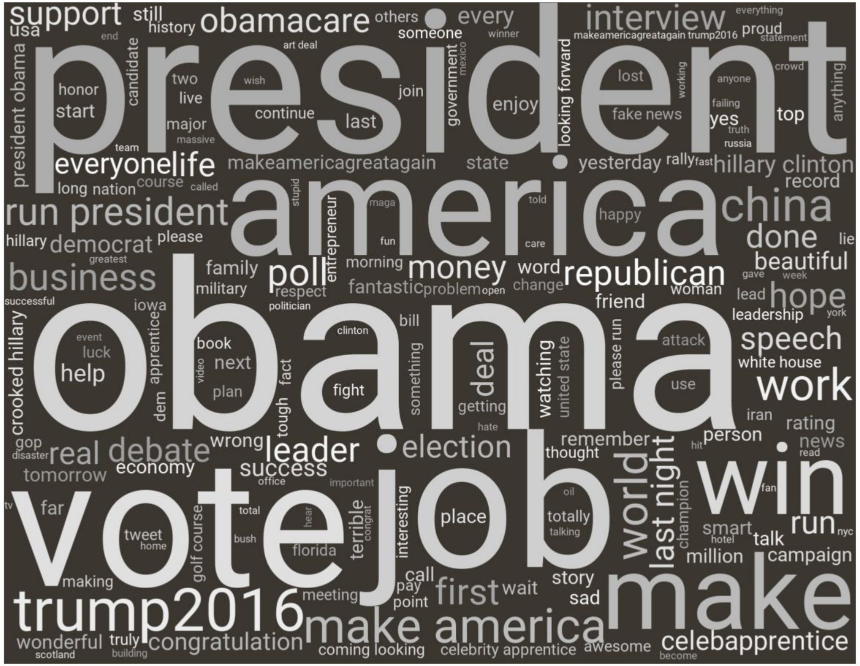
Patterns & Trends in Trump Tweets

Jason Snouffer
Data Mining I
Spring 2018

Introduction

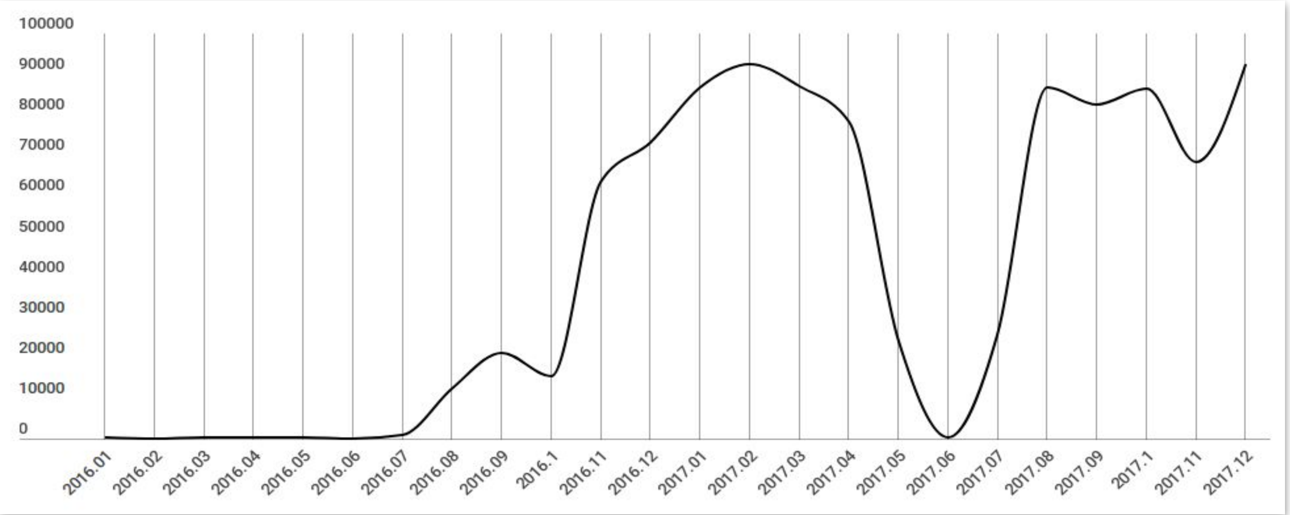
- Investigate patterns / trends in tweets related to and by President Trump
- Investigate what topics are being discussed in the tweets



Datasets

- 1 million Trump-related tweets (early 2016 - Feb 2018)
    - Random sample of 40 million tweets from Dr. Breitzman
- Profile data for 300K users who authored Trump-related tweets
- Profile data for 1.3M @realDonalTrump Twitter followers
- 33K @realDonaldTrump tweets

Trump-related tweets per month



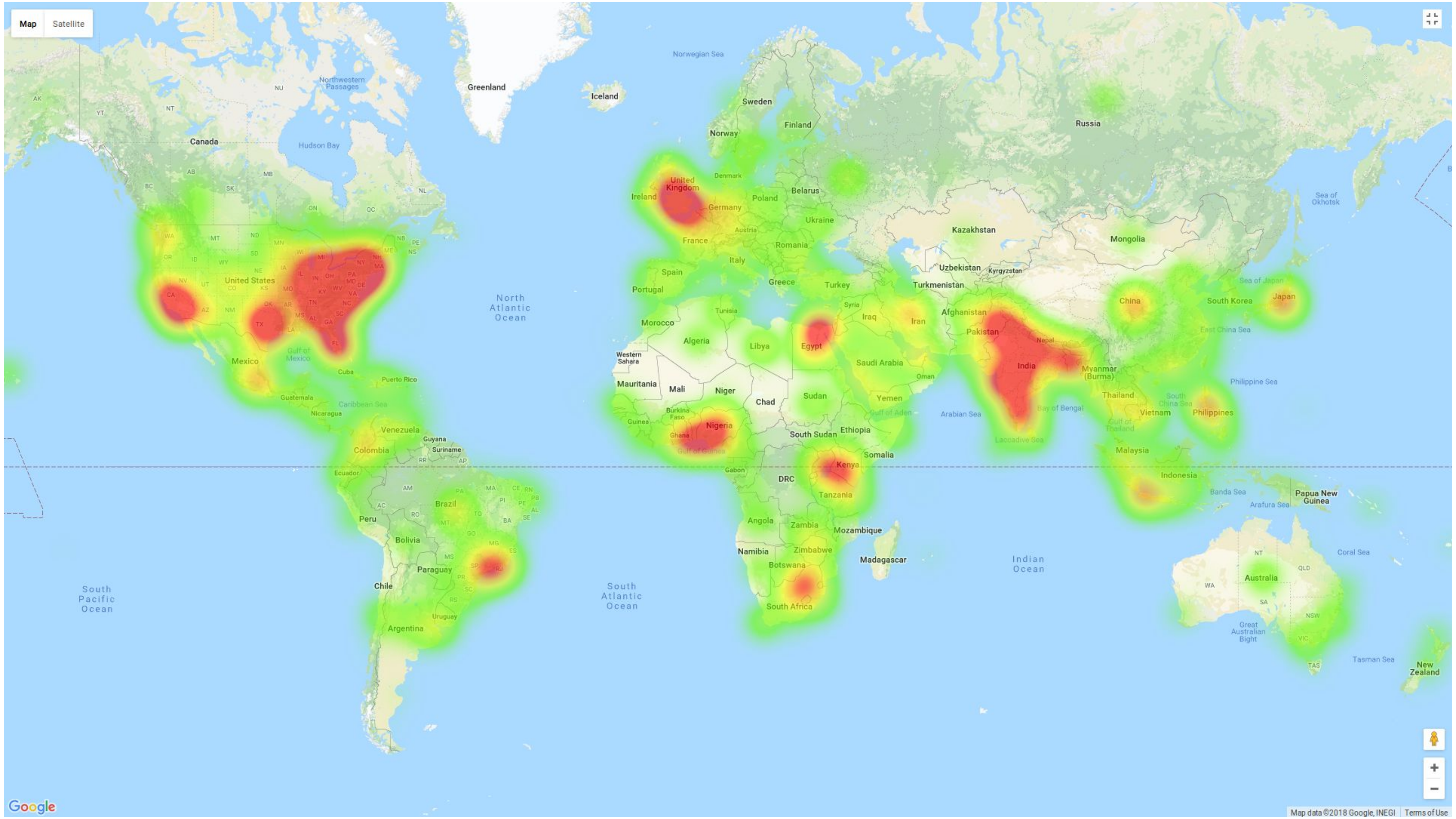
Analysis Overview

- Geolocation
- Naive Bayes classification
  - Pro-Trump vs Anti-Trump
  - Trump vs Staff
- Gender classification
- Sentiment analysis
- K-means clustering of tweets

