Determining Political Sentiment of Politicians' Twitter Data Utilizing Python

Adrian Vasu

Data Science is a multi-disciplinary field and as such it makes it very difficult to narrow down a single project with reasonable scope. The project chosen to demonstrate the programming skills acquired in the aforementioned PHY 299 course was the development and implementation of a program which analyzed all 100 United States Senators' recent tweets to determine their relative sentiment. This was done by combining the most recent 100 tweets for each senator and breaking this down by political party. This data was then visualized, and two interesting trends were noted. The first trend which is worth mentioning is the peak inversion of the two major political parties (republican and democratic). This means that as the republican party's relative sentiment increased the democrats decreased and vice versa. The second trend is the overall decrease in peaks and leveling out of the sentiment once the COVID-19 stay at home order was issued mid-March. Both major parties' sentiment leveled out towards neutral very quickly once the order was announced. Overall, this project was a successful demonstration of the programming skills acquired and is an interesting application of politicians' twitter data.

1. Introduction

"Big Data", Analytics, Data Science, Information Analysis, Business Intelligence. These are all filed that can trace their roots to a specialized area of mathematics and computer science which has arisen within the past few years. This area out of which all of these fields arise has the goal to adapt and utilize information to develop knowledge. This is done through the establishment of a baseline. This baseline is the ground state in which the aforementioned information is compiled through human interaction to determine patterns or trends which can be analyzed through compute applications. This baseline is then utilized to teach algorithms how to understand and predict outcomes based on similar input data. This is the fundamental basis of this area of computer science. For this project these data science and information analysis fundamentals will be utilized to create an algorithm which will determine the political sentiment of 100 United States senators and report the sentiment as a function of time.

a. Data Science

Data science is defined as the inter-disciplinary field that utilizes varying methods, algorithms, systems, and applications to extract knowledge and comprehension from both structured and unstructured data [1]. This statement clearly captures that data science is not an easily identifiable concept. Its roots are so entangled with the various disciplines from which it is made up. This does not mean that data science is this large mystical being with egregious amounts of power that cannot be tamed. When correctly utilized, the principles of data science allow access insights from information and data that would otherwise be inaccessible.

Since data science involves the collection, study, analysis, and utilization of large data sets it has a heavy basis in statistics. In its most fundamental approach, data science makes use of the Law of Large Numbers (LLN). The LLN states that the average of the results obtained from a

large number of trials will be close (within acceptable error depending on data set size) to the expected value and the error in this measurement decreases as the number of trials increases [2]. This idea is fundamental in the field because it demonstrates that what can be perceived as a series of random events when a small sample is taken normalizes as the sample size increases. This idea has held true throughout the years. According to the Ramsey theory there is no such thing as true or perfect randomness; especially for large data sets [3]. This further demonstrates that once the pattern or algorithm is deduced from studying enough data it can be applied to new data sets and similar deductions can be made about the new data set with relative confidence.

The tricky part is determining or building these deduction or prediction algorithms. One method for creating this algorithm is through the application of Machine Learning (ML). ML is the process of creating algorithms that computers can utilize to complete tasks without an explicit set of instructions.

The idea that the computer can complete these tasks without this explicit instruction set is the key fundamental to machine learning. For example, a machine learning algorithm will begin as a blank slate. This blank slate is merely a box of code which has not been specialized yet. What this means is that it will accept any task or instruction and attempt to complete it but, in these beginning steps, the ML algorithm needs help. This is the training step. The key is that the algorithm is given access to its own code which allows it to continually improve itself without human interaction. Therefore, throughout the training steps this algorithm is constantly improving and iterating upon itself creating new pathways and recognition steps.

Once the training step is complete the ML algorithm is then finalized. All of the permissions to modify itself are then revoked and what is left is a black box. This black box is now highly specialized in deductions or predictions based on the training data set. Say the data utilized to train this algorithm is a set which contained carbon dioxide emissions data for 20 of the 50 states. Once the algorithm has been trained on what the environmental impact was corresponding to the emissions levels of these specific states it can be applied to data from the other 30 states. The algorithm would then predict the environmental impact for these states with whatever confidence was established throughout the training step. This example is one of the numerous applications where ML and data science are utilized to make predictions and analyze data.

