

Tweet Times and Vocabulary Sentiment of Donald J. Trump

Adrian Sandoval-Vargas

Department of Computer Science

Oakland University

Abstract

This study aims to figure out Donald J. Trump's popular tweet times, topics in which he was active on, and his sentiment on tweeted topics. The data was extracted using Twitter's API [1] and Tweepy [2] across 16 days by searching the user @realdonaldtrump and filtered out any retweets. In this study we will look into data preprocessing, data visualization to extract key events, and perform sentiment analysis on the filtered data.

Introduction

This report will focus on how the data was extracted, cleaned, processed, and analyzed. It will include visuals of the data collected and what tools were used to obtain the data and visualized. The data are tweets from Donald J. Trump from 11/20/2019 – 12/05/2019. These tweets were extracted from Twitter using Twitters API Calls and Tweepy as the interface. The tweets were extracted in UTF-8 format to keep integrity of the data for when we process the data. The data was filtered for https links, '@, #, & amp', stop words using the NLTK library [3], Visualizations were done by matplotlib, and Sentiment Analysis was done using TextBlob [4].

The paper will be laid out in 5 sections. In section 1, I will introduce the data, discuss and visualize key features, and perform data cleansing techniques to get a clean dataset. In section 2, I will discuss on the numerical side of the tweets (which is Date and Time) and investigate Trump's tweets throughout the time period the data was collected. Section 3 will be similar to section 2, but I will be doing sentimental analysis on Trump's tweets. Section 4 will have a specific case study on what was Trump tweeting about on his most active day. Section 5 will discuss the overall results of the project.

1. Data Extraction, Data Cleansing, and Preprocessing

The data was extracted from Twitter by using Tweepy's library along with Twitter's API keys. The extraction code:

```
auth = tweepy.OAuthHandler(key, sec)
auth.set_access_token(at, atc)
api = tweepy.API(auth, wait_on_rate_limit=True)
with open('realdonaldtrump2.csv', 'a', encoding='utf-8', newline='') as file:
    writer = csv.writer(file)
    for tweet in tweepy.Cursor(api.user_timeline, id='realdonaldtrump',
                               lang="en", include_rts = False):
        print (tweet.created_at, tweet.text)
        writer.writerow([tweet.created_at, tweet.text])
```

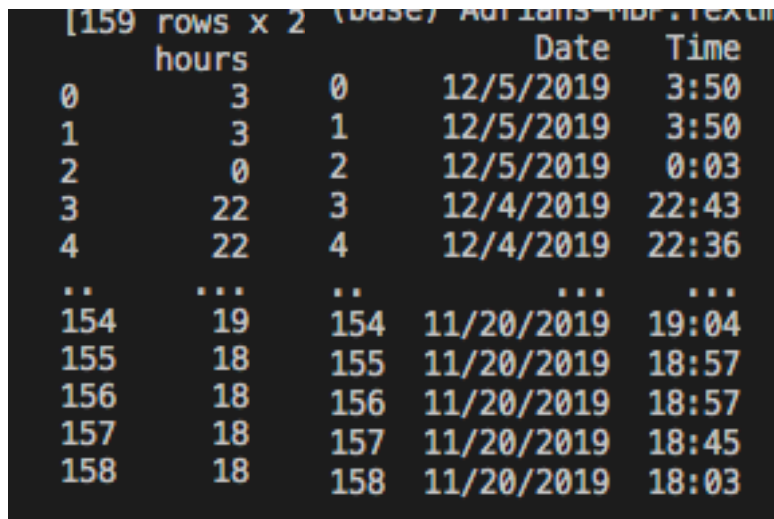
This generated a csv file that we can now import into our project for data analysis. Once we import that data using pandas, we get a data frame that looks like so:

