Election Analysis Based on Twitter

Fariha Moomtaheen
University of South Florida
fmoomtaheen@mail.usf.edu

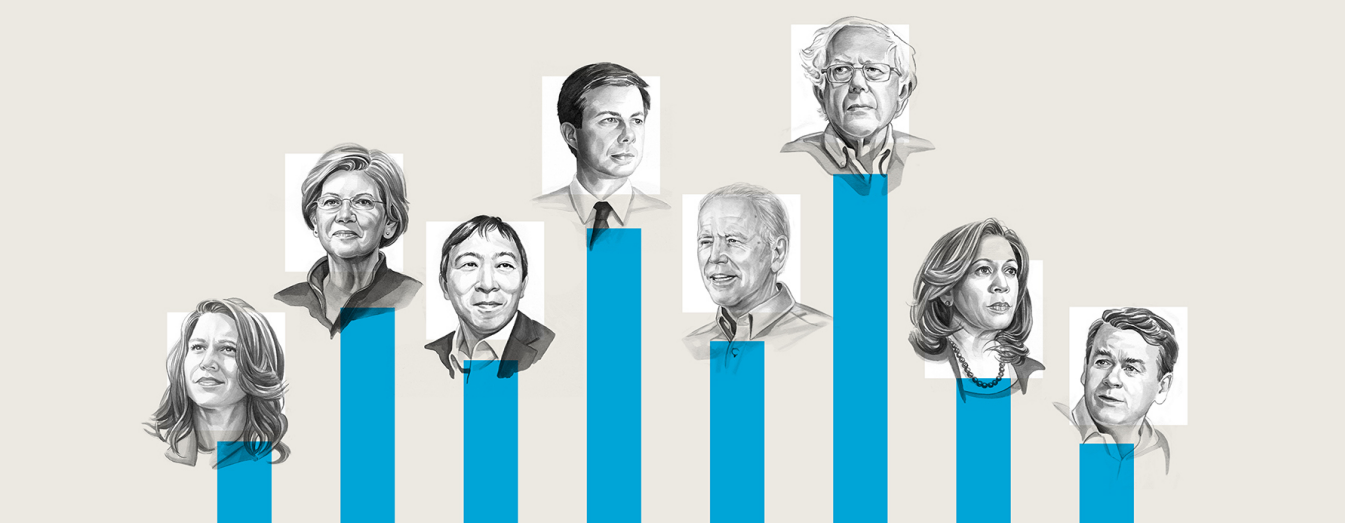

Figure 1: Democratic Party Candidates for 2020 Presidential Election.

ABSTRACT

The role of Social Media in predicting real world phenomena is getting increasingly popular. Tracking people's activity provides a spontaneous way to guess their preferences. Using proper metric and algorithm in datasets found from social media can eliminate the need to conduct polls by directly interacting with humans. This papers analyses data found in one of the most popular social media platform Twitter and compares them with real world data. The comparison is based on several criteria and focuses on exploring the most significant one for future works. This is a groundwork for further exploration into this area.

CCS CONCEPTS

• **Information systems** → **Web and social media search**; • **Networks** → **Social media networks**; **Online social networks**; • **Human-centered computing** → **Social media**; • **General and reference** → *Evaluation*; Empirical studies.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

KEYWORDS

datasets, social networks, poll prediction, data mining

1 GITHUB REPOSITORY

https://github.com/FarihaUpoma/Social-Media-Mining-Project-Election-Analysis-using-Twitter

2 INTRODUCTION

In this modern world, people are getting more absorbed in the online world rather than the real world. Almost all the interactions a person has in his day to day life are conducted on social medias. They are more comfortable and expressive in virtual world. They don't have to face other people, and they can reach a larger population with their opinions without leaving the comfort of their own space. These advantages are making it possible to get to know about a person more easily. Everything a person does online has footprints and records. It is possible to know almost everything about a person only by his social media posts without even knowing him personally, let alone ever talking to him. Aside from the security side of this aspect, this can provide great tool to observe and form social ideologies from people's online activities.

Now-a-days there are groups or discussions about every possible area of interest like food, hobby, holiday, destinations, and etc. These vast amounts of data that are available online do not only give us

insight about a person, but also provides us a view of his role in the social structures and group behaviors. These social structures and dynamics have been discovered by studying the information flow from one individual to another[1]. The underlying network structure influences how the news or information disperses in the community. Different social media platforms adopt different structure to offer connectivity among population. Facebook, Twitter, Youtube, LinkedIn these are some of the most popular social media platforms.

