

2020 Final Project

Is social media a better predictor for election outcomes?

The Problem

Polling

The current polling system continues to miss ongoing voting outcomes and fail to capture voter sentiment and deliver accurate representations of voter intentions and eventual 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 Bolsonaro and British 2019 Parliament,

Problem statement

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

Challenges deep-dive

Challenge 1

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

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

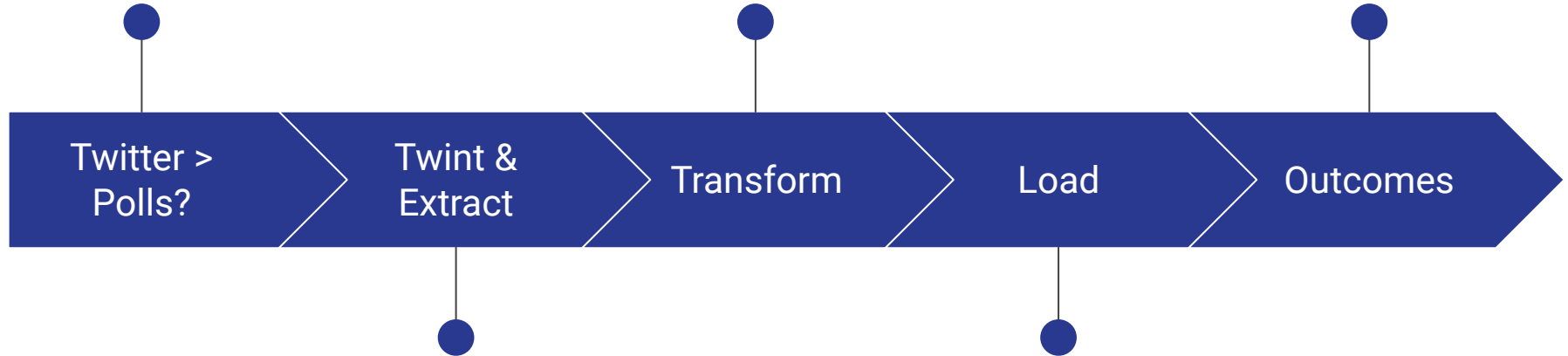
How to Effectively Transform the data for use?

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 and sentiment extraction?



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, we will need to Find(Extract the Data from a program called TWINT which pulls Twitter data), Transform(Structure the Twint Data and tokenize Key Trump and Biden names) and Load the data into a prediction model(Using Random Forest) train and predict for outcomes based on the number of likes and retweets of the two candidates

Machine Learning and NLP Update

The group decided to observe and present NLP on 3 key events: First Debate, Town Hall and Second Debate. The following slides only detail the First Debate surrounding Biden tweets. We will update and provide NLP/Machine Learning data findings and sets for 3 total events, with two subsets for each candidate, a total of 6 detailed presentations.

Implementation

Database Mongo DB Code Import

MongoDB Import Data

Import Dependencies

```
In [5]: from pymongo import MongoClient
import os
import pandas as pd
```

Connecting with MongoDB

```
In [2]: # Creating a connection with MongoDB
client=MongoClient('localhost', 27017)
```

```
In [3]: # Providing list of collections under database called 'test1'
db=client.test1
collect_names=db.list_collection_names()
collect_names
```

```
Out[3]: ['trump3_20_df']
```

```
In [6]: # Open collection in python
data=db.trump3_20_df
h_list=data.find()
trump3_20=pd.DataFrame(list(data.find())) # Creating new name for database in python
trump3_20.head()
```

```
Out[6]:
```

	_id	id	conversation_id	created_at	date	time	timezone	user_id	username	name
0	5f9f437262b1a9ea2baac1b6	1319790707717136384	1319761576996573186	2020-10-23 19:59:59 Eastern Daylight Time	2020-10-23	1900-01-01 19:59:59	-400	781454486	redmazuratii	redmazurati

Export Mongo DB

MongoDB Export Data

Import Dependencies

```
In [1]: import pandas as pd
import datetime as dt

# Import Dependency to create functions to export cleaned datasets to MongoDB
import pymongo
```

