Fundamentals of Data Science: Twitter case

Marit Beerepoot[10983430], Delano de Ruiter[11422874], Jessy Bosman[11056045], Jeffrey Bosman[12108693], and Bastiaan Timmeman[12052477]

Universiteit van Amsterdam, Science Park. 904, 1098XH Amsterdam, The Netherlands

Abstract. In this paper sentiment and topic analyses are being used to predict the outcome of the U.S. presidential elections.

Keywords: Sentiment analysis · Topic analysis · 2016 US presidential elections · twitter.

1 Introduction

People have used the internet for years to express their opinions on a wide variety of topics like religion, politics and much more. Traditionally these opinions would be posted on personal websites like blogs or message boards. Since the growing popularity of microblogging services however, people have started more and more to voice their opinions on these platforms. One of the biggest microblogging services today is Twitter. Since its inception in 2006 the number of tweets posted per day has risen to 500 million in 2013 [1]. This provides researchers with a fast pool of opinions and sentiment on numerous topics. The combination of textual messages with hashtags and mentions makes this data a good source for sentiment and topic analysis. A probable use of these types of analysis is to predict the outcome of elections based on the sentiment people feel towards participating candidates. The main research objective of this paper is to see if similarities can be found between tweets send prior to the 2016 U.S. presidential elections and the actual results. This is done based on 600.000 tweets from the weeks prior to the election. All these tweets are geotagged, therefore the state the tweet was posted in is known. Furthermore the findings of this paper have been correlated with state level data, like demographic and average income data. Another aim of this research was to extract the most prominent topics discussed in the tweets. This topic analysis was done in order to compare our results with the most prominent topics in the elections according to other traditional research. This paper is comprised of the methods that were employed to process and analyze the tweets, the results that were gathered and the processing method. Finally a conclusion to our main research question will be provided.

2 Methodology

This research can be divided into 4 sections: the preprocessing, comparison of the tweet results and the real results, the comparison of a topic analysis and the real important topics during the election and a correlation analysis. The methods for each section will be discussed separately.

2.1 Data preprocessing

The programming language Python was used for reading, processing and analysing the data. Before the data could be used, the data was read by using the JSON library and the result was stored in a Pandas Dataframe. After analysing the Dataframe it was noticed that there were a lot of bot tweets. The bot tweets can be distinguished from the other tweets by the composition of the tweets, because they only contain a mention or hashtag and a link. Since it is not desirable that the bot tweets have influence on the results, the bot tweets were removed.

After the bot tweets were removed a new column called clean text was added to the dataframe. This column contains lowercase text that is free of mentions, hashtags, hyperlinks and emojis. There was also a column added containing only the hashtags and one containing only the mentions, to make them easier accessible. The resulting Dataframe was saved as a new json file, so these steps didnt have to be executed every time the code was executed.

2.2 Does the ratio of Trump/Hillary voters per state correspond to the ratio of the tweets about Trump and the tweets about Hillary?

In order to find an answer to this question, it first had to be determined when a tweet can be seen as a tweet for Hillary or a tweet for Trump. The assumption was made than when a person is not voting for Trump, the person is voting for Hillary. So when a someone tweeted something negative about Hillary, it was assumed that the user voted for Trump, and the other way around. To know whether a tweet contains a positive or a negative content a sentiment analysis was executed. To execute this analysis, a tagged dataset with tweets has been downloaded [2] from the computer science domain of Stanford University. This dataset was used as a training and a test dataset for the classifier. After the classifier was trained and tested (with 75.8% accuracy), the classifier was used to determine whether the given tweets were positive or negative. The result was saved in a newly added column in the Dataframe. The result of this sentiment analysis and the usage of mentions and hashtags were used together to determine who gets the vote. When a user used the hashtag dumptrump, nevertrump or imwithher it was assumed that the tweet could be seen as a vote for Hillary and when maga, croockedhillary or neverhillary were used it could be seen as a vote for Trump. After checking the obvious hashtags, the more neutral hashtags and the mentions were checked. When a user mentioned @HillaryClinton or used the hashtags hillaryclinton and hillary and the tweet was labeled as positive in the

sentiment analysis, it was seen as a vote for Hillary. If the tweet was labeled as negative it was seen as a vote for Trump. The same was done the other way around with tweets that mentioned @realDonaldTrump or used the hashtag trumppence16, trump, donaldtrump. When both were mentioned/hashtagged the tweet was labeled as both. There were also tweets in the dataset that did not contain a mention or a hashtag. These tweets were replies and quotes of tweets with the mentions and the hashtags, but didnt contain a hashtag or mention themself. These tweets were hard to categorize to a specific person, therefore these tweets were disregarded in this analysis.

