Exploring the Change in Tweets on COVID-19 One Year Later

Adeel Ali

September 29, 2021

CSE-632-50

Abstract: The topic of COVID-19 contains many subtopics that have trended and disappeared over the last year. To capture key differences of trends between this year and last, a data set of COVID-19 related tweets taken in 2020 was compared to a similarly captured data set taken recently. After performing pre-processing techniques such as lowercasing, tokenization, lemmatization, stemming, and removing stopwords/punctuation, both data sets were prepared for visual analysis. Through bar charts, specifically chosen trends were compared by frequency and by sentiment. The results of visual analysis showed certain trends such as ‘Biden’ and ‘mandates’ from this year and ‘lockdown’ from last year were not as common as expected, but were both viewed in net positive sentiments. Meanwhile ‘vaccine’ and ‘test’ were as common as expected from last year and were viewed in positive sentiment as well. The trend of positive sentiment suggests people are receptive to changes in the COVID-19 situation by policy and leadership.

Project: Exploring the Change in Tweets on COVID-19 One Year Later

Description:

Now that more and more citizens are getting vaccinated, the world grows closer to reducing deaths and severe hospitalizations due to COVID-19 to 0. It took a long time to get to this point, however, and there were a lot of changes in sentiment and topics about the virus throughout these one and a half years. I wanted to explore exactly how these topics changed over time to learn how much vaccines had an impact on the conversation.

Data Collection:

Using Twitter's API and a Python script, a stream of live tweets was initiated between Sunday September 12, 2021 at 5:57 PM PST up until Sunday September 12, 2021 at 10:27 PM PST. The live stream filtered out specific keywords of "coronavirus" and "COVID-19" to stay consistent with a previously collected dataset from last year. This data was stored in a .txt file with each data point being a JSON object. During the parsing phase, specific attributes of the data would be extracted from each JSON object. After parsing was complete, there was roughly 87,381 usable data points.

Data Exploration:

For both the dataset I collected and the previously gathered dataset, I noticed there were several attributes that had missing data including location, url, time_zone, etc. While some amount of missing data would be acceptable, these three specific entries had too few usable entries for analysis, so these and other attributes of similar missing data amounts would need to be removed.

Most attributes have a lot of noise. The name of users will greatly vary as will their id numbers and date joined. The attribute with the most noise, however, would be the user description and the text (tweet message) because there are a wider range of acceptable responses for both of these attributes. Adding to the range was the ambiguity coming from various languages the tweets were written in as well as spelling errors and emoji potentially misconstruing a message. The noise and ambiguity could be dealt with during preprocessing, however, since we will further need to break down the tweet message to phrases and words.

There were not too many numerical values in this dataset, however, followers and friends were two instances in which normalization and discretization seemed useful. There were a wide range of both followers and friends, so standard deviation normalization may be a good start to measure the spread of followers and friends with respect to each data point. Additionally, discretizing by 10 or 100 friends/followers may also prove useful as opposed to creating a separate category for 1 friend vs 2 friends for example.

An attribute that may need to be created from unstructured data is frequency of tweet. This would involve determining how many tweets arrived during the duration of data collection and creating a ratio from this information. The frequency of the previously gathered dataset can be compared to my data during the analysis portion.

As a final thought before pre-processing would begin, I decided the main focus would be on the text portion of the tweets because of the limitation of the amount of data and times both data were collected.

Hypothesis:

There are three aspects of the data I intend to explore between tweets this year and last year. I first hypothesize the discussion will be centered on vaccines this year, while last year will have more discussion on lockdown enforcement. While the time the historical data was taken last year suggests there was no lockdown occurring, I believe people may have been talking about how lockdowns could return. In this year, vaccines are still relevant to discussion as booster shots are being recommended and there is still urgency to reach the unvaccinated. This is an important hypothesis to investigate how seriously people took lockdowns compared to vaccines.

My second hypothesis is that there will be equal discussion of Trump last year compared to Biden this year, but both will be among the most discussed. From this hypothesis I want to evaluate whether a change in leadership resulted in more, less, or same amount of discussion about COVID-19. My theory is that COVID-19 may have been discussed more in 2020 which may cause more tweets about Trump's handling of the pandemic; however, Biden recently imposed a mandate about vaccination which may cause an increase in discussion about Biden's handling of the pandemic as well. Ultimately, this hypothesis is important to discover whether a change in president had a positive, negative, or neutral impact on the COVID-19 fight.

