Project Presentations - 1

CMSC 498J: Social Media Computing

Department of Computer Science
University of Maryland
Spring 2016

Hadi Amiri
hadi@umd.edu
Project Titles

- **G2**: Link Prediction between Candidates and Twitter Users
  - Matthew Demers, Katherine Kelly, Jake Roeland

- **G3**: Detecting Politically Influential Individuals via Twitter
  - David Kaminsky, Jared Rushin, Kyle Kelley

- **G7**: Calculating Trending Subreddits
  - Azam Abdulkadir, Nicolas Chao, Kaval Patel

- **G10**: Comparison of Political Beliefs of Presidential Candidates on Twitter
  - Rudmila Rahman, Taylor Moore, Siddharth Dave

- **G14**: Cyberbulliyng from a Linguistic Perspective
  - Ana Gabriela Larios, Anna Benitez, Matthew Marsandi
Link Prediction between Candidates and Twitter Users

Matthew Demers, Katherine Kelly & Jake Roeland
Initial Problem...

Can we predict the primary winners from twitter?
Can we predict the primary winners from twitter?

Issue: We cannot confirm the twitter world will provide the same results as the real world
Revised Problem...

...To predict if a twitter user is following a presidential candidate based off of the their tweets and the other users they follow.
Why is it relevant?

Similar methods could predict if a user is going to follow anybody based off of various factors.

Show how important online presence is to a candidate’s campaign

One method for mapping the twitter world to real world
Step 1 - Collect Users and their tweets

Using Tweepy library in conjunction with the Twitter API we were able to create our own initial dataset consisting of the followers of the various presidential candidates’ Twitter accounts

Collected 4151 unique users

Collected up to 200 most recent tweets for each user
Step 2 - Extract User Information

Check the number of tweets the user has made referencing a candidate

Methods: Retweeting, Mentions, and Hashtags

Only 2535 out of 4151 users with tweets available

For each user in our dataset, check what percent of their friends follow the candidates

Due to time constraints, we only used 100 users, and only their first 50 friends
Step 3 - Train a learner

We used the PyBrain library to train a backpropagation neural network to predict if a user follows each candidate.

Ran two separate tests:

a. Using all of the users with just the data collected from their timeline.

b. Using a subset of users who we collected demographics about people they are following from.
Results
Results

Following prediction for BernieSanders

Following prediction for HillaryClinton
Results using demographics of the user’s friends

Red line: Using friend’s data
  Average error = 48.26%

Blue line: Without using friends data
  Average error = 49.24%
Correlation of Unique Users to tweets

Mentions

Retweets
Insights

Demographics of young voters/twitter users.

A third of twitter users are ages 18-29

Sanders polls well with young voters.

Twitter results inconsistent with Primary results

Sanders receives overwhelming support on twitter compared to Clinton

Over a 3 minute span, 124 support Sanders hashtags were used and only 51 support Clinton hashtags were used.

But only received 33.2% of the vote in the Maryland Primary compared with Clintons 63.0%

As of 2014, only 20% of young americans (18-29) voted
More insights

Most people don’t make political tweets

Trump, who has the highest online presence, only had mentions in 1.9% of tweets

With the average candidate mention in only .79% of tweets

Candidates with higher online presence have better polling numbers

Kasich has 8.6K tweets and Trump has over 31.8K tweets

Over 3 minutes span, Kasich had 104 mentions where Trump had 2,099 mentions

Kasich had 23% and Trump had 54.4% of Maryland Primary votes
How to improve...

Build a bigger dataset

Use more context from user timeline

Filter users (Remove spam and private accounts)

We only collected data from people already following candidates, not any that aren’t following anyone
References

- http://www.pewinternet.org/2015/08/19/the-demographics-of-social-media-users/
Detecting Politically Influential Individuals via Twitter

By: David Kaminsky, Jared Rushin, Kyle Kelley
Young people don’t vote

How to reach young people?

Social Media

How to convince young people to vote?

Reach out via Twitter

Find influential twitter users → Determine political alignment → Recruit as volunteers for campaigns to get out the vote
Technology Stack

- neo4j
- Python
- Twitter
Neo4j - Database Content
Method

1. Collect tweets via Twitter Streaming API (using Tweepy)
   a. Filter on location (MD), and Keywords

2. Import tweets into Neo4j for analysis

3. Calculate scores for 4 metrics:
   a. $\frac{\text{followers}}{\text{max(followers)}} \times 100$
   b. $\frac{\text{following}}{\text{max(following)}} \times 10$
   c. $\frac{\text{retweets}}{\text{max(retweets)}} \times 100$
   d. $\frac{\text{likes}}{\text{max(likes)}} \times 10$

4. Aggregate scores into composite Influence Score

5. Evaluate?
Results

![Influence Score Graph]

- Abdelaziz848
- JoeTrippi
- StoolBlock
- DeatheTrap
- educationweek

The Influence Score varies among the individuals, with JoeTrippi having the highest score.
Evaluation

How good are Twitter users at spreading info?

