Understanding and Predicting US elections with Twitter

Renuka Gurung, Pranshu Bahl, Yi Ning Liang, Zheya Feng, and Georgios Sotos
Team 1

University of Amsterdam, 1090 GH Amsterdam, The Netherlands

Abstract. To understand how twitter data can be used to predict US elections different methods were applied. Processing the twitter data into a readable format; exploring different features of the tweet such as language, Country, User Mentions, User Replied to; compare with US Demographics data and using techniques like sentiment analysis, topic modeling, text classification and network analysis to get useful insights. No comparison could be made with the education, income data as the twitter user profiles did not have comparable information. The population estimates (2017) data is almost perfectly positively correlated with the tweet counts of the users per state. The users talked mostly about Donald Trump (~63%) as compared to when they were talking only about Hillary Clinton (~22%). Sentiment analysis on the data shows that ~50% tweets from in the United States have neutral sentiment when talking about either or both candidates. Moreover, we apply topic modeling paired with sentiment analysis and conclude that trump has overall more positive tags than Hillary. We further apply text classification and conclude that ~38% tweets talk about the politics. Hence, we see that topic modeling does not appropriately determine the essence of the tweets. On analyzing the reply and hashtag network, we observe that Donald Trump was able to connect and influence with the twitter audience more than Hillary Clinton. These results show that the results of the 2016 US Presidential Elections could have been predicted using techniques like sentiment analysis, topic modeling, and network analysis.

Keywords: Sentiment Analysis, Topic Modeling, Text Classification, Networking Analysis, Twitter, Politics, U.S. Election.

Political parties across the globe are learning that along with advertising on traditional mediums such as television and newspapers, they must invest in digital marketing if they want to compete with their rival parties. Over the past few years, we've seen an uptake in UK and US parties, in particular using social media campaigns to defeat their opponents.

This increase has led to a massive database of the likes/dislikes of the various userbases which can be used to further analyze and to better understand behaviors.

The 2016 U.S. election was a major event, which affected the whole globe, wanting to predict such a huge global event, is only natural and understandable in order to avoid uncertainties and create contingency plans.

For this assignment we analyze a twitter dataset to see whether it could have predicted the 2016 US Presidential Election results. We use 657,307 tweets collected from

the period 12/08-12/09 2016. We mainly look for tags like-" Donald Trump", "Hillary Clinton" when looking for tweets. We parse the twitter data in a table to compare and analyze the textual content in the tweets. The geographical location of the users is mostly but not limited to the United States.

We apply different techniques like sentiment analysis, network analysis, text classification, and topic modeling to analyze the twitter data. These techniques are detailed in the following sections and the success of them in helping predicting results mentioned thereof.

1 Methodology

To properly gauge and predict the winner of the US elections via tweets, firstly some preprocessing on the data had to take place. We filter the @mention and the URL links, even though these both carry certain information, but this information doesn't add value to identify sentiment. Concerning hashtags in the tweets, the "#" will be removed, as the remaining text adds value. Lastly, we remove all non-letter characters including numbers and lower-case all the words. Then we will conduct four different analyses, which will be discussed in the following paragraphs (Fig. 1).

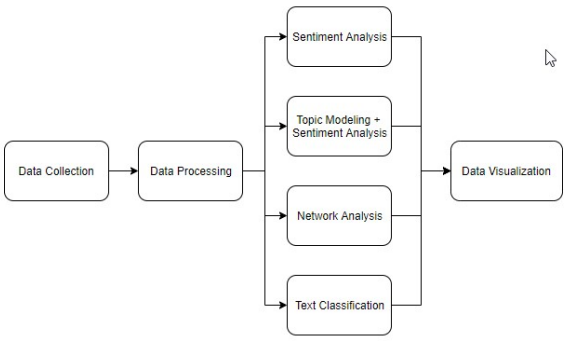

Fig. 1. Research Framework

1.1 Sentiment Analysis

Sentiment analysis can be defined as the prediction of emotions in a word, sentence or corpus of document. It is intended to serve as an application to understand the attributes opinions behind certain topics. Therefore, sentiment analysis has power of finding the opinions and affinity of people towards specific topics of interest.