Reading CSV Files

```
In [6]: # Reading a test csv file to demonstrate operability
trump3_20_df=pd.read_csv("../2020_Data/trump3_2020.csv")

C:\Users\Greg\anaconda3\envs\PythonData\lib\site-packages\IPython\core\interactiveshell.py:3063: DtypeWarning: Columns (9,22) have mixed
types.Specify dtype option on import or set low_memory=False.
interactivity=interactivity, compiler=compiler, result=result)
```

Function to Clean Datasets

```
In [7]: # Creating function to clean datasets(this is just an example)
def drop_cols(db):
    col_list=['place', 'retweet', 'quote_url', 'thumbnail', 'near', 'geo','source',
              'user_rt_id', 'user_rt', 'retweet_id', 'retweet_date', 'translate',
              'trans_src','trans_dest']
    db.drop(col_list, axis=1, inplace=True)
    db['date'] = pd.to_datetime(db['date']) # Converting to datetime
    db['time'] = pd.to_datetime(db['time'], format='%H:%M:%S') # Converting to datetime

    return ('Dataset Cleaned')
```

Function to Export Datasets to MongoDB

```
In [8]: def export_collection(data):
    name=[x for x in globals() if globals()[x] is data][0] # Assigning name of DataFrame to variable 'name'
    a_dict=data.to_dict('records') # Creating a Dict() to send to MongoDB
    client=pymongo.MongoClient('mongodb://localhost:27017/') # Making a connection to MongoClient
    db=client['test1'] # Using the database called 'test1'
    user_info_table=db[f'{name}'] # Creating a collection (or table)
    user_info_table.insert_many(a_dict) # Inserting collection to dictionary created above
    return ('Collection Inserted to Mongo Database')
```

Cleaning Dataset

```
In [9]: drop_cols(trump3_20_df)
```

```
Out[9]: 'Dataset Cleaned'
```

Loading Dataset into MongoDB

```
In [10]: export_collection(trump3_20_df)
```

```
Out[10]: 'Collection Inserted to Mongo Database'
```

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	dailyphoei	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	maya7386	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	woodrow6	Woodrow		Pres. you missed the op to mention HBCU. @JoeBiden obviously doesn'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	SADIE	2020	@Jillbiden46 @JoeBiden â€œVoteBidenHarris2020 â€œ
7	9/30/2020	19:59:59	-400	3.4E+07	reenechel	Renee		@CNNPolitics @JoeBiden @TheView @JoyVBehar @AnnCoulter @FoxNews @kanyewest @KevinHart4real @ABCNetwork THIS IS ANTIFA
8	9/30/2020	19:59:59	-400	2.4E+08	monkjonk	Monk	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 @lonlyPlayDumb4U @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€œ
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_joni	Ed W. Jones		@JoeBiden #TrumpTrain2020 https://t.co/XYQXxOBnCa
18	9/30/2020	19:59:58	-400	3.1E+09	geordie_p	Paul		@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								

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)
```

Code Used to Clean Tweet Column Data from Special Characters (https: links and #'s) Step 2

- Clean up special characters ([^\\w\\s#@/:%._-])

```
biden_1_debate_df['cleaned_tweet'] = biden_1_debate_df['cleaned_tweet'].str.replace('[^\\w\\s#@/:%._-]', '', flags=re.UNICODE)
```

- Add "NA" to cleaned up new "cleaned_tweet" column

[19]:

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

- Keeping only Non "NA" data in "cleaned_tweets"

[21]:

