STAT 4559 Final Project:

Xinyue Qiu(xq8zu)

Introduction

Dataset Background

This ‘Covid 19 tweets’ data set is from Kaggle. This is a data set drawn from twitter database using covid19 relevant hashtags. The dataset contains 178683 tweets, and basic tweet metrics like usernames, user locations, user profile, date posted, number of user followers. The dataset covers information of tweets from July 4th, 2020 to August 30th, 2020.

Potential Insights and Questions of Interest

This dataset contains all tweets relating to covid19. Given that coronavirus is a widely discussed issue that prevails for months, this dataset can be used for analysis of sentiments and most common words for discussion of this pandemic. By completing this study, I aim to give visions upon which of the hashtags in covid19 tweets, aside from covid19 hashtags, are among the most popular ones. Other insights like twitter users’ sentiment towards specific hashtags, illustrated by tweet text, could also be investigated in this dataset. An additional exploration of word frequency for this dataset should provide a general idea of how different hashtags related tweets address keywords differently. Specific questions of interest would be restated in later sections, after exploratory data analysis.

Ideal Data and Existing Data

To answer my questions of interest, I would like to use a combination of covid19 tweets data and associated hashtags that are posted during July and August. The reason why this particular data set is chosen is because this dataset contains a substantial amount of tweets that ensures size of statistically significant data analysis. Given the large sample size, insights gained during this study can elicit reasonable inferences. However, for a wider generalization of findings in this study, another dataset that covers a longer time length would be ideal to use. This is because the discussion of covid19 on twitter has persisted for over a year, and having data that covers two months is not ideal. Additionally, worldwide twitter users are affected by and engaging in discussion of the pandemic. Consequently, word frequencies in text of tweets are sensitive to uncertainties. If the covered time length is short, any abrupt incidents relating to the covid19 pandemic could alter the word frequencies of tweet texts. If the covered time length is longer, such effects could be diluted due to the larger size of the dataset. Out of the purpose of drawing more general inferences from the statistical analysis of the data, an ideal collection data for my questions of interest would be composed of tweets posted since the beginning of this pandemic till now.

Data Cleaning and Processing:

Data Cleaning

As the dataset used contains raw text from tweets with lots of factors that are irrelevant to our words of interest, data cleaning is required. Using python, I removed url links that are characterized by having ‘http’ as prefix, and other irrelevant usernames characterized by symbol ‘@’, special symbols, punctuations, and stop words using functions that search for regular expressions. In addition to cleaning the text in tweets data set, I also cleaned the hashtag variable by stripping symbol ‘#’ from all hashtags, and made all hashtags lowercase.

Data processing

Subsequently, given the goals of this study is to study word frequencies and sentiments, I create several variables from text and hashtags that contain needed information for further analysis. The first variable created is hashtags in lists. After stripping ‘#’ symbol from all hashtags, I tokenize the hashtags and get rid of all empty tokens. The remaining tokens represent hashtags, and each tweet has a list of hashtags.

Following the creation of lists of hashtags, the second variable constructed is co-hashtags. This variable is created by a function that takes in the list of hashtags for each tweet, and gives all the combinations of two hashtags. This written function is capable of creating co-hashtags and putting the list of co-hashtags into the dataframe of tweets as an element. If a list of hashtags like (‘covid’, ‘coronavirus’, ‘maskon’) is implemented to the function, it will output all co-hashtags in a list (‘ covid, coronavirus’, ‘covid, maskon’, ‘coronavirus, maskon’). The list is composed of unique pairwise combinations of hashtags. Third variable is created based on the same list of hashtags, counting the length of hashtags in the list and recording it as hashtag count. This variable provides indexes of the co-hashtag extracting function, and allows subsequent dataset subsetting.

Another pair of variables, polarity and sentiment were created for sentiment analysis. These two variables aim to classify the contents of all text by assigning a polarity and level of sentiments, among neutral, positive and negative, to the tweet. I used the package ‘textBlob’ to assign the polarity to each tweet text, and classify the tweets based on polarity. If the tweet has a greater than 0 polarity, it was classified as positive tweet; if the tweet has a less than 0 polarity, it was classified as negative tweet; if the tweet has a polarity equal to 0, it was classified as neutral.