My last hypothesis is that a third major discussion will be on mandates this year compared to last year's focus on testing. Since vaccination was not an option until the end of last year, the precautions taken involved getting tested if someone was potentially exposed to the virus, then quarantining if they were confirmed to be exposed. Hence I imagine testing should be more commonly discussed last year. As for this year, as previously mentioned, the mandate should be more commonly discussed because it is the most recent change in policy regarding the pandemic. This final hypothesis is worth investigating to analyze the effectiveness of a pandemic policy change.

Data Preprocessing:

Step 1:

The first preprocessing step was to remove extra spaces read from the CSV file. This was easy to fix with my own collected data since Python could remove newline characters during parsing. For the historical data, however, I needed to use the drop null method for a Pandas Data Frame to remove NULL lines and then reset the index back to sequential after removing those lines. This set the stage for me to use Python specifically through Jupyter notebook to import various libraries such as Pandas to help with the remainder of the project.

At this stage, there are a lot of features that do not contribute any useful information. So I applied data reduction through feature selection, making a decision based on a priori knowledge that certain features do not have an impact on any of the hypotheses I have. As planned in the data exploration phase, all the features that were not the main tweet text would be removed. These steps were important to optimize preprocessing of the text which was obviously relevant for my hypotheses.

Step 2:

Next, I took several steps to clean the text data at once. I first used the preprocessor library which was designed to filter tweets. I chose settings to remove URLs, emoji, smiley faces, numbers, Twitter reserved words, and mentions. This immediately removed irrelevant information for my hypothesis. Subsequently, I lowercased all letters, then used regular expressions and the string library to replace punctuation (as defined in the string library) with an empty string wherever non-whitespace characters were found to equal one of the defined punctuation marks. Note that this function is applied for every tweet individually to preserve a form that can have sentiment analysis completed later. This step as a whole makes the data easier to process since uppercase and lowercase words won't be duplicated, and punctuation does not contribute to analyzing my hypothesis.

Step 3:

The most important preprocessing step will be tokenization. This is breaking down each tweet text message into individual words. Tokenization is best paired with stemming (which is breaking down a word into its root and dropping unnecessary inflections) and lemmatization (which is converting a word to its base form with context of word placement in the sentence). As an example, lemmatization would convert the word "mandates" to "mandate". Stemming would convert "mandate" to "mandat". Even though "mandat" is not an actual word, this result is the same as one from the word "mandatory". So we are able to group "mandatory" and "mandate" under the same category which preserves the meaning and helps my hypothesis. I specifically used PorterStemmer and WordNetLemmatizer from the NLTK library to achieve stemming and lemmatizing respectively.

At the same time stemming and lemmatization is taking place, the stopword removal process can occur where words that add no additional meaning to a sentence (such as "a", "the", "then", etc.) are removed. Again this is best during tokenization since filtering through unnecessary words

will slow down the analysis process. I specifically used the stopwords list from the NLTK corpus library. Removing stopwords during this phase is also beneficial since I can add additional words to the list of stopwords as I see them. The way I noticed candidates for additional stopwords was by finding the most common words between all tweets and displaying them in descending order of counts. Meaningless letters such as “brt” were removed as well as non-English stopwords such as “que”, “de”, etc. through trial and error. I used the assistance of Google Translate to avoid removing words in another language that were not stopwords.

Step 4:

As final preprocessing steps for the tokenization portion, I grouped words together that discussed the same topic, variations of spellings on the same word, or the same word after being translated from Google Translate. For instance, “covid-19” can be treated the same as “coronavirus” for my hypothesis. This part needed to be done carefully since I did not want to double-count or group falsely. To elaborate, if one tweet were to contain the string ‘Donald Trump’ and I grouped the occurrence of ‘Donald’ and ‘Trump’, then I would have double-counted the occurrence of Trump as a discussion topic. Similarly, a false grouping can occur if I assume the word ‘president’ is referring to only the president of the US, when in reality someone could be tweeting about the president of their respective country. As a result, all occurrences are more conservative estimates of a general topic because of this restriction. The graphs obtained from this step will be discussed in the final portion.

Step 5:

Sentiment analysis was relevant for my hypothesis to expand upon the meanings of discussions. To be able to be specific about whether a change in presidency had positive or negative impact, for instance, needs to be quantifiable. A similar analogy can be made for seriousness of vaccines, lockdowns, mandates, and testing.