- Likes per retweet
  - Do people care about what you have to say?
- Retweets per follow
  - Do you followers agree with what you have to say?

Compute these values for each user

Compare rank based on Influence Score with LikesPerRetweet + Retweets per Follow

Rank Difference / # Users = Error

Average error Influence Score = 23.37%
Power Laws / Rich Get Richer

Few accounts get very high follower numbers.

Most accounts have < 500, but some have as many as ~ 1 million.
Sentiment Analysis

What do Marylanders think of the 4 main politicians left in the primaries?

MPQA Subjectivity Lexicon
Calculating Trending Subreddits

Azam Abdulkadir, Nicolas Chao, Kaval Patel
Motivation

• Trending subreddits feature does not include subreddits from default subreddits

• Our algorithm calculates trending subreddits (with the default subs) to find the recently active subreddits and compared it to Reddit’s popular/recent activity list
Approach

• Select 100 subreddits by collecting 300 of the top posts in the past 24 hours. Then determine which subreddit those posts originate from.

• Examine all comments for a post and search for comments that link to other subreddits.
Approach

• We then create a graph where each subreddit selected is a node.
• A directed edge \((n_1, n_2)\) exists between subreddits in the graph, if in a post to subreddit \(n_1\) there exists a comment that links to \(n_2\).
Approach

• Using the previously described approach, we construct the entire graph and run the PageRank algorithm.

• Sort the resulting subreddit ranks by descending order, and the resulting list is the list of “trending” subreddits.
Data Collection

- Dataset was created by interfacing with reddit’s api.
- Multiple requests had to be made for each individual post
  - Limits on how many comments received on individual request
  - Time consuming if using a single thread
- Spawned a thread to collect data on each subreddit, then spawned additional threads to collect comments in a post in those subreddits.
Data Collection Results

- 132,678 Comments Collected
- 2500 had comments that linked to some subreddit
  - Of these comments, 147 were unique (didn’t create a edge that already existed in the graph), did not point to itself and pointed to one of the 100 subreddits we were collecting from
Graph created from data (excluding nodes w/o any in or outlinks)
Evaluation

- Initially we used the simple PageRank Algorithm:
  - PankRank Algorithm:
    - All Nodes receive value of $1/N$, where $N$ is the total number of nodes.
    - Run PageRank Update Rule $K$ times, where $K$ is 10,000
  - PageRank Update Rule:
    - All nodes give current PageRank/Number of outlinks to each node it outlinks to
    - Do the previous step for all nodes
    - All nodes sum the values up into new PageRank score
  - After running $K$ times a equilibrium is reached
Simple PageRank Algorithm Issues:

- Nodes with no inLinks did not have any value
- Cycles where two nodes only linked to each other “trapped” the score between the two
- Some node did not have any outLinks which also caused a “trapping” issue
Issues
Solution

- **Scaled PageRank Algorithm**
  - Use Simple PageRank Algorithm
  - For each iteration 0 to K:
    - Multiply the PageRank scores by S, the scale factor, in this case .8
    - Take the (1-S) value, in this case .2, and divide by the total number of nodes, 73
    - Take that value, (1-S)/N, and add it to all nodes’ PageRank Score
  - The end result is a scaled PageRank Score to correct for cycles
# Expected Results

Reddit provides a list of popular/recently active subreddits

<table>
<thead>
<tr>
<th>Position</th>
<th>Subreddit</th>
<th>Position</th>
<th>Subreddit</th>
<th>Position</th>
<th>Subreddit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AskReddit</td>
<td>7</td>
<td>todayilearned</td>
<td>13</td>
<td>news</td>
</tr>
<tr>
<td>2</td>
<td>The_Donald</td>
<td>8</td>
<td>videos</td>
<td>14</td>
<td>hockey</td>
</tr>
<tr>
<td>3</td>
<td>politics</td>
<td>9</td>
<td>AdviceAnimals</td>
<td>16</td>
<td>gaming</td>
</tr>
<tr>
<td>4</td>
<td>funny</td>
<td>10</td>
<td>worldnews</td>
<td>17</td>
<td>soccer</td>
</tr>
<tr>
<td>5</td>
<td>nba</td>
<td>11</td>
<td>leagueoflegends</td>
<td>18</td>
<td>SandersForPresident</td>
</tr>
<tr>
<td>6</td>
<td>pics</td>
<td>12</td>
<td>gifs</td>
<td>19</td>
<td>movies</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>20</td>
<td>relationships</td>
</tr>
</tbody>
</table>
## Actual Results