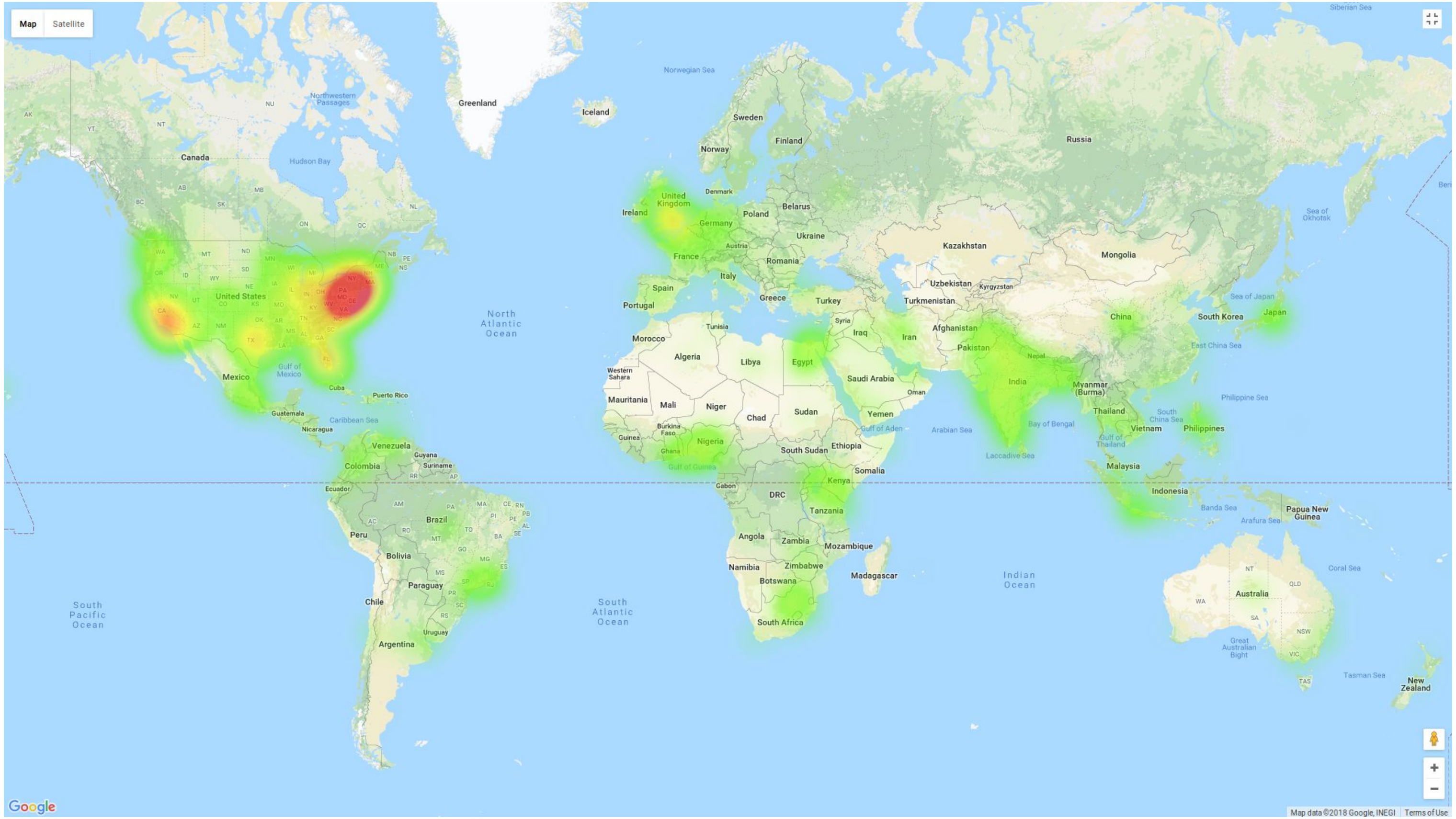
Tweet Analysis - Geolocation

- Less than 0.1% of Trump-related tweets are geotagged with latitude/longitude
- For remaining tweets, coordinates predicted using location (if, exists) from Twitter user profile
- Location field is cleaned and passed into Java OpenStreetMap api
    - Returns location description and coordinates (if query is successful)

Trump followers worldwide



Trump tweeters worldwide



@realDonaldTrump account

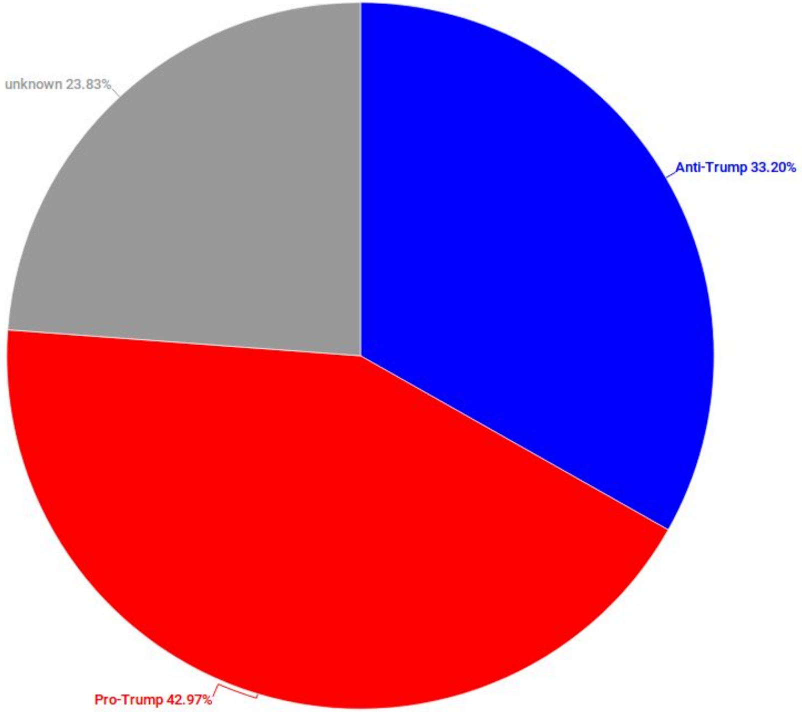
- 7.5% of Trump's tweets are geotagged
  - 1535 - Trump Tower, Manhattan, NYC
  - 273 - Mar-a-Lago Club, Palm Beach, FL
  - 269 - Trump National Golf Club, Bedminster, NJ
  - 105 - Trump National Doral Miami
  - 37 - various LA golf courses
  - 21 - Golf courses in UK
  - 16 - The Ritz-Carlton, Moscow
  - 13 - various locations in Mumbai
  - 13 - various locations and airports
  - 11 - Downtown Marriott, Des Moines, IA

Tweet Analysis - Naive Bayes

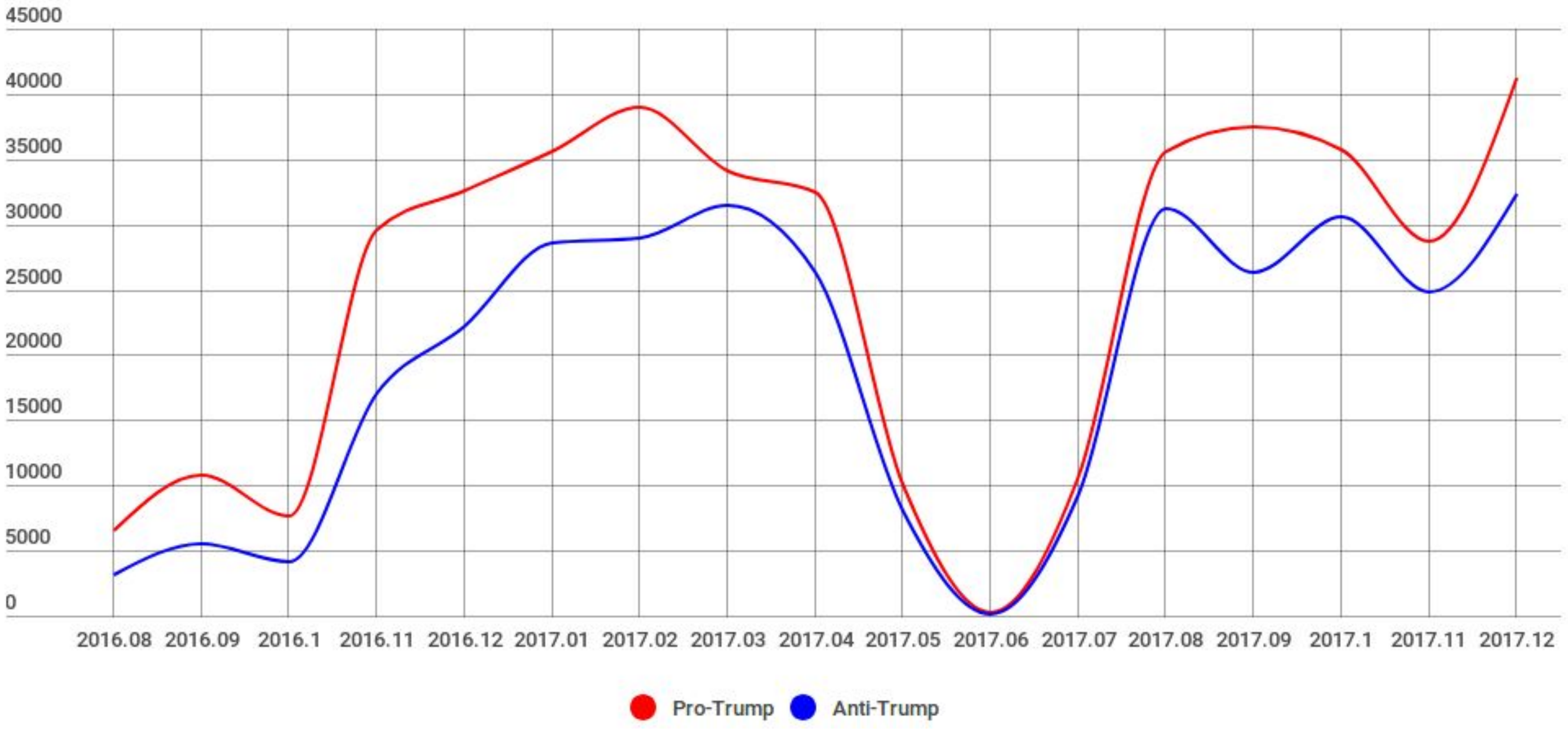
- Naive Bayes used to classify:
  - Trump-related tweets as either Pro-Trump or Anti-Trump
  - @realDonaldTrump tweets as authored by Donald Trump or his staff
- Used Java LingPipe api for NB classification
- Multiple NB models trained and then vote on classification
  - Increased prediction accuracy by 5% on test data
- Each classifier constructed with different parameters
  - bag of words
  - bag of words w/ hashtags removed
  - n-grams of size 3, 4, 5
  - n-grams of size 3, 4, 5 w/ stop words removed
- If votes are tied, tweet is left unclassified

