreddit embeddings from Step 2 to categorize subreddits into groups.

Method II: Bag of Communities Approach

- 1) Download Reddit comments, remove comments with negative scores, and preprocess data.
- 2) Train supervised classifier (Fasttext) to predict which subreddit each comment was posted in.
- 3) Use subreddit embeddings from Step 2 to categorize subreddits into groups.
- 4) Train supervised classifier (Fasttext) to predict which group of subreddits each comment was posted in.

Method II: Bag of Communities Approach

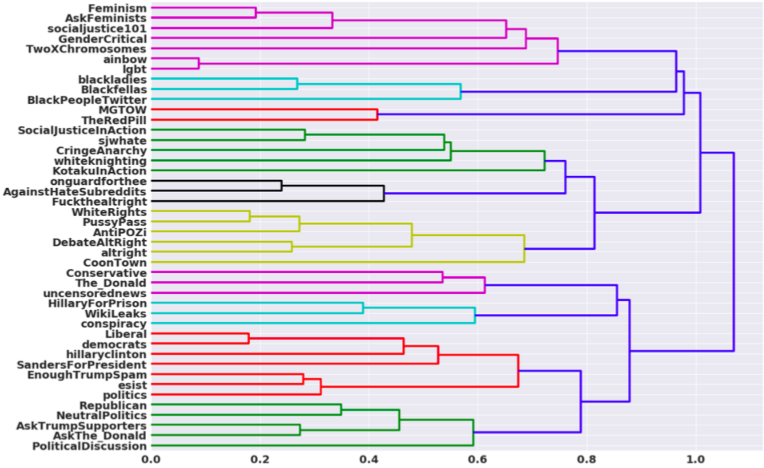
- 1) Download Reddit comments, remove comments with negative scores, and preprocess data.
- 2) Train supervised classifier (Fasttext) to predict which subreddit each comment was posted in.
- 3) Use subreddit embeddings from Step 2 to categorize subreddits into groups.
- 4) Train supervised classifier (Fasttext) to predict which group of subreddits each comment was posted in.
- 5) Apply trained classifier from Step 4 on Twitter data.

Method II: Bag of Communities Approach

- 1) Download Reddit comments, remove comments with negative scores, and preprocess data.
- 2) Train supervised classifier (Fasttext) to predict which subreddit each comment was posted in.
- 3) Use subreddit embeddings from Step 2 to categorize subreddits into groups.
- 4) Train supervised classifier (Fasttext) to predict which group of subreddits each comment was posted in.
- 5) Apply trained classifier from Step 4 on Twitter data.
- 6) Calculate daily average predicted probability that tweets are classified as belonging to a group of alt-right subreddits.

Validation of Method II: Hierarchical Clustering

Figure 1: Validity Check: Hierarchical Clustering of Subreddits



Validation of Method II: Classifying Twitter Accounts

Accounts classified as **Sport**:

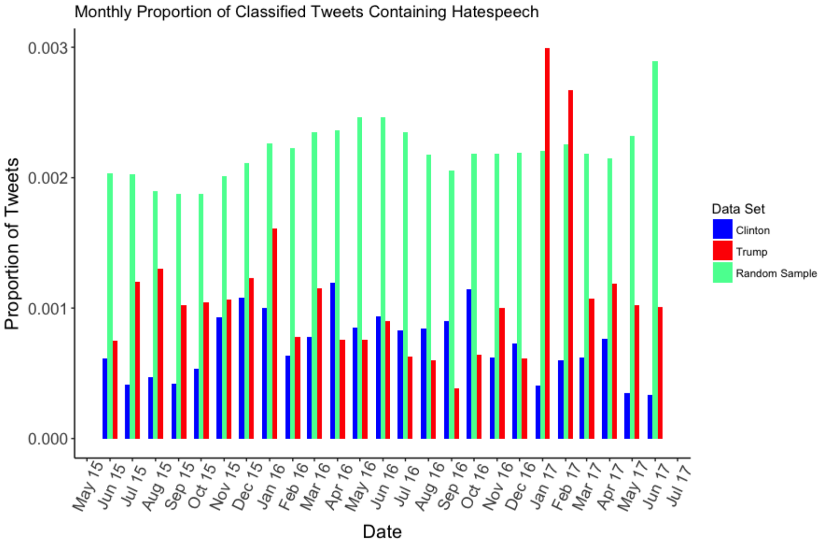
Validation of Method II: Classifying Twitter Accounts