	Date	text
0	2019-12-05 03:50:35"I would like to have the Attorney General...
1	2019-12-05 03:50:35	When I said, in my phone call to the President...
2	2019-12-05 00:03:14	Just read the best Maureen Dowd column, in the...
3	2019-12-04 22:43:19	.@NATO has now recognized SPACE as an operatio...
4	2019-12-04 22:36:31	The Fake News Media is doing everything possib...
..
154	2019-11-20 19:04:30	I WANT NOTHING! https://t.co/KKUfwSIRAi
155	2019-11-20 18:57:07"I WANT NOTHING! I WANT NOTHING! I WANT NO...
156	2019-11-20 18:57:06	Impeachment Witch Hunt is now OVER! Ambassador...
157	2019-11-20 18:45:36	We join families of Kevin King & Tim Weeks...
158	2019-11-20 18:03:40	https://t.co/HbgEgZsPZ9

Figure 1: Raw data captured by Tweepy

As we can see we have many unfavorable characters and text in our data. Our first step is to separate our Date column into Date and Time. To do so I simply looped the dataframe['Date']

column and split the text by the whitespace in between. Since we won't be looking into too much detail on the exact time he tweeted, we round the time to the hour.



hours		Date		Time
0	3	0	12/5/2019	3:50
1	3	1	12/5/2019	3:50
2	0	2	12/5/2019	0:03
3	22	3	12/4/2019	22:43
4	22	4	12/4/2019	22:36
...
154	19	154	11/20/2019	19:04
155	18	155	11/20/2019	18:57
156	18	156	11/20/2019	18:57
157	18	157	11/20/2019	18:45
158	18	158	11/20/2019	18:03

Figure 2: Hours data frame on the left, Date/Time on the right

Now that we have the time separated, we have to preprocess the text. First, we have to extract the urls from our tweets. This was done by using the UrlExtract Library [5]. The code is:

```
w = []
for i in trump['text']:
    if extractor.has_urls(i):
        url = extractor.find_urls(i)
        for k in url:
            i = i.replace(k, " ")
            w.append(i)
    else:
        w.append(i)
```

This code checks to see there is a url in the tweet and replaces the url with whitespace. Next we have to remove the following symbols ['@', '#', 'ɪmp;']:

```
#remove '@' and '#'
for i in range(len(w)):
    if '@' in w[i]:
```

```

w[i] = w[i].replace('@','')
if '#' in w[i]:
    w[i] = w[i].replace('#','')
if '&' in w[i]:
    w[i] = w[i].replace('&', '')

```

This generates an array of all the tweets free of unwanted data. Since we will be using the NLTK library for the stop words and the Sklearn library for the CountVectorizer [6] function we need to make the 'w' array into a single array:

```

w = ' '.join(w)
z = []
z.append(w)

```

Now, to prepare the data for some analysis we will be making a boolean matrix with the following code:

```

x = vec.fit_transform(z)
y = [i for i in vec.vocabulary_.keys()]
y.reverse()
bM = pd.DataFrame(x.toarray(),index=['count'],columns=y).T

```

This puts our data at our disposal to further our knowledge on the data we mined.

2. The Time Data

In this section we'll dive into the timestamps of the tweets made by Donald J. Trump and point out key features. Firstly, I would like to take a look at Trump's overall tweets per hour

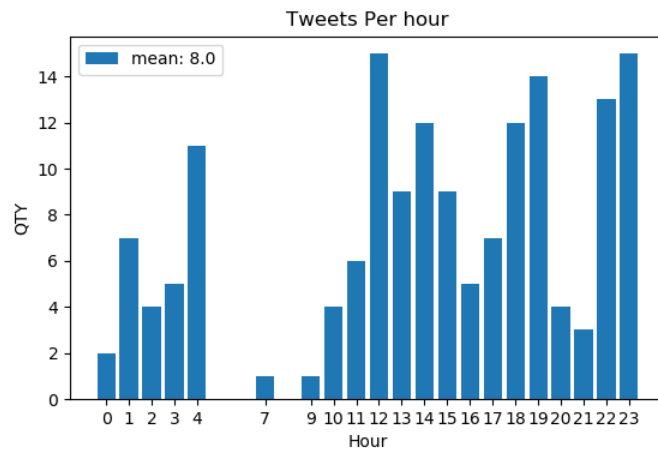


Figure 3: Trump's Tweets per hour

As shown in figure 3, trump is super active during 12:00, 19:00, and 23:00. He averages 8 tweets per hour. Continuing,