So for the sentiment analysis portion, I needed to backtrack from tokenization. The TextBlob sentiment analysis tool builds upon research done by linguists to categorize words based on sentiment and subjectivity depending on the relative location of that word in a sentence. This tool is not perfect, but can give at least a rough estimate of sentiment by assigning polarity (positive meaning good sentiment, negative meaning bad sentiment, and zero meaning neutral sentiment). Since I needed full sentences, I needed to return to a state where stopwords were not removed and I needed to ungroup translated words since sentiment analysis would not work on non-English words. After returning to this state, I grouped positive, negative, and neutral sentiments for a filtered word of interest relevant to my hypothesis. Note I needed to perform tokenization, lemmatization, and stemming again, however, it was done line by line instead of all at once. The sentiments would be stored in their respective arrays by filtering to be displayed in a bar graph. The sentiment analysis graph will also be discussed in the next section.

Data Visualization and Analysis

All visualizations were created using the matplotlib library. By taking counts from the tokenized list and storing sentiment values from the preprocessing section, bar graphs could be generated to compare various keywords to each other when relevant to the hypotheses. Each type of visualization is given its own section.

Most Discussed Topics:

The first pair of visuals worth examining are the most discussed topics in both years. These visualizations are important for my hypotheses to determine whether vaccines were most discussed this year and lockdown was most discussed last year. Figure 2 suggests my first hypothesis was wrong, and that lockdown enforcement was not among the top nine discussed topics of 2020. Similarly, my second and third hypotheses were wrong according to Figure 1 since 'Biden' and 'mandate' were not among the top nine discussed topics of 2021.

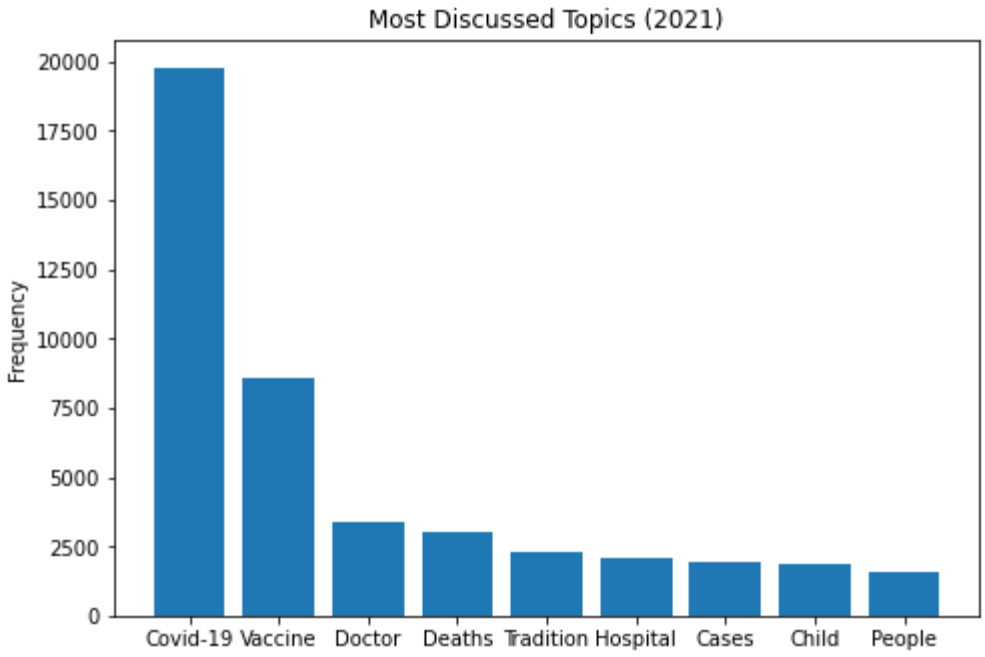

Figure 1. Data from this year showing most discussed topics (grouped by common words)

