2020 Final Project

Is social media a better predictor for election outcomes?

The Problem

Polling

The current polling industry continues to recently miss ongoing voting outcomes and fail to capture voter sentiment and deliver accurate representations of election outcomes.

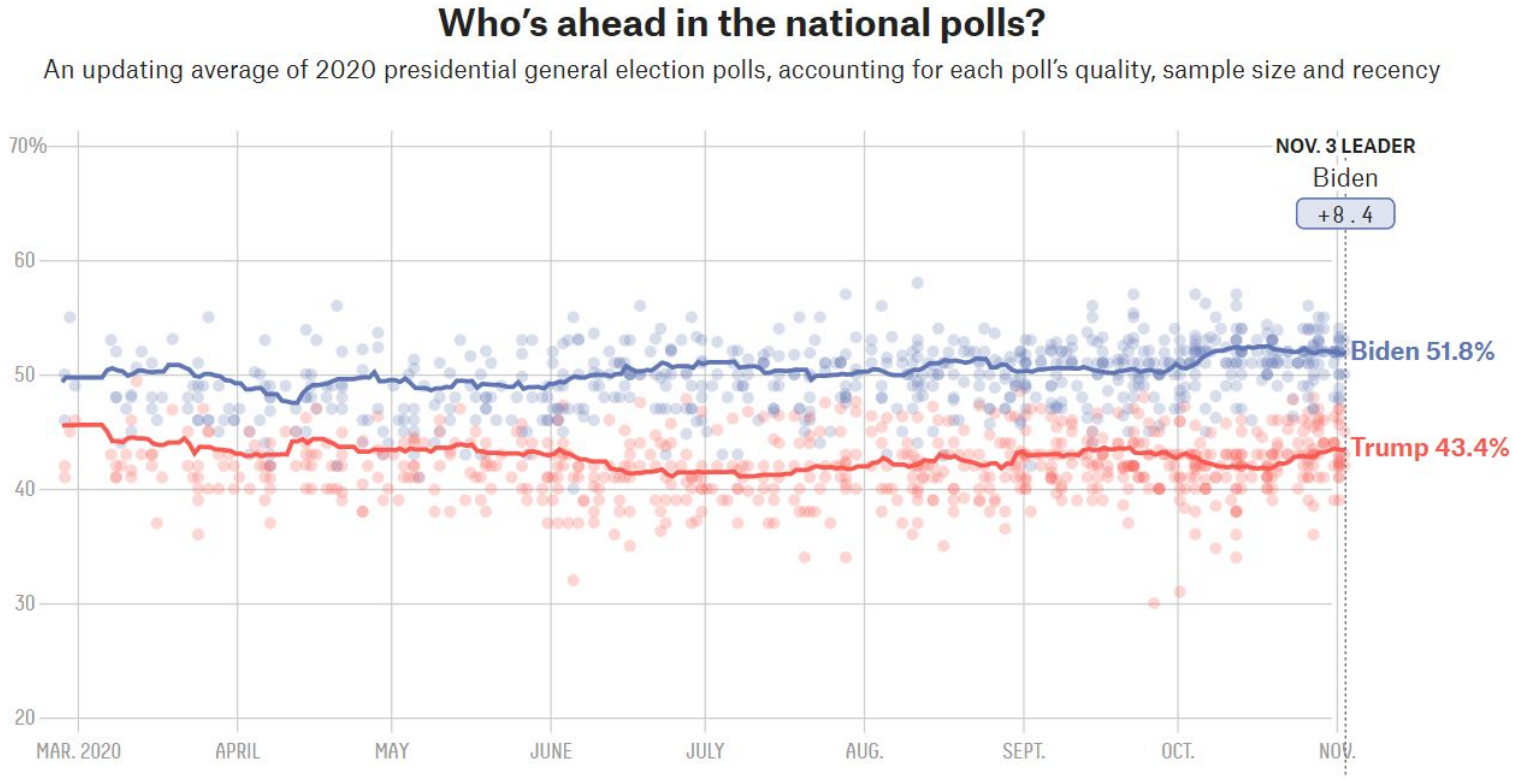
Context

Large Polling institutions such as 538, CNBC, Reuters and other high level polling institutions missed key election outcomes such as Brexit 2016, Trump 2016, Brazil's Bolsanaro and British 2019 Parliament. Will They Miss 2020?

