Tweets of #election2016

Anurag Prasad & Jarrod Lewis
{anuragp1, jl101995}@bu.edu

Overview

Twitter acts as a canvas for effortless, raw emotional expression. Unlike most social platforms, it attracts users with its simplicity of content production rather than a breadth of features. Impulsive thoughts are often our most honest thoughts, and Twitter is a massive space where people speak their minds spontaneously. The world had much to say about the 2016 election, and Twitter's content offers opportunities for exploration.

Data

The dataset was gathered using the Twitter Streaming API over a three day period, starting the day before the 2016 Presidential Election and ending the day after the election. Tweets were loaded in as a collection of JSON objects. Each tweet contains user information, geolocation (if enabled), hashtags used, tweet body text and additional fields pertaining to the tweet and user.

The data was collected from tweets containing at least one of the following trending keywords and/or hashtags:

```
'election', 'donald', 'trump', 'hillary', 'clinton', 'debates', 'vote',
'politics', 'ballot', 'obama', 'equality', '#election2016', '#electionday',
'#ivoted', '#imwithher', '#makeamericagreatagain', '#2016election',
'#lockherup', '#deleteyouraccount', '#crookedhillary', '#nevertrump',
'#feelthebern', '#blacklivesmatter', '#imvotingbecause', '#thirdparty',
'#garyjohnson', '#electionfinalthoughts'
```

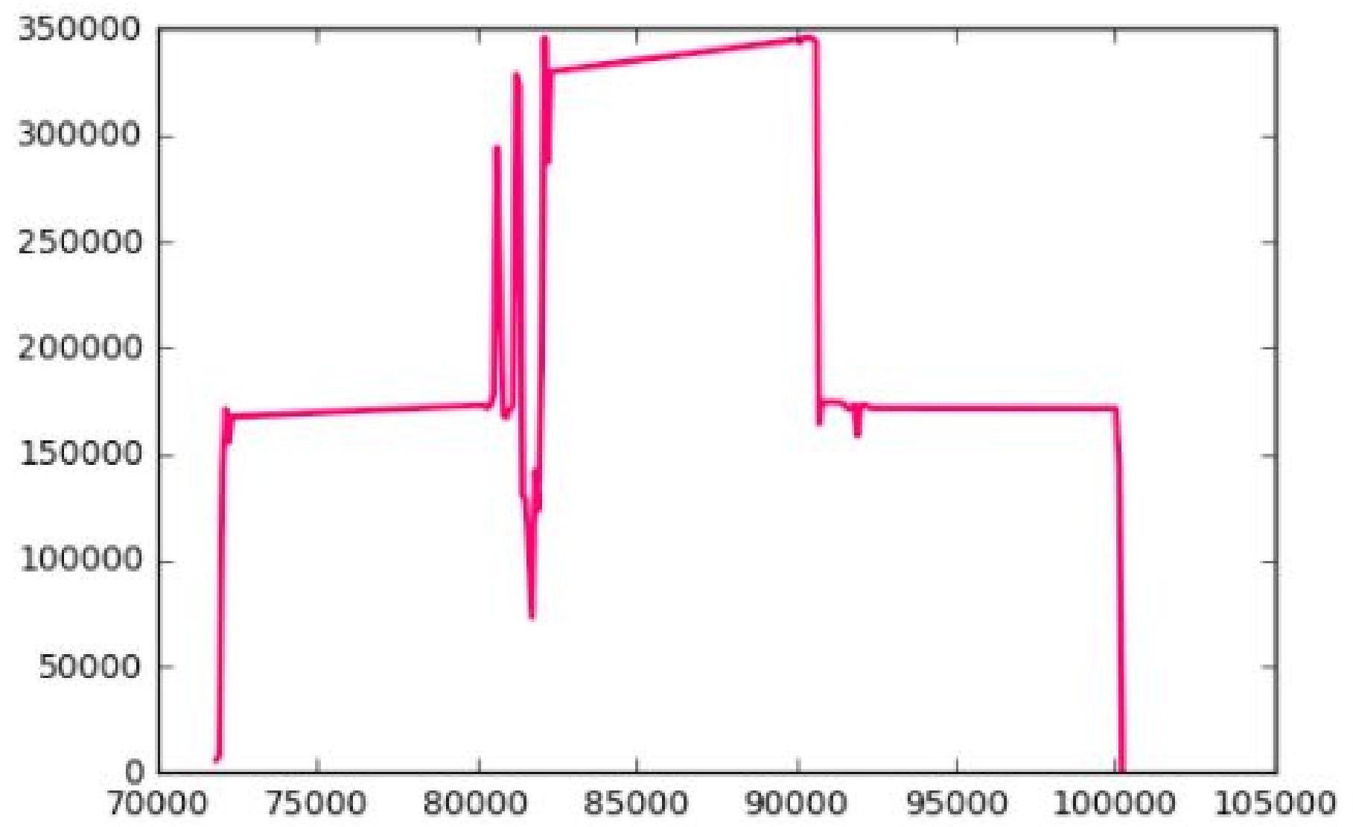
Over the course of the three days we gathered approximately 12 million tweets. Because the original dataset was too large to be used on a personal computer, 300,000 tweets were sampled using the following reservoir sampling algorithm, Algorithm R:

```
ReservoirSample(S[1..n], R[1..k])
  for i = 1 to k:        // Initially fill R
       R[i] := S[i]

  for i = k+1 to n:
       j := random(1, i) // randomly replace items in R
       if j <= k:
           R[j] := S[i]                              (1)
```