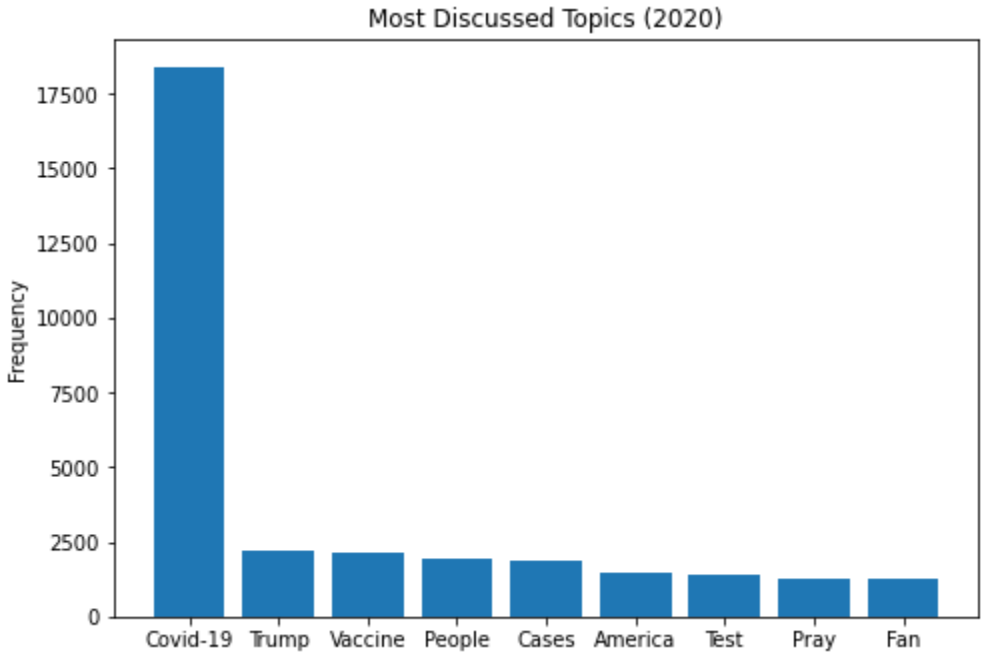

Figure 2. Data from 2020 showing most discussed topics (grouped by common words)

Frequency of Relevant Words for Hypotheses:

Next I wanted to do a direct comparison of word occurrence for each hypothesis. Starting with hypothesis one, I expected the word ‘vaccine’ from 2021 to occur as commonly as ‘lockdown’ from 2020. Figure 3, however, shows ‘vaccine’ from 2021 has a much larger percentage of occurrence compared to ‘lockdown’ in 2020, proving my hypothesis incorrect.

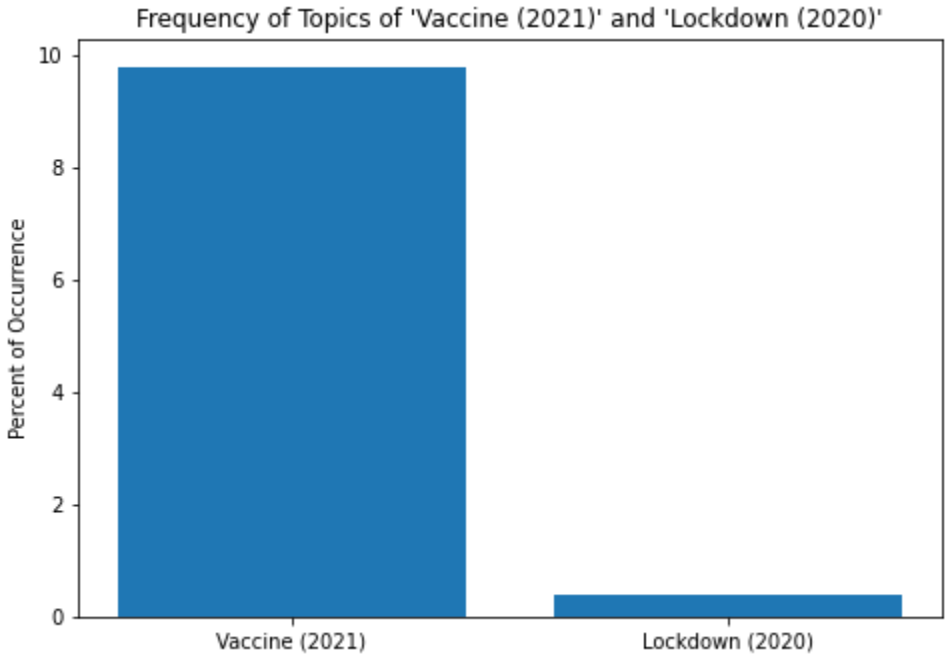

Figure 3. Word ‘Vaccine’ occurrence in 2021 compared to ‘Lockdown’ occurrence in 2020.

