Fundamentals of Data Science - Assignment 1 Did Twitter data predict Trump's election?

Daphnee Chabal[1], Max Crous[1], Savvina Daniil[1], Albert Folch[1], and Elena Gkerpini[1]

[1]University of Amsterdam, Amsterdam, The Netherlands

Group 15

Abstract. A collection of tweets written between August-September in 2016 related to the US presidential election of that year were analyzed. We investigated whether political discourse on Twitter matches the real-world and whether Trump's victory could have been predicted by the political discourse on Twitter. Twitter users were more male than the general population and some states were overrepresented in number of tweets. Among other metrics studied, Trump had more trending hashtags on his side, and tweets overall expressed more positive sentiment towards him while proportions of Clinton voters was correlated with proportions of positive sentiments expressed towards her. We also provide a network of hashtag co-occurrence revealing tag clusters with some overarching theme or sentiment. A post-hoc finding revealed an overwhelming debate about Trump's Mexican wall the day of Trump's visit with the president of Mexico.

Keywords: Twitter · Sentiment Analysis · Topic Analysis · US Presidential Election 2016 · Hashtag Networks.

1 Introduction

1.1 Context

Twitter's central role in the US Presidential campaign of 2015-2016 was unmistakable [1], with more than a billion tweets written and stored about the election in just three months (August-November 2016)[2]. A great deal of the political debates between candidates and the general public resided on Twitter[3]. Many American adults reported getting their weekly updates on the campaigns from Twitter[4].

One may therefore expect to extract revealing statistics, reflecting the political opinion landscape and ultimately the results of the elections[5][6]. However, recent research shows that Twitter users do not represent the general population [7][8]. There is also a discrepancy between Clinton's win of the popular vote by more than 2.8 millions votes and Trump's notorious omnipresence and alleged popularity on social media[9]. Moreover, Trump's ultimate victory surprised a vast majority of the American and international community, off- and online.

1.2 Research Questions

This report therefore aimed at investigating the relationship between Twitter user data and population-wide candidate preference by investigating two main questions.

Question 1. Are Twitter users who wrote about the election representative of the general population, especially among US voters?

Question 2. Do tweet sentiment, topic, and engagement predict the result of the election and if so, which metrics revealed a candidate's advantage over the other in the public opinion?

2 Methods

2.1 Dataset

For this project, a dataset of 657,307 tweets from August 12th to September 12th, 2016 was used. This dataset was extracted by monitoring the official Twitter accounts of Hillary Clinton and Donald Trump (@HillaryClinton, @realDonaldTrump) and the following election-related hashtags: #maga, #trumppence16, #hillaryclinton, #hillary, #crookedhillary, #donaldtrump, #dumptrump, #nevertrump, #imwithher, #neverhillary, #trump. The dataset was converted from a JSON format to a Python Pandas dataframe.

2.2 Preprocessing

Preprocessing was applied in tasks that rely on the semantic content like topic analysis and sentiment analysis. This was done by extracting the text-body of each tweet along with hashtag-words (i.e. keeping the words directly following hashtags) by using the NLTK library and regular expressions. Regular expressions were also used to remove links, mentions (i.e. @ followed by usernames), hashtags, other non-alphabetical characters and emojis. Punctuation marks and English stop words were removed as they are deemed irrelevant due to their frequent occurrence. Only the root of each word is kept by using lemmatization. All characters were converted to lower case and words shorter or equal to 3 characters were removed.

2.3 Tweet Classification

2.3.1 Text-body A sample of 593,268 tweets which were written by US-based users and written in English was compiled for general sentiment and topic analyses of the text-body of tweets. The sentiment analysis used the Naive Bayes Classifier which is a probabilistic classifier based on applying Bayes' theorem with strong independence assumptions between the features, words in this case. Topic analysis used a Latent Dirichlet Allocation (LDA) topic model.