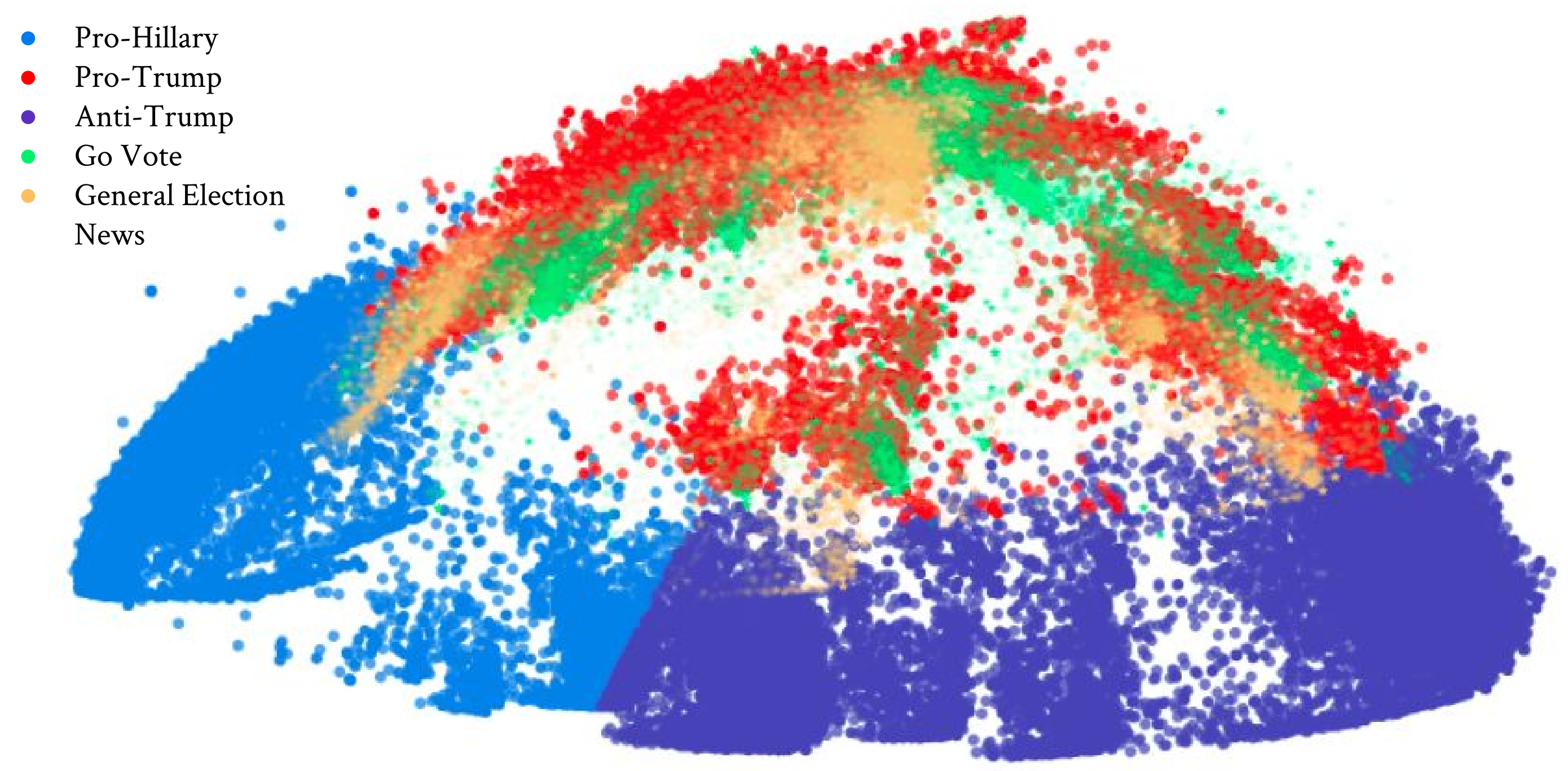
Tweet Stream Time Series

On Election Day, tweets related to the election more than doubled for over twelve straight hours. The tweets likely peaked above 350,000 per 30 minutes but our count was limited by the Twitter Streaming API.