A sentiment analysis will be conducted based on the sentence rather than whole tweet. This level will determine that each sentence shows a positive, negative or neutral sentence level sentiment. We used TextBlob[1]. which is a python library for processing textual data. It provides a simple API for diving into common natural language

processing (NLP) [2] tasks such as part-of- speech tagging, noun phrase extraction, sentiment analysis, classification, translation and more.

The sentiment function of Textblob returns two properties, polarity and subjectivity. Polarity is float which lies in the range of [-1,0,1] where 1 means positive statement, 0 means neutral and -1 means a negative statement. Subjective sentences generally refer to opinion, emotion or judgment whereas object refers to facial information.

1.2 Topic Modeling

Topic modeling is an unsupervised Machine Learning algorithm, which allows for the identification of latent topics within text data.

With the cleaned data from the data processing stage, the tweets will be pooled together as mentioned by [2] Sanner et. al. to improve accuracy of the generated topics. Further the aggregated tweets will be processed via the CountVectorizer of SciKit-Learn [4]. The aggregation will be applied per candidate. Then the Latent Dietrich Allocation (LDA) with ten planned topics will run across the whole aggregated dataset. Not more than ten topics are defined, as each topic will consist of multiple tokens and sorting it by the most relevant 20 token and analyzing those will already be sufficient at this stage. Each token inside the ten topics then will be further analyzed with sentiment analysis, allowing ultimately for a sentiment classification per topic. The aim is to grasp overall what the topic sentiment score is per candidate, to provide possible features for future election prediction.

1.3 Text Classification

Text classification is used to classify text in certain predefined categories based on a trained dataset.

In order to analyze topics in the tweet, a stemming tokenizer function is applied[4]. The Porter Stemmer and the tokenizer function from the NLTK package will be used. Due to time constraints, existing classified data sources are planned to be used to train the classifier algorithm, namely the news data source provided by SciKitLearn. The provided data source is based on posts for newsgroups and has a resemblance with tweets. Further it categorizes the posts in 20 topics. Proceeding the data will randomly be split up in a training set of 25 % of the data and a test set of 75 %.

Afterwards the classifier will be trained by firstly vectorizing, then tf-idf-transforming and lastly applying a Naïve Bayes classifier. The trained classifier then will be run across all test data. To understand if the classification made sense, ethnographic research will be applied. Since 20 topics are available, 2 of the most occurring non-related topics to politics will be analyzed, to understand mis-classification.

1.4 Network Analysis

A network can be defined as a graph in which nodes and/or edges have attributes (e.g. names, degree). In our case, the twitter reply network would be a directed graph with nodes and edges represent users and connection respectively. The hashtag networks are

undirected graphs with nodes and edges represent hashtags and the co-present relation of hashtags respectively.

The graph visualization platform Gephi will be employed to do network analysis. Firstly, the data will be organized into a standard format in python. Then in Gephi, the visualization of the networks is done through displaying nodes and edges in various layouts. For the layout algorithm, since the networks in our case have more than 10000 nodes, the most appropriate one would be force Atlas2 which is a force-directed algorithm applies to up to 1 000 000 nodes[6]. Then with Louvain community detection algorithm, we will be able to classify the underlying communities of the nodes and attribute colors, size and other advanced properties to nodes.

Since the whole dataset is too large for Gephi to process, the network analysis will be done on subsets of the original dataset. Specifically, the first 10000 tweets that replied to someone as the dataset for reply network analysis; For hashtag network, the two datasets are hashtags that used to reply to Clinton and Trump respectively.