2.3.2 Hashtags A third of all tweets (244,082 out of 657,307) used hashtags. Out of these hashtags a list was compiled of the most popular tags and of tags that contained sentiment. We sampled 50 tweets per popular hashtag to make sure we correctly classified the sentiment associated with each hashtag. The co-occurrence of hashtags was investigated by using a Python library that implements the Louvain cluster detection algorithm and by plotting each unique tag as a node in a graph, with edges running between tags that appeared together in tweets.

2.3.3 Mexican-topic analysis A sample (n = 52,636) of all English and Spanish tweets written on August 31st and September 1st 2016 was compiled, regardless of user-location, after we noticed a peak in Mexico-based users tweets about the election, related to Trump's meeting with the Mexican President on August 31st. Trending topics over those days were also analyzed using an LDA model.

3 Results

3.1 Twitter users

There was a difference in the amount of tweets written per person in different states. Individuals in Washington DC, California, or Texas tweeted significantly more than inhabitants in Wyoming and South Dakota among other states (when controlling for population differences). We combined this index-metric of Twitter engagement per state (divided by population) with actual results of the elections per state (red/blue) and yielded the image seen in Figure 1.
Washing DC residents were disproportionally engaged[10]. Though its population is relatively small, Twitter users there seemed to be highly interested in the topic of US Elections. Meanwhile, only 30% of states that eventually voted for Trump had have an index higher than the median of the indexes, while 76% of Clinton's winning states did.
Gender assigning can be done on a tweet or unique users bases. In the first case, the gender_guesser Python library was assigned a gender to the first names of 415,077 tweets, for which 62.37% were found to be male and 37.63% female (Figure 2). For unique users, 66.68% were male and 33.32% were female. These statistics are in disagreement with the US twitter population, which was 53% male and 47% female[11]. This could either mean that there was significantly more male participation in election related tweets, or that many female names were not correctly recognized by the gender_guesser. After inspecting 100 of the unrecognized names, we found that very few names could be assigned a gender. There were 8 clearly male names and 7 clearly female names. This supports the greater participation hypothesis. Using the census data with the Python ethnicolr library to compute probabilities for ethnicity based on the second names, yields the following results (out of 101k tweets that could be assigned an ethnicity) White: 45,272, Black: 14,840, Asian: 29,654, Native American: 4,754, Two Races: 7,044.

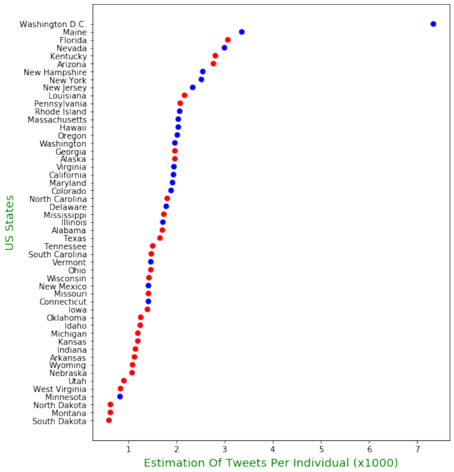

Fig. 1: Estimation of Tweets per Individual in Each State. Blue represents Clinton Voters, Red represents Trump voters

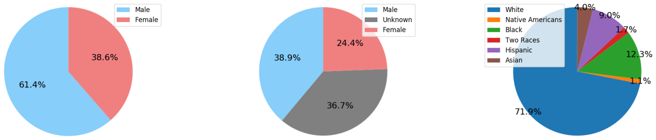

Fig. 2: Gender and Ethnicity of Usernames

When normalizing for the number of tweets posted (unique users are counted once), the percentages remain almost identical (Figure 2).

3.2 Hashtag Distribution

Hashtags followed a power law distribution as most hashtags were only used a few times but the 10 most popular tags account for 29.95% , the first 100 for 49.94% and the first 500 for 65.54% (Figure 3). This means that focusing on sentiments of the top 100 hashtags will cover a large parts of our hashtag data.
The top 100 list was skewed. It contained 20 Pro-Trump hashtags (appearing 74,321 times) expressing either negative sentiment towards Clinton or positive sentiment towards Trump. Those Pro-Trump hashtags also trended more frequently overall. Meanwhile, only 12 hashtags were Pro-Clinton (appearing 47,221) with either negative sentiment towards Trump or positive sentiment towards Clinton. There also were 11 neutral hashtags which simply referred to the candidates (thus the sentiment of the tweet depends on the text-body and emojis). Here we present a subset of those lists.