Twitter offers a great way to study how social media helps propagating an idea or information. Bases on one's like and predilection people form connections with others and by observing these connections valuable conclusions can be drawn. These massive amount of data is a gold mine for researchers to do empirical analysis. Politically active internet users Apart from providing researchers ample resource for studying social structures, social media also helps mold public opinions. Social media shapes the networked public sphere and facilitates communications among different politically oriented communities[2]. The major contributions of this paper are-

3 MOTIVATION

In our interest to explore effect or validity of social media to mimic or predict real world events I decided to look into the 2020 presidential election of the USA. Recently, there are nationwide primary elections going on to select the candidate for the democratic party. The elections will begin on February 3rd and end on June 16. The final nomination will be held at the Democratic National Convention in July. A total of twenty eight candidates started their journey for this contest and some of them have already withdrawn. This is the largest field ever where candidates' profession ranges from US senators, House Representative, mayors, entrepreneurs to even a self-help guru[3].

These primary elections are creating enough anticipations for generating massive amount of online data. People are talking about them, discussing and criticizing their every policy and moves and thus expressing their own preferences. If there is a proper way to mine these data important information can be found. Instead of conducting polls to know about elections tides, a dive into this online world could help us draw valuable conclusions. This paper tries to explore the data generated from twitter based on this presidential election to see if they parallel to real world data, which could in time predict even future events.

4 DATA COLLECTION

4.1 Twitter Platform

Twitter was created by Jack Dorsey, Noah Glass, Biz Stone, and Evan Williams in 2006, and since then it has rapidly grown to become one of the ten most visited websites in the world[4]. It has been described as the SMS of the Internet[5] as it has over 321 million active users as for the year 2018. The numbers don't stop there as these are growing everyday.

This large volume of data is getting increasingly recognized by data science companies to measure public opinion as a proxy to the real world[6].

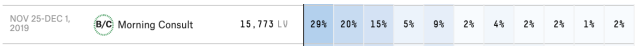

Figure 2: Morning Consult poll result found on fivethirtyeight website.

4.1.1 Structure of Twitter. Twitter allows users to post upto 140(later increased to 280) character texts, called "tweets", for the audiences who follow the user. These tweets can also contain URLs or hashtags. Users have two types of directed connections with other users. One is directed towards him, which are the people who follow this user. Other kind is directed from him to the accounts that this particular user follows. Users also share their views using mentions and retweets. Mentions are usually directed towards a person, in reply to one of his posts or comments. By mentioning his name, he gets notified about the mention and can get involve in direct communication. When someone shares someone else's opinion about something and wants to broadcast the original posts to his followers also, he generally retweets the tweet. This is common in celebrity posts or about any news or posts from public forums.

Hashtags allows the users to gather in a common cause without interacting with one another. If a user wants to talk about a particular topic, he can choose the hashtag that's been in use for that topic and use it in his post. That way when other users use that hashtag, they all can see posts where the hashtag has been used. This is a clever way of tracking public opinions and finding trending topics in any interval.

4.2 National Poll

To compare twitter data with real world results, I used a website called fivethirtyeight https://projects.fivethirtyeight.com/2020-primaries/democratic/national/. This website focuses on opinion poll analysis, politics, economics, and sports blogging[7]. It gathers polls/surveys conducted by various other sources and present them with analytical results. I followed the "National 2020 Democratic Presidential Primary Polls" and noticed the poll results. These polls are conducted by either phone or online or other formats. The sample populations are classified as A [adults], RV [REGISTERED VOTERS], V [VOTERS] and LV [LIKELY VOTERS]. I picked the poll conducted by Morning Consult as it has the highest number of samples among the other recent polls. The samples number is 15,773 Likely Voters. It is conducted during the period November 25, 2019 to December 1, 2019. According to this poll, Joe Biden is well ahead in the elections with 29 percent of support.

I use this poll result to compare with the results found from Twitter to see if there's any correlation in them. Figure 2 shows the poll.

5 DATA EXTRACTION TOOLS

5.1 Github Repository

To collect data from twitter I first used twitter search API. But the default API from twitter developer account only gives us data from the past 7 days. This is an obstacle for successfully getting the data from the time period I needed. I found the Github repository called GetOldTweets created by Jeferson Henrique for this purpose. This repository allows us to determine the time period along with search

Table 1: tweet object returned by GetOldTweets.