2 Results and Discussion

2.1 Sentiment Analysis

Candidate	Number of tweets
Only Donald Trump	369,428
Only Hillary Clinton	131,200
Both	84,357

Fig. 2. Numbers of tweets that candidates were mentioned

As we can see in Fig. 2, we observe that Donald Trump was mentioned almost 3 times more than Hillary Clinton. In order to gain more insight, we performed sentiment analysis to determine the attitude or the emotion of the users about Donald Trump and Hillary Clinton.

Majority of the tweets are originating from the United States, further analysis will be focused on this country (90.06%). Fig. 3 shows the opinion of twitter users toward Donald Trump and Hillary Clinton for the top 5, states which have the greatest number of tweets, 219,584 tweets in total and takes up to 33% of the whole dataset. We can see that in these 5 states Donald Trump was much more discussed than Hillary Clinton, for Hillary Clinton the number of positive tweets were 18% more than the negative tweets , on the other hand, the number of positive tweets were 37% more than the negative tweets for Donald Trump.

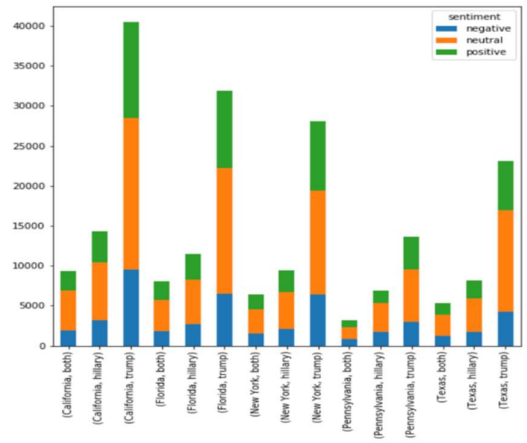

Fig. 3. Sentiment analysis for the tweets in the United States, by state.

2.2 Topic Modeling

Tweets were pooled together in the categories 'Trump', 'Hillary', or 'both'.

10.9 % of the tweets could not be categorized and were filtered out. For each of the ten topics 20 tokens were identified. After applying on the generated token, the aforementioned sentiment analysis procedure, it turned out that the twitter pool 'Trump' had overall out of 10 generated topics overall, 8 positive topics, deducting one negative topic, and ignoring one neutral topic, his overall topic sentiment score is at +7 (ranging from -10 to + 10). 'Hillary' instead had an overall topic sentiment score of +1, whereas the grouped tweets mentioning 'both' candidates were identified with an overall topic sentiment score of +5 (Fig. 4).

		Candidate Mentioned		
	Sentiment of 10 topics per	Trump	Hillary	Both
Sentiment	Positive (+1)	8	5	7
	Neutral (0)	1	1	1
	Negative (-1)	1	4	2
	Total	7	1	5

Fig. 4. Topic Sentiment Score per candidate mentioned

2.3 Text Classification

After training the classifier on the training set, and testing it on the test data, an accuracy of 91% was observed. Given the observed accuracy the classifier was applied on all 657.307 tweets. Fig. 5 shows only that 37.39 % of all tweets belong to the categories related to politics. These were: "talk.politics.guns","talk.politics.misc", and "talk.politics.mideast". The unexpected category being classified as the second most occurring

topic was “rec.sport.hockey” with 14.1%. This was unexpected, because the data itself is related to politics. It also appears that tweets mentioning trump (21.4%), have more in common with sports, than it does with Hillary (17.1 %). Further, the 5th most occurring topic was soc.religion.christian with 10.16 %. The remaining 38.35 % were classified in the other 15 remaining topics.

	candidate_mention				
Classifier	Neither	both	Hillary	Trump	Grand Total
talk.politics.guns	10.64%	22.37%	14.00%	15.54%	15.57%
rec.sport.hockey	18.11%	9.76%	13.90%	14.37%	14.10%
talk.politics.misc	7.75%	18.29%	14.55%	9.57%	11.49%
talk.politics.mideast	9.60%	8.10%	11.02%	10.75%	10.33%
soc.religion.christian	13.81%	6.99%	10.12%	10.19%	10.16%
sci.crypt	7.17%	6.51%	7.86%	7.58%	7.46%
rec.sport.baseball	4.80%	5.27%	3.20%	7.06%	5.81%
rec.autos	3.67%	1.97%	2.77%	4.75%	3.88%
rec.motorcycles	4.18%	2.90%	2.98%	3.65%	3.48%
sci.med	3.83%	2.64%	4.41%	3.17%	3.42%
alt.atheism	2.90%	4.18%	4.14%	2.66%	3.18%
sci.electronics	2.45%	3.59%	3.06%	2.96%	3.01%