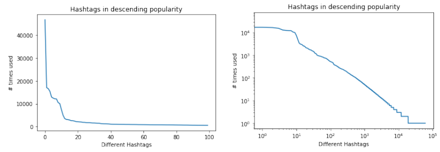

Fig. 3: Power Law in Hashtag Popularity (using log())

Pro-Trump/Anti-Clinton: MAGA, TrumpPence16, TrumpTrain, NeverHillary
Pro-Clinton/Anti-Trump: NeverTrump, ImWithHer, Hillary2016, UniteBlue.
Neutral: Trump, DonaldTrump, Hillary, Clinton, TrumpPence.

To expend our list of sentiment-strong hashtags, we looked for co-occuring tags which often occurred in the same tweets as a tweet would not use pro and against candidate x tags at the same time. We only included tags that appeared in at least 10 tweets and had been mentioned 90% of the time in one of the pro candidate hashtag lists, to avoid selecting neutral hashtags. This method proved successful. The number of pro-Trump hashtags increased from 20 to 417 and from 12 to 214 for Clinton. Although the skew only worsened, the hashtags did in fact belong in the candidate camp, as can be seen from a random sample from the added sets. **New Trump:** TrollingHillary **New Hillary:** TurnSenateBlue.

We also investigated whether we could find clusters of hashtags with similar meanings. We plotted each hashtag in the top 100 as node and drew edges between the hashtags that appeared together. With each co-occurrence the weight of theses edges increased and the more clustered these hashtags were plotted. To unpack the crowded resulting network, we removed insignificant edges (e.g. edges between nodes that albeit only co-occur once) and using backbone extract algorithms found in [12] and [13](Figure 4). The clusters that are found do in fact represent different overarching themes or sentiment. For example, cluster 3 contains state names and abbreviations and cluster 6 contains radically anti-Hillary tags.

3.3 Sentiment analysis: text body

The comparison of the insights from the tweets to the real data can be done in different ways. On the left hand-side of Figure 5, the ratio of positive to negative tweets for each candidate per state was calculated. Red states mean that Trump had a greater ratio than Clinton, while Blue states mean that Clinton had a higher ratio. States in white (North Dakota, South Dakota and Vermont) are those with too little data to make a concrete conclusion. In the middle of Figure 5, we can see the actual results from the elections, with red/blue signifying which candidate (Trump/Clinton) received the most electoral votes. Another way to correlate the findings to real data is by using the estimate income as seen in the right of Figure 5. Some of the state colors match the state colors of

