Analyzing Twitter Tweets Emotions with NLP Based of 2020 Elections

Samuel Jones, Corbin Whitton, Enoch Levandovsky

COMP4340 Adv Databases Design/Programming

Introduction

This project analyzed the american peoples reactions to the elections by using Natural Language Processing (NLP) to analyze the emotions tweets from the dataset before and after the elections took place. In the following sections we will demonstrate how we used Python to glue the Twitter API and a NLP API to build a dataset of tweets, and how we used Pandas and MatPlotLib to process and display the results.

Background

To do (for the milestone we focused on generating visualizations).

Methods

Collecting Data - Enoch

This Project used Python to connect to and retrieve public Twitter tweets and analyze them with the IBM NLP. We set up a stream for twitter with python and then fed the tweets into IBM's Natural Language Processor, which generated values for each of the major emotions (sadness, joy, anger, disgust, and fear). This data was then stored into separate CSV files to follow the RDBMS database designs principles.

Twitter API

A twitter developer account was requested and created from developer.twitter.com. Once access to the API was given, the API keys were generated. After the authonication was configured, a sample code for streaming tweets was used and applied. At this point the data is ready to stream data, but the search query needed to be changed to match our research. The following queries were created for each candidate. Note that a '-' symbol implies omission.

```
{"value": "trump -biden -has:links -has:images -has:media -has:videos -is:quote lang:en sample:1", "tag": "Trump"}
{"value": "biden -trump -has:links -has:images -has:media -has:videos -is:quote lang:en sample:1", "tag": "Biden"}
```

FINISH

IBM Natural Language Understanding

To do (for the milestone we focused on generating visualizations).

Uploading Data into Hadoop - Corbin

Uploading data to Hadoop. I looked over the data to ensure that it was a proper CSV file. Once I did that, I uploaded the CSV file into HDFS tmp/data/twitter-elections folder. I then ensured that all permissions were given to the file for all users.
I then opened the Data Analytics Studio interface, and tried to create a table from this CSV file. It read the metadata from it fine, such as the data-types of the columns, and the column names. However, when I tried to import it, it said that the user hive does not have access to the. I thought this was strange, so I logged in as the admin, hive, and my personal account, to check the permissions of the file, and repeat the steps above. Still nothing. I then restarted HDP entirely and stopped/started the VM. I then tried to import it into Hadoop that is on my system, which worked flawlessly. Here are some screenshots of the metadata and a brief look at the table.

COLUMN NAME	DATA TYPE
ID	BIGINT
Date	TIMESTAMP
Candidate	STRING
Sadness	DOUBLE
Joy	DOUBLE
Fear	DOUBLE
Disgust	DOUBLE
Anger	DOUBLE

ID	DATE	CANDIDATE	SADNESS	JOY	FEAR	DISGUST	ANGER
1323052849874542594	2020-11-01 16:02:48	Trump	0.395454	0.005437	0.065566	0.570942	0.335284
1323052852810493952	2020-11-01 16:02:50	Biden	0.182015	0.047664	0.101418	0.828351	0.047899
1323052854068670466	2020-11-01 16:02:51	Trump	0.238678	0.144076	0.096391	0.507965	0.231494
1323052857004859394	2020-11-01 16:02:52	Biden	0.182015	0.047664	0.101418	0.828351	0.047899
1323052859102007298	2020-11-01 16:02:53	Trump	0.592663	0.009135	0.190862	0.356536	0.221334
1323052867910045696	2020-11-01 16:02:55	Trump	0.239174	0.051921	0.180257	0.728585	0.059709
1323052869168279552	2020-11-01 16:02:56	Biden	0.169176	0.154784	0.061783	0.083	0.048059
1323052877556994050	2020-11-01 16:02:58	Trump	0.090914	0.012558	0.130612	0.563769	0.563969
1323052892237025280	2020-11-01 16:02:59	Trump	0.029179	0.789288	0.033703	0.168471	0.013685
1323052895592501248	2020-11-01 16:03:00	Biden	0.519236	0.12791	0.191532	0.186832	0.272063

Generating Visualization - Crobin

I started with trying to connect Python to Spark using PySpark, which I was able to get working after extensive testing, however, since I was only recently able to get the data into Hadoop, I

abandoned connecting to Hadoop via Python. The best way to get Python to connect to Hadoop was https://stackoverflow.com/a/26061902 (for my environment).

To work with the CSV files, I decided to use Pandas, it works similarly to SQL, and allows us to import and manipulate the data quickly. I started by importing Pandas as pd, and reading the CSV file. Once the CSV file was read into the dataframe, I parsed the date so it can be manipulated easier. Then I split the single dataframe into two, one for each candidate (this helped plot the data points).

