DS 501 CASE STUDY 1

TWEETING FOR HILLARY

September 22, 2016

Matthew Beaulieu
Yousef Fadila
Meng Li
Monica Tlachac

Contents

MOTIVATION

More and more people are turning to social media for political political news. As many people get their political news from social media as local TV, both of which are more popular than any news source other than cable TV (Gottfried). Thus, social media can yield important insights into politics. PromptCloud, a data service, emphasises the importance of a compelling campaign and claims "the more compelling campaign is a direct result of better data collection, analysis and smart decision making" (PromptCloud). Given the popularity of social media for political news, social media could contain important data useful for political campaigns.

For instance, every hour there are 333 unique tweets containing #hillaryclinton (RiteTag). These tweets contain a wealth of information that could be useful in for Hillary Clinton's presidential campaign. In fact, Peter Greenberger, the director of global political sales for Twitter, says "In 2012, Twitter was the real-time spin room for politics and drove the narrative with reporters" and this "will make Twitter a core component of list building, fundraising, persuasion, and mobilization for political campaigns in 2014" (Williams).

Even without complicated analysis twitter can yield impactful results about campaigns. The 2016 presidential primaries could have been predicted by the amount of tweets about each candidate from the top six news networks (Simms). With more complicated analyses, information gathered from tweets can not only predict but also guide political campaigns. As such, campaign managers should take advantage of the information captured by Twitter.

This report will contain information on how data from collected tweets can be useful to Hillary Clinton's presidential campaign. Simple analysis, such as word count, can highlight popular topics and issues that should be addressed. A sentiment analysis could shed light on how different voters view Clinton. Combined with geographic locations, this indicates where campaigning efforts should be focused. The tweets provide a great amount of real time data from the public and, if used correctly, could be very influential in campaign strategy.

The Data

The first task is to gather tweets about the presidential candidate Hillary Clinton. However, there are a lot of people named “Hillary” and “Clinton” could refer to an entire family of political figures. Also, given the language of twitter, finding many tweets that contains the text “Hillary Clinton” is very unlikely. Thus, we filtered out tweets specifically mentioning “@HillaryClinton”. While some tweets about Hillary Clinton might be missed, this ensures that every tweet gathered is about Hillary Clinton.

The tweets useful to a United States presidential campaign are from the United States. Also, while not all tweets from the United States about Hillary Clinton are in English, language translations of tweets are very inaccurate due to slang used in twitter so only tweets in English are considered for the analysis.

In all, ~ 15500 tweets were collected from the Twitter streaming API over the span of September 15, 2016 to September 20st, 2016.

The Text

The text from the collected tweets about Hillary Clinton was isolated and all words were changed to lowercase. Then additional filters were run to remove punctuation, @, #, RT, and hyperlinks.

Word Frequency

Word frequency can yield useful information. An analysis of the text that are in tweets about Hillary Clinton gives insight into concerns of the general population. This can allow the campaigns to address specific topics important to potential voters.

The following table shows the frequency of the top 30 words . The frequency was calculated using the function ‘freq’ from the Python package Wikiwords. The ratio is the value divided by the frequency, where value is the amount of each given word. Ratio is preferred over frequency for word analysis because it yields words not as frequent in the human language. The below table is sorted by the ratio.

Top 10 Most Popular Words

Words	Counts
trump	1240
hillary	915
like	679
campaign	589
dont	541
vote	538
people	536
poll	487
us	462
u	454

Obviously this is not a perfect method for gathering the most important words in the tweets; as wikiwords did not recognize ‘amp’ and so assigned it the highest frequency. On the positive side, sorting by ratio yields words not as frequent in the human language so they are useful for analysis purposes.

From this table, it appears as though Donald Trump, Hillary Clinton‘s adversary, is a very popular topic, even in tweets about Clinton. Trump‘s name (appearing as ‘trump’ or ‘trumps’) features in 2.4 percent of the gathered tweets.

A very troubling part of this table that the campaign should address is the high quantity of negative words. It is important to ensure that potential voters don’t think Clinton is ‘unfit’ for presidency, a ‘liar’, or is a hypocrite (‘hypocrisy’). Additionally, words such as ‘cant’, ‘doesnt’, ‘didnt’, ‘wont, ‘dont’ and ‘isnt’ all suggest that people may have a lack of confidence in Clinton.

Lastly, there are key words pointing to specific issues: ‘bodyguards’, ‘benghazi’, ‘poorest’, ‘blackmail’, ‘pneumonia’, and ‘audiobooks’. These suggest that these specific things are important to and on the minds of the people who tweet about Clinton, and thus should be addressed.