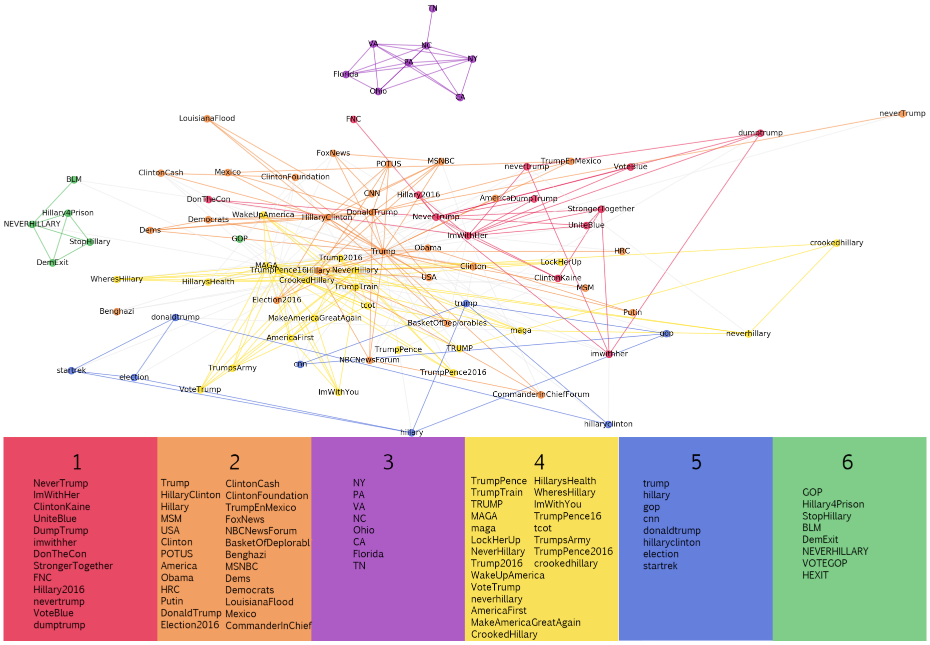

Fig. 4: Co-occurrence Network (Pruned)

the sentiment analysis in Figure 5.This map in conjunction with the real map of the elections suggests that the higher the estimate income is, the higher the likelihood is to vote for Donald Trump. Figure 6 displays the correlation between the findings and the actual results of the elections. This metric is critical to acknowledge the level of accuracy of the calculations depending on the tweets compared to reality in terms of predicting the elected president per state when users express themselves positively. There was a positive correlation between the proportion of Clinton voters per state and the proportion of pro-Clinton tweets per state. However, proportion of Trump voters per state was not associated with proportion of pro-Trump sentiment in Trump-referring tweets per state.

3.4 Topic analysis of text bodies

One way to analyze the topics the tweets contain is by separating them depending on who they draw attention to. The analysis of these two groups of tweets yielded different topics, of which the most significant were the following.

Tweets referring to Donald Trump:

Topic 1 is 0.019*"america" + 0.014*"great" + 0.012*"deplorable" + 0.012*"make" which points out the nickname Hillary Clinton used for Trump's supporters (de-

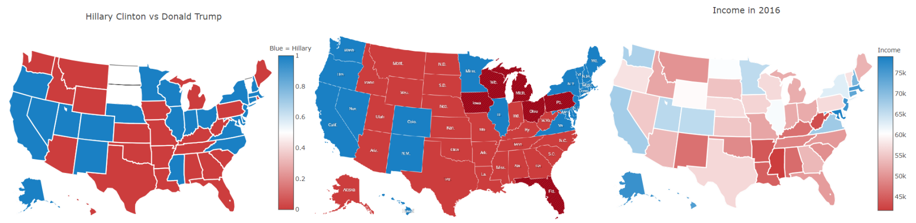

Fig. 5: Findings vs real data of the elections [14] vs estimate income in 2016 from real data [15]

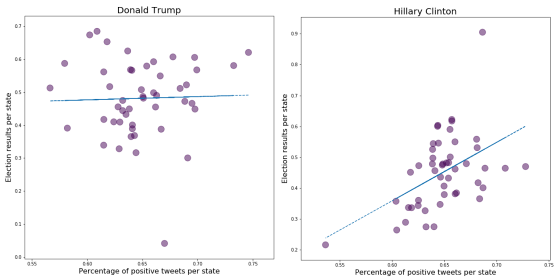

Fig. 6: State candidate preference in tweets vs actual voting results

plorable) and also cites Trump's slogan.

Tweets referring to Hillary Clinton:

Topic 1 is 0.016*"like" + 0.015*"pneumonia" + 0.011*"know" + 0.010*"sick" which is a reference to the time when Hillary Clinton got diagnosed with pneumonia which also caused her early departure from 2016's

3.5 Mexico

Considering Trump's controversial statements (and subsequent actions) regarding Mexico and Mexican immigrants, an analysis of tweets from Mexico seemed appropriate. After tracking relatively extreme action during two specific days, topic analysis was applied to correlate this to the meeting between Donald Trump and the Mexican President Enrique Peña Nieto.