I wanted to go one step further to examine whether the word ‘vaccine’ was more common than ‘lockdown’ in 2021 while ‘lockdown’ was more common than ‘vaccine’ in 2020. Figure 4 shows vaccine was more common than ‘lockdown’ in 2021, however, ‘vaccine’ is more commonly discussed compared to ‘lockdown’ in 2020 which is also against what I expected. It is important to note the scale of Figure 4 because the difference between the occurrence of ‘lockdown’ and ‘vaccine’ in 2020 is only about 0.4%.

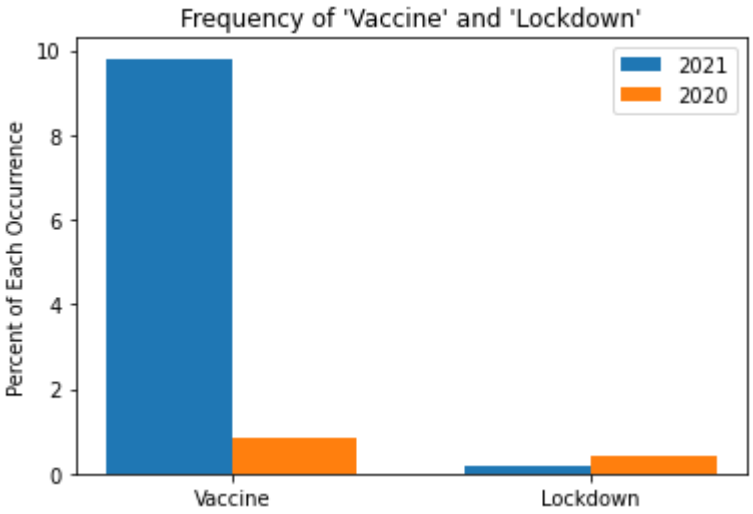

Figure 4. Word ‘Vaccine’ compared to ‘Lockdown’ occurrence in 2020 and 2021.

Doing a direct comparison for hypothesis two, I found yet again my hypothesis to be incorrect. I expected ‘Biden’ and ‘Trump’ to be discussed equally, however, Trump in 2020 was discussed significantly more than Biden in 2021 as seen in Figure 5.

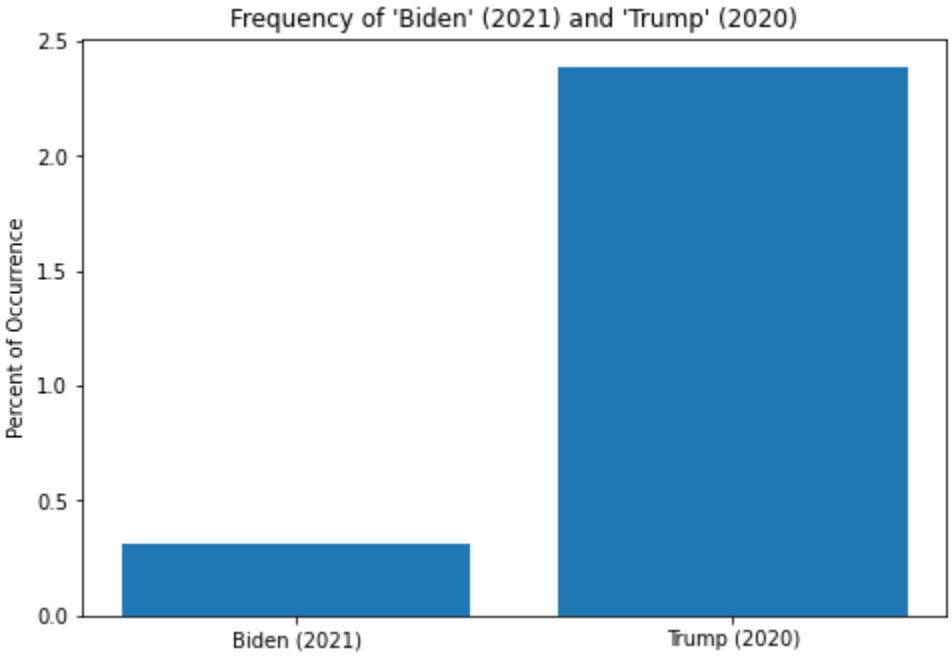

Figure 5. Word ‘Biden’ occurrence in 2021 compared to ‘Trump’ occurrence in 2020.