There are many applications for these ML algorithms such as computer vision (utilizing cameras and compute power to analyze and determine visual characteristics of objects i.e. quality control on an assembly line) or email filtering (outlook's focused inbox utilizes ML algorithms to sort out emails from real people). This truly shows that machine learning is one of the many applications of the field of data science.

b. Natural Language Processing

Natural Language Processing (NLP) is a field which combines aspects from computer science, artificial intelligence, data science, and linguistics. This field was established to outline how computers can interact with human languages (natural languages). The main focus of the field is how to program compute power to process and analyze this natural language data while understanding connotation [4].

NLP is an extremely interesting field because it is possibly one of the most complex fields of informatics and data science. Human language is constantly evolving is littered with hidden meanings and subtones which are very hard to distinguish because the same word can mean two

very different things depending on the surrounding sentence. This field will be extremely important in the coming years because as humans further their interaction with computers and artificial intelligence becomes more prominent understanding will be a big factor to ensure that miscommunications and interpretations do not occur.

This being said, NLP on a small scale is also extremely difficult because of two main reasons. One the implementation has not been open sourced by any major contributors. Because of this all major NLP open source code is severely behind what is commercially available. Secondly even small amounts of NLP require extreme amounts of compute power due to the large databases and libraries required to store the connotation and how words are perceived depending on their associations.

c. Twitter

Twitter is a social media platform developed by Jack Dorsey which was founded on March 21, 2006. Since then 14 years have passed and Twitter has become an application used daily by more than 145 million people across the world. The differentiating factor between Twitter and its competitors is that Twitter only allows for the sharing of small 280-character messages (including photos or videos sometimes) to a network of followers which one would accumulate over time. These messages are known as "tweets" and are sent out instantaneously when one user decides to post. While Twitter has a set of guidelines which one is supposed to follow regarding what they are able to post, within the past few years Twitter has become a political hotbed with many people engaging in heated discussions about controversial topics. This recent influx of political material into the platform is also motivated by many politicians taking to the platform to share their messages with the "world".

d. Python

Python 3 is a programming language which many scientists and programmers take advantage of on a regular basis. It has become so popular for a few reasons. Python is extremely readable especially when following the Pythonic programming guide. The fact that it is an open language (i.e. it is free and accessible) which can run on any machine which supports c (programming language supported by all Windows, Linux, macOS, Unix operating systems) further adds to its popularity. Finally, Python has numerous packages which can be utilized in development which make it extremely agile. It is an extremely important programming language which is gaining popularity continuously despite already being one of the most popular programming languages in the world.

One package which is extremely important; especially in the field of data science (personal use) is SciPy. SciPy stands for Scientific Python and has various ML algorithms built in that one can use to analyze and report trends in data sets. Another few important packages are: Numpy (Numerical Python), MatPlotLib (Mathematical Plotting Library) and tweepy (Twitter API for Python). When MatPlotLib and Numpy are used in combination this allows for the visualization of complex data. tweepy is the API utilized for this project to interface with the Twitter backend and acquire all of the data [5]. Taking advantage of Python 3 and all of these packages allowed for the creation and completion of this project.

e. In Combination

Programming, Data Science, Machine Learning, Twitter. How do all of these seemingly disjointed topics come together to produce a cohesive project? Well, everything comes together

to create an application implementing data science fundamentals to analyze politicians' Twitter data and determine their sentiment. This sentiment analysis allows the deduction of the overall positivity of a Twitter user and the generation of a graphic which demonstrates this relative positivity and how it changes over time.

f. Project Motivation

The motivation for this project is largely due to the aforementioned recent influx of political messages that have become prevalent on Twitter. There are a lot of conversations about very important topics and not all of them are positive. Whether a politician actually cares about a topic and what their sentiment is should be quantifiable. This program aims to quantify this idea and report the results by political party.

2. Results

The program proposed was written, the data collected and analyzed, and here are the results.

a. Project Outcomes

The overall project outcomes were three sets of results. One for each political party represented in the United States Senate. The Republican dataset had 53 senators' tweets, the Democratic 45, and the Independent 2. To see the official breakdown utilized please refer to the "senators" file in Appendix B. These datasets can be seen reported below in sections B through D with a comparison in section E.

b. Republican Results

Please find the results for the last 100 tweets for each republican senator below in Figure 1.