NB classification - Pro-Trump vs Anti-Trump

- 1500 hashtags categorized as either Pro-Trump or Anti-Trump
- Hashtags used to classify 30K tweets
- Classified tweets used to train NB models
- 20% of training data saved for validation
- Accuracy: 95.6%

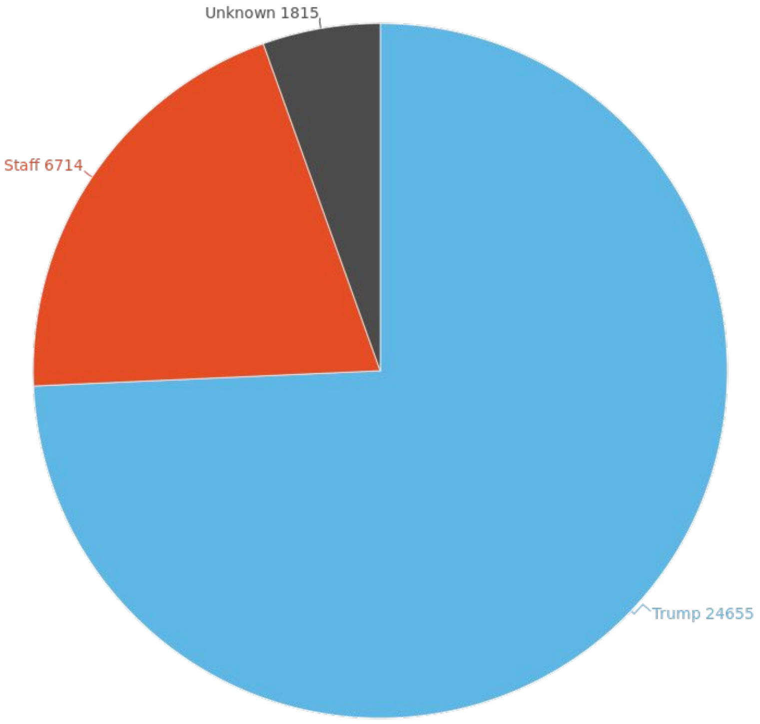


Pro-Trump vs Anti-Trump Time Series

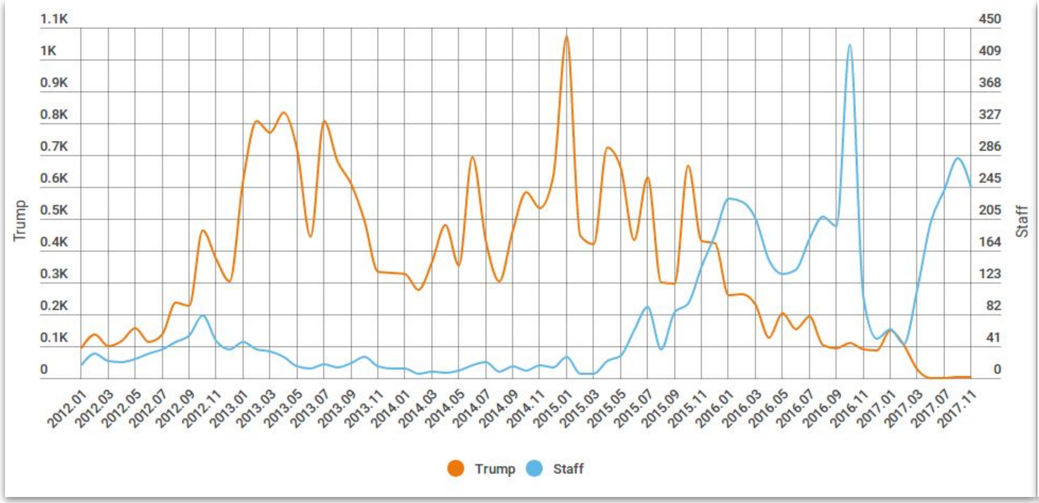


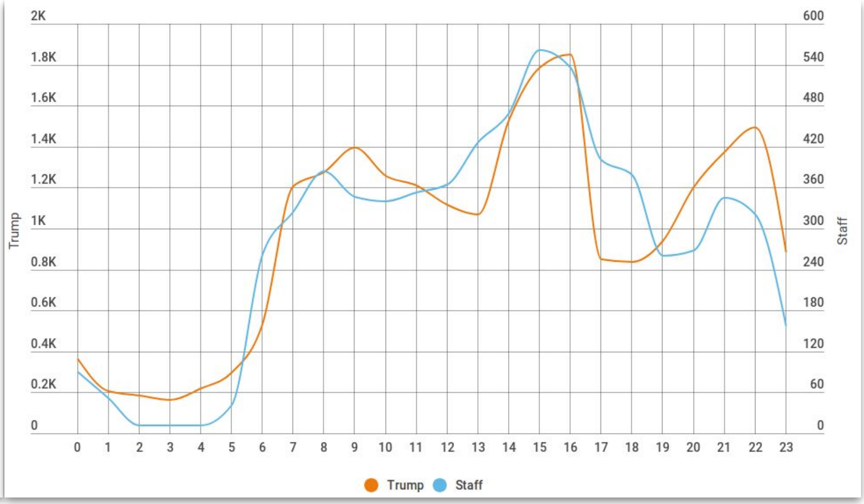
NB classification - Trump vs Staff

- Study from DZone indicates tweets from Android are from Trump and iPhone tweets are from staff
  - Study did not analyze tweets from other devices/sources
- Trump stopped tweeting from his Android on 3/25/2017
- 19K Android/iPhone tweets used to train NB models
- 10% of training data saved for validation
- Accuracy: 92.2%