<table>
<thead>
<tr>
<th>1 - 10:</th>
<th>PageRank:</th>
</tr>
</thead>
<tbody>
<tr>
<td>thathappened</td>
<td>0.125793821</td>
</tr>
<tr>
<td>kenm</td>
<td>0.079810793</td>
</tr>
<tr>
<td>me_irl</td>
<td>0.049670799</td>
</tr>
<tr>
<td>imgoingtohellforthatis</td>
<td>0.045371661</td>
</tr>
<tr>
<td>subredditsimulator</td>
<td>0.037433174</td>
</tr>
<tr>
<td>circlejerk</td>
<td>0.032106623</td>
</tr>
<tr>
<td>tumblrinaction</td>
<td>0.028683519</td>
</tr>
<tr>
<td>meirl</td>
<td>0.026533431</td>
</tr>
<tr>
<td>oldpeoplefacebook</td>
<td>0.025535233</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>11 - 20:</th>
<th>PageRank:</th>
</tr>
</thead>
<tbody>
<tr>
<td>soccer</td>
<td>0.024295871</td>
</tr>
<tr>
<td>animalsbeingbros</td>
<td>0.022675324</td>
</tr>
<tr>
<td>4chan</td>
<td>0.022601187</td>
</tr>
<tr>
<td>globaloffensive</td>
<td>0.021705922</td>
</tr>
<tr>
<td>science</td>
<td>0.020803094</td>
</tr>
<tr>
<td>tifu</td>
<td>0.019677967</td>
</tr>
<tr>
<td>funny</td>
<td>0.019391723</td>
</tr>
<tr>
<td>jokes</td>
<td>0.018052954</td>
</tr>
<tr>
<td>food</td>
<td>0.017792952</td>
</tr>
<tr>
<td>mildlyinfuriating</td>
<td>0.016545503</td>
</tr>
</tbody>
</table>
PageRank of Subreddits

PageRank Score

Sorted Indices
Standard Deviation

![Bar chart showing standard deviation between computed position and actual position.](chart.png)
Why We were Wrong

• Users tend to link to more specific subreddits instead of the larger, broader subreddits

• Limited Data Set:
  • 24 hours worth of data from a website over a decade old
  • Some Stats on Reddit as of 2015:
    • Total Number of Posts: 190,227,552
    • Total Number of Subreddits: 853,824
    • Total Number of Comments: 1,715,454,785
  • We collected .0076% of all comments on Reddit
Sources


http://redditlist.com/
Comparison of Political Beliefs of Presidential Candidates on Twitter

Rudmila Rahman, Taylor Moore, Siddharth Dave
In the light of the election season, it’s always good to know where your nominees stand on various issues. It’s also important to know where our current President stands on the same issues. If you like the current President’s stance on the issues you will want to vote for candidate that agrees.
Our Goal

* To show you where our current President stands on five issues that we have taken into consideration.
* Get the views of the current Presidential hopefuls viz, Cruz, Sanders, Clinton and Trump.
* Show how their views differ so you know how your favorite candidate compares.
Issues in consideration

- Abortion
- Gun Control
- Gay Marriage
- Universal Healthcare
- Mass deportations/Immigration Reform
Why do I care?

* Because you’re going to vote (or at least should).
* Do you really want almost nearly a decade of someone who believes global warming is a hoax?
Used platform
Process
Collecting data

- Twitter crawler streaming API
- NLTK
- Sentiment analysis
Datasets

- Extracting tweets from POTUS
- Extracting tweets from the candidates
- Filtered tweets
While you could get your information on TV, remember that the channels do alter and twist facts.

These days politicians (like some real estate tycoon) use social media to connect to the people.

Twitter reflects opinions over a longer period of time than a published statement.

All of our subjects are twitter friendly.
For our analysis we take in a politician's twitter ID as input for a given politician and the POTUS.

We pass in keywords relevant to an issue eg: #Obamacare, #ProChoice, #BUILDTHEWALL

We determine the similarity of that politician's tweets to that of the POTUS.
Results – Obama

Abortion

Sentiment Analysis Results

The text is **pos**.

The final sentiment is determined by looking at the classification probabilities below.