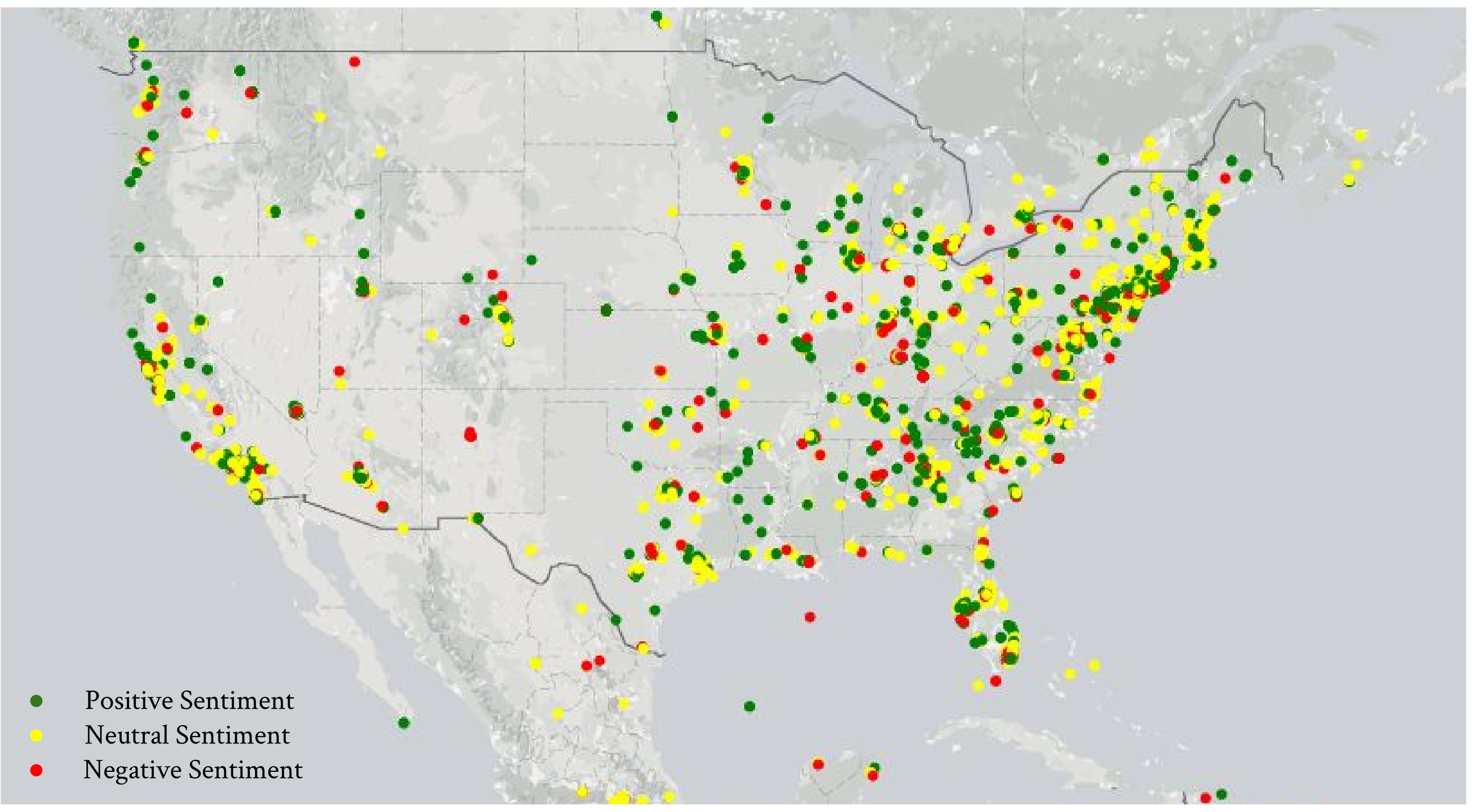
Clusters

Tweets were clustered based on **tf-idf** scores. Tf-idf scores allowed us to re-weight the counts of words in tweets so we could focus on words that are more likely to be relevant for differentiation. After calculating tf-idf scores, we stemmed and vectorized the terms. We then applied **Principal Component Analysis** to reduce the dimensionality of the vectors. We needed a tool that could help with multivariate data and with many topics to be considered in our determination of tweet similarity. An appropriate algorithm was **K-means** clustering, which we ran to find the following inherent groups in the tweets.