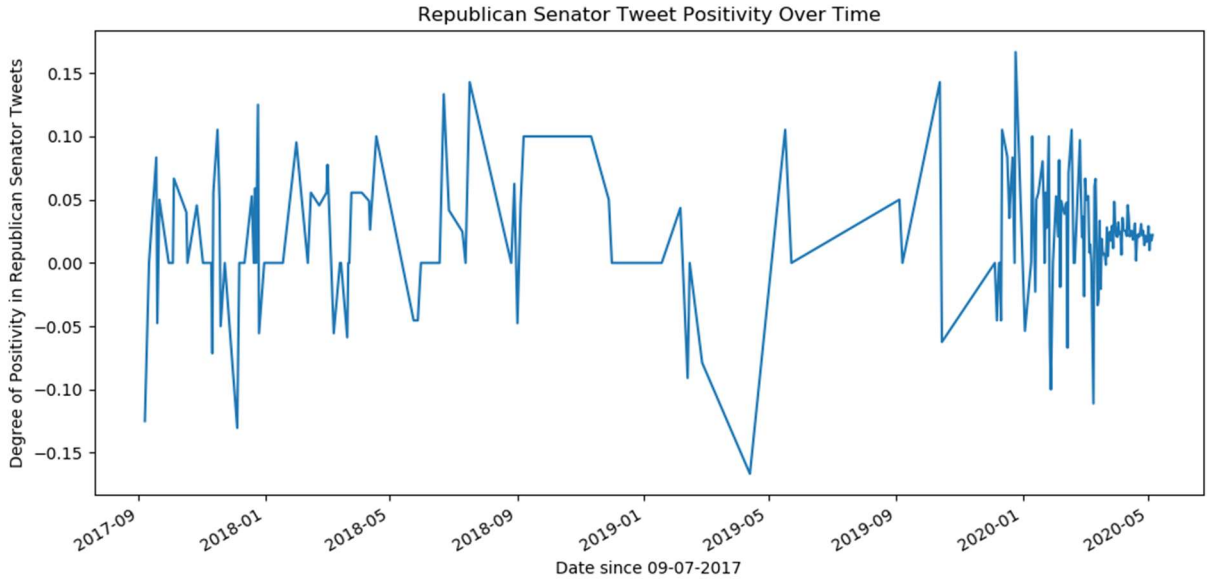

Figure 1: Republican Senator Tweet Positivity Over Time

In this figure note that the relative positivity is displayed on the y axis. The maximum range on this is from -1 to +1 theoretically with -1 being pure negativity in the tweet sentiment and +1 being pure positivity in the tweet sentiment. This scale is normalized to make comparison between the data sets easier. Note that this data set ranges from September of 2017 through the

present. This implies that the Republican senators do not tweet as much as the Democratic or Independent senators since the exact same number of tweets was gathered from each senator.

c. Democratic Results

Please find the results for the last 100 tweets for each democratic senator below in Figure 2.