```
biden_1_debate_df = biden_1_debate_df[biden_1_debate_df['cleaned_tweet'].notna()]
```


Cleaned Data Set 3

*Key Data point is cleaning up special characters in Tweets to arrive to Clean Data in the Final Column

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	id	conversat	created_a	date	time	timezone	user_id	username	name	tweet	language	mentions	urls	photos	replies_cc	retweets	likes_cou	hashtags	cashtags	link	video	thumbnail	reply_to	cleaned_tweet
2	1.3E+18	1.3E+18	2020-09-13	19:59:59	-400	1.29E+18	dailyphoe	Daily Pho	@kathyhcn	['kathyhol']	'https://v			1	0	2	[]	[]		https://tw	0		{'user_id':	Should we have CRT in schools We need
3	1.3E+18	1.3E+18	2020-09-13	19:59:59	-400	1.28E+18	maya738l	maya	@RockettEn	['rockett54']					0	0	0	[]		https://tw	0		{'user_id':	Literally trump but ok
4	1.3E+18	1.3E+18	2020-09-13	19:59:59	-400	5.5E+08	woodrowi	Woodrow	Pres. you	en	['joebider']				0	0	0	['blackvot']		https://tw	0		{'user_id':	Pres. you missed the op to mention HB
5	1.3E+18	1.3E+18	2020-09-13	19:59:59	-400	3.4E+07	reneechel	Renee ðŸŒ	@CNNPol	en	['cnnpolit']				0	0	0	[']		https://tw	0		{'user_id':	THIS IS ANTIFA
6	1.3E+18	1.3E+18	2020-09-13	19:59:59	-400	2.4E+08	monkjona	Monk Aed	#JoniMit	und	en	['morning']	'https://v		0	0	0	['jonimite']		https://tw	0		{'user_id':	-
7	1.3E+18	1.3E+18	2020-09-13	19:59:59	-400	1.7E+09	ikechukwi	Rafe Miya	@Imageci	en	['imageci']				2	0	0	[']		https://tw	0		{'user_id':	Yeah silence is oppression especially
8	1.3E+18	1.3E+18	2020-09-13	19:59:59	-400	1.12E+18	kempson_	Luke Kemp	@Scottd1	en	['scottdd1']				0	0	0	[']		https://tw	0		{'user_id':	I couldnt give a fuck about Mansfall...
9	1.3E+18	1.3E+18	2020-09-13	19:59:59	-400	2E+07	modeka	Modeka	@realDor	en	['realdons']				0	0	0	[']		https://tw	0		{'user_id':	sure knows how to pull in a crowd doe
10	1.3E+18	1.3E+18	2020-09-13	19:59:58	-400	1.26E+18	lpryanovi	Desdemor	@Ramboj	en	['rambop']				0	0	2	[']		https://tw	0		{'user_id':	Exactly. And I admit from when my pho
11	1.3E+18	1.3E+18	2020-09-13	19:59:58	-400	1.7E+08	kevinsteir	Dza	Biden sell	en	['joebider']				0	0	0	[']		https://tw	0		{'user_id':	Biden selling shirts with trump face or
12	1.3E+18	1.3E+18	2020-09-13	19:59:58	-400	4.5E+07	s1lentone	Ivan Pere;	@Plu9to	en	['plu9to']				1	0	2	[']		https://tw	0		{'user_id':	ANTIFA is not an organization you dop
13	1.3E+18	1.3E+18	2020-09-13	19:59:58	-400	3.1E+09	geordie_p	Paul ðŸŒ	@Martinf	en	['martinbi']				0	0	0	[']		https://tw	0		{'user_id':	And youve spilled beer down your shin
14	1.3E+18	1.3E+18	2020-09-13	19:59:58	-400	2.8E+09	martinem	Emily Mai	@umdbul	en	['umdbull']				0	0	0	[']		https://tw	0		{'user_id':	Me too
15	1.3E+18	1.3E+18	2020-09-13	19:59:58	-400	1.29E+18	rome6722	Rome	@JoeBide	en	['joebider']				0	0	0	[']		https://tw	0		{'user_id':	Best president ever
16	1.3E+18	1.3E+18	2020-09-13	19:59:57	-400	1.30E+18	too_survi	SoulSurviv	@Chrissi	en	['chrisserie']				0	0	2	[']		https://tw	0		{'user_id':	He knows no shame.
17	1.3E+18	1.3E+18	2020-09-13	19:59:57	-400	9.50E+17	browntow	Aaron an	@knallfa	en	['knallfall']				3	0	0	[']		https://tw	0		{'user_id':	That would fail at first instance and y
18	1.3E+18	1.3E+18	2020-09-13	19:59:57	-400	1.30E+18	beoverloc	Don't Be	C @Joebide	en	['joebider']				0	0	0	[']		https://tw	0		{'user_id':	I didnt know that a idea could be ar
19	1.3E+18	1.3E+18	2020-09-13	19:59:57	-400	1.16E+18	popsixtye	POPSIXTY	@JoeBide	en	['joebider']				0	0	0	[']		https://tw	0		{'user_id':	Will you disavow BLACK supremacists
20	1.3E+18	1.3E+18	2020-09-13	19:59:57	-400	2.8E+07	freitweete	ðŸŒ@ðŸŒ	@DSunkle	de	['dsunkler']				0	0	0	[']		https://tw	0		{'user_id':	Mit Selbstreflektion haben die Politike
21																								

Machine Learning Observations and Answers

1. Preprocessing of preliminary data is importing TextBlob to provide feedback on sentiment, utilizing two metrics, polarity and subjectivity.
2. Our model, sentiment, is unsupervised machine learning,. We do not use previous/past results as we don't go back and measure sentiment of twitter ID's previous tweets(No actual training and testing.
3. Limitations: This analysis only provides unique, daily representations of twitter user sentiment. We could further delve into how sentiment changed from the first debate for each twitter user over the 4 weeks of data that we pull from Twint. Furthermore, we don't have much understanding of why a Twitter ID is the way they are, or speak out that way. But the one benefit is we capture unique, user data to provide observations about how people actually react and respond to today's pressing issues and elections.

NLP Code: Establish Sentiment Feedback