Following a similar pattern to my previous hypothesis, I wanted to see if ‘Biden’ was discussed more than ‘Trump’ in 2021, while ‘Trump’ was discussed more than ‘Biden’ in 2020. Figure 6 supports this hypothesis. Again, it is important to note the margin by which ‘Biden’ is discussed more in 2021 is very narrow with barely 0.2% more discussion for ‘Biden’. This could be because of the conservative estimates I used for keywords relating to Biden.

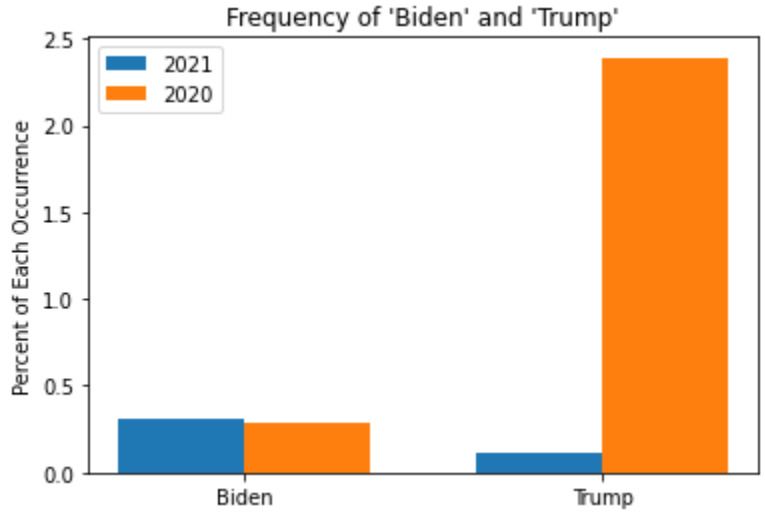

Figure 6. Word ‘Biden’ compared to ‘Trump’ occurrence in 2020 and 2021.

Conducting the same analysis for my final hypothesis, I was proven incorrect for a third time when the topic of ‘test’ was discussed more in 2020 than ‘mandate’ was discussed in 2021 as seen in Figure 7. It is worth elaborating that there is a possibility the word ‘test’ could encompass tests given at school for instance, however, it is unlikely that a significant portion of people were tweeting about this along with the COVID-19 tag. Further evidence is given by the time of collection in mid-August when most schools have not started yet.

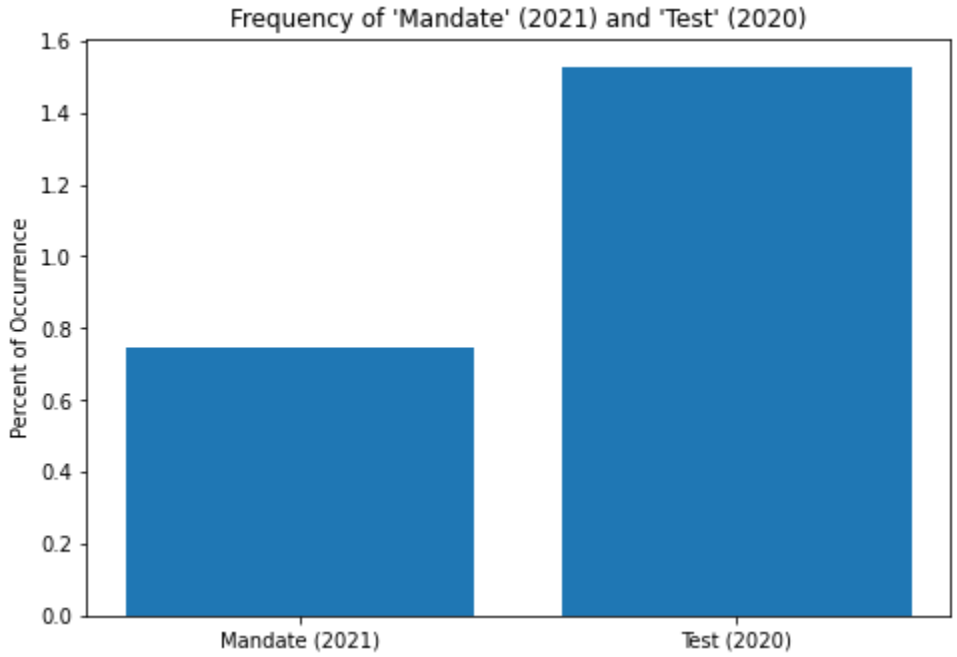

Figure 7. Word ‘Mandate’ occurrence in 2021 compared to ‘Test’ occurrence in 2020.