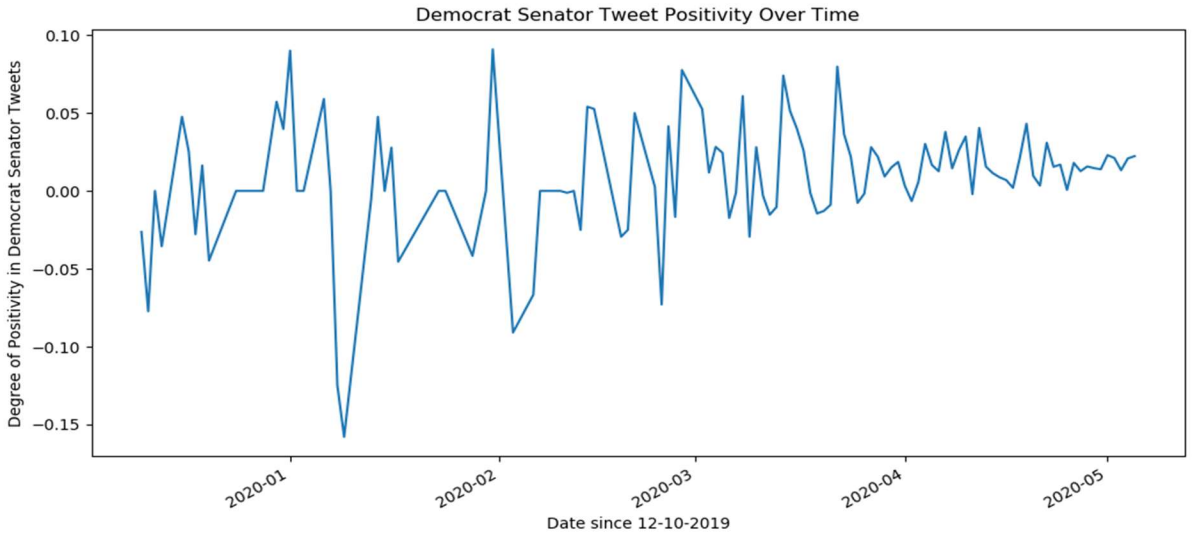

Figure 2: Democratic Senator Tweet Positivity Over Time

In this figure note that the theoretical maximum and minimum of the normalized positivity are the same as in the republican figure. It is worth noting that the maximum value for the republican senators was 0.16 whereas the maximum value for the democratic senators was 0.09 thus implying that the republicans had a higher maximum degree of positivity. Also note that the democratic senators tweet much more often with their date range spanning from December 2019 through the present.

d. Independent Results

Please find the results for the last 100 tweets for each democratic senator below in Figure 3.

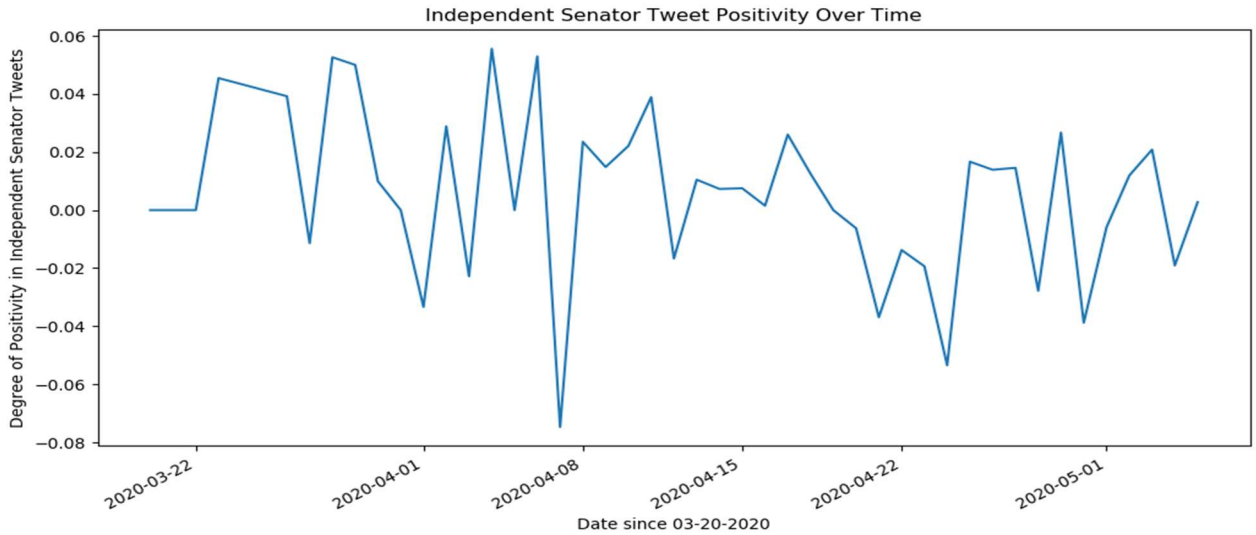

Figure 3: Independent Senator Tweet Positivity Over Time

Note that with the date range comparison it is the independent senators which tweet the most often. Their tweet frequency ranges from late March 2020 through the present implying the highest tweet rate out of the three political parties represented. Also note that this graph also maintains the same theoretical maximum and minimum values for the positivity but is much narrower than either the Democratic or Republican graphs implying that the Independent senators are more neutral than either of the larger two parties.

e. Comparison

Since all of the results did not mean much on their own two graphs were created to compare the results of the three political parties. Please find these results below in Figures Figure 4 and Figure 5.