Trump vs Staff Time Series





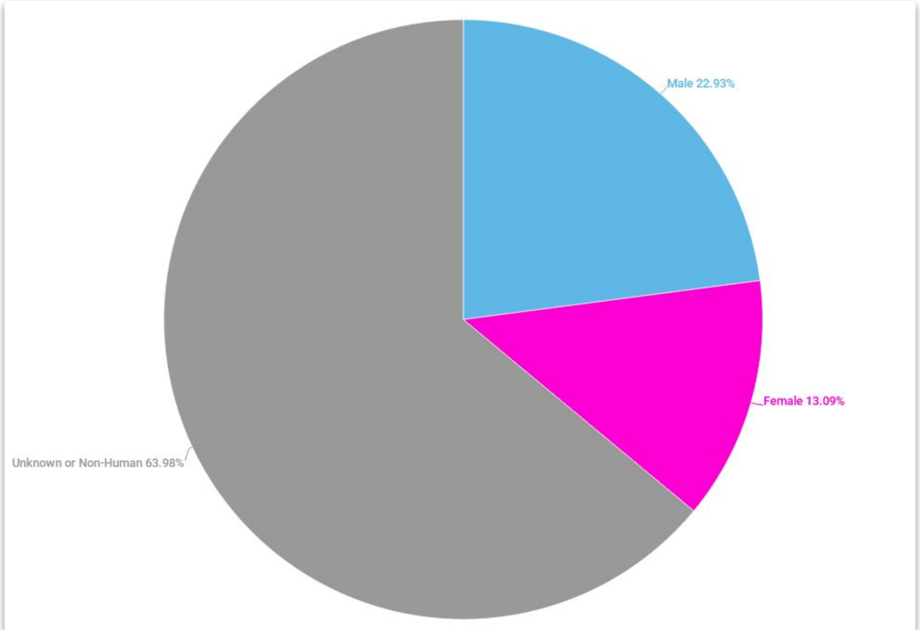
vs Month / Year

vs Time of Day

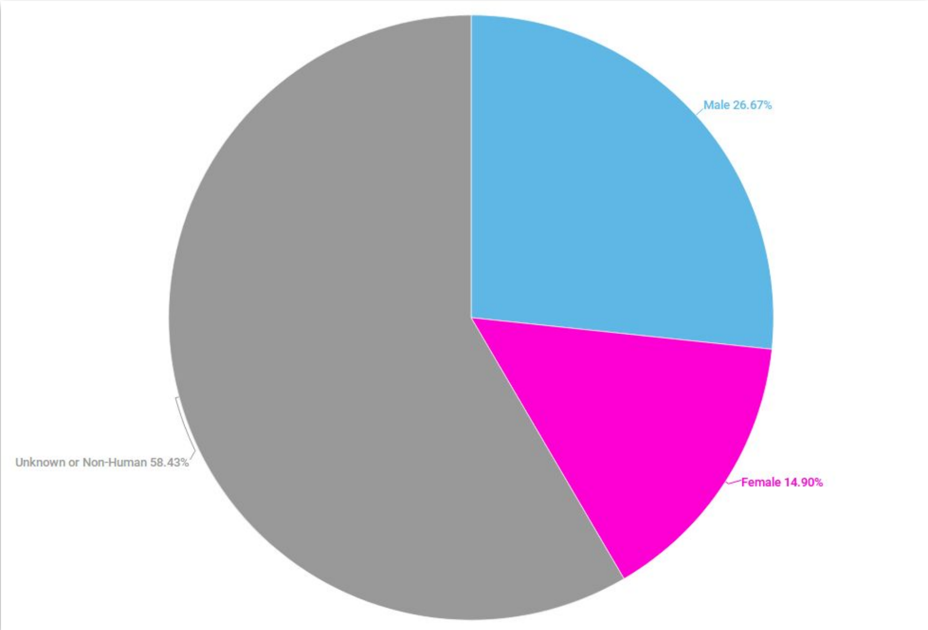
Tweet Analysis - Gender Classification

- Gender predicted for @realDonaldTrump followers and Trump-related tweeters.
- Name field from Twitter user profile used to predict gender
- Name passed into Python ProbablePeople library to guess if name represents a person or an organization/brand
- Given name returned by ProbablePeople is passed into Python GenderGuesser library

Gender Classification Results

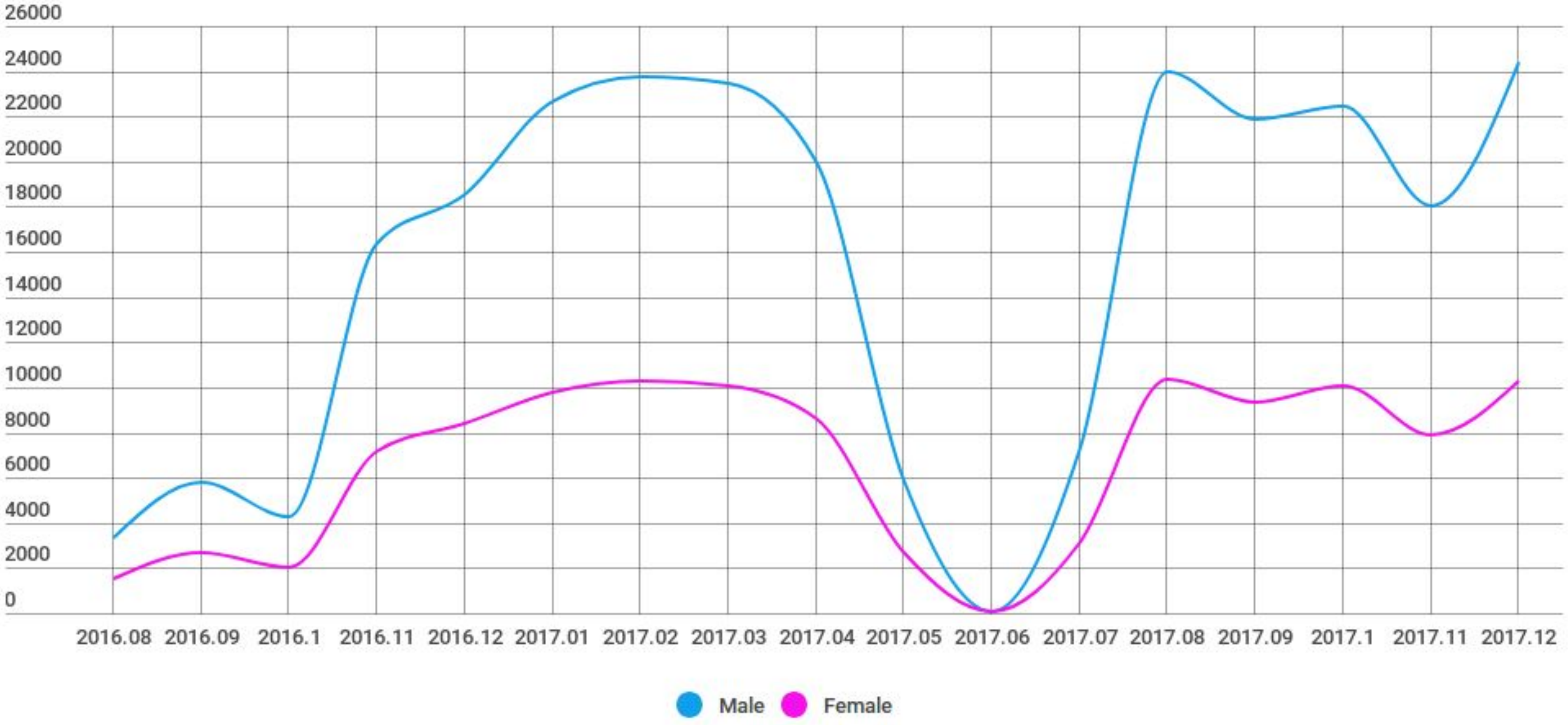


@realDonaldTrump Followers



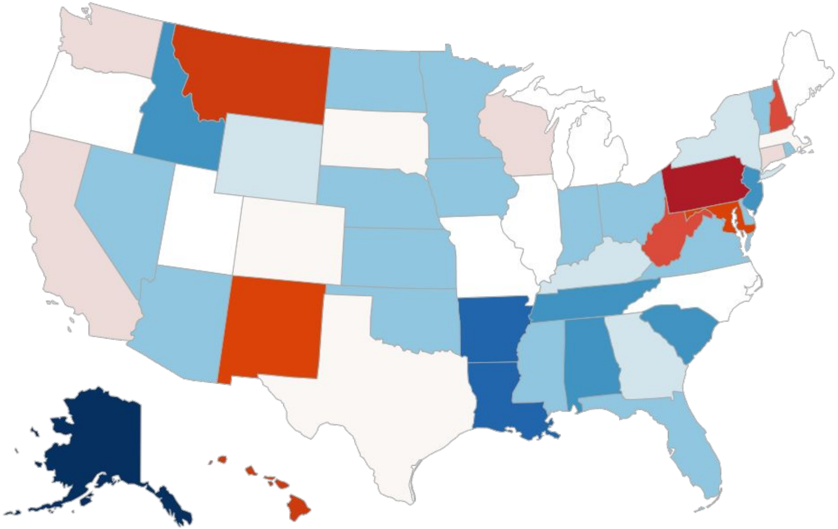
Trump-related tweeters

Gender Time Series



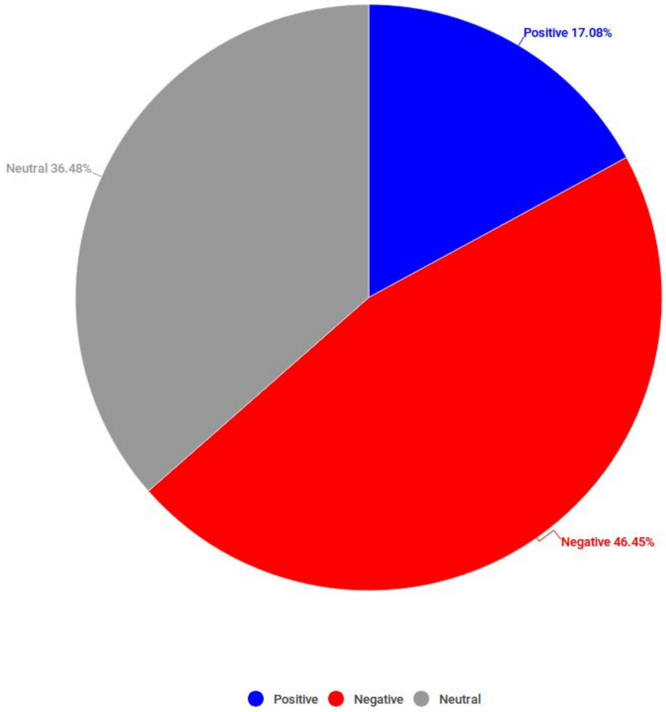
Tweet Analysis - Sentiment

- Sentiment scores calculated for all tweets using Syuzhet package in R
  - Uses Word-Emotion Association Lexicon (National Research Council of Canada)
  - List of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive)