```
In [32]: from textblob import TextBlob
```

```
In [33]:
```

```
Feedback1 = "KamalaHarris Yeah for the worse We need a productive President one who stands for us all The ONLY MAN FOR THE JOB IS  
Feedback2 = "realDonaldTrump The 2020 election is a matter of life and death. So vote like your life depends on it BidenHarris2020  
Feedback3 = "realDonaldTrump Hell no VoteBlueToEndThisNightmare VoteBidenHarris2020 VoteBidenHarrisToSaveAmerica"  
Feedback4 = "Contrary to what the pundits in the mainstream media would have you believe, President Donald Trump obliterated Joe  
Feedback5 = "Donald Trump is amazing. He is the greatest president ever, Joe Biden is a racist, downright stupid, ignorant human"
```

```
blob1 = TextBlob(Feedback1)  
blob2 = TextBlob(Feedback2)  
blob3 = TextBlob(Feedback3)  
blob4 = TextBlob(Feedback4)  
blob5 = TextBlob(Feedback5)
```

```
print(blob1.sentiment)  
print(blob2.sentiment)  
print(blob3.sentiment)  
print(blob4.sentiment)  
print(blob5.sentiment)
```

```
Sentiment(polarity=-0.2, subjectivity=0.8)  
Sentiment(polarity=0.0, subjectivity=0.0)  
Sentiment(polarity=0.0, subjectivity=0.0)  
Sentiment(polarity=0.2, subjectivity=0.2)  
Sentiment(polarity=0.20000000000000004, subjectivity=0.75)
```

```
In [34]:
```

```
import nltk  
import string  
import pandas as pd  
  
from collections import Counter  
import matplotlib.pyplot as plt
```


NLP Code: Isolate Tweets and Find Subjectivity and Polarity

```
In [36]: df=pd.DataFrame(biden_1_debate_df, columns=['cleaned_tweet'])
df.head()
```

Out[36]:

	cleaned_tweet
0	Should we have CRT in schools We need a leader.
1	Literally trump but ok
2	Pres. you missed the op to mention HBCU. obvio...
3	THIS IS ANTIFA
4	-

```
In [37]: # Create function to obtain subjectivity
def getSubjectivity(text):
    return TextBlob(text).sentiment.subjectivity

# Create function to obtain polarity
def getPolarity(text):
    return TextBlob(text).sentiment.polarity

# Create Columns for Subjectivity and Polarity
df['Subjectivity']=df['cleaned_tweet'].apply(getSubjectivity)
df['Polarity']=df['cleaned_tweet'].apply(getPolarity)
df.head()
```

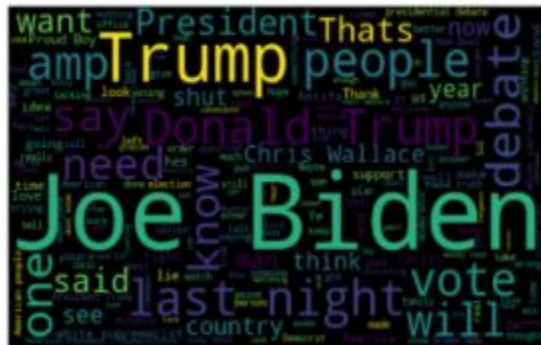
Out[37]:

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

NLP Code: Generate WordCloud