Problem statement

Can we disrupt the Status Quo Polling by using social media to measure and predict voter enthusiasm coupled with election outcome for the 2020 Presidential Election?

Expected Final Polls



Final 2020 General Election Outcome

Polling Data						
Poll	Date	Sample	MoE	Biden (D)	Trump (R)	Spread
Final Results	--	--	--	50.8	47.4	Biden +3.4

Challenges deep-dive

Challenge 1

Can Twitter be used to predict the U.S. Presidential Election Outcomes?

We want to focus on one source of social media data, Twitter, to help gauge voter enthusiasm and eventual election outcome but where can we find the data.

Challenge 2

Which Twitter info can and should be used but where do we find it?

Even though there is large amounts of data, which data should we use and how can we clean up the data for effective and efficient use?

Challenge 3

How to Effectively Transform the data for use?

Even after locating the data and locating key information, which prediction model will effectively load the data and present a cohesive structured answer?

Which is a better election predictor?

Can we tokenize and transform the data into actual reliable sources?

What is our expected outcome using NLP sentiment extraction & clustering?

Where can we find the data?

MongoDB database with python and import the datasets from MongoDB straight into Jupyter notebook

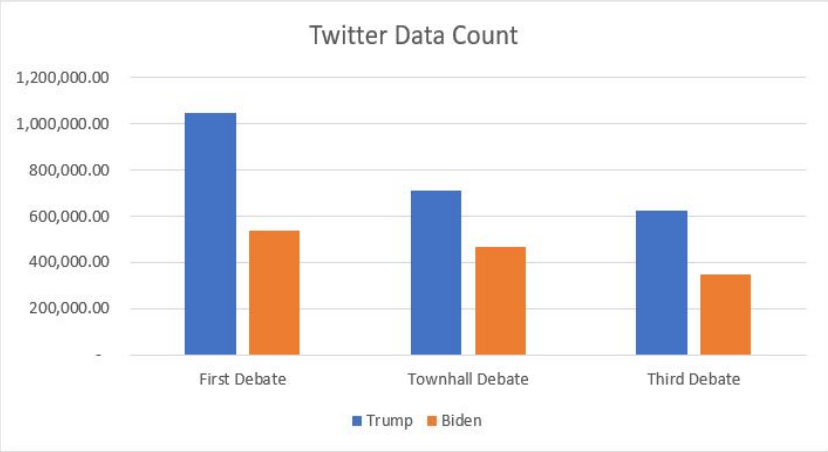
Solution

ETL, Tokenize & Random Forest

In order to perform effective data analysis and prediction models:

ETL:

1. We will need to Find(**Extract** the Data from a program called TWINT which pulls Twitter data),
2. **Transform**(Structure the Twint Data and tokenize Key Trump and Biden names) and
3. **Load** the data into a Subjective and NLP Sentiment machine and gauge twitter sentiment for outcomes based on the number of likes, retweets of the two candidates and tweet length



	Trump	Biden
First Debate	1,046,429.00	536,261.00
Townhall Debate	713,142.00	465,376.00
Third Debate	622,608.00	347,930.00