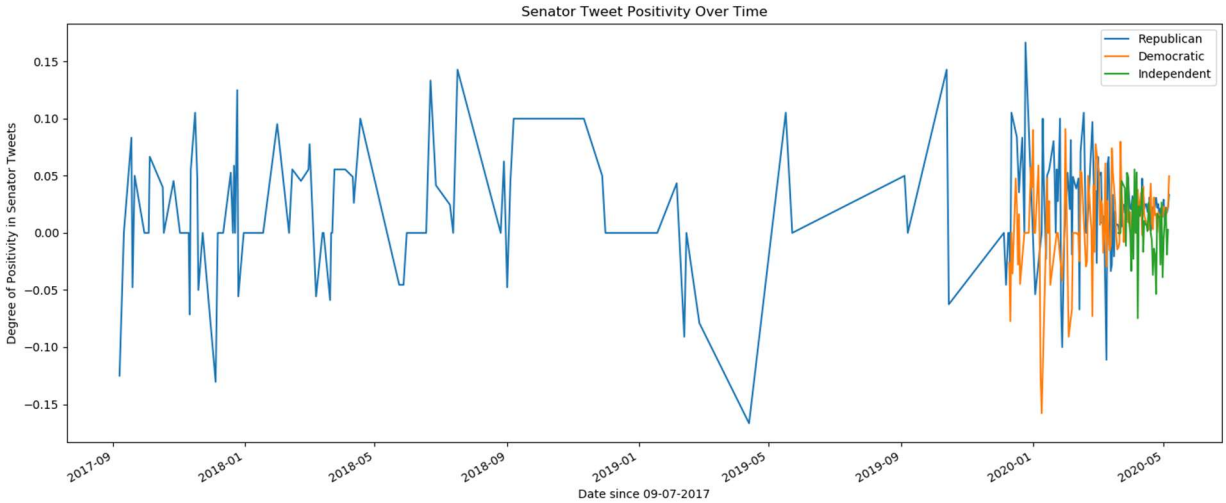

Figure 4: Senator Tweet Positivity Over Time Complete Data

In the figure above it is clear to note the tweet frequency between the senators. The republican senators (blue line) span a much larger area than the democratic (orange line) and they are both beat by the independent (green line). This figure does not provide any useful information because all of the overlap is condensed into such a small portion. Therefore Figure 5 was created.

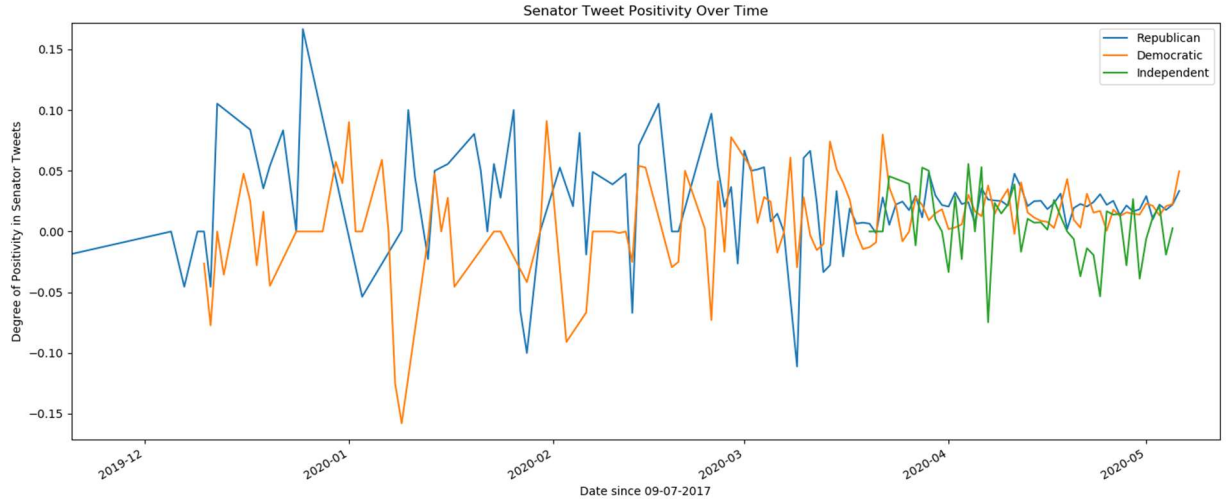

Figure 5: Senator Tweet Positivity Over Time Since December 2019

Figure 5 is a much more useful figure in comparing the tweet sentiment between the political parties. One thing to note on this figure are the relative peaks between the democratic and republican parties. While it might seem to be common sense that when the republicans are happy the democrats will be unhappy and vice versa this figure demonstrates this. Note the highest peaks of the republican senators are coincided by the lowest peaks for the democrats and vice versa. Another thing to note in this figure is that after mid-March (when the COVID-19 stay at home order started) there were very few peaks in either direction. The senators have stayed relatively neutral suggesting that they are trying to stop the wide swings that were occurring much more frequently before the pandemic affected most of the United States.