To visualize the data, I used MatPlotLib, I first set the title of the graph, then plot multiple lines on the graph, and saved/displayed the graph. I didn't know how to change the X axis tick labels dynamically (based on the data in the table), so I didn't want to hardcode the solution.

Generating the Quartiles

Apache SQL language had a percentile_aprox() function that I was able to use to create the data. This was different from Corbins Method above as I used the SQL language to process the data and used matplotlib to display the data. The following SQL command below was used to generate the quartiles for each emotion.

```
SELECT day(`date`) as day,
    percentile_approx(`sadness`,0.5,100000) as sadness_50th,
    percentile_approx(`sadness`,0.25,100000) as sadness_25th,
    percentile_approx(`sadness`,0.75,100000) as sadness_75th,
    percentile_approx(`joy`,0.5,100000) as joy_50th,
    percentile_approx(`joy`,0.25,100000) as joy_25th,
    percentile_approx(`joy`,0.75,100000) as joy_75th,
    percentile_approx(`fear`,0.5,100000) as fear_50th,
    percentile_approx(`fear`,0.25,100000) as fear_25th,
    percentile_approx(`fear`,0.75,100000) as fear_75th,
    percentile_approx(`disgust`,0.5,100000) as disgust_50th,
    percentile_approx(`disgust`,0.25,100000) as disgust_25th,
    percentile_approx(`disgust`,0.75,100000) as disgust_75th,
    percentile_approx(`anger`,0.5,100000) as anger_50th,
    percentile_approx(`anger`,0.25,100000) as anger_25th,
    percentile_approx(`anger`,0.75,100000) as anger_75th

FROM emotionsFile
WHERE `candidate` = 'Biden'
GROUP BY day(`date`)
ORDER BY day
```

Since there were a lot of columns, the data was grouped into colors by emotions and the 50th percentile was given a solid line and the other quartiles were given a dotted line. An interactive legend was also added to give the ability to hide and show lines as shown in this form. The data line graphs were then screenshotted.