Data Source Information

1. First Debate(September 29th, 2020)
2. Townhall (October 15th, 2020)
3. Third Debate (October 22nd, 2020).

Implementation

Raw Data from Twint/Pre-Cleaned

	D	E	F	G	H	I	J	K
1	date	time	timezone	user_id	username	name	place	tweet
2	9/30/2020	19:59:59	-400	1.3E+18	dailyphoe	Daily Phoenix		@kathyhoffman_az @JoeBiden Should we have CRT in schools? We need a leader. https://t.co/xMbU1CIVb3
3	9/30/2020	19:59:59	-400	1.3E+18	maya738	maya		@Rocket54441 @JoeBiden Literally trump but ok
4	9/30/2020	19:59:59	-400	5.7E+07	wesatkins	Big Wes		@JoeBiden @MonicaLewinsky https://t.co/Wni5F0WJHt
5	9/30/2020	19:59:59	-400	5.5E+08	woodrow	Woodrow		Pres. you missed the op to mention HBCU. @JoeBiden obviously doesnâ€™t care about AA community. He and a black president could not care less after receiving the #blackvote
6	9/30/2020	19:59:59	-400	3.1E+08	sadie_75r	ðŸ‘“·SADIEðŸ’™2020		@Jillbiden46 @JoeBiden â¤ï¸ #VoteBidenHarris2020 â¤ï¸
7	9/30/2020	19:59:59	-400	3.4E+07	reneechel	Renee ðŸ‡¨ðŸ¦ ðŸ‡		@CNNPolitics @JoeBiden @TheView @JoyVBehar @AnnCoulter @FoxNews @kanyewest @KevinHart4real @ABCNetwork THIS IS ANTIFA
8	9/30/2020	19:59:59	-400	2.4E+08	monkjona	Monk ÃedÃ¡n Shaun		#JoniMitchell - #SexKills https://t.co/NiRYkTOBr0
9	9/30/2020	19:59:59	-400	1.7E+09	ikechukwu	Rafe Miyagi		@Imagecaptured @JoeBiden Yeah silence is oppression especially from the president that why the activist still never harmed anyone but vandalize, Iâ€™m not approving of all that but when you are the presider
10	9/30/2020	19:59:59	-400	1.1E+18	kempson_	Luke Kempson		@Scottd1885 @FormerLiberal @IonlyPlayDumb4U @JoeBiden I couldn't give a fuck about Mansfail.. Nice try to have a little dig though.. Bless ya.
11	9/30/2020	19:59:59	-400	2E+07	modeka	Modeka		@realDonaldTrump @JoeBiden sure knows how to pull in a crowd doesn't he don? He did a great job making you look like a fool. You lost the debate according to reliable polls and will lose the election by a lar
12	9/30/2020	19:59:59	-400	1.2E+18	donaldjste	Donald J Stephens		@Jillbiden46 @JoeBiden â¤ï¸
13	9/30/2020	19:59:58	-400	1.3E+18	lpryanovic	Desdemona Rose Ga		@Rambopolitan @glomad128 @JoeBiden Exactly. And I admit from when my phone echos my own voice back at meâ€"I canâ€™t think at all when there is gibberish in my ear.
14	9/30/2020	19:59:58	-400	1.7E+08	kevinstein	Dza		Biden selling shirts with trump face on it , kind of sus @JoeBiden you like looking at him or what ?
15	9/30/2020	19:59:58	-400	1.2E+18	texans445	Matt		@JoeBiden #fooked
16	9/30/2020	19:59:58	-400	4.5E+07	s1lentone	Ivan Perez		@Plu9to @deanna5266 @PattyArquette @SpeakerPelosi @RepAdamSchiff @JoeBiden ANTIFA is not an organization you dope, it's an ideology lol. And BLM is a movement not an organization either lol.
17	9/30/2020	19:59:58	-400	2.8E+09	ed_w_jon	Ed W. Jones ðŸŒŠðŸ		@JoeBiden #TrumpTrain2020 https://t.co/XYQXxOBnCa
18	9/30/2020	19:59:58	-400	3.1E+09	geordie_p	Paul ðŸŒŸðŸŒŸðŸŒ		@MartinBiddulp12 @appropriatepro2 @JoeBiden And you've spilled beer down your shirt.
19	9/30/2020	19:59:58	-400	2.8E+09	martinemi	Emily Martin		@umdbulldogs93 @MeidasTouch @JoeBiden Me too!!
20	9/30/2020	19:59:58	-400	1.3E+18	rome6722	Rome		@JoeBiden Best president ever
21								

Twitter Examples for both Biden and Trump

@marklevinshow @realDonaldTrump In about 1 week added 11,000 deaths..

@BuckSexton @realDonaldTrump Keep stroking his ego. Lol https://t.co/GCVSWYcPiU

@sofiafte @GOPChairwoman @realDonaldTrump Think about this. #WhoBuiltTheCagesJoe or #AllTalkNoActionJoe

The Data Issue in column “tweet”

tweet
@kathyhoffman_az @JoeBiden Should we have CRT ...
@Rocket54441 @JoeBiden Literally trump but ok
@JoeBiden @MonicaLewinsky https://t.co/Wni5F0WJHt
Pres. you missed the op to mention HBCU. @Joe...
@Jillbiden46 @JoeBiden ❤️ #VoteBidenHarris2020...