Popular Tweets

Unfortunately, the tweets were pulled from the twitter streaming API, meaning they were saved as they appeared. Thus, none of the tweets were retweeted or favorited by any users. A tenth of the tweets were pulled again, this time from the REST API; the whole set was not pulled due to rate limits. The below table show the most popular tweets of this subset.

The Most Popular Tweets of The Subset

Retweet Count	Text
17540	RT @YoungDems4Trump: Show this to those who don’t fully grasp the severity of @HillaryClinton’s email breach. Because the media won’t.
6426	RT @bfraser747: BREAKING NEWS \n\nNobody is buying it anymore Lying’ @HillaryClinton &amp; why should they? \n#Trump2016 #MAGA
2149	RT @SenGillibrand: Love this.\n \n"@HillaryClinton has spent her life fighting for children—here are 8 ways she’s changed their lives.”
2076	RT @Harlan: Whoa, this is a BRUTAL take on @HillaryClinton courtesy of Colin Powell.\n\nYes it’s real.\n\nh/t @bennyjohnson
1804	RT @RealBenCarson: @HillaryClinton hiding her diagnosis of pneumonia may be a mistake from which there is no recovery.
1758	RT @DrJillStein: Obama admin including @HillaryClinton auctioned offices to the highest bidder - even FCC chair. #DNCleak #PayToPlay
1587	RT @FoxNews: Fox News Poll: @realDonaldTrump leads @HillaryClinton on ’Trust to do a better job on the economy.’ #SpecialReport
1586	RT @FoxNews: Fox News Poll: @realDonaldTrump leads @HillaryClinton on ’Trust to do a better job on the economy.’
1392	RT @Jorge_Silva: Celebrating #HispanicHeritageMonth with @HillaryClinton’s quotes read by @DNCLatinos.
1385	RT @ImWithYou010: @hillaryclinton talking about building a WALL and deporting MEXICANS LOL! please make this go viral! #Trump2016

The more times a tweet is retweeted, the more potential voters are exposed to the sentiments within the text. Unfortunately for Clinton, 8 of the top 10 texts are negative. There are some different concerns presented in retweeted tweets over word frequency, such as the ‘email breach’ and ‘economy’. However, some issues such as ‘pneumonia’ are present in both. This table also shows that Clinton may have the support of Hispanic voters and should highlight her work with children.

Screen Name	Count
HillaryClinton	15421
realDonaldTrump	2718
FoxNews	1532
POTUS	503
CNN	481
politico	283
timkaine	263
FLOTUS	245
MSNBC	244
USAneedsTrump	235

Hashtag	Count
#MAGA	385
#ImWithHer	351
#SpecialReport	209
#NeverHillary	178
#DNCleak	177
#HispanicHeritageMonth	163
#tcot	156
#Trump	149
#TrumpPence16	125
#HillaryHealth	102

Popular Entities

These are the top ten most popular screen names and hashtags that appear in the collected tweets. This shows the most common screen names and hashtags associated with Clinton. Given the word frequency, it is not surprise that Trump's name appears in the screen names and hashtags multiple times. In fact, Trump's slogan (Make America Great Again) is the most popular hashtag in tweets about Clinton. Overall the hashtags, similar to the word frequency, are concerningly negative towards Clinton. Additionally, there are many news sources in the list of screen names so it shows that Clinton is mentioned by news sources.

SENTIMENT ANALYSIS

A sentiment analysis is useful because it reveals what potential voters like and dislike about Hillary Clinton. When all tweets are analyzed together, it is only a guess on whether things are mentioned positively or negatively. With knowledge about the sentiment behind tweets, it can allow for a campaign to better target their efforts. Specifically, the campaign can find top concerns in each area.

Before starting the sentiment analysis, the data needs to be cleaned by adding additional filters. All tweets containing mentions of Donald Trump were removed. This was done because then any positive or negative sentiment could have applied to either Hillary Clinton or Donald Trump. There will still be outliers where the sentiment extracted could be aimed at another politician or idea, but by removing tweets that mention Trump, it reduces the amount of noise in the data.

Python's NLTK text classifier was used to perform the sentiment analysis on the text portion of the tweets; though this classifier was not designed specifically for tweets. This sorted the tweets into three categories: positive, negative, or neutral. For the following results, only the positive and negative tweets were used. This results in a much smaller dataset than the original.