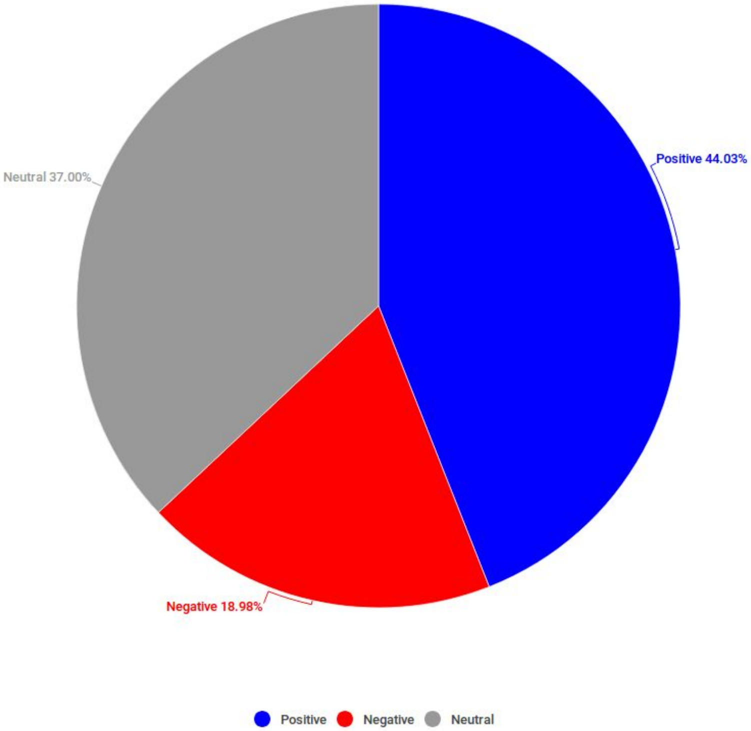


Trump vs Staff Sentiment

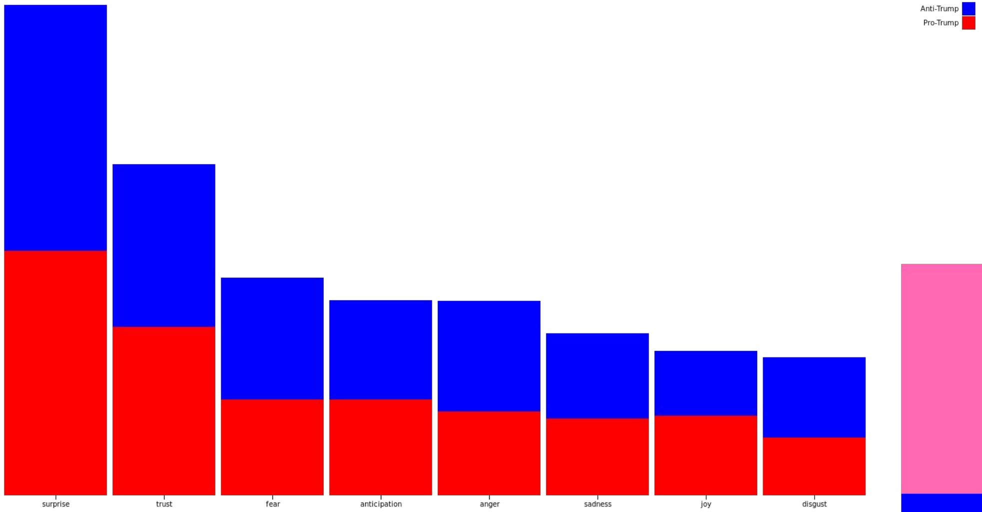
Trump



Staff

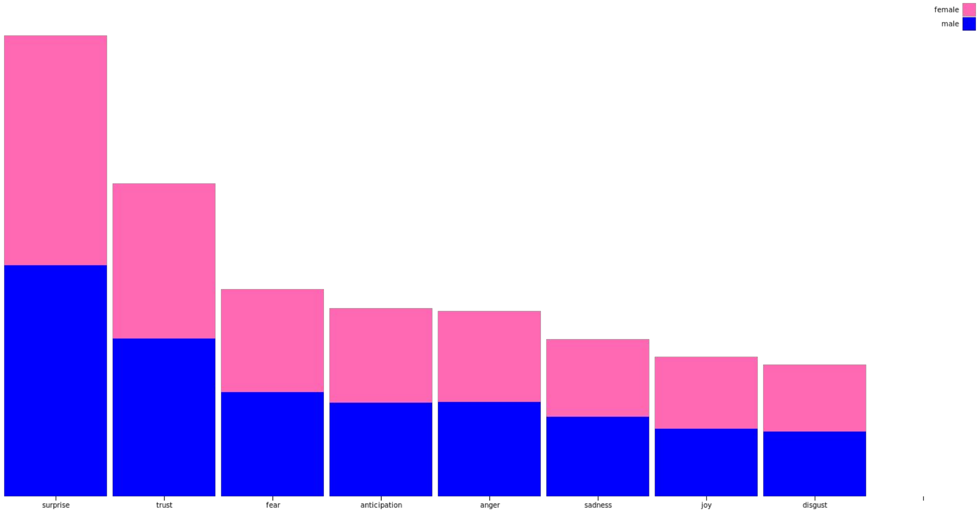


Emotion scores of Trump Related Tweets

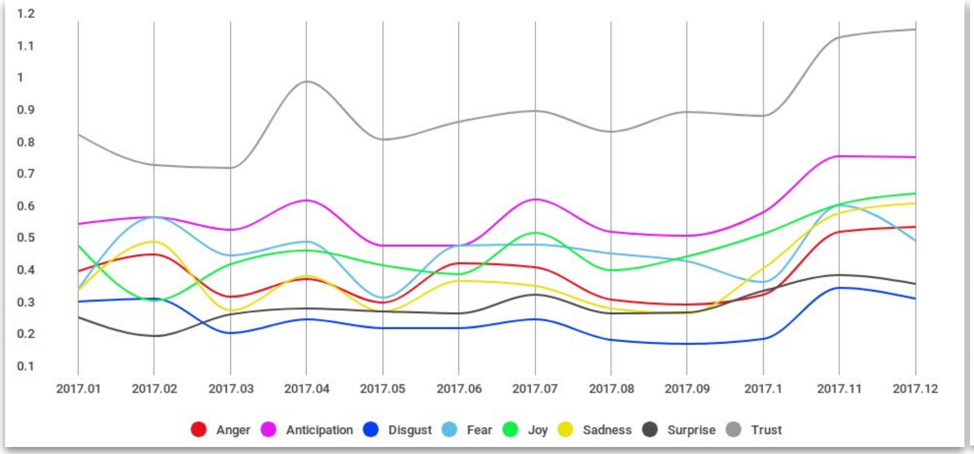


Pro-Trump vs Anti-Trump

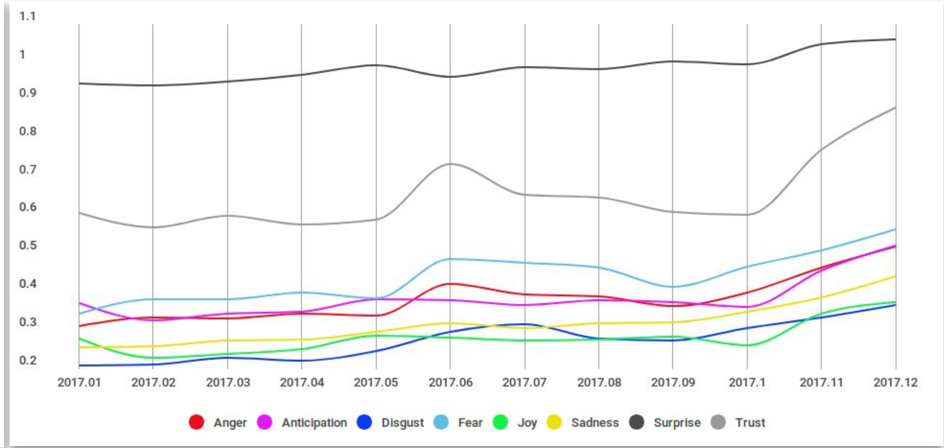
Male vs Female



Emotion Time Series (2017)