Cleaning Process:

Code Used to Clean Tweet Column Data from Special Characters (https: links and #'s) Step 1 To Clean up Raw "tweet" Column and insert new column called "cleaned_tweet"

In [11]:
```
def cleaned_tweet (row):
    clean_tweet=row["tweet"]
    s = []
    for word in clean_tweet.split():
        if '@' not in word and 'https' not in word and '#' not in word:
            s.append(word)
    return (' ').join(s)
```

```
def label_na (row):
    if len(row['cleaned_tweet'].strip())==0:
        return np.NaN
    else:
        return row["cleaned_tweet"]
```

```
biden_1_debate_df["cleaned_tweet"]=biden_1_debate_df.apply (lambda row: cleaned_tweet(row), axis=1)
```

Tableau

For the following DashBoard, we will use Tableau to represent the clean data findings concerning tweet lengths, likes and retweets.

NLP Code: Isolate Tweets and designate Positive, Negative or Neutral Tweets

In [39]:

```
# Obtaining Polarity Analysis
def getPolarityAnalysis(score):
    if score < 0:
        return 'Negative'
    elif score == 0:
        return 'Neutral'
    else:
        return 'Positive'

df['Sentiment']=df['Polarity'].apply(getPolarityAnalysis)
df.head()
```

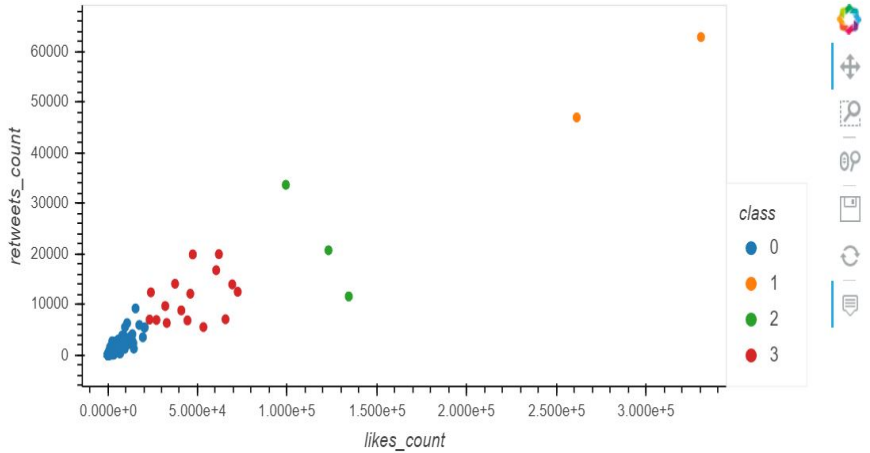
Out[39]:

	cleaned_tweet	Subjectivity	Polarity	Sentiment
0	Should we have CRT in schools We need a leader.	0.000000	0.000000	Neutral
1	Literally trump but ok	0.500000	0.500000	Positive
2	Pres. you missed the op to mention HBCU. obvio...	0.333333	-0.111111	Negative
3	THIS IS ANTIFA	0.000000	0.000000	Neutral
4	-	0.000000	0.000000	Neutral