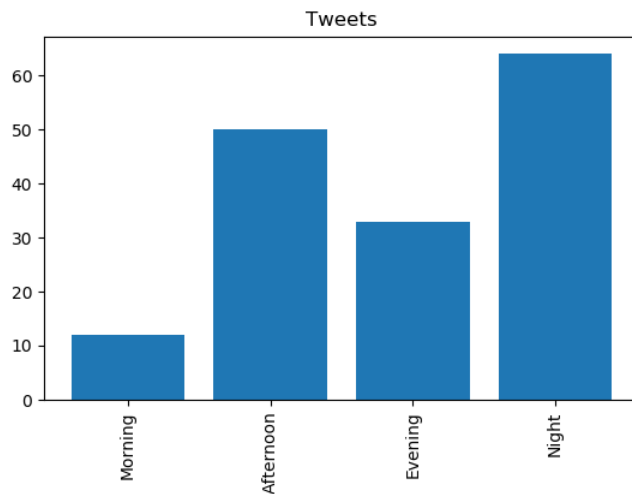


Figure 4: Trump popular time of day to tweet

Trump seems to tweet more active between 8pm – 6am as apposed to any other time of day.

Next I'll be looking at the tweets per day information:

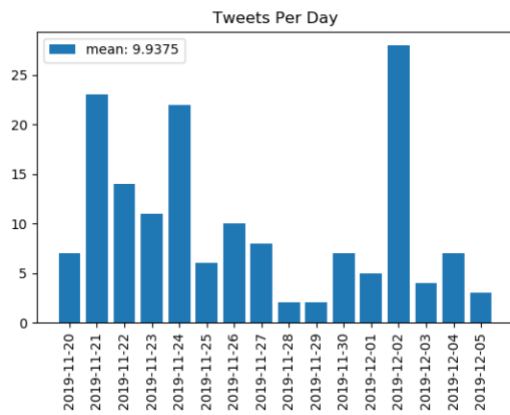


Figure 5a: Trump's tweets per day

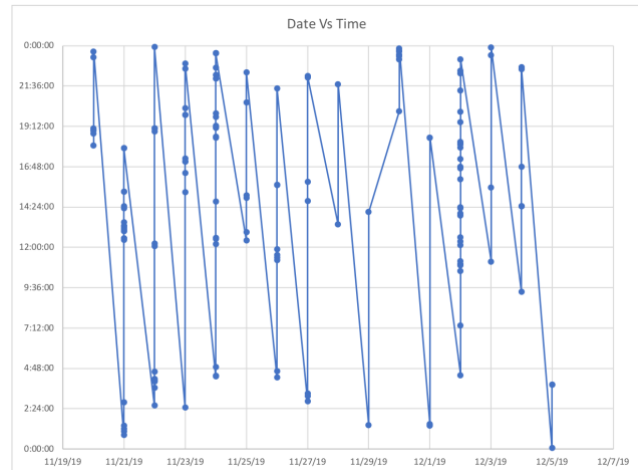


Figure 5b: Trump's Time Vs Date tweets

As show in figure 5a, He had a significant twitter day on 2019-12-02 and he averages around 10 tweets per day. Also in figure 5b, we can see the relation bewteen the quantity of tweets and the time of day he is most active. Mornings are pretty dry and evenings are heavy with tweets.