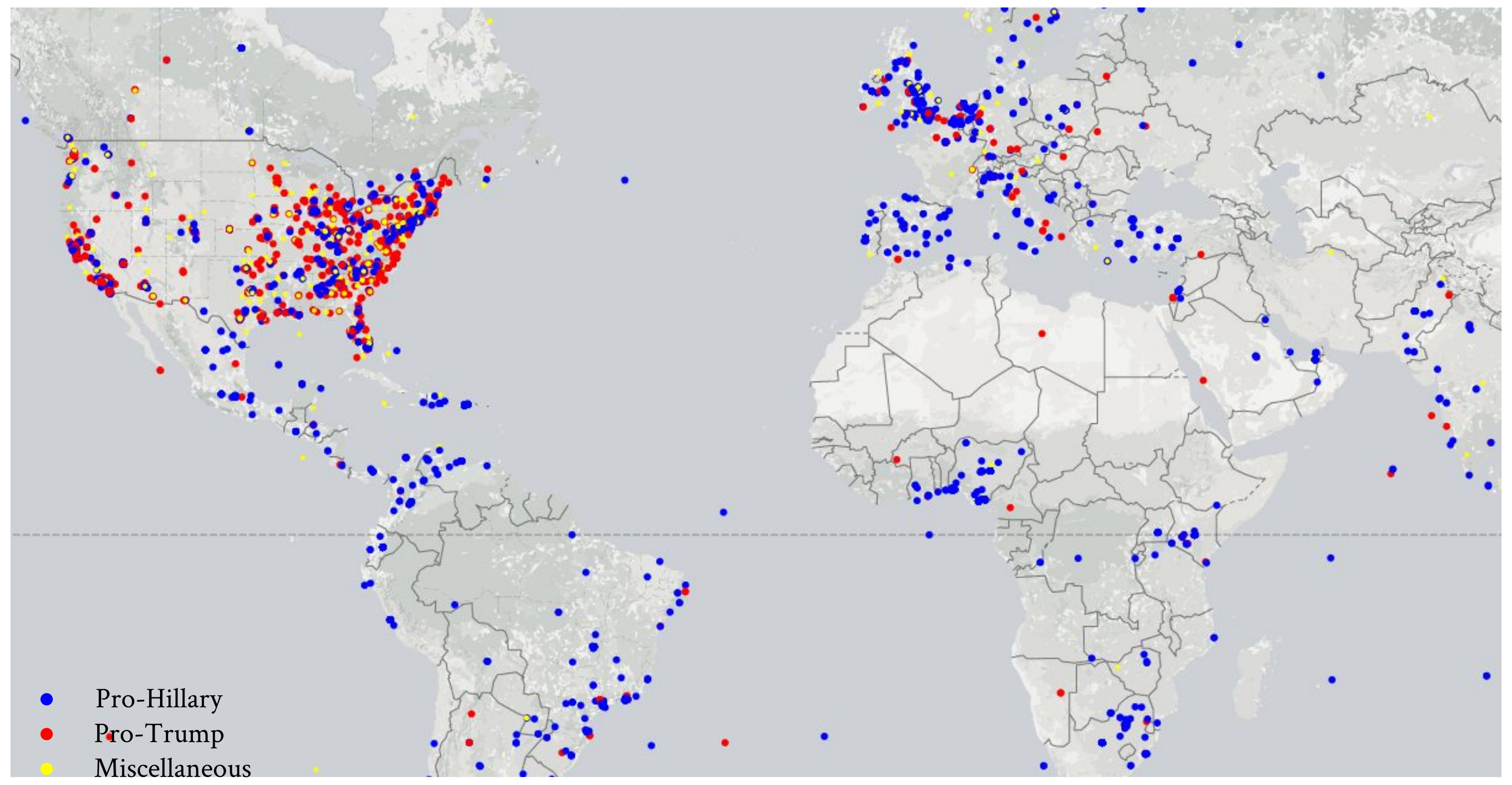
Sentiment Analysis

Sentiments of users mapped by geolocation shows that most tweets were neutral, with more positive than negative overall. Normalized, weighted sentiment scores from -1 to 1 (negative to positive), were calculated using NLTK's Vader, a rule-based model for sentiment analysis. Trump supporters had a more positive sentiment overall compared to Clinton supporters. The correlation coefficient between sentiment score and cluster is .18, with a p-value of 2.4e-41; this shows low correlation but has high significance to be considered meaningful. We suspect that the difference in sentiment is tied to how each candidate ran their campaign, with Trump being much more emotional, leading to more excitement around his campaign. The overall sentiment was much more neutral than we expected, especially considering how fervent people were throughout this election cycle.



Geolocation

We mapped each tweet that contained geolocation, colored by the cluster that that the tweet falls in. For this visualization we combined pro-Hillary and Anti-Trump into one cluster and we combined the two neutral clusters. The United States was much more red than expected, which shows that more users fell into a pro-Trump cluster. Every other country shown on the map is overwhelmingly blue, with most of the red coming from Western Europe, where other populist movements have been emerging in the Right Wing.



Conclusion

The five principal clusters from K Means were Pro-Hillary, Pro-Trump, Anti-Trump, Go-Vote and General Election News. There is nothing immediately surprising about the results of these clusters. The two major candidates in the 2016 Election dominate the clusters, as would be expected, with three out of five clusters being defined by their names. Considering the strength of Donald Trump's online following it was surprising to see that the size of the pro-Hillary and anti-Trump clusters far outweighed the pro - Trump cluster.

The sentiment of the tweets around the election was surprisingly neutral. However, we suspect this has more to do with the limitations of the sentiment analysis algorithm's capability to detect subtle nuances in tweet language. his election was extremely emotionally charged and left very few people across the United States feeling neutral, so improvement to this algorithm is necessary.

The geolocation corresponding to the clusters produced a fascinating result. We expected only to zone in on the United States, until we realized the polarization from the rest of the world. Every country outside the US was overwhelmingly colored blue to represent their place in a Pro-Hillary or Anti-Trump cluster. The only real Trump support can be seen within the United States.

References: (1) Vitter, Jeffrey S. (1 March 1985). "Random sampling with a reservoir" (PDF). *ACM Transactions on Mathematical Software*.

CS505: Computational Tools for Data Science

Instructor: Mark Crovella (crovella@cs.bu.edu)