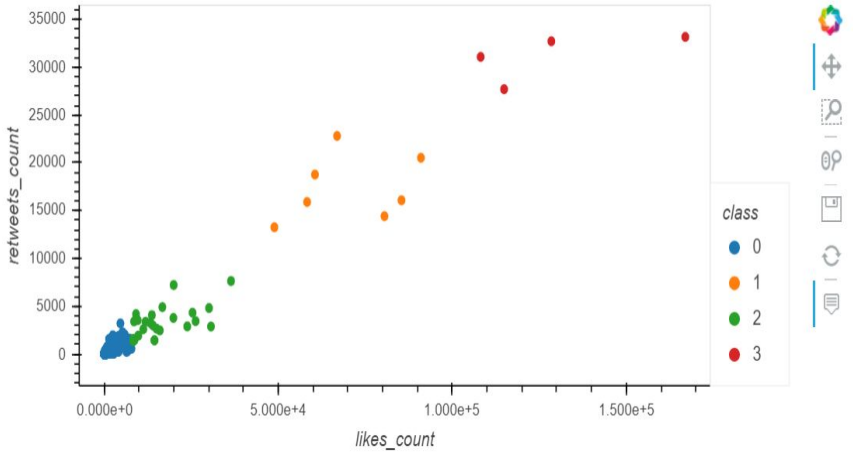
Unsupervised Learning

Likes and Retweets

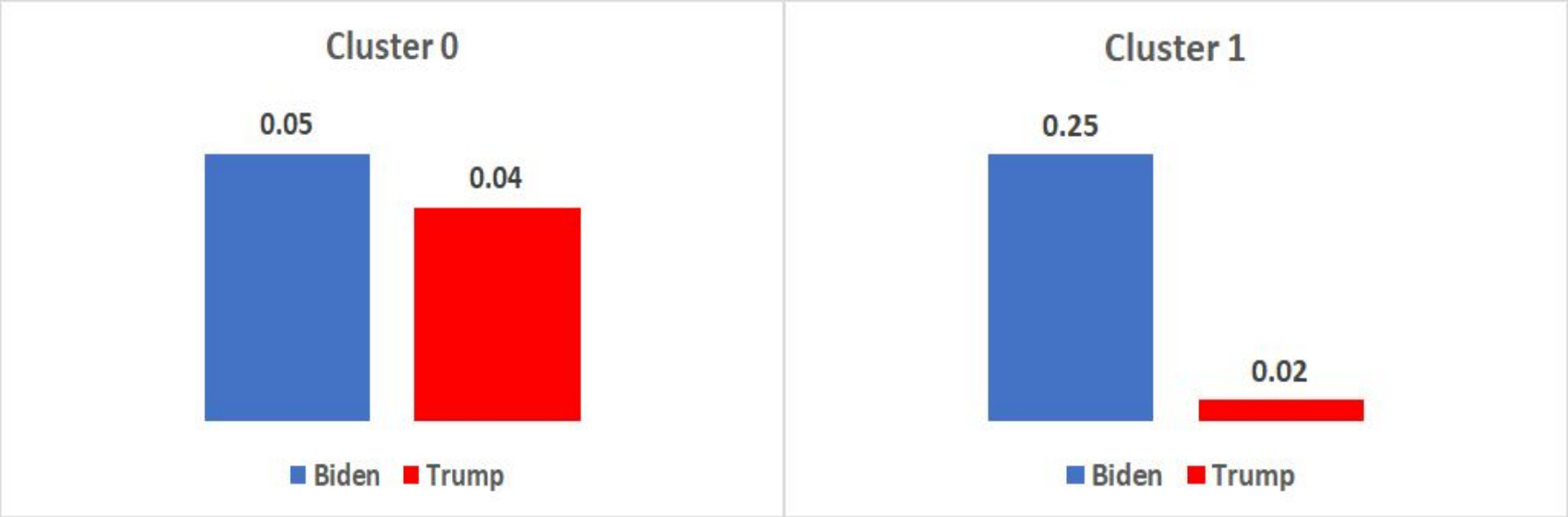
Biden



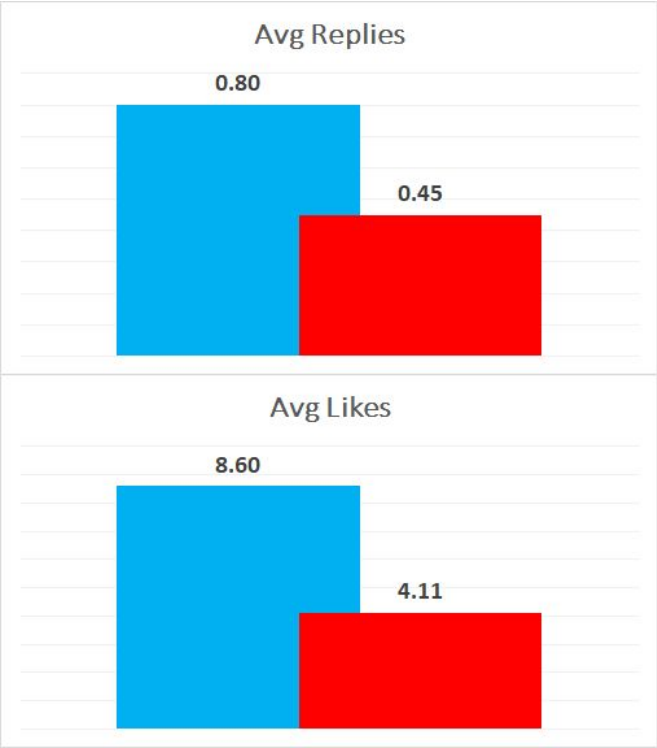
Trump



Average Sentiment



Cluster 0


Avg Replies
0.80
0.45
Avg Likes
8.60
4.11
Biden
Trump
Avg Retweets
2.06
0.84
Avg Tweet Length
18.82
17.18

Cluster 1


Biden
Trump
Avg Replies
18278
9354
Avg Retweets
54892
17400
Avg Likes
295880
70232
Avg Tweet Length
23.50
18.43