3. The Text Data

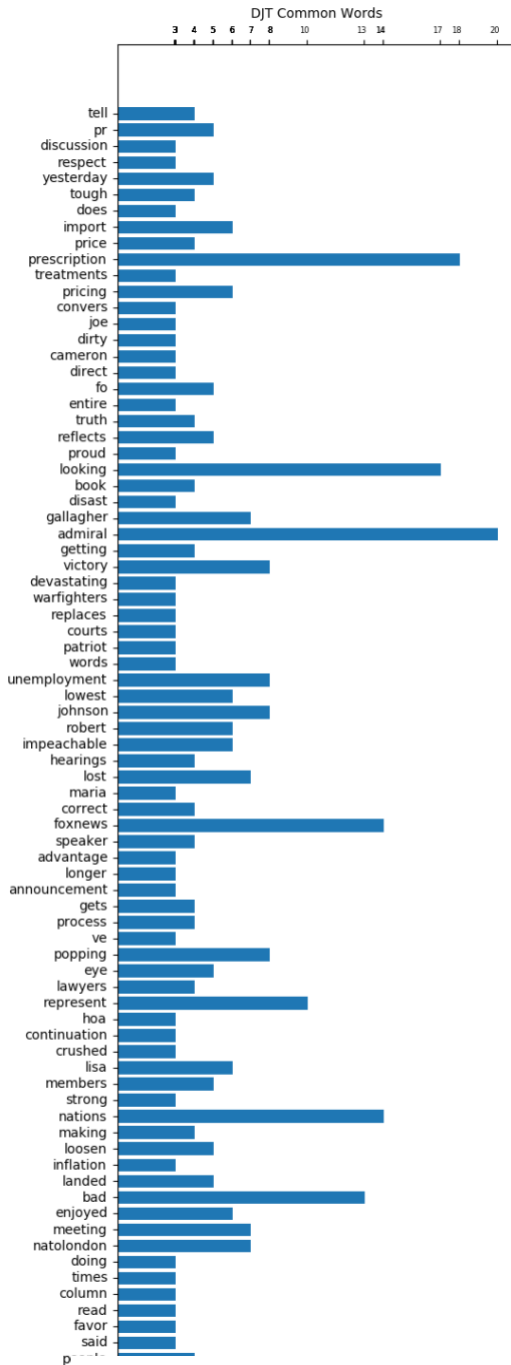


Figure 6: Common words of data

In this section we will take a look into the tweets alone for sentiment analysis. I used the libraries NLTK, TextBlob, and Sklearn to apply stop words, sentiment, and vecotrize the count of words respectively. In section 2 we obtained constructed a data frame 'bM' that contained data of all the words in the tweets and the count of the words. To avoid a long list words, I have selected words in which Trump mentioned more than twice.

We can see in Figure 6 the most common words used by Trump in the time interval the dat was collected. Some notable words are bad, nations, represent, foxnews, admiral gallagher, looking, and prescription. These alone don't give us a huge amount of knowledge but if you look at the neighbors words we can have an idea of what Trump was talking about. For example, around 'Perscription' we collected 'treaments' and 'pricing' which we can infer that Trump is talking about the pricing for perscriptions and treaments. Another area we can see this cohesivness is by the word 'Admiral'. Right

before 'Admiral' we see 'Gallagher' and after 'Admiral' we see 'Getting', 'Victory', 'Devastating', and 'Warfighters'. So throught out his tweets he congradulates Admiral Gallagher

who is a war fighter in on getting a victory. In conjunction with this, I conducted a sentiment analysis on the tweets.