Comparing ‘Mandate’ and ‘Test’ in 2021, I found a thin lead for the word ‘Test’ by about 0.1% seen in Figure 8. The negligible difference makes sense because the mandate policy comes with an option to get tested every week as well. I could not generate a graph for ‘Mandate’ in 2020 because there was no vaccine mandate at the time, and if there was discussion about a mandate, it would have been a mask mandate, which is not the same.

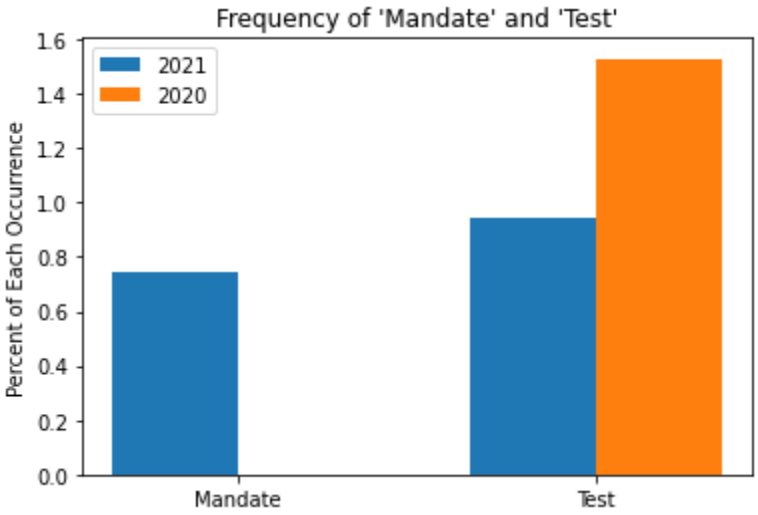

Figure 8. Word ‘Mandate’ occurrence compared to ‘Test’ occurrence in 2020 and 2021.

Sentiment Analysis Visuals:

The last of the visualizations involved sentiment analysis comparisons. I specifically chose to compare the same words from the Frequency of Relevant Words section; however, instead of viewing frequency, I chose to visualize the percentage of words that were marked as positive, negative, or neutral by the Sentiment Analysis tool from TextBlob. Looking first at hypothesis one for Figure 9, it appears vaccines are viewed significantly more neutrally than positively with very few negative sentiments. Lockdowns had the same ranking of sentiments, however, the results were less spread out. This figure suggests most people had net positive views of vaccines and lockdowns, providing evidence that people took lockdowns and vaccines seriously.

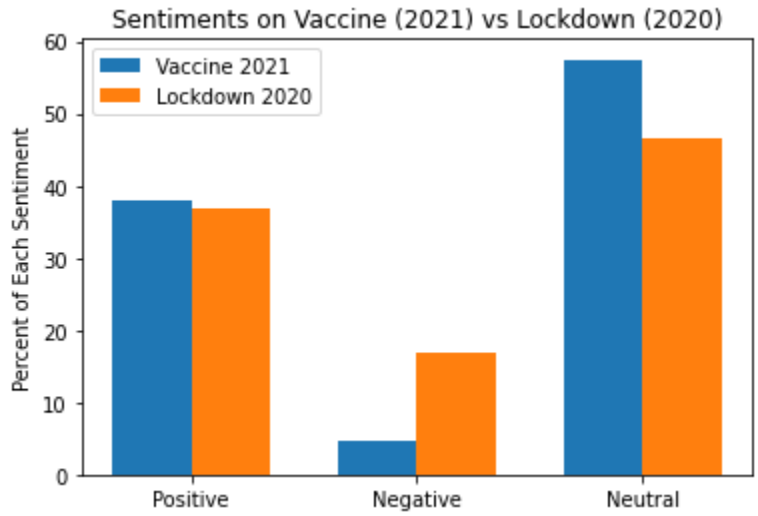

Figure 9. Sentiment Analysis for ‘Vaccine’ in 2021 compared to ‘Lockdown’ in 2020.

For hypothesis two, Figure 10 suggests Biden has almost half of all tweets about him as positive sentiments regarding Covid-19 specifically, while Trump has almost half of all tweets about him as neutral regarding Covid-19 specifically. This shift from neutral to positive along with a decrease in negative sentiments from last year suggests people are more satisfied with the change in leadership related to Covid-19.