Geographic Analysis

Knowing the sentiment of potential voters by geographic area could be very useful. It would allow campaigns to know where to focus their campaigning efforts. It could also offer an additional perspective to polling data. However, most tweets do not include geographical location in a useable format. Thus, the below table and map are not representative; if the positive ratio and negative ratio are both 0, that state had no tweets with location data and sentiment. The usefulness of this geographic analysis could be improved with a greater number of tweets and more advanced location mining techniques.

Even accounting for some states having no data, this map does not represent known state sentiment about Hillary Clinton (light blue represents a negative ratio and purple represents a positive ratio). Ideally, if a sentiment analysis by state was accurate, it would look like the following map from New York Times, which depicts current polling data. Thus, we must conclude that a sentiment/geographic analysis is not useful for predicting poll data. However, with more tweets and more location information, this approach could still yield good results.

Hashtags

After the text of the tweets were sorted by analysis, the most popular hashtags for both the positive and negative tweets were analyzed. The hashtags used more than 30 times are displayed in tables below.

Top Hashtags in Positive Tweets	Count
#HispanicHeritageMonth	118
#ImWithHer	107
#MAGA	72
#tcot	65
#Democrats	50
#RedNationRising	46
#WakeUpAmerica	43
#NeverHillary	32
#HillaryClinton	31

Top Hashtags in Negative Tweets	Count
#ImWithHer	74
#LatinosWithTrump	51
#AmericansUnitedForTrump	49
#MAGA	42
#NeverHillary	39
#CrookedHillary	38

Clearly there is some error in how the tweets were classified, as there are negative hashtags in the positive tweets and positive hashtags in the negative tweets. Additionally, some hashtags, such as ‘ImWithHer’ and ‘MAGA’ feature at the top of both categories, despite ‘ImWithHer’ being positive and ‘MAGA’ being negative.

Given these results, the sentiment algorithm needs to be retrained. This was done manually based on the hashtags to yield better results. ‘ImWithHer’, ‘TurnNCBlue’, ‘StrongerTogether’ are the chosen positive hashtags. ‘WakeUpAmerica’, ‘NeverHillary’, ‘ImInHer’, ‘MAGA’, ‘CrookedHillary’, and ‘AmericansUnitedForTrump’ are the chosen negative hashtags. The other hashtags were sorted based on their appearance with these chosen hashtags. Thus, the below tables are the results after being trained with select hashtags.

Top Hashtagein POSITIVE Tweets - VERSION B	Count
#ImWithHer	293
#StrongerTogether	78
#Hillary2016	38
#GOPdebate	25
#imwithher	23
#TurnNCBlue	23
#Vote	14

Top Hashtags in Negative Tweets	Count
#MAGA	385
#NeverHillary	178
#CrookedHillary	100
#NYPD	83
#LatinosWithTrump	70
#AmericansUnitedForTrump	70
#WakeUpAmerica	65

Conclusion

Twitter data showed great promise as being helpful to campaign managers. They permit for a large amount of data to be gathered quickly, and live. Given the large amount of people who obtain political news from Twitter, having access to that information is important. Individual types of analyses vary in their successfulness.

Word frequency of the the text in the entire dataset was helpful because it highlighted topics about Hillary Clinton that are being talked about. For instance, given the word 'pneumonia', potential voters are clearly concerned about Clinton's health. Similarly, 'benghazi' is a topic that the campaign may want to focus on more.

Viewing the most retweeted tweets gives an idea of what sort of information about Clinton that many people are exposed to. Overall, these tweets were negative, something that the campaign may want to try to combat with more positive tweets. The most common hashtags in the tweets were also negative, though clearly 'ImWithHer' has gained a following, a fact that a campaign can utilize to spread more positive messages that will reach a larger audience.

The sentiment analysis was limited due to the amount of tweets available that were classified as positive or negative. For that reason, and the lack of geographic information, tweets are not the proper tool to determine geographic locations to focus campaigning efforts. However, there is great potential for using sentiment analysis using hashtags on word frequency, though a better trained classifier and more tweets would yield more impactful results.

In conclusion, political campaigns should utilize the large amount of data available in Twitter. While some of the analyses used in this report were successful and some were unsuccessful, Twitter has great potential for changing the future of how political campaigns gather data about potential voters.

REFERENCES

The Electoral Map Looks Challenging for Trump Available at:<http://www.clinchem.org/content/56/10/1649.full.pdf>

Sentiment Analysis with Python NLTK Text Classification Available at: <http://text-processing.com/demo/sentiment/>

Twitter predicted the results of the Presidential Primaries. Could it predict the general election, too? Available at: <https://blog.parsehub.com/what-27000-tweets-can-tell-you-about-the