After the sentiment analysis and the labeling were finished, some plots were created to give some insight into the data. The data of the tweets was compared with the real election results data. First, the percentages of republican voters based on the tweets and on the real result were plotted in a map. Hereafter these percentages were plotted in a bar graph per state. Next a new column was added to the tweets dataframe that consisting of who would win the election based on the tweets. Lastly a confusion matrix was created to see how many states winners were estimated correctly.

2.3 Do the topics found in the tweets correspond with the topics discussed in the election?

Finding the topics discussed in the tweets helps us determine which topics voters find most important. The sentiment analysis already helped to determine whether a tweet is from a Hillary voter or a Trump voter. This question deals with finding these most discussed topics for both Hillary and Trump voters. Our findings can be compared with alternative literature to see if the most important topics are the same to most discussed topics found here. This might give some insight into the topics that made people choose a candidate. The preprocessing and sentiment analysis already cleaned the data somewhat, including a column that indicates whether the tweet is for Hillary or for Trump. Some additional cleaning had to be done, including the removal of stopwords, the removal of punctuation and lemmatizing the words.This was mostly done using the NLTK library. After that, the LDA model was made. The number of topics was set to 20. This seems like an appropriate amount for this case. The aim was to find some general topics like: education, health and military. Some experimentation showed that when using too little topics these dont become clear. When using too many topics these general topics get spread out. The amount of passes for the LDA model is set to 5 for both Trump and Hillary tweets. More passes could result in a stronger model, however this would require a large amount of time. An adequate insight of topics was obtained by using only 5 passes. After the different topics were found, a comparison was made with alternative literature, to further specify the found topics.

2.4 Are there correlations between demographic information about the states and the ratio Trump/Hillary tweets?

First an extra dataset was downloaded with demographic information about the states. After that the tweets were filtered, to only select the tweets that were sent from within the United States. Scatterplots were created to visually check if there is a correlation. A new Dataframe was created with the states and the information about the states. Percentages of tweets for Hillary and Trump were calculated per state using the Trump/Hillary rating (created as described in 2.2). After that scatter plots were created to find a correlation between the percentage of tweets for Trump and the immigrant population, the median of the family income, the median of the household income and the percentages of different skin colours. Lastly, the correlation coefficients were calculated to support the findings from the scatterplots.

3 Results and discussion

3.1 Does the ratio of Trump/Hillary voters per state correspond to the ratio of the tweets about Trump and the tweets about Hillary?

Based on the sentiment analysis, a prediction of voters could be made. The following ratio between Trump and Hillary voters per state was found:

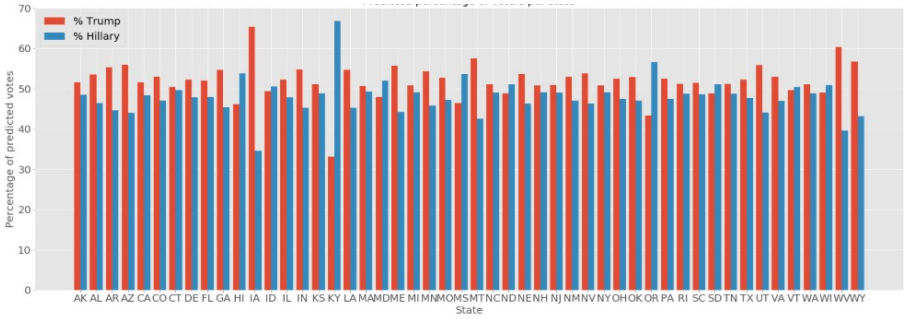


Fig. 1: Predicted percentage of voters per state

Most of the ratios are close to each other, with a few exceptions. 82% of the outcome the sentiment analyse predicted a higher voters percentage for Trump. Comparing to the real outcomes of the elections the following ratio is shown:

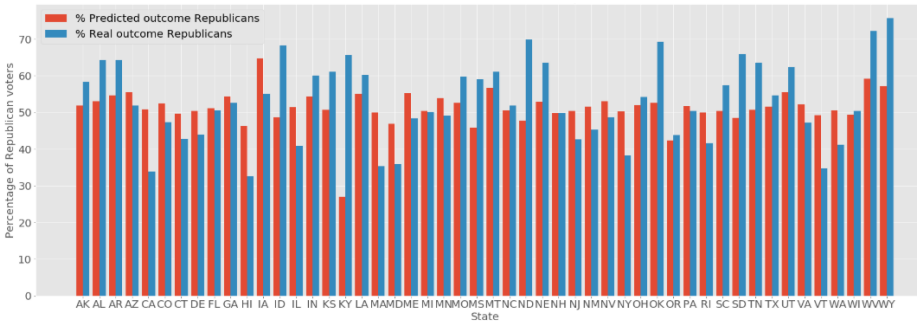
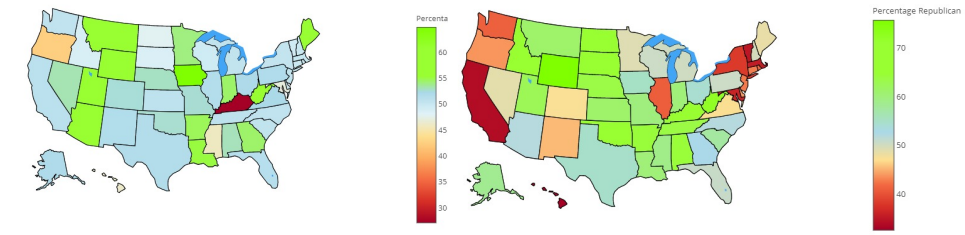


Fig. 2: Predicted percentage of voters per state

On the map the distribution of Republican voters, the first map is the outcome of the real elections the second is the predicted based on our sentiment analysis.



Predicted percentage Republican voters

Real percentage Republican voters

Fig. 4: Real and predicted percentage Republican voters

Looking at the images there is not an overwhelming difference between the right predicted states. The only visible difference is that the real election outcome per state is more pendulous, because the percentages are less in the mid range. 30 out of 50 states were predicted correctly. This resulted in an accuracy of 60%.

3.2 Do the topics found in the tweets correspond with the topics discussed in the election?

Our LDA model came up with 20 topics for each candidate. Two notable word clouds per candidate are shown below. The first two are for Hillary voters. The word cloud on the left shows a topic that includes healthcare and people, this topic seems very positive. The topic on the right is about Trump and his connection to Russia. This topic seems very negative.

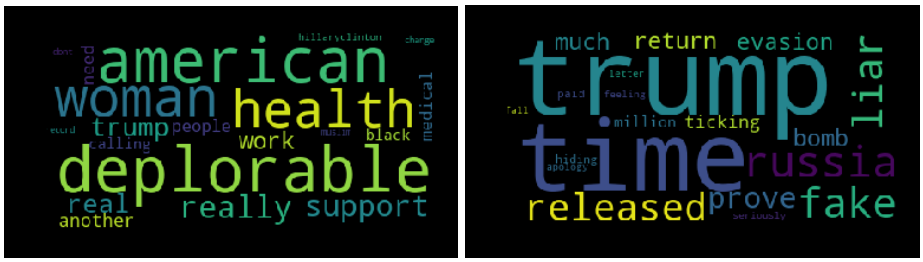


Fig. 6: Topic analyses wordclouds for Hillary

The topics generated for the trump tweets are shown below. Here, a topic about Russia and trump is visible again, but in a more positive light. The second word cloud shows a topic about Hillary and emails, which is more negative.

Fig. 8: Topic analyses wordcloud for Trump

From these word clouds it became clear that both sides talk about some of the same topics. Also both sides make accusations of the other candidate being a liar. The topics can now be compared with topics from a survey of Pew Research Center about top issues [3]. Only a few of our found topics in the LDA are in the list of top issues. Word clouds were found with information about terrorism, foreign policy, health care and immigration. Which where respectively The second, third, fourth and sixth most important topic. However, the other top issues are not addressed in our topics.