Fig. 5. Distribution of identified topics across observed tweets, split by candidates

Ethnographic research on the two most occurring non-related political topics showed that by picking some random tweets, that both categories can be explained. The tweets which got classified as “rec.sport.hockey” with 14.1% representation, have on a first look in common to deal with more vulgar and emotional language and discuss win and loss (Fig. 6) , whereas tweets classified as soc.religion.christian have been using words which could be seen related to religion (Fig. 7).

Classifier	Clean Text
rec.sport.hockey	He's begging because he cannot win without our support and right now he has about 1%. Not happening #NeverTrump
rec.sport.hockey	@realDonaldTrump u suck#noballs
rec.sport.hockey	@Lagartija_Nix @BarbMuenchen @LouDobbs @ScottBaio @misterdish69 @realDonaldTrump @roycan79 @myGianLuca @joyreaper @soniafarace
rec.sport.hockey	Where's King Barry? Oh wait he's golfing; and I forgot Bush broke the Levies during Katrina #Trump
rec.sport.hockey	Just when you thought #Trump couldn't look more ridiculous, he puts on a billed cap…
rec.sport.hockey	@HaroldWNelson @realDonaldTrump @FoxBusiness
rec.sport.hockey	Where's King Barry? Oh wait he's golfing; and I forgot Bush broke the Levies during Katrina #LiberalHypocrisy #Trump

Fig. 6 Rec.sport.hockey classification examples

Classifier	Clean Text
soc.religion.christian	#Catholic #Pope #Trump #Republicans
soc.religion.christian	@realDonaldTrump on the basis of that last phony medical documentation you gave, nobody would believe anything you release.
soc.religion.christian	@realDonaldTrump Money Transfers
soc.religion.christian	@realDonaldTrump Always keep God in your mind daily He will inspire you to say the right things..you will win
soc.religion.christian	@realDonaldTrump God give you the strength to endure the vitriol coming from the Left.FIGHT THE GOOD FIGHT for us the people God Bless.
soc.religion.christian	@suediamond11 @MarcoGutierrez @Latinas4Trump @LatinosForTrump @immigrant4trump @DineshDSouza @seanhannity @realDonaldTrump Amen. A wise man.
soc.religion.christian	@mike_pence @mike_pence is a whore to @realDonaldTrump 's hatred of modern America. They're both alt.right, and hate us all.

Fig. 7 Soc.religion.christian classification examples.

2.4 Network Analysis

For the reply network, we yield the following network graph which contains 907 communities.

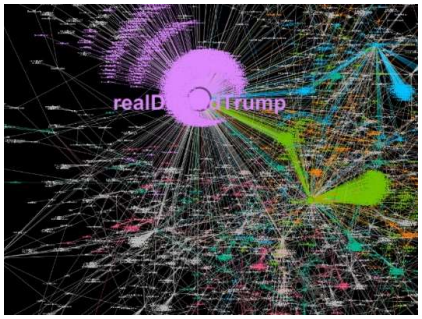

Fig. 8. Reply network graph based on 10000 tweets

The most noticeable three communities are realDonaldTrump, HillaryClinton and FoxNews. In the community with realDonaldTrump. Kayleigh McEnany, the national Spokesperson for the RNC, is the second largest node, and the third is a user who has the hashtag #AmericaFirst in his description, he seems to be big fan of Trump. Obviously, the network of Trump is much larger than Clinton, and this fact leads us to believe that Trump is more influential in social media.