**Subjectivity**
- neutral: 0.1
- polar: 0.9

**Polarity**
- pos: 0.7
- neg: 0.3

Gay Marriage

Sentiment Analysis Results

The text is **pos**.

The final sentiment is determined by looking at the classification probabilities below.

**Subjectivity**
- neutral: 0.2
- polar: 0.8

**Polarity**
- pos: 0.7
- neg: 0.3
<table>
<thead>
<tr>
<th>Subjectivity</th>
<th>Polarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>neutral: 0.2</td>
<td>pos: 0.7</td>
</tr>
<tr>
<td>polar: 0.8</td>
<td>neg: 0.3</td>
</tr>
</tbody>
</table>

The final sentiment is determined by looking at the classification probabilities below.
Results – Clinton

Abortion

Sentiment Analysis Results

The text is pos.

The final sentiment is determined by looking at the classification probabilities below.

Subjectivity
- neutral: 0.2
- polar: 0.8

Polarity
- pos: 0.6
- neg: 0.4

Gay Marriage

Sentiment Analysis Results

The text is pos.

The final sentiment is determined by looking at the classification probabilities below.

Subjectivity
- neutral: 0.3
- polar: 0.7

Polarity
- pos: 0.7
- neg: 0.3
Clinton

Universal Healthcare

Sentiment Analysis Results

The text is **pos**.

The final sentiment is determined by looking at the classification probabilities below.

**Subjectivity**
- neutral: 0.4
- polar: 0.6

**Polarity**
- pos: 0.5
- neg: 0.5

Gun Control

Sentiment Analysis Results

The text is **pos**.

The final sentiment is determined by looking at the classification probabilities below.

**Subjectivity**
- neutral: 0.1
- polar: 0.9

**Polarity**
- pos: 0.7
- neg: 0.3
Results – Cruz

Abortion

Sentiment Analysis Results

The text is **neg.**

The final sentiment is determined by looking at the classification probabilities below.

**Subjectivity**
- neutral: 0.1
- polar: 0.9

**Polarity**
- pos: 0.2
- neg: 0.8

Gay Marriage

Sentiment Analysis Results

The text is **neg.**

The final sentiment is determined by looking at the classification probabilities below.

**Subjectivity**
- neutral: 0.2
- polar: 0.8

**Polarity**
- pos: 0.2
- neg: 0.8
Cruz

Gun Control

Universal Healthcare

Sentiment Analysis Results

The text is **neg**.

The final sentiment is determined by looking at the classification probabilities below.

**Subjectivity**
- neutral: 0.4
- polar: 0.6

**Polarity**
- pos: 0.4
- neg: 0.6

Sentiment Analysis Results

The text is **neutral**.

The final sentiment is determined by looking at the classification probabilities below.

**Subjectivity**
- neutral: 0.5
- polar: 0.5
Results – Sanders

Abortion

Sentiment Analysis Results

The text is **pos**.

The final sentiment is determined by looking at the classification probabilities below.

**Subjectivity**
- neutral: 0.3
- polar: 0.7

**Polarity**
- pos: 0.7
- neg: 0.3

Gay Marriage

Sentiment Analysis Results

The text is **pos**.

The final sentiment is determined by looking at the classification probabilities below.

**Subjectivity**
- neutral: 0.3
- polar: 0.7

**Polarity**
- pos: 0.7
- neg: 0.3
Universal Healthcare

Sentiment Analysis Results

The text is neutral.

The final sentiment is determined by looking at the classification probabilities below.

Subjectivity
- neutral: 0.5
- polar: 0.5

Gun Control

Sentiment Analysis Results

The text is pos.

The final sentiment is determined by looking at the classification probabilities below.

Subjectivity
- neutral: 0.1
- polar: 0.9

Polarity
- pos: 0.7
- neg: 0.3
Result – Trump

Abortion

Sentiment Analysis Results

The text is **pos**.

The final sentiment is determined by looking at the classification probabilities below.

**Subjectivity**
- neutral: 0.4
- polar: 0.6

**Polarity**
- pos: 0.5
- neg: 0.5

Gay Marriage

Sentiment Analysis Results

The text is **neutral**.

The final sentiment is determined by looking at the classification probabilities below.

**Subjectivity**
- neutral: 0.5
- polar: 0.5
Trump

Universal Healthcare

Sentiment Analysis Results

The text is neg.

The final sentiment is determined by looking at the classification probabilities below.

Subjectivity
- neutral: 0.1
- polar: 0.9

Polarity
- pos: 0.1
- neg: 0.9

Gun Control

Sentiment Analysis Results

The text is neg.