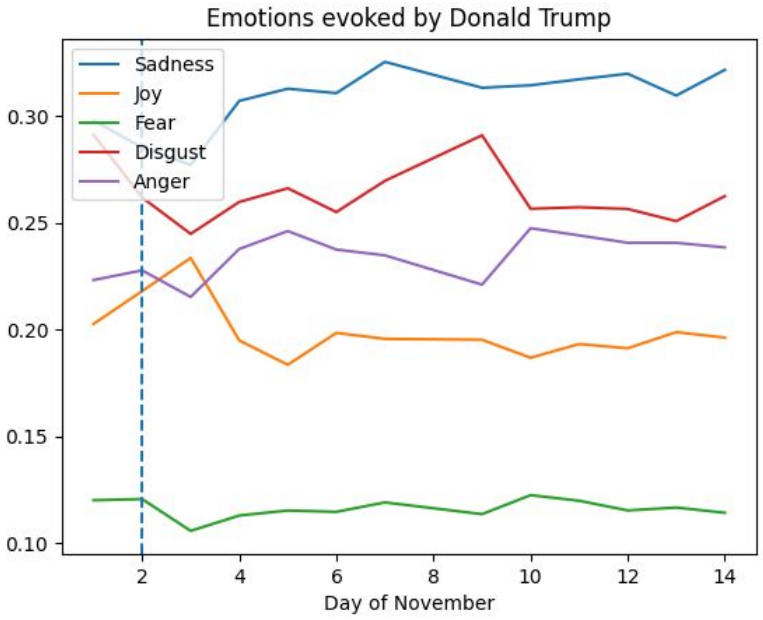
Results/Analysis

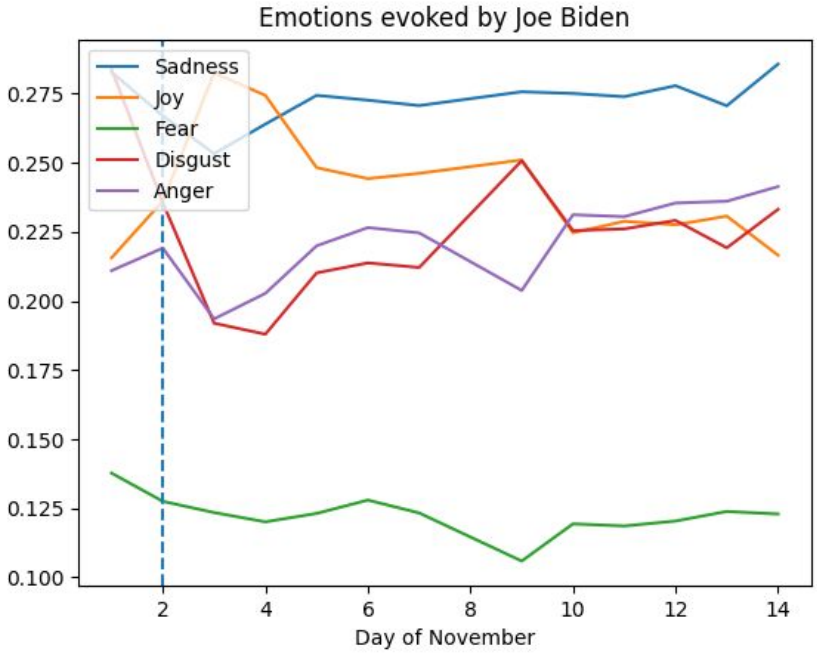
Our original hypothesis was that the emotions of sadness, anger, and disgust would slowing degrade after the elections assuming that the populations would slowly accept the next candidate. Along with this the emotion of joy would rise on the winning candidate right after they declaration of victory and inversely on the losing candidate. We also expected the Q1 and Q3 to deviate further from each other around the elections and slowly deviate closer after the elections, under the assumption of a semi peaceful transition.

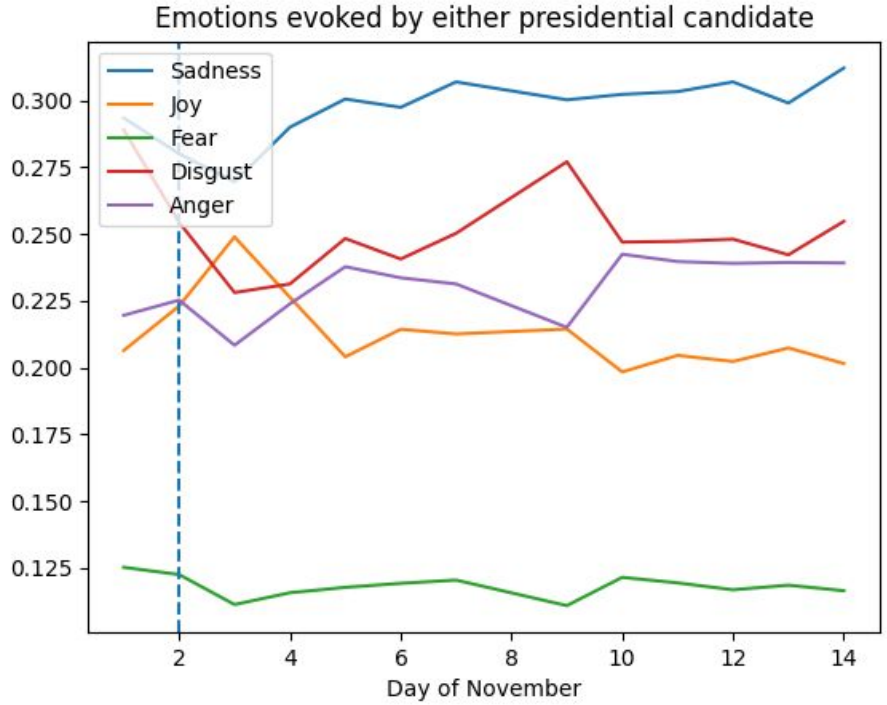
We originally planned on making graphs depicting tweets captured per time period, and 2nd quartile perception of candidates per time period, however, I (Corbin) struggled to learn pandas thoroughly enough to generate these graphs.

We generated 3 different graphs depicting the average emotions per day, per presidential candidate. One for Joe Biden, one for Donald Trump, and an overall graph that depicts the overall emotions felt during election season.