3. Methods

To complete this project there were two attempts made. Before the sentiment analysis was settled on the program was supposed to take Twitter data and determine a person's political party. This was difficult and therefore the sentiment analysis was conducted, and these are the methods through which this study and analysis was carried out.

a. Original Attempt

As mentioned, there was a previous attempt made which aimed to take users' Twitter data and determine their political affiliation. This attempt was going to be carried out through many of the same methods utilized for the sentiment analysis, but the major difference was that an ML algorithm was to be implemented and trained based on the senators' tweet dataset. This algorithm once trained was then to be applied to other Twitter users and their tweets fed into the algorithm to determine their political affiliation. This was going to be conducted to assist in the determination if a person can completely be purely democratic or republican. This attempt was shut down due to the difficulties outlined in the following section.

b. Difficulties Encountered

It was mentioned that it is difficult to determine political affiliation based on the Twitter data and this was due to one main reason. This reason is that there is no direct affiliation between words and their political meaning. For example, both the republican and democratic senators will use the word "Trump" in their tweets, but it is the connotation that matters in determining the political affiliation. This invokes the requirement for natural language processing into this project. This would have created numerous issues which have been highlighted in the introduction section on NLP.

c. Alternative Approaches

Since the implementation of the NLP algorithms to determine the political affiliation were deemed too difficult for the project scope another solution was needed. Enter the sentiment analysis of politicians' tweets. This approach required a small refactoring of some code to conduct and removed the need for both ML and NLP. The removal of the ML is unfavorable but necessary to remove the NLP requirement from the original attempt.

d. Selected Approach

Through consultation with the professor it was determined that this was a suitable substitution and deviation from the original project scope. Therefore, the project was pivoted from developing a program which would determine a user's political affiliation to determining

the sentiment of politicians' tweets. This would be conducted on all 100 United States Senators and then broken down by political party.

e. Implementation

The implementation of this idea can be broken down into five steps seen in the list below.

1. Acquire Twitter data and create initial CSV file
2. Preprocess the data
3. Analyze the data
4. Postprocess the data
5. Visualize the data

These 5 steps all correspond to their own python file within the program. The five corresponding python files can be seen in the list below.

1. createCSV.py
2. preprocess.py
3. analyze.py
4. postprocess.py
5. visualize.py

For the complete code breakdown and line by line explanation of the implementation please see Appendix B.

To determine the relative positivity of a tweet and create the CSV file for the visualization an algorithm was devised. The output from the preprocessing step was taken and fed into the analyze.py program. This program had the goal of determining the overall sentiment of a tweet. This was done through a word by word analysis of every tweet. The algorithm is extremely simple. The value of the tweet starts at 0. For every word in the tweet that is located on the positive wordlist (Appendix B wordlists folder) one is added to the value and vice versa for every word that is found in the negative wordlist. At the end, the value is divided by the total number of words in the tweet giving the relative positivity of one tweet.

To prepare this data for plotting it was decided to create the plot on a day by day basis. Therefore the postprocess.py program takes the analyzed output with the relative positivity of every tweet and sums the values for every tweet that was sent on a particular date. Note that the separation of political parties is maintained throughout. Once the total positivity was determined for a day this value was then divided by the total number of tweets that were sent on this day to maintain the value between -1 and +1. This data was then taken and plotted to create the visualizations in Figures 1 through 5.

4. Discussion

Overall, this project was a success. While there were failures at the beginning in the attempt to create a program, which would determine the political affiliation of Twitter users every large project which is undertaken will have roadblocks and difficulties. What matters is the steps taken to resolve these difficulties and the end product.

a. Project Recap

Recapping, a program was created which collected and analyzed Twitter data for sentiment. This analysis was done on all 100 United States Senators and the results were broken down by

political party to determine the overall positivity of a political party as it evolves over time. The data was visualized and reported in Figures 1 through 5 in the Results section. Analysis was realized on the data and demonstrated two very interesting trends which are discussed further below in section B.

b. Results Summary

The results for this project were visualized and reported in the Results section of this report. Further discussion about the two interesting data trends can be seen below in sections I and II.

i. Inverse Peaks

The first interesting data trend which was visualized is the inverse peaks of the republican and democratic senators tweet positivity. In Figure 6 below two notable areas are circled. Note that in both of these instances this inverse relation is apparent. As the positivity of one political party increases the positivity of the other party decreases sharply. This relationship seems second nature to those who follow politics closely, but it is extremely interesting to have hard data and numbers attributed to this almost childish behavior.