The final sentiment is determined by looking at the classification probabilities below.

Subjectivity
- neutral: 0.3
- polar: 0.7

Polarity
- pos: 0.3
- neg: 0.7
In our research we found out that candidates from similar political parties tend to have the same opinion on most major issues. (Homophily)

Hillary Clinton agrees most with President Obama

While Donald Trump agrees the least.

In the future, could expand to show

- Politicians with similar views as candidates in local races
- Suggest the candidates that have the closest views to a given Twitter user
Social Media and Bullying

Who is Affected?

As many as 25% of teenagers have experienced cyberbullying at some point. Online bullying is very similar to the bullying that happens in school: Both behaviors include harassment, humiliation, teasing and aggression. “The use of electronic communication to bully a person, typically by sending messages of an intimidating or threatening nature.”

Why it Matters

Presents unique challenges in the sense that the perpetrator can attempt to be anonymous, and attacks can happen at any time of day or night. A recent study in the journal JAMA Psychiatry suggests that both victims and perpetrators of bullying can feel long-lasting psychological effects.
Why Twitter?
71% of Teenagers use Twitter

56% of Teenagers use it daily

84% of Teenagers post once a day
Natural Language Processing
Our main focus was collecting data from the DMV.

An overall of 7,889 tweets were collected for initial analysis and classification.
Plan of Attack

Checked for certain words

Determined Sentiment

Determined what type of Attack occurred
Step 1: Collecting the Tweets

The 7889 Tweets

➔ The tweets were collected using the Tweepy API. We created a library of “bad words” to filter and only collect tweets that contain at least one of the words.

➔ We kept track of various attributes, such as location, user ids, and the tweet itself.
Predictions

From the list of “bad words”

➔ Out of the seven categories, the list for sexual bad words was the longest!

➔ Rate at which tweets were collected

➔ Top categories: Insults, Sexual, and Defamation
Step 1: Collecting the Tweets

The Method

➔ We used and incorporated functions from the Tweepy API to listen to live tweets and limited the search to be within the DMV area.

➔ The initial tweets we got with the code made sure another user was mentioned. "Focus on user mentions with tweets."

➔ We added all the tweets into a dictionary to store them for the next step.
Step 2: Determining The Sentiment

The Method

➔ We used the Aylien API client to determine the sentiments from the tweets.

➔ We categorized each tweet as Positive, Neutral, or Negative.
Step 2: Determining The Sentiment

Results

20% of Tweets were Positive
57% of Tweets were Negative
23% of Tweets were Neutral
Step 3: The Categories

7 Categories

- Insult
- Sexual
- Rejection
- Defamation
- Threat
- Religious
- Intelligence

We then categorized negative and neutral tweets into common forms of bullying.
Step 3: The Categories

The Method

➔ From researching we found that in order to determine bullying we needed the tweet to match one of the below cases:

+ BadWord!, Pronoun...
+ You BadWord...
+ I BadWord Pronoun...
+ Pronoun BadWord...
+ Pronoun ... BadWord ...

A personal pronoun appearing near profanity.

➔ “you” and “yourself” important pronouns
## Step 3: The Categories

### Results

<table>
<thead>
<tr>
<th></th>
<th>Insults</th>
<th>Sexual</th>
<th>Rejection</th>
<th>Defamation</th>
<th>Threat</th>
<th>Religious</th>
<th>Intelligence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Virginia</strong></td>
<td>593</td>
<td>381</td>
<td>36</td>
<td>209</td>
<td>403</td>
<td>85</td>
<td>59</td>
</tr>
<tr>
<td><strong>Maryland</strong></td>
<td>316</td>
<td>210</td>
<td>20</td>
<td>136</td>
<td>230</td>
<td>52</td>
<td>24</td>
</tr>
<tr>
<td><strong>DC</strong></td>
<td>246</td>
<td>263</td>
<td>87</td>
<td>153</td>
<td>207</td>
<td>86</td>
<td>54</td>
</tr>
</tbody>
</table>
Step 3: The Categories

Results

30% of Tweets were Insults
22% of Tweets were Sexual
4% of Tweets were about Rejection
13% of Tweets were about Defamation
22% of Tweets were Threats
6% of Tweets were about Religion
3% of Tweets were about Intelligence
By creating software that easily detects and categorizes these attacks, we can later incorporate them into every profile:

- Start monitoring if a specific profile receives a significant amount of attacks
- Notify a person’s closest connections
- Develop safety precautions to handle possible situations based on the category flagged