3.3 Are there correlations between demographic information about the states and the ratio Trump/Hillary tweets?

To answer this question different parameters of the states were collected [4–6]. These parameters then were correlated with the found predicted percentage of trump voters. It was expected that, because of Trumps negative stand on immigration and ethnicity, states with a higher population of immigrants or a large variety of races would correspond with a lower vote rate for Trump. The correlations are shown in the figure below:

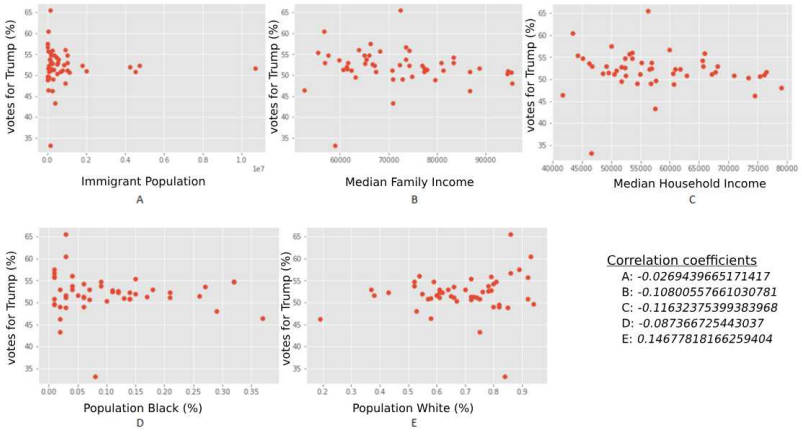


Fig. 9: Correlations between different demographics

However in contrast to the expectation, no correlations between Trump and ethnicity, immigration or income were found. With correlation coefficients ranging from -0.12 to 0.15, no support for correlations were found between state data and the percentage of Trump voters.

4 Conclusion

The main objective of this paper is to see if similarities can be found between tweets send prior to the 2016 U.S. presidential elections and the actual results. This was researched using three sub-questions. First the ratio of Trump/Hillary voters in the elections was compared with the ratio of the tweets for Trump/Hillary, and based on that it was estimated in how many states Trump would win and in how many Hillary would win. 30 of the 50 states were estimated correctly, which resulted in a accuracy of 60%. After that a topic analysis was executed and compared with topics that were important in the election.

Only a few of the topics found in the important election topics were found in the LDA. It might be that these top issues arent actually tweeted about that much and that voters find other topics more reportable. Further research can be done to confirm this. Finally, it was investigated whether there are correlations between the demographic information and the ratio of Trump/Hillary voters of the tweets, that could support why Trump/Hillary won or lost in certain states. However, no correlations were found between the tweets and the demographic information, so no conclusive argumentation could be used to support the outcome of the predicted election votes. Based on the results of these sub-questions, it is concluded that, with the use of around 600.000 tweets, only minor similarities were found between real election data and predictions of election data based on tweets. Only 60% of the states were estimated correctly, which is not high enough to conclusively state that twitter data can be used to predict the outcome of elections. Alternatively, focusing on the topics, although some prominent topics of the elections were found in the LDA model, some other topics were not. Therefore it is inconclusive whether or not the LDAs accuracy is sufficient enough to use in the topic analysis of the tweets.

References

1. Internet live stats homepage, https://Internetlivestats.com. Last accessed 7 Oct 2018
2. Stanford sentiment analysis training data, http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip. Last accessed 5 oct 2018
3. Author, N. title: The Top Issues For Voters In The 2016 Presidential Election https://www.forbes.com/sites/niallmccarthy/2016/07/11/the-top-issues-for-voters-in-the-2016-presidential-election-infographic/#163ff66b23fd. Last accessed 5 oct 2018
4. Income in US states https://www.deptofnumbers.com/income/states/. Last accessed 5 oct 2018
5. Data and Statistics about the U.S https://www.usa.gov/statisticsitem-36857. Last accessed 5 oct 2018
6. Poverty rate in the United States in 2017, by state https://www.statista.com/statistics/233093/us-poverty-rate-by-state/. Last accessed 5 oct 2018