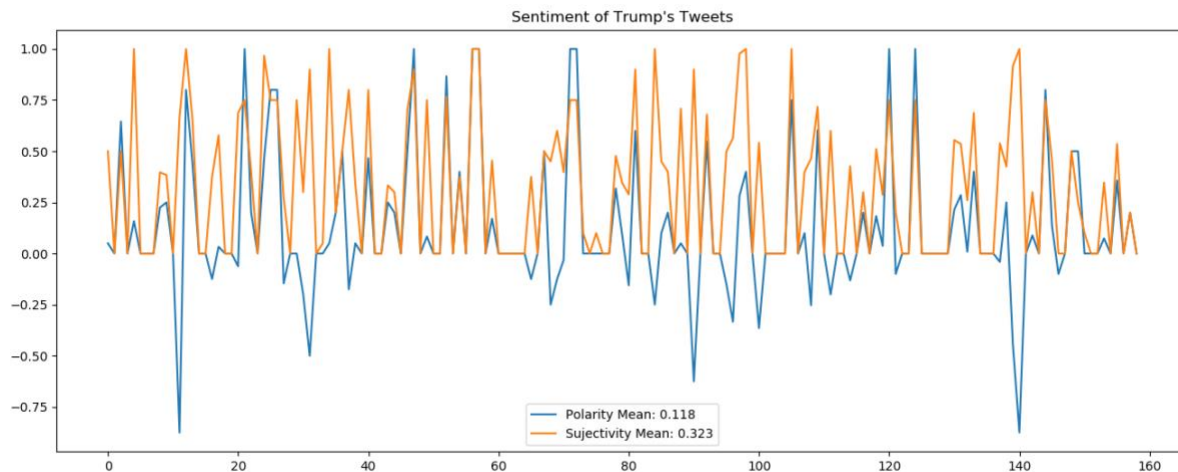


Figure 7: Polarity and Subjectivity of Trumps Tweets

Donald Trump's Tweets had an average Polarity of 0.118 which means he was slightly positive, but his tweets had an average subjectivity of 0.323 which means that he was more subjective to his polarity. I would like to finally mention some people he brought up during this time period:

['Maureen Dowd', 'Mini Mike Bloomberg', 'Lisa Page', 'Peter Strzok', 'Adam Schiff', 'Speaker Pelosi', 'Trump Economy Breaks', 'Doug Wead', 'Donald Trump', 'Tim Scott Says Trump', 'Nancy Pelosi', 'Robert Johnson', 'Oliver North', 'Ukraine', 'Mike Pompeo', 'Rick Perry', 'Mick Mulvaney', 'John Bolton', 'Sam Dewey', 'Ken Braithwaite', 'Richard Spencer', 'Navy Seal Eddie Gallagher', 'Elise Stefanik', 'TRUMP', 'Daniel Cameron', 'Susan Rice', 'Joe Concha', 'Mitch McConnell', 'Ken Starr', 'Trident Pin', 'Bob Mueller', 'Eric Swalwell', 'Gordon Sondland', 'TuckerCarlson', 'ZELENSKY', 'Kevin King', 'Tim Weeks']

So that gives us a list of people that he often talked about in his tweets. Which to no surprise he mentions Mike Bloomberg (Presidential Runner (Democrat)) with a gray insult, Adam Schiff (head of Intelligence Committee for the Impeachment), Nancy Polosi (Speaker of the House), Zelensky (President of Ukraine), and other Fox News Reporters and Republicans.

4. 12/02/2019 Case

In this section I will take a deeper dive into what caused Trump to tweet a lot on 12/02/2019.

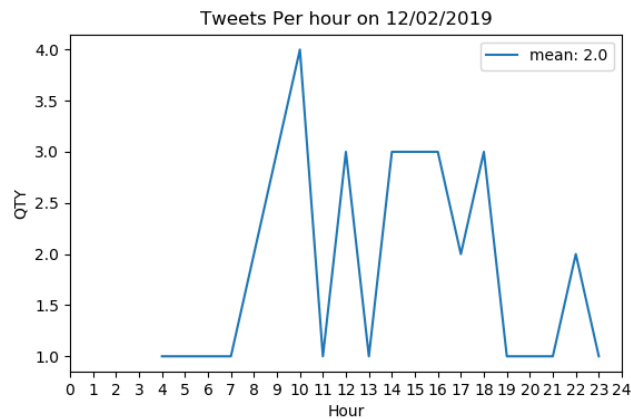


Figure 8: Hour Vs Tweet Qty

Firstly, on 12/02/2019 Trump averaged 2 tweets per hour with max of 4 tweets at around 9:00 AM.