3.5.1 Peak Day Figure 7 depicts the number of tweets from Mexico per day throughout the whole time frame of the data. Two out of the numbers are obvious outliers: on Wednesday the 31st of August and on Thursday the 1st of September,

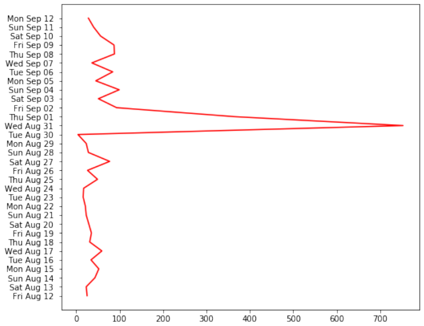

Fig. 7: Tweets from Mexico per Day

tweets from Mexico were 20 and 10 times the median, respectively. This indicates that during those two days certain events may have caused an intense reaction from Mexico's part. After some research, it was discovered that on Wednesday the 31st of August, Trump was invited in Los Piños, the residence of the Mexican president for his first meeting with him after announcing his candidacy.

3.5.2 Mexico Topic Analysis Topic analysis was used on all tweets of those two days written in either English or Spanish language (not just the ones from Mexico) to determine whether the meeting was trending on Twitter at that point. It yielded the following two topics:

Topic 1 is 0.039*"mexico" + 0.029*"wall" + 0.026*"trump" + 0.021*"president" + 0.015*"great", a reference to Trump's plan to build a wall on the border between Mexico and US.

Topic 2 is 0.037*"xico" + 0.017*"muro" + 0.017*"mexicano" + 0.014*"presidente" + 0.013*"visita" which alludes to Trump's meeting with the Mexican president and their discussion about the infamous wall.

It is obvious that the meeting provoked reactions from Tweeter's part and was the main topic of discussion during those two days. More specifically, the wall that Trump said would be built on the border of Mexico and US is the main point of the tweets.

Mexico's massive reaction is justified considering the Mexican President's refusal to directly criticize the racially charged comments made by Trump throughout his campaign. After all, "Enrique Peña Nieto entered the meeting as Mexicos most unpopular president since polling on presidential approval started in the mid-1990s"[16].

4 Conclusion

4.1 Findings

We conclude that Twitter users are not a representative sample of the general population, neither worldwide (i.e gender), nor US-based (i.e. ethnicity). The difference in amount of tweets per person in each state gave a disproportionately louder voice to Democrate-voting states on the platform. In real life, however, everyone gets one vote.

Trump's victory could have been predicted by a) our sentiment analysis of hashtags which revealed that Pro Trump/Anti Clinton hashtags were the most trending, b) our sentiment analysis of text-bodies where tweets talked more often about Trump, positively or negatively, than Clinton, with overall more positive sentiment expressed towards Trump, and c) our topic analysis which revealed a focus on Trump's political message (i.e. make America great again) and on Clinton's physical sickness episode. Our state-wise sentiment analysis showed that generally people expressed more positive opinions about Clinton in states which gathered the most vote for her, but the same is not true for Trump. This leads us to conclude that Twitter was somewhat reliable in predicting Clinton's popularity among voters but Trump's overall dominance in the campaign and election process. Finally, a post-hoc finding that Twitter overwhelmingly responded to Trump's visit to Mexico with our topic-analysis, showing further the impact of Trump's political message (i.e. building a wall) on the platform.

4.2 Our limitations

There isn't a one-one correspondence between state-wise sentiment analyses and election results, as people expressed proportionally more positive opinions about Clinton than they did for Trump even in states which ended up winning the electoral college vote (e.g. Nebraska). However, our sentiment analysis indeed predicted election results in swing states such as Florida, Iowa, Pensylvania. The low volume of data for certain states may have flawed the results, making it harder to draw an accurate picture for states with little twitter users.

Tweets from Donald Trump himself were included when applying sentiment analysis. However, they are believed to only mildly have affected the results due to their small number (6).

Finally, potential tweets from news agencies were not excluded throughout the process. This limitation possibly caused an increase in the differentiation between the results of the analysis and the actual public opinion.