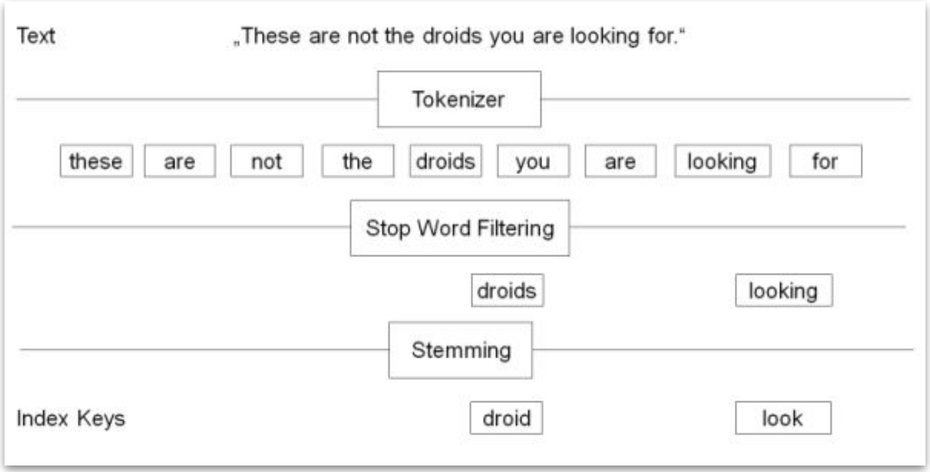
@realDonaldTrump



Trump Related

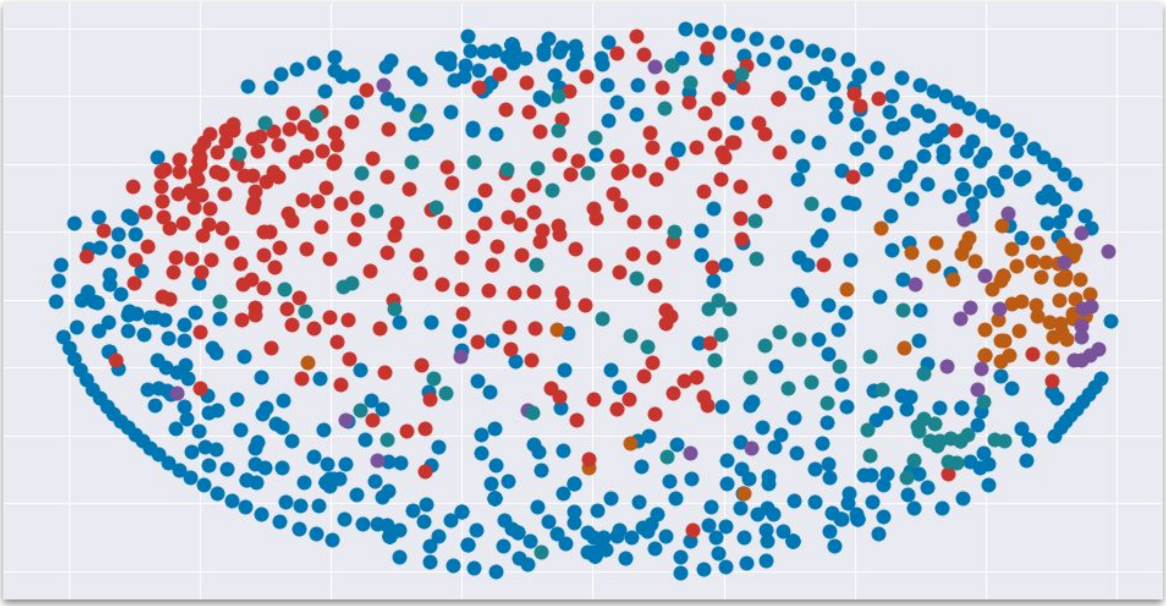
Clustering Tweets

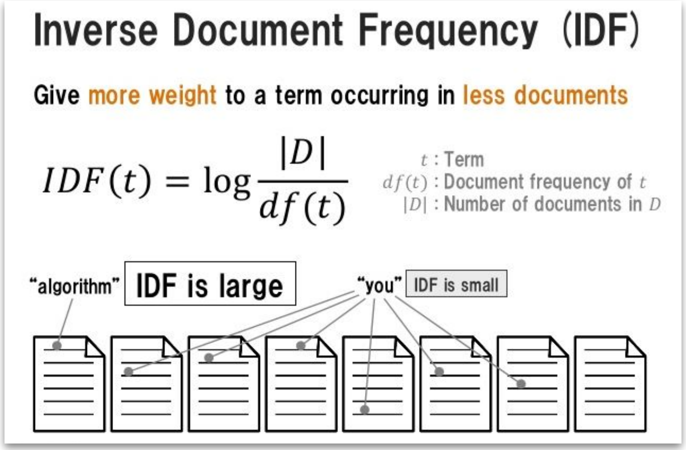
- Zoomed in on tweets from Oct 2017 through Dec 2017
  - About 190K tweets
- Used Python 3.6
  - scikit-learn, pandas, nltk, numpy
- Tweet pre-processing
  - Tweets are tokenized
  - Stop words are removed
  - Remaining words are stemmed



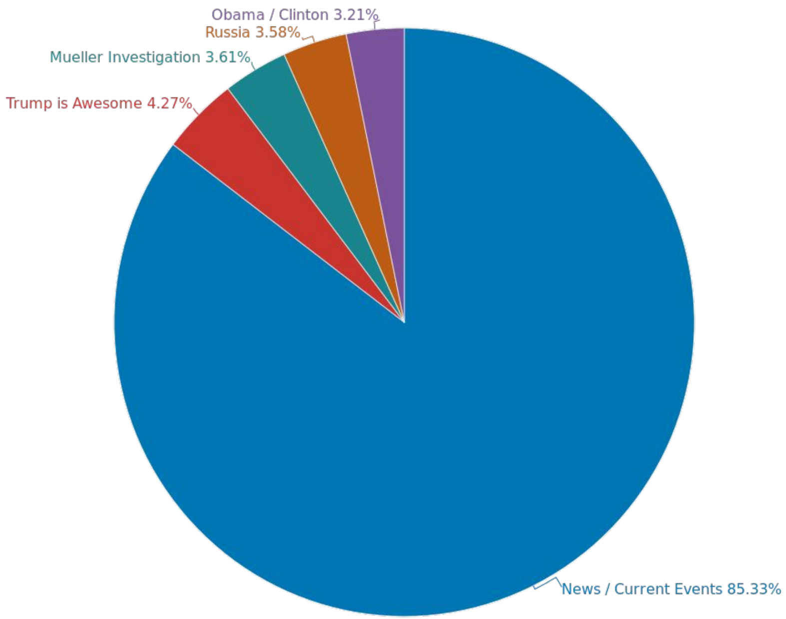
Clustering Tweets (continued)

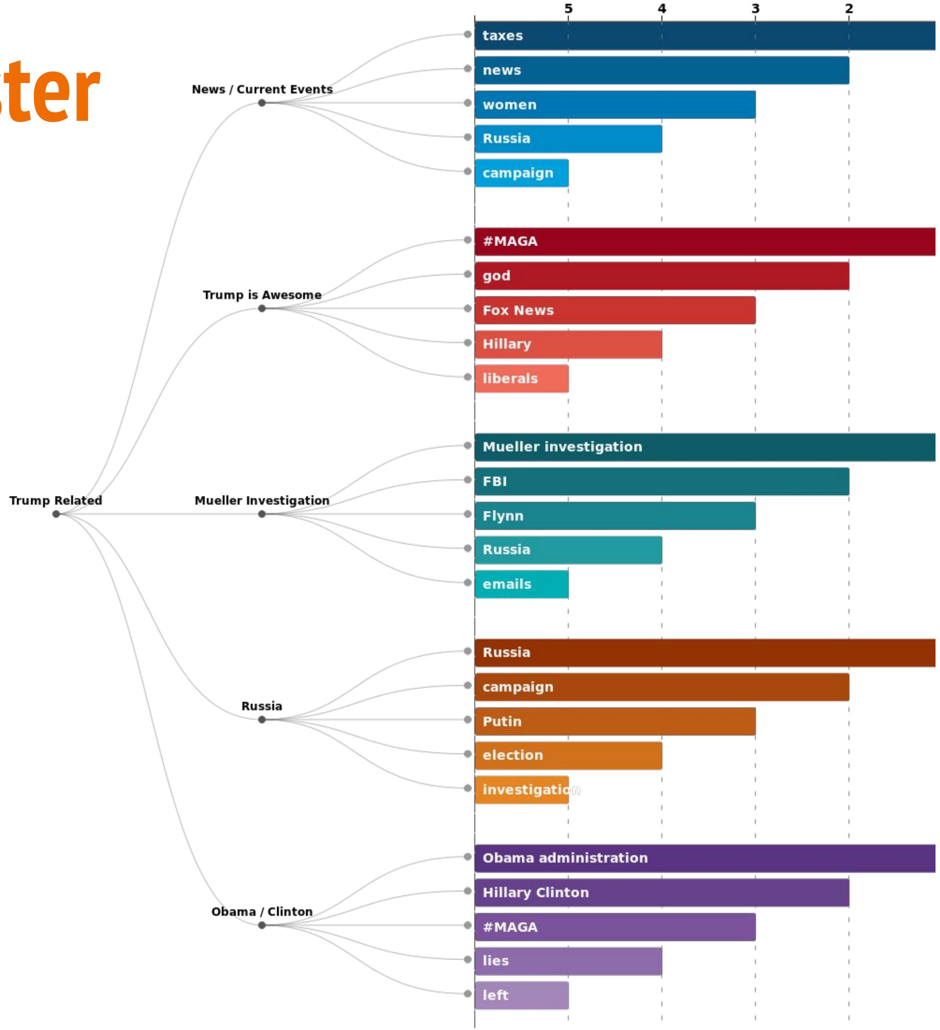
- Term Frequency - Inverse Document Frequency (TF-IDF) calculated
- K-means clustering performed (k = 5)