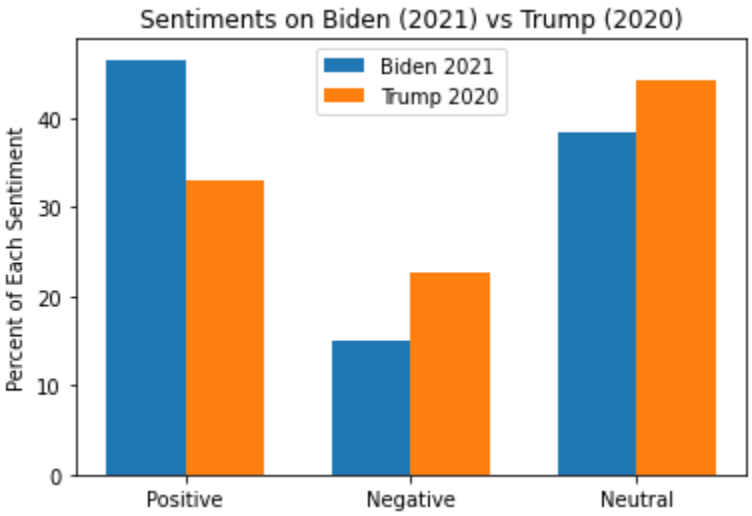

Figure 10. Sentiment Analysis for ‘Biden’ in 2021 compared to ‘Trump’ in 2020.

For the final hypothesis, Figure 11 suggests the word ‘mandate’ follows very closely to the pattern of ‘vaccine’ in that most sentiments are neutral, followed by positive, then a distant negative. This similarity, however, makes sense considering the mandate has to do with vaccines so the sentiments should be linked. The sentiment of ‘test’ closely resembles ‘lockdown’ as well which also makes sense because these were the two modes of fighting against the spread of the virus last year. Once again a net positive from both ‘mandate’ and ‘test’ sentiment provides evidence people took these measures seriously.

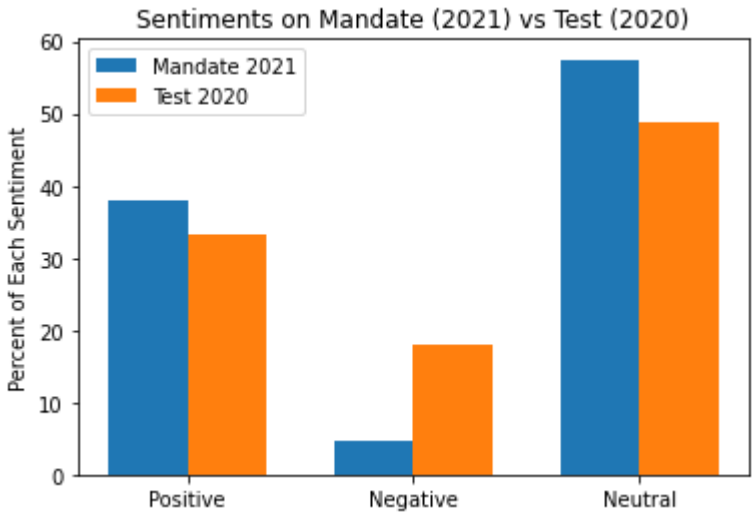

Figure 11. Sentiment Analysis for ‘Mandate’ in 2021 compared to ‘Test’ in 2020.

Conclusion

I had several misconceptions about what most people would be tweeting about. I expected far more tweets about mandates, Biden, and lockdowns than were actually tweeted which was why my hypotheses were so off. The best explanation I can come up with was that the time and duration of my data collection as well as the given data did not encompass the full scope of possible discussions.

Additionally, some word frequencies may have been inflated by individuals who were continuously sending out the same tweet. In my design, I did not filter repeated individuals by id or filter by repeated tweet, however, this is something I hope to implement in the future for even better accuracy. So far I have discussed limitations in data collection design, not of the analysis itself, however.

From the findings of the data, there is agreement in news and this data in so far as most people have taken COVID-19 seriously. In tweets regarding vaccines, lockdowns, mandates, and testing, the net majority of people had positive sentiments about such policies. Additionally, a change in presidential leadership has been seen in a positive light as well with regards to COVID-19. All this information is certainly re-assuring as the fight against this virus slowly comes to an end.