Attribute	Type
id	str
permalink	str
username	str
text	str
date	date
retweets	int
favorites	int
mentions	str
hashtags	str
geo	str

Table 2: Number of tweets for each candidate.

Candidate	Tweets
Biden	2189
Booker	71
Buttigieg	961
Harris	277
Sanders	495
Warren	945
Yang	678

query, and even number of maximum tweets. Using this repository I successfully collected all data related to the elections.

The github repo can be found in here- https://github.com/Jefferson-Henrique/GetOldTweets-python

5.2 Twitter User Lookup

Twitter provides some APIs for developers, among which I used the user lookup API. Using this I searched for user information with only their screen names.

The returned user object has all the information associated with that particular username. More about the returned user user object can be found here - https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/user-object.

5.3 NLTK package

For sentiment analysis, I used a python library called Natural Language ToolKit. It has built in functions for finding the polarity and subjectivity of a given text along with some other great tools.

6 METHOD

GetOldTweets takes as parameter a querysearch text which it matches with all tweets of a given period. I used the names of the candidates one by one for the time period November 25, 2019 to December 1, 2019. The command to use in the terminal looked like this-

python3 Exporter.py –querysearch "biden" –since 2019-11-25 –until 2019-12-01 –maxtweets -1

As the returned tweets would perform an 'and' operation of all its querysearch words, I used the name of the candidates one by one to generate the data. I counted the number of tweets and based on that created a bar graph of the candidates.

Table 3: Number of positive tweets for each candidate.

Candidate	Positive Tweets	Negative Tweets
Biden	1761	428
Booker	58	13
Buttigieg	802	159
Harris	231	46
Sanders	413	82
Warren	783	162
Yang	561	117

Table 4: Number of positive tweets for each candidate.

Candidate	Followers Exposed to tweets
Biden	3598737
Booker	2945725
Buttigieg	784432
Harris	13128807
Sanders	1835577
Warren	199643
Yang	3468228

We know that if someone talks about someone, it doesn't necessarily mean that he's talking positively about that person. Negative tweets could also create a bias in the result. To filter out the negative tweets I used NLTK and counted the number of positive tweets and negative tweets.

Since, we are trying to predict the number of likely voters I ignored the number of negative tweets and plotted the graph using only the positive tweets. Though negative tweets may also have great significance, in this particular area that was neglected.

To approach the problem from another angle I searched for the users information next who generated those tweets. I used the twitter user lookup API which returned with all the information. Initially I just focused on the number of followers to generate another bar graph.

7 ANALYSIS

We see from the bar graphs that twitter data has some similarities with the real poll result. While twitter data corresponds with the popularity of a candidate this doesn't properly convey the supporter number of that candidate. This is where the differences start. A candidate can be famous across Twitter, but that doesn't mean he has the highest number of voters on his side. This is what we observe from the Yang and Buttigieg bars. Due to Yang's unique characteristics he stands out among other candidates and even gets to be the trending topic. Buttigieg's controversy regarding his past jobs also got a large number of people talking about him. I tried to eliminate this discrepancy by eliminating the negative tweets.

This still doesn't correlate with the polls. But if we focus on the number of negative tweets of each candidate, we can see that Biden has the lowest percentage of positive tweets about him. Though the

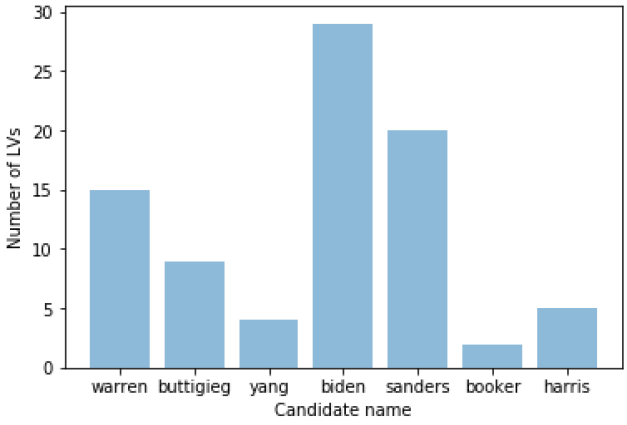

Figure 3: Bar graph corresponding to Morning Consult.