References

1. Kapko, M. 03/11/2016: Twitter's impact on 2016 presidential election is unmistakable, CIO Homepage, `https://www.cio.com/article/3137513/social-networking/twitters-impact-on-2016-presidential-election-is-unmistakable.html`. Last accessed 28 September 2018
2. Levy, G. 08/11/2016: Twitter Wins Big in 2016 Campaign, US News Homepage, `https://www.usnews.com/news/politics/articles/2016-11-08/more-than-1-billion-tweets-were-sent-about-the-election`. Last accessed 28 September 2018

3. Isaac, M., Ember, S.. 08/11/2016: For Election Day Influence, Twitter Ruled Social Media, The New York Times Homepage, https://www.nytimes.com/2016/11/09/technology/for-election-day-chatter-twitter-ruled-social-media.html. Last accessed 28 September 2018
4. Unknown Author 18/07/2016: Election 2016: Campaigns as a Direct Source of News, Pew Research Center Journalism & Media Homepage, http://www.journalism.org/2016/07/18/candidates-differ-in-their-use-of-social-media-to-connect-with-the-public/. Last accessed 28 September 2018
5. Bright, J., Hale, S.A., Ganesh, B., Bulovsky, A., Margetts, H., Howard, P.: Does Campaigning on Social Media Make a Difference? Evidence from candidate use of Twitter during the 2015 and 2017 UK Elections. (October 19, 2017).
6. Dizekes, P., 26/09/2016: Does Campaigning on Social Media Make a Difference? Evidence from candidate use of Twitter during the 2015 and 2017 UK Elections, Massachusets Institute of Technology News Homepage, https://news.mit.edu/2016/how-twitter-explains-the-2016-election-0926. Last accessed 28 September 2018
7. Mellon, J., Prosser, C.: Twitter and Facebook are Not Representative of the General Population: Political Attitudes and Demographics of Social Media Users. (February 29, 2016).
8. Mislove, A., Lehmann, S., Ahn, Y.Y., Rosenquist, J., N.: Understanding the Demographics of Twitter Users. Proceedings of the Fifth International Conference on Weblogs and Social Media, (January, 2011).
9. Beckwith D. - 02/02/17: United States Presidential Election of 2016, Encyclopaedia Britannica Inc., https://www.britannica.com/topic/United-States-presidential-election-of-2016. Last accessed 26 September 2018
10. Unknown Author - 09/11/2016: DC Voters Elect Gray to Council, Approve Statehood Measure, NBC Washington Homepage, https://www.nbcwashington.com/news/local/DC-Election-Statehood-Council-Seats-400275901.html. Last accessed 28 September 2018
11. Joyce G. - 25/01/2017: A Data Portrait of Tweeters in the USA, Brandwatch Analytics Homepage, https://www.brandwatch.com/blog/react-tweeters-in-the-usa/. Last accessed 28 September 2018
12. Serrano, M., Bogu, M., and Vespignani, A.: Extracting the multiscale backbone of complex weighted networks. PNAS **106**(16) 6483-6488 (April 21, 2009). doi:https://doi.org/10.1073/pnas.0808904106
13. Blondel, V., Guillaume, J.L,, Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. P10008 (March 4, 2008). doi: 10.1088/1742-5468/2008/10/P10008
14. Leip D., Wasserman D. - 09/08/17: Presidential Election Results: Donald J. Trump Wins, The New York Times Homepage, https://www.nytimes.com/elections/results/president. Last accessed 26 September 2018
15. Guzman G. - 09/2017: Household Income 2016, U.S. Department of Commerce, Census Boureau Homepage, https://www.census.gov/content/dam/Census/library/publications/2017/acs/acsbr16-02.pdf. Last accessed 26 September 2018
16. Jacobs B., Agren D. - 01/09/2018: Mexican president contradicts Trump's account of border wall discussion, The Guardian Homepage, https://www.theguardian.com/us-news/2016/aug/31/

`donald-trump-mexico-meeting-president-pena-nieto-immigration`. Last accessed 28 September 2018