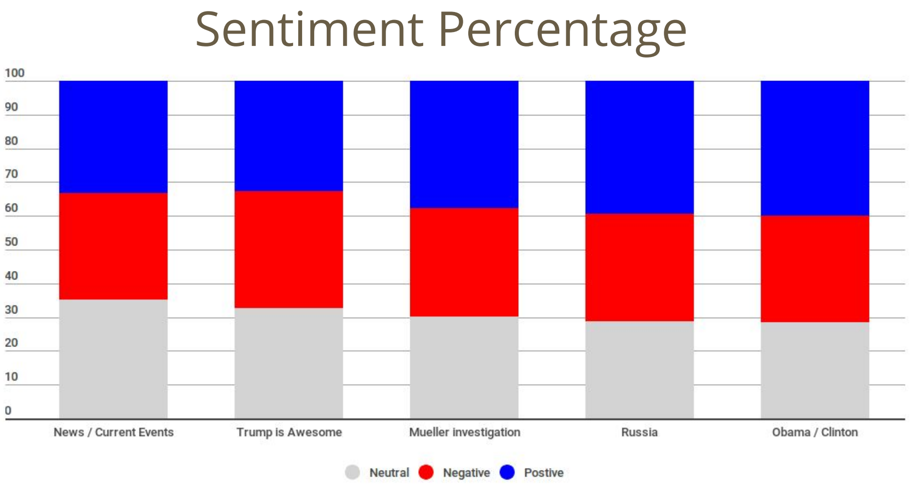


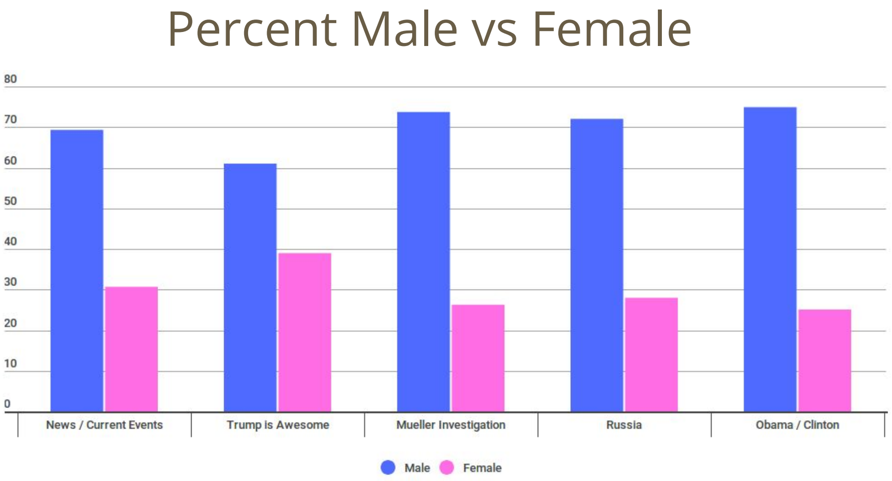
Dominant topics per cluster

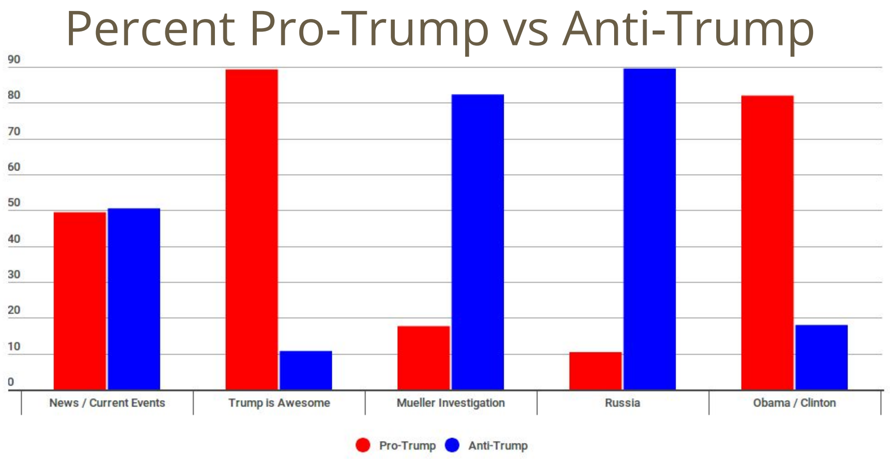




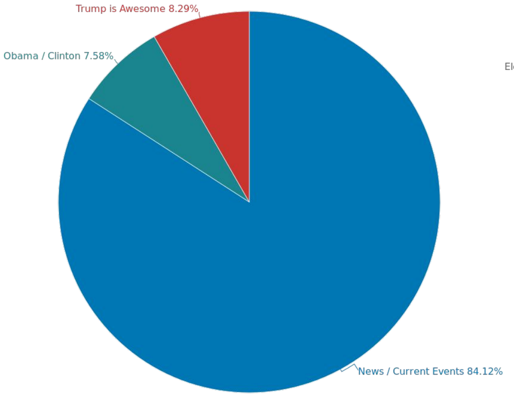
Demographics per cluster

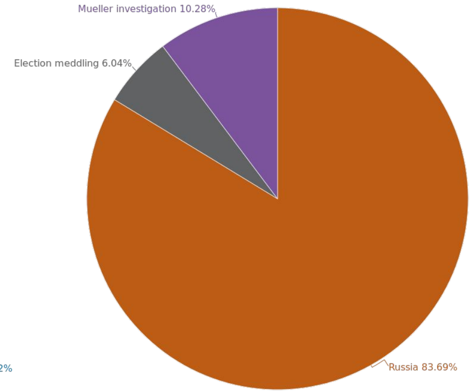


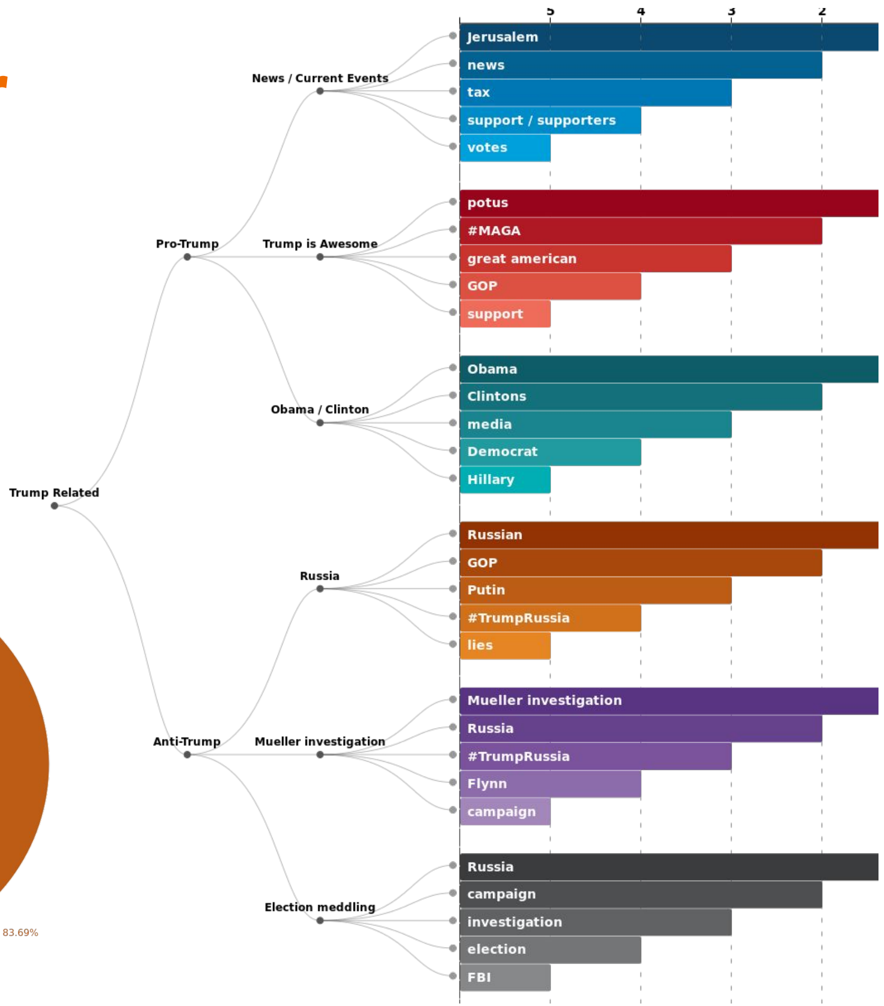




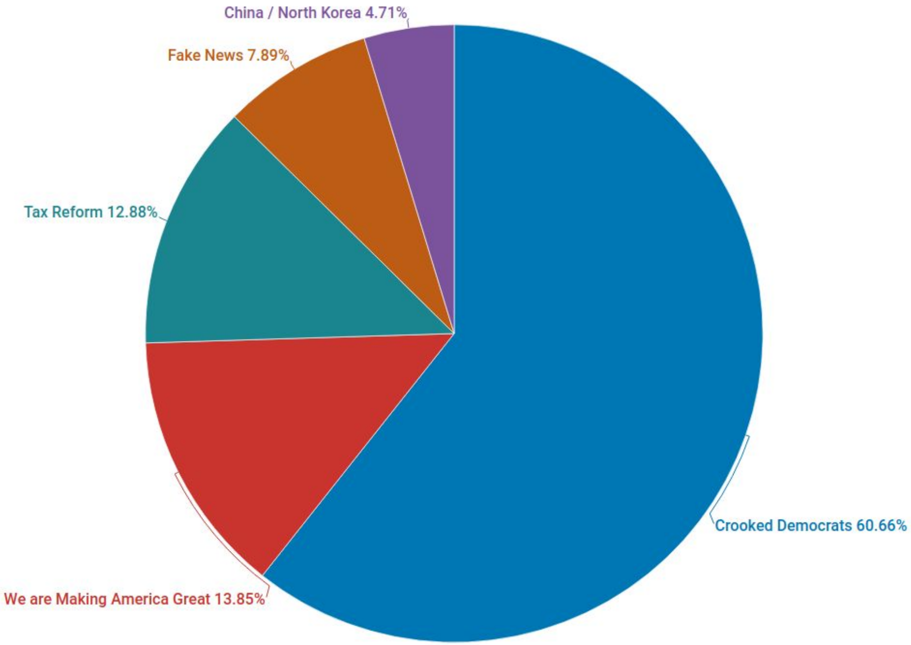
Dominant topics per cluster
Pro-Trump vs Anti-Trump

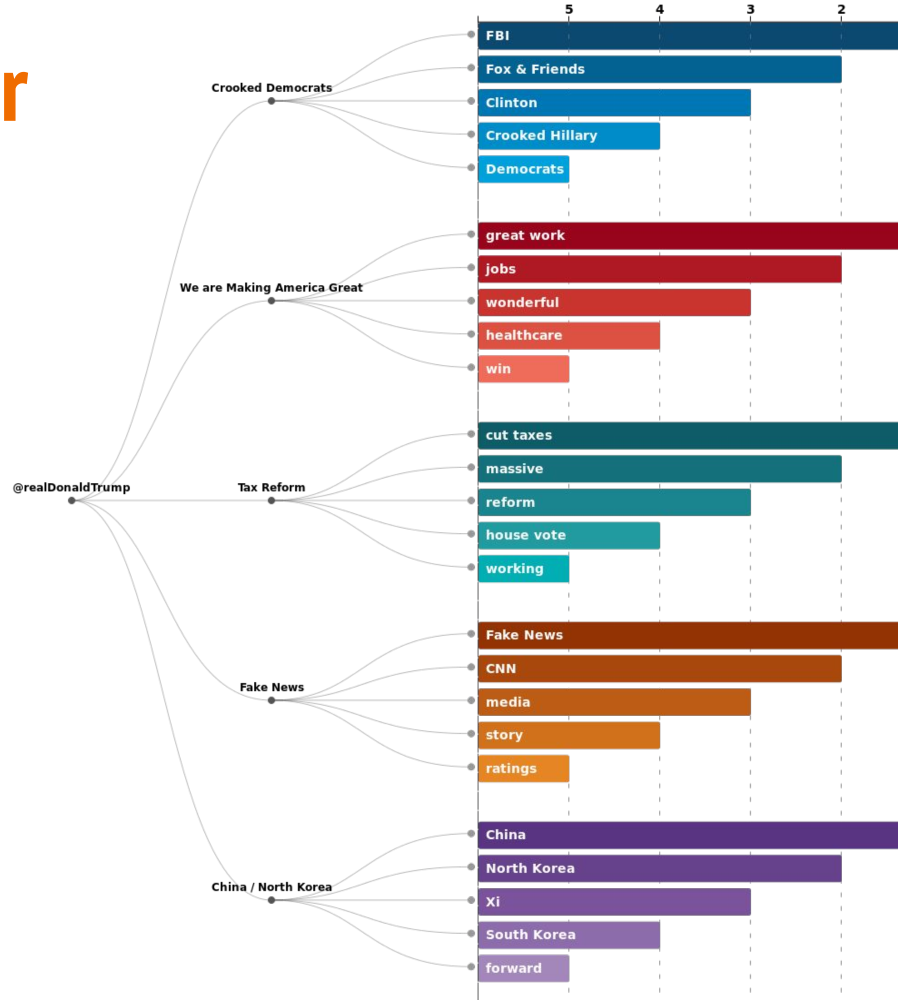




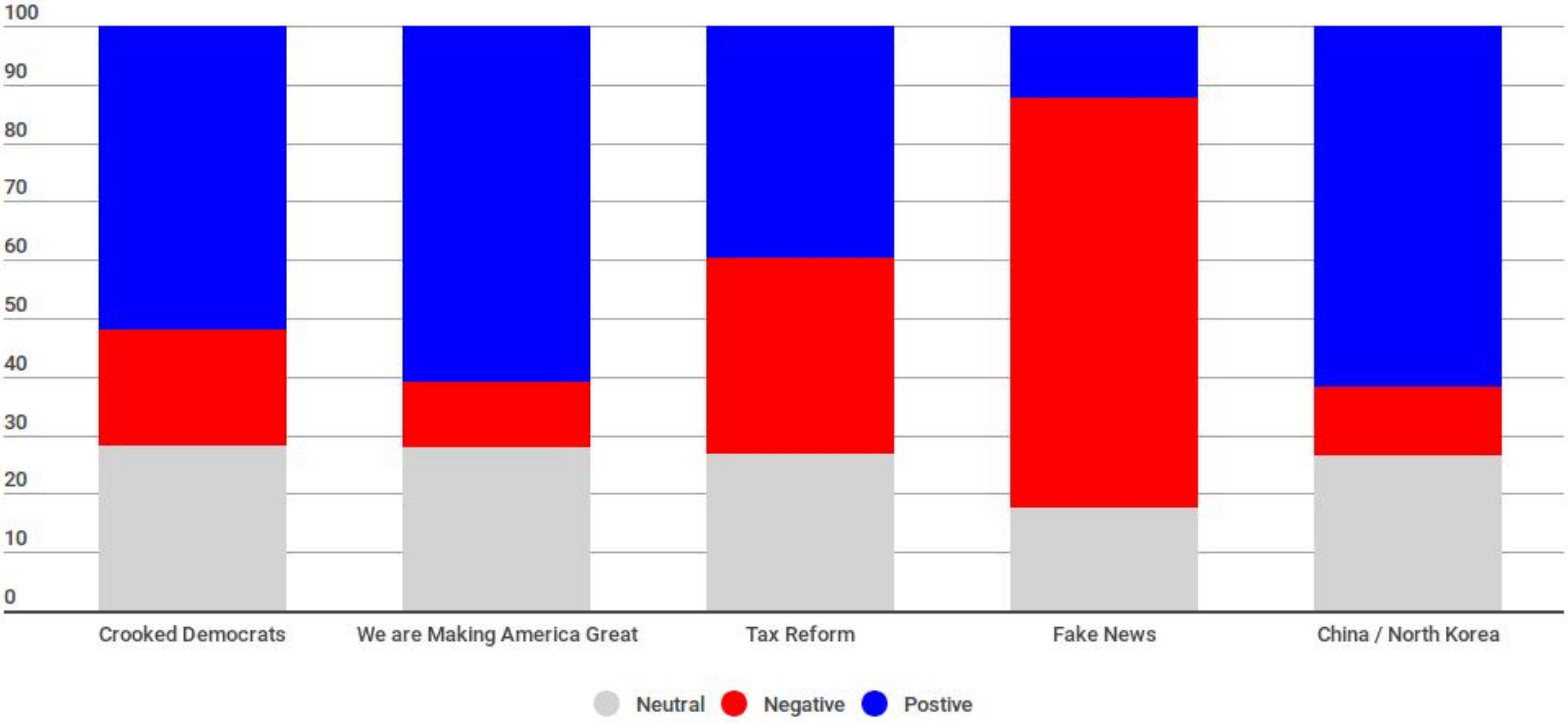


Dominant topics per cluster @realDonaldTrump





Sentiment per cluster



Future analysis

- Correlate sentiment scores and tweet counts with other datasets, such as polls, approval ratings, market performance, current events, weather, etc.
- Personality analysis of Trump, using his tweets, plus interview and speech transcripts
- Personality analysis of Trump followers using Twitter user profile data
- Investigate techniques for extended demographic analysis (ie: political party, nationality/race, age, etc.)
- Analysis of Russian Troll tweets dataset from Kaggle

Questions?