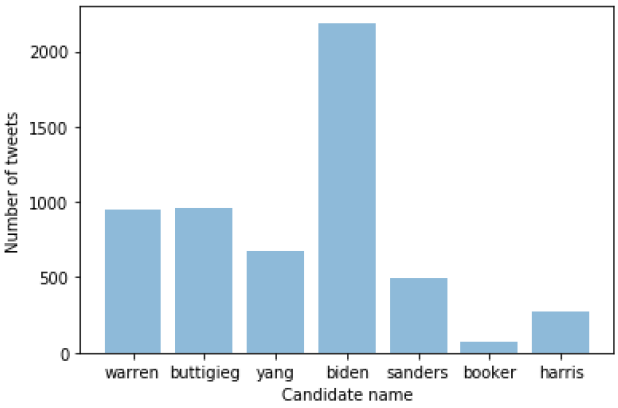

Figure 4: Bar graph based on number of tweets.

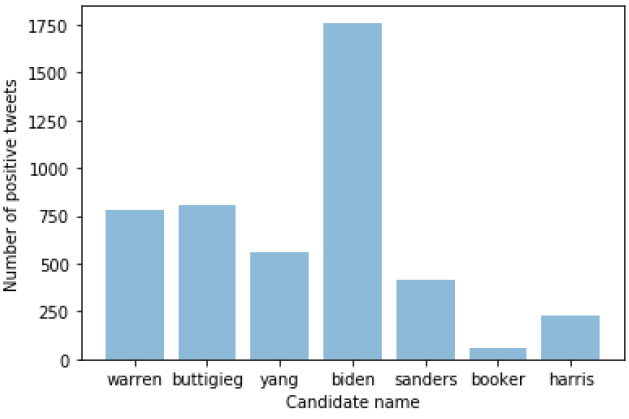

Figure 5: Bar graph based on positive tweets.

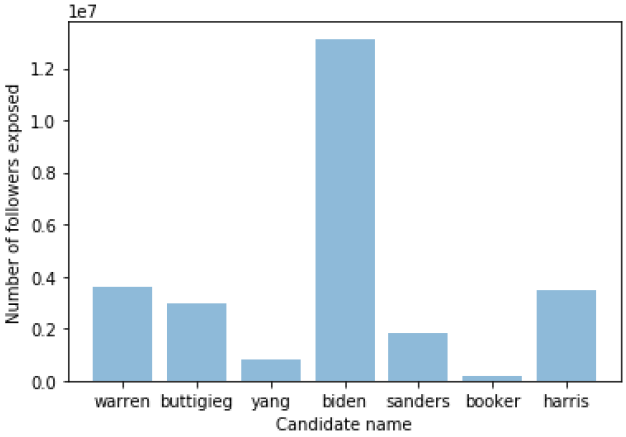

Figure 6: Bar graph based on number of users exposed.

Table 5: Percentage of positive tweets for each candidates.

Candidate	Postive Tweets Percentage
Biden	0.8%
Booker	0.817%
Buttigieg	0.835%
Harris	0.834%
Sanders	0.834%
Warren	0.829%
Yang	0.827%

difference isn't that much prominent, still we can say that Biden has got the largest negative fan base as opposed to others. Apart from Biden, if people are talking about some candidate they are mostly talking about them positively.

Another interesting thing to explore is the exposure that each candidate got through tweets about them. The basic way to find exposure is by the number of followers that sees a tweet. I counted the total number of followers of all the users that tweeted about a candidate and ranked them. This ranking could allow us to explore more about effects of social media on political events.

8 FUTURE WORKS

This work is just the beginning for finding proper method to employ twitter data for election prediction. Initially, I planned to build a linear model for finding the best feature to correctly predict poll result. This machine learning technique can help us find the proper weights and importance of the features in predicting elections. Also, to find the proper exposure and see if that has any effect on the popularity of the candidates, we can focus on some other characteristics of the users. The retweet counts of the tweets and number of favorites can also be include for this purpose. Using the subjectivity field of the sentiment analysis, we could filter more tweets based on the context of the tweets.

9 REFERENCES

(1) https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1509/1839
(2) https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2847/3275
(3) https://projects.economist.com/democratic-primaries-2020/?fsrc=scn/tw/te/bl/ed/whoisaheadinthedemocraticprimaryrace
(4) "Top Sites". Alexa Internet. Retrieved May 13, 2013.
(5) D'Monte, Leslie (April 29, 2009). "Swine Flu's Tweet Tweet Causes Online Flutter". Business Standard. Retrieved February 4, 2011. "Also known as the 'SMS of the internet', Twitter is a free social networking service".
(6) https://www.toptal.com/data-science/social-network-data-mining-for-predictive-analysis
(7) https://en.wikipedia.org/wiki/FiveThirtyEight