```
In [38]: # Plotting a word cloud
import matplotlib.pyplot as plt
from wordcloud import WordCloud

allWords= ' '.join([twts for twts in df['cleaned_tweet']])
wordCloud=WordCloud(width=500, height=300, random_state=21, max_font_size = 119).generate(allWords)
plt.imshow(wordCloud, interpolation='bilinear')
plt.axis('off')
plt.show()
```



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]:

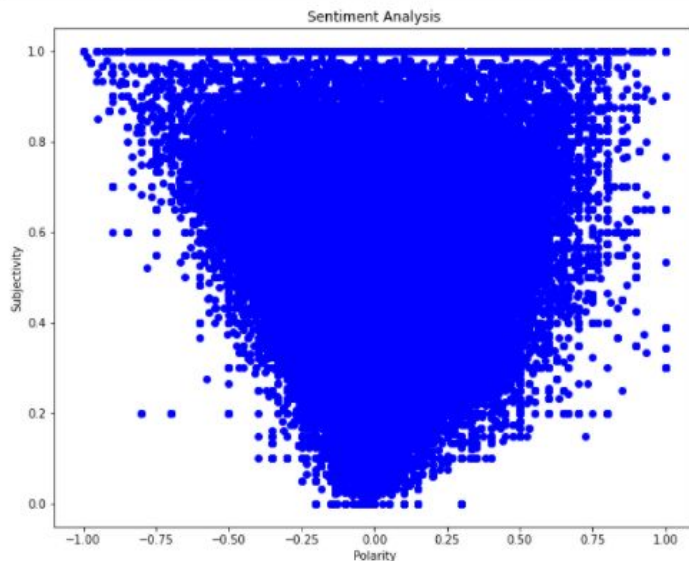
	cleaned_tweet	Subjectivity	Polarity	Sentiment
0	Should we have CRT in schools We need a leader.	0.000000	0.000000	Neutral
1	Literally trump but ok	0.500000	0.500000	Positive
2	Pres. you missed the op to mention HBCU. obvio...	0.333333	-0.111111	Negative
3	THIS IS ANTIFA	0.000000	0.000000	Neutral
4	-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

NLP Code: Tweet Sentiment Analysis

In [40]:

```
# Plotting polarity and subjectivity
plt.figure(figsize=(10,8))
plt.scatter(df['Polarity'], df['Subjectivity'], color='Blue')

plt.title('Sentiment Analysis')
plt.xlabel('Polarity')
plt.ylabel('Subjectivity')
plt.savefig('Resources/sentiment_test.png')
plt.show()
```



Tableau

For the DashBoard, we will use Tableau to represent data findings

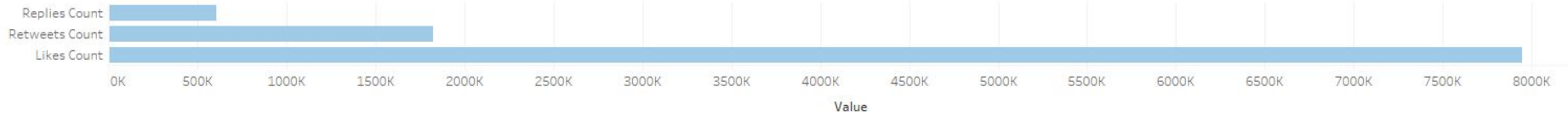
Trump Tableau Dashboard

<Trump>




Biden Tableau Dashboard

Sheet 5



Flaws in model

- Social Media is a conglomerate of data inputs from billions of people and we are only using one key piece of the entire Social Media universe. Therein lies the issue with accuracy and can be fine tuned by incorporating more sources of Social Media Data (Google analytics, Facebook, Instagram etc)
 - Do certain tweets/words carry for effective influence over readers? Hard to quantify if a certain tweet persuades more voters to actually vote or suppress the opposition/readers
 - How can we be certain social media is even a reliable source for election prediction on its own but may in fact just be a part of a social web of influence where one's ability to predict election outcomes
- 

Observations