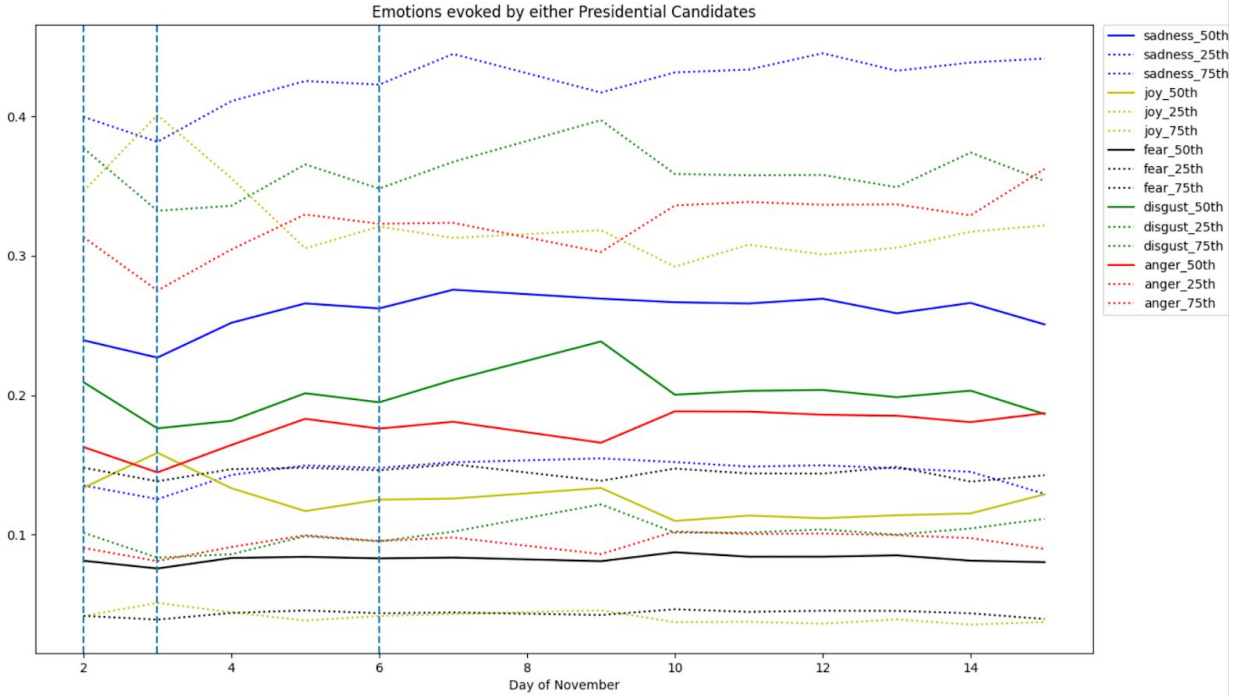
According to our results, it appears that messages that appear the day after election day, are very joyful towards Joe Biden. We placed a line at #2 to signify election day.