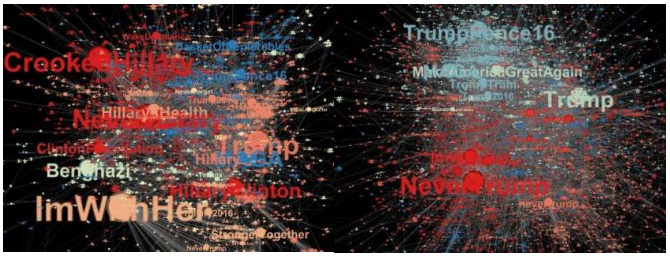

Fig. 9. Hashtag networks of the hashtags people used to reply to Clinton(left) and Trump(right)

The hashtag network for Clinton and Trump has 324 and 622 communities respectively. In the Clinton network, the most frequently used hashtag “ImWithHer” is in the same community with “StrongerTogether”, “NeverTrump”, these are all positive hashtags. On the contrary, “CrookedHillary” is with “ NeverHillary”, “HillaryForPrison. In Trump network, most frequently used negative hashtags are “NeverTrump”, “ImWithHer” and “DumpTrump” and these are in the same community. “TrumpPence16”, “MAGA”, “TrumpTrain” are the biggest ones within the community that displays positive attitude. Based on these facts, we have reasons to believe that the hashtags in the same community are of similar sentiment. The average degree and network diameter of the two hashtag networks are (2.788, 12)(Clinton) and (5.565, 11)(Trump), which implies that the hashtags used to reply to trump are more connected than to Clinton.

3 Conclusion

Based on the result of sentiment analysis, we can draw a conclusion that Donald Trump was discussed more than Hillary Clinton. 137,164 tweets discussed only Donald Trump versus 50,148 tweets discussed only Hillary Clinton. And the ratio of positive tweets for Donald Trump was also higher than Hillary Clinton.

Topic modeling alone does not help us understand what is happening inside tweets without ethnographic research. However, pairing it with a sentiment analysis, allows for a quicker automated approach. Factors affecting the maximum feature size, most relevant token per topic can influence the topic sentiment score when using LDA. Further experimentation and additional research is necessary to properly utilize this feature for predictions. However, it did yield interesting results with Trump leading in positive topics.

Regarding the text classification the conducted methodology on classifying the tweets showed that the trained classifier needs more related context towards tweets, and that it is difficult to gauge the effectiveness of a trained classifier on a new kind of dataset (from newsgroup postings to twitter posts). Hence it is recommended to either identify ways to optimize data collection for future research, to apply a different classifier, or to identify measures to gauge topic classification on new datasets more efficient. The classifier at this stage could not be utilized to help predict the elections.

Based on the results of reply network analysis, Trump built a stronger connection with twitter users and enjoyed a higher popularity on the social media.

Hence, we can conclude that sentiment analysis, topic modeling paired with sentiment analysis and network analysis, help us in identifying steps in predicting which candidate is going to win the elections. However, at this stage, it is not possible yet with high accuracy to predict the actual winner. The identified results might lead to a successful prediction in the future.

References

[1] TextBlob: Simplified Text Processing. https://textblob.readthedocs.io/en/dev/
[2] Natural language processing. https://en.wikipedia.org/wiki/Natural_language_processing
[3] Mehrotra, R et. al. (2013). Improving LDA topic models for microblogs via tweet pooling and automatic labeling, *Proceedings of the 36th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR 2013)* (pp. 889 - 892). New York NY USA: Association for Computing Machinery (ACM).
[4] Fabian Pedregosa, et al. Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825-2830 (2011)
[5] Edward Loper and Steven Bird. 2002. NLTK: the Natural Language Toolkit. In Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics - Volume 1 (ETMTNLP '02), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 63-70. DOI: https://doi.org/10.3115/1118108.1118117
[6] Gephi Tutorial Layouts. https://gephi.org/users/tutorial-layouts/