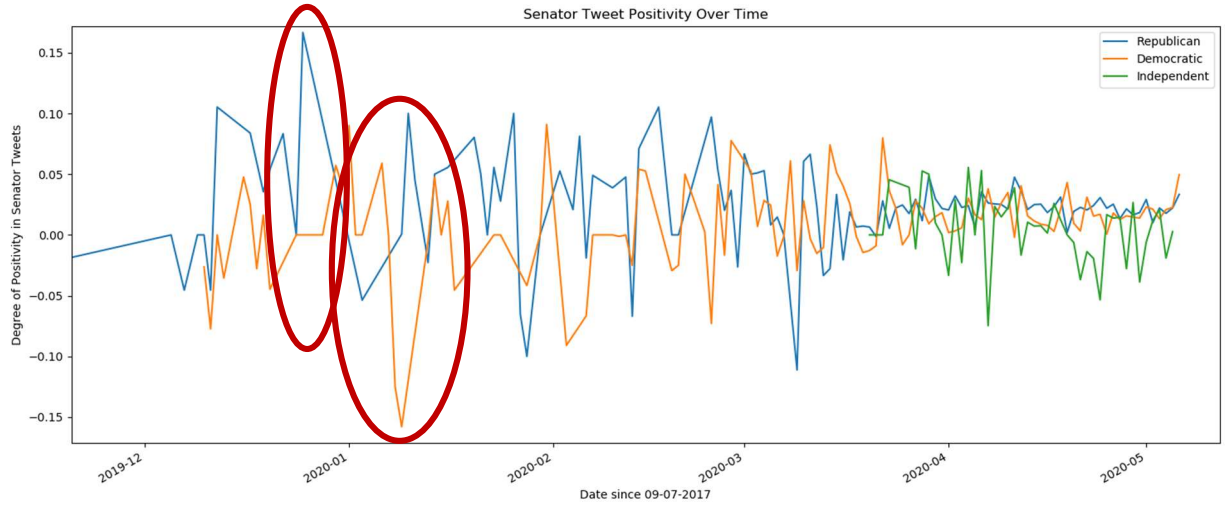

Figure 6: Senator Tweet Positivity Over Time with Peak Highlighting

ii. COVID-19 Stay at Home Order

The second interesting data trend is what occurs in the data after the COVID-19 stay at home order was issued for most states in the United States around mid-March. Figure 7 showcases this area. Note that there are many fewer peaks in the republican and democratic tweets. This implies that overall, the tweets trend much more towards neutral or even a tad positive. This demonstrates that both major parties are finally in agreeance over the state of affairs. Note that this is the area where the Independents have their largest peaks. This is due to Senator Bernie Sanders' displeasure with the government's handling of the COVID-19 pandemic. Overall, this is an interesting note and showcases how much of an impact large politically charged global events can have an impact on politicians' approach to outreach.