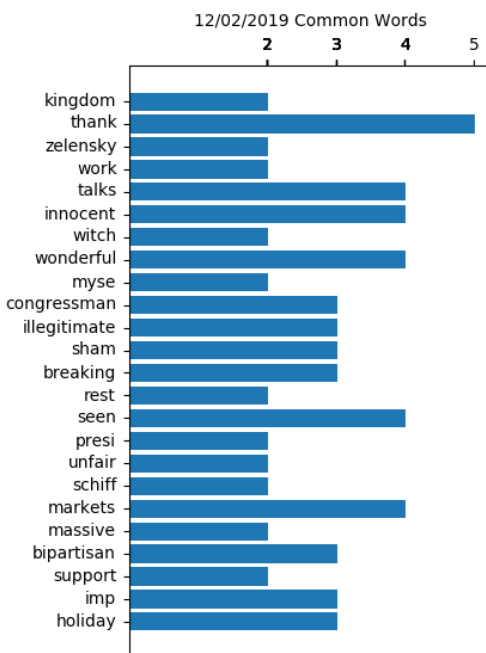


Figure 9: Common on 12/02/2019

Now by extracting words that occurred more than 2 times in his tweets on this day we can group these words into two sections. From 'Zelensky' to 'Schiff', He mentions words like 'innocent', 'congressman', 'illegitimate', 'sham', and 'unfair'. We can infer that something is going on in congress that he isn't very happy about. The other section is 'Markets' to 'Support'. From these section we can infer that he talked about Economies and that bipartisan support is necessary.

Now digging into the sentiment of the tweets on this day.

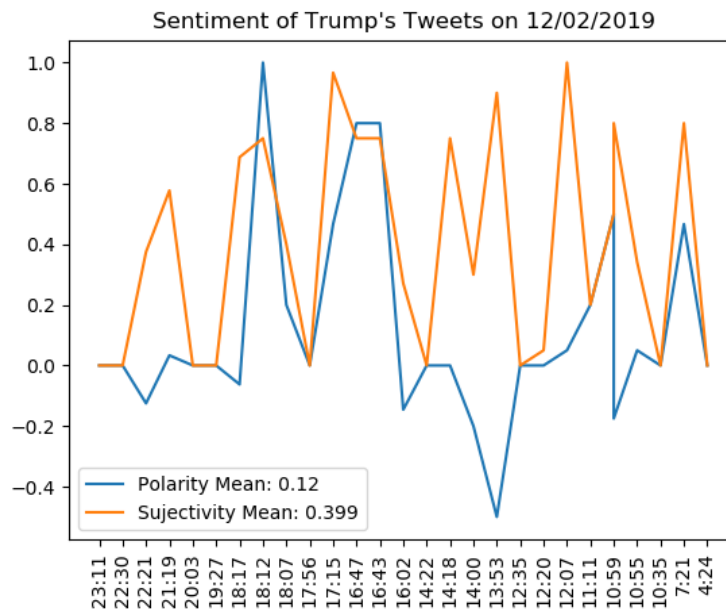


Figure 10: Sentiment of 12/02/2019

On this day Trump had a positive polarity and a high subjective score which mirrors the Overall scores as discussed in section 3.

In conclusion, on this day two major events happened in actuality; Nato Summit in London and Justiciary Committee started hearings from top law professors for Articles of Impeachment. Our data shows that what we extracted from his tweets were fairly accurate on events that truly did happen. Where he talked about impeachment being unfair, a sham, a hoax, and illegitimate. This primarily caused his polarity score to drop as he bashed Congress on the impeachment. What helped him gain some polarity points was the Nato Summit and he talked about the economies. He was overally subjective in his tweets wether it was of the impeachment, Nato, or the economy.

5. Conclusion

The results for this project are that Donald Trump tends to tweet at night more than any time of day. In this time period his sentiment was slightly positive with 3 times more subjectivity. What this tells us is that when Trump is actively tweeting, he is being positive about him self before going to bed.

References

- [1] Twitter API - <https://developer.twitter.com/en/docs>
- [2] Tweepy - https://tweepy.readthedocs.io/en/latest/getting_started.html
- [3] NLTK - <https://www.nltk.org/>
- [4] TextBlob - <https://textblob.readthedocs.io/en/dev/>
- [5] urltextextract - <https://pypi.org/project/urltextextract/>
- [6] CountVectorizer - https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html