Examples of accounts classified as **Anti-Trump**:

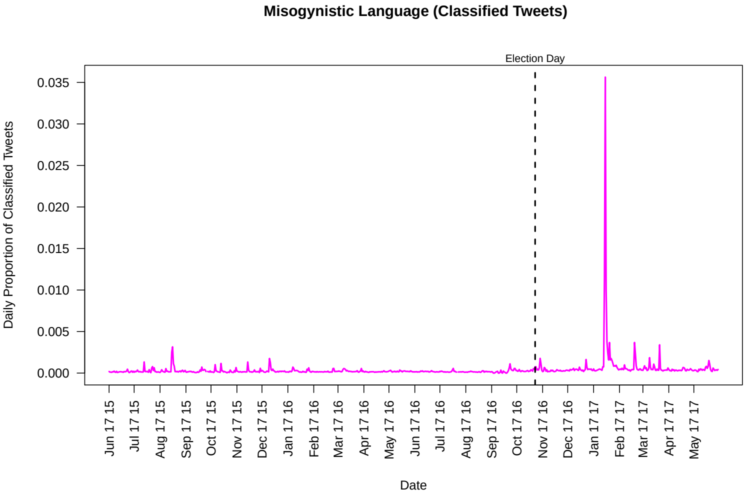
Validation of Method II: Classifying Twitter Accounts

Accounts classified as **Alt-right**:

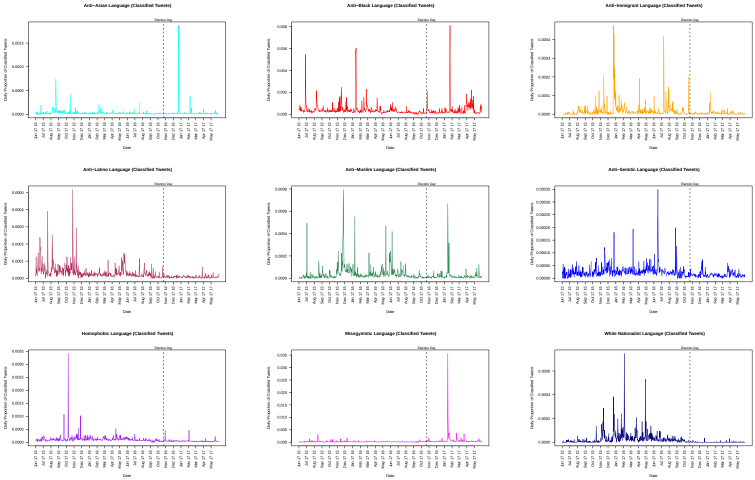
What can we learn from Twitter data?



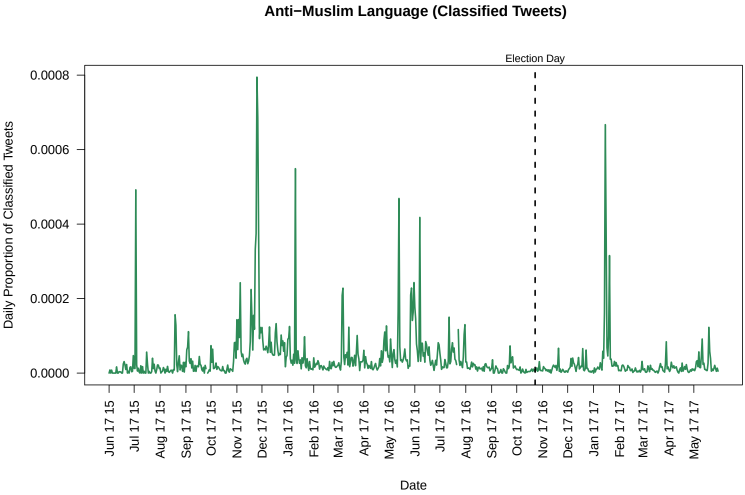
What can we learn from Twitter data?