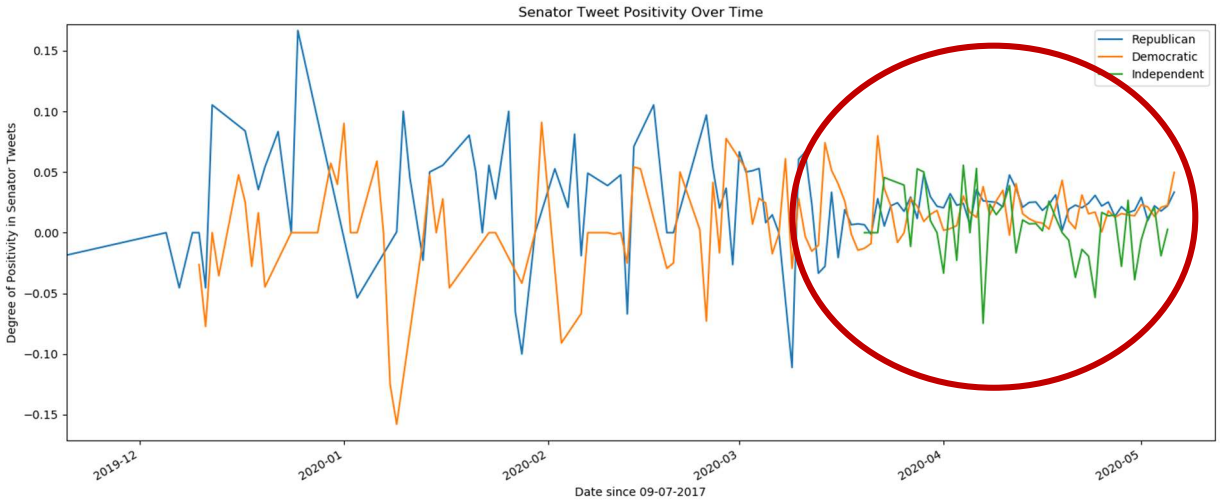

Figure 7: Senator Tweet Positivity Over Time with COVID-19 Section Highlighting

c. Conclusion

Wrapping up, this project was extremely interesting and fun to work on. There were data trends exposed that were both interesting and eye opening from a political standpoint. The code written will be useful as long as Twitter does not change the methods utilized in their API access and I believe that I will maintain the program and look at the results periodically. This project demonstrated what I have learned and more in PHY 299 Scientific Modeling with python. I have utilized many of the programming tools I have accumulated throughout the years along with the analysis tools learned in PHY 299.

Finally, I would like to thank Dr. Campbell for advising this project and allowing me to take this direction. I have learned a lot about python and programming in general throughout this project course. If there are any questions please reach out to me through GitHub whose link can be found in Appendix B.

5. References

[1] V. Dhar, Data science and prediction, ACM, 2013.

[2] G. R. Grimmett and Stirzaker, D. R., Probability and Random Processes, Oxford: Clarendon Press, 1992.

[3] B. M. Landman and A. Robertson, "Ramsey Theory on the Integers," Student Mathematical Library, 2004.

[4] T. Winograd, "Procedures as a Representation for Data in a Computer Program for Understanding Natural Language," 1971.

[5] A. Samuel, How can computers learn to solve problems without being explicitly programmed?, Dordrecht: Springer, 1959.

6. Appendices

a. Sample Twitter Data

twitter_handle	tweet_id	tweet_text	tweet_date	hashtags
senatorloeffler	1257805043673320000	rt @govkemp: on #nationalteachersday, i want to thank all of georgia's incredible teachers. your hard work and willingness to adapt in unce...	5/5/2020 22:51	nationalteachersday
senatorloeffler	1257680887459830000	#ppp is delivering for georgians. thanks to round 1 + round 2, georgia has received so far...	5/5/2020 14:37	ppp
senatorloeffler	1257444000400840000	rt @gadoenutrition: this week we celebrate teachers for the work they do to support the education of all georgia students and for their par...	5/4/2020 22:56	
senatorloeffler	1257431865570610000	absolutely abhorrent.	5/4/2020 22:08	
senatorloeffler	1257407542424190000	rt @senatemajldr: the senate is back in session because we have critically important work to do for the nation. our bosses are the american...	5/4/2020 20:31	
senatorloeffler	1257350636460480000	more than ever before, this year's #mentalhealthawarenessmonth reminds us how important it is to pay close attentio... https://t.co/uprdi1t7l0	5/4/2020 16:45	mentalhealthawarenessmonth
senatorloeffler	1257051101687390000	this month and every month, let us remember: freedom isn't free.	5/3/2020 20:55	
senatorloeffler	1256994783463190000	with over 30 million americans unemployed, we need to revitalize our economy.	5/3/2020 17:11	
senatorloeffler	1256654815314220000	#americastrong https://t.co/iiocqelkak	5/2/2020 18:40	americastrong
senatorloeffler	1256629636999580000	rt @blueangels: #atlanta, your blue angels and @afthunderbirds should be overhead in about 45 minutes!	5/2/2020 17:00	atlanta, americastrong, healthcarheroes

This is a sample of the tweets.csv file which stores the Twitter data. It is organized into these 5 columns and stored in CSV format. To view the file and interact with it please view the complete code base on GitHub.

b. Complete Code Base

The complete code base can be found on GitHub. The link is below.

https://github.com/adrianvasu/Politicians-Sentiment-Analysis-Research-Public.git