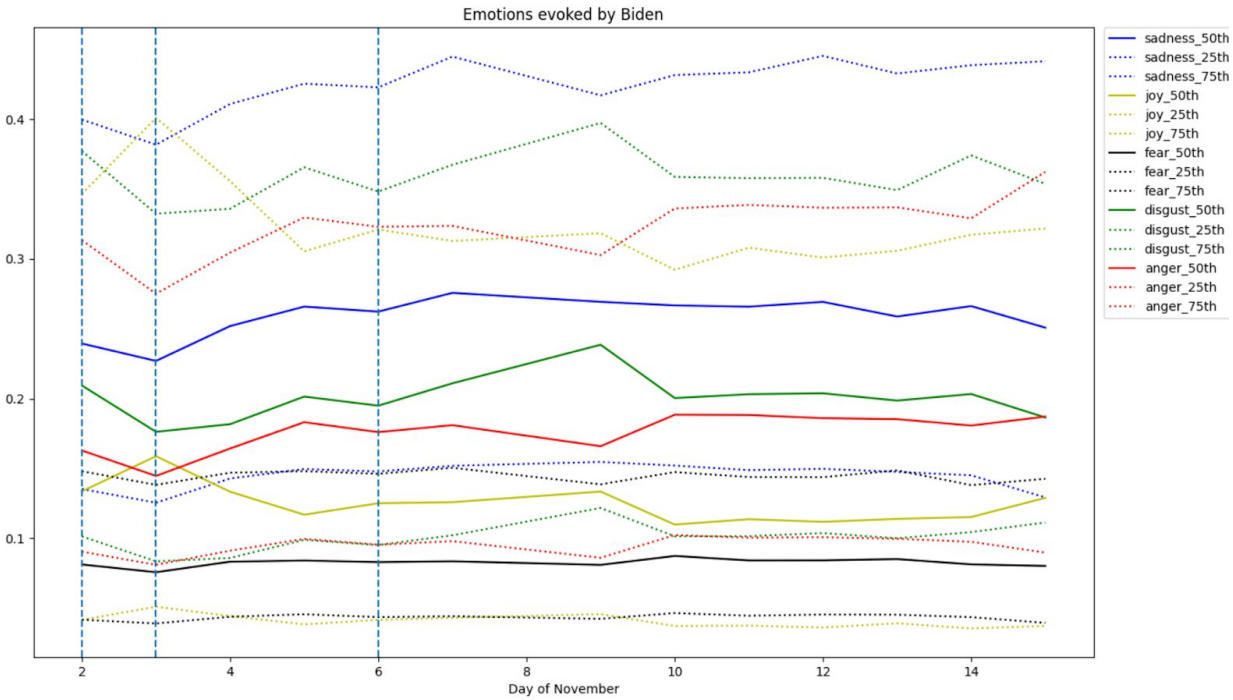


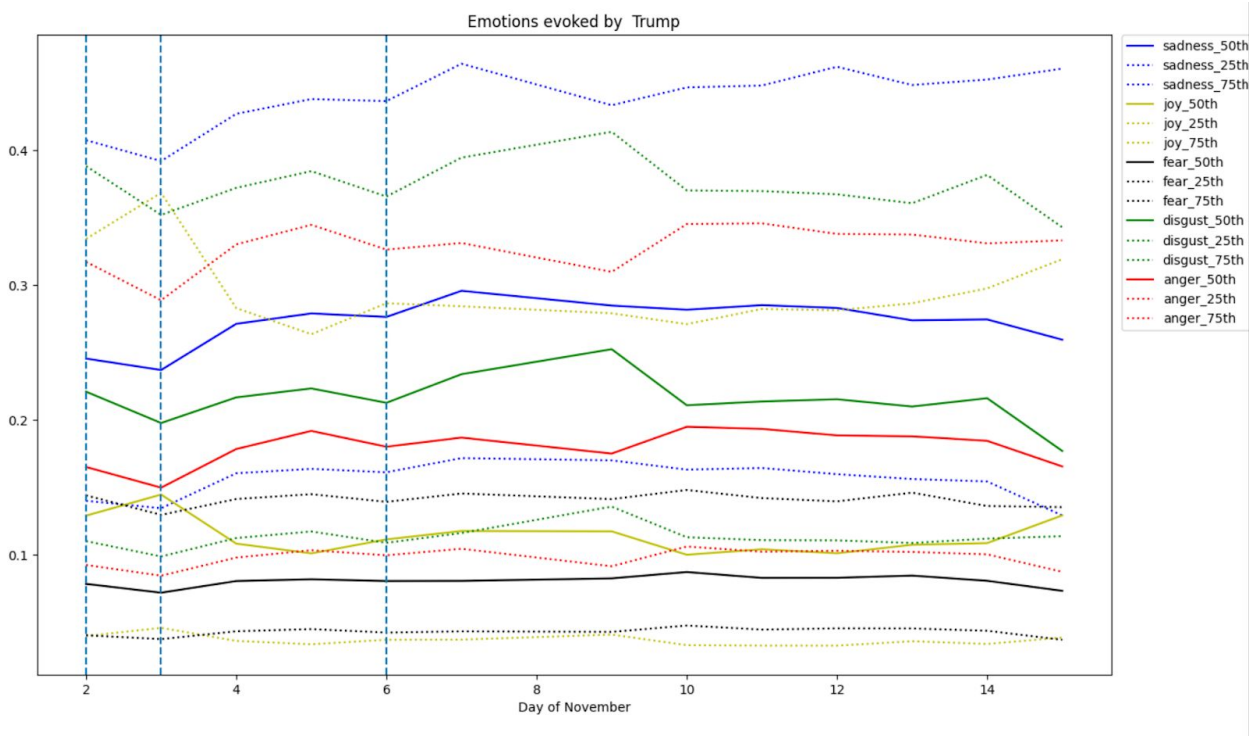

Emotions evoked by Joe Biden
Sadness
Joy
Fear
Disgust
Anger
0.275
0.250
0.225
0.200
0.175
0.150
0.125
0.100
2
4
6
8
10
12
14
Day of November


Emotions evoked by either presidential candidate
Sadness
Joy
Fear
Disgust
Anger
0.300
0.275
0.250
0.225
0.200
0.175
0.150
0.125
2
4
6
8
10
12
14
Day of November

Further and Continued Research by Enoch Levandvosky (time stamp 5:27am)







Things noticed

- Q1 & Q3 sadness quartiles are diverging
- Fear emotion is small and decreasing
- Disgust was decreasing every since before the elections
- Anger towards Biden is increasing while as anger towards Trump is decreasing

The following pie graph was obtained from a simple count(*) query and excel data displayer.