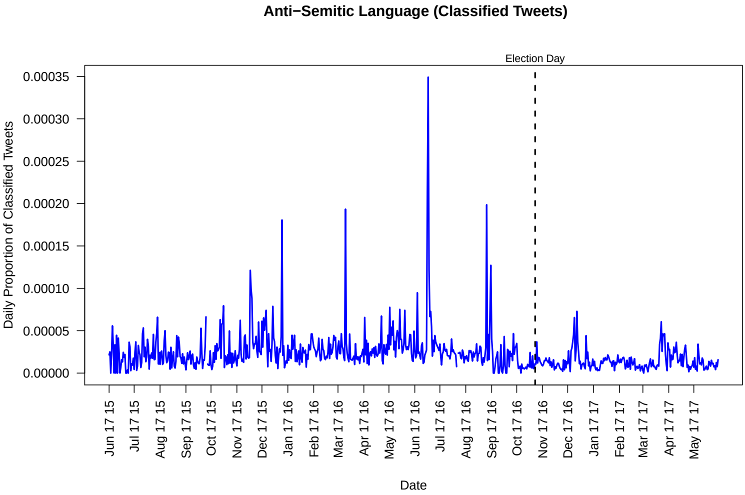
What can we learn from Twitter data?



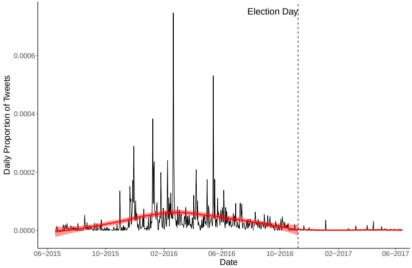
What can we learn from Twitter data?

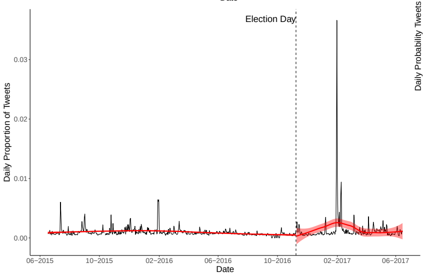


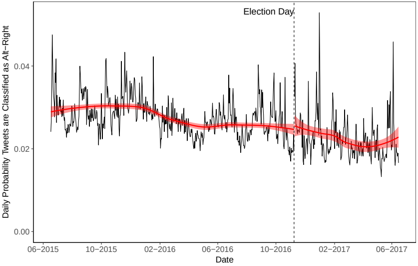
What can we learn from Twitter data?



Dictionary vs. Subreddit Analysis







So what does this tell us about analyzing social media data?

So what does this tell us about analyzing social media data?

- Social media data...

So what does this tell us about analyzing social media data?

- Social media data...
  - are really big!

So what does this tell us about analyzing social media data?

- Social media data...
  - are really big!
  - are optimized for search

So what does this tell us about analyzing social media data?

- Social media data...
  - are really big!
  - are optimized for search
- Need for more systematic approaches to measuring online behavior

So what does this tell us about analyzing social media data?

- Social media data...
  - are really big!
  - are optimized for search
- Need for more systematic approaches to measuring online behavior
- Multiple methods & data sources & iterative validation increase our confidence that we're measuring what we think we are

So what does this tell us about analyzing social media data?

- Social media data...
  - are really big!
  - are optimized for search
- Need for more systematic approaches to measuring online behavior
- Multiple methods & data sources & iterative validation increase our confidence that we're measuring what we think we are
- Even better, combine offline and online measures!

Some Takeaways

- Social media data opens up new measurement opportunities for social scientists.

Some Takeaways

- Social media data opens up new measurement opportunities for social scientists.
- But key challenges remain:

Some Takeaways

- Social media data opens up new measurement opportunities for social scientists.
- But key challenges remain:
    - Representativeness

Some Takeaways

- Social media data opens up new measurement opportunities for social scientists.
- But key challenges remain:
    - Representativeness
    - Reproducability

Some Takeaways

- Social media data opens up new measurement opportunities for social scientists.
- But key challenges remain:
    - Representativeness
    - Reproducability
    - Temporal Validity

Some Takeaways

- Social media data opens up new measurement opportunities for social scientists.
- But key challenges remain:
    - Representativeness
    - Reproducability
    - Temporal Validity
    - Researchers are at the mercy of the platforms

Some Takeaways

- Social media data opens up new measurement opportunities for social scientists.
- But key challenges remain:
    - Representativeness
    - Reproducability
    - Temporal Validity
    - Researchers are at the mercy of the platforms
- Like any data source there are pros and cons and approaches are constantly evolving...

Thank You!

alexandra.siegel@colorado.edu
@aasiegel
alexandra-siegel.com