Variables of Interest Table

Variable	Variable Description	Data Type
Text	Cleaned, lower case and tokenized tweet text, free of url, emojis, special symbols and punctuations.	string
hashtags_list	Tokenized lower case hashtags without '#'	list
co-hashtags	Pairwise combinations of all hashtags in tweet	list
hashtag_count	Hashtag counts for each tweet	int
polarity	Sentiment scores of tweet text, lies in the range of [-1,1] where 1 means positive statement and -1 means a negative statement	float
sentiment	Classified sentiments based on polarity. Contains three levels: negative(polarity<0), positive(polarity>0), neutral(polarity=0)	string

Exploratory data analysis

Before proceeding to hypothesis testing, several visualizations of the distribution of co-hashtags and hashtags were conducted. By subsetting the data set based on hashtag count, a dataset with hashtag count larger than 1 is created. This subset contains a total of 67127 tweets. Then using matplotlib.pyplot.bar, I plotted the top 40 most common co-hashtags from the dataset.

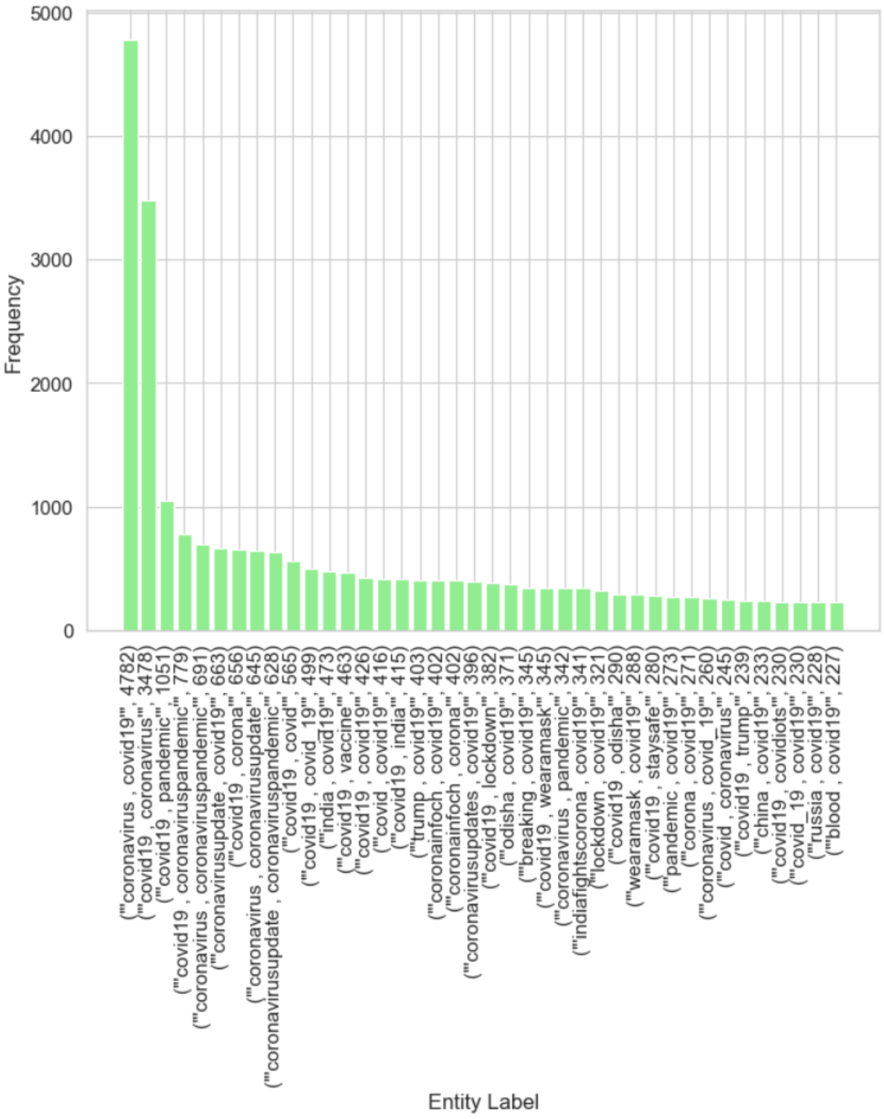


Figure 1: top 40 common co-hashtags for all tweets that have more than 2 hashtags

Looking at figure 1, it is noticeable that most co-hashtags contain information about covid and coronavirus. Given that hashtags like covid19, covid, and coronavirus were all used as the hashtags for extraction of tweets from twitter dataset, this result is not surprising. Other than the co-hashtags that contain only covid19 hashtags, other frequent co-hashtags contain information regarding trump, india, lockdown, wearamask. This elicits an initial idea into the most commonly referred topics accompanying twitter users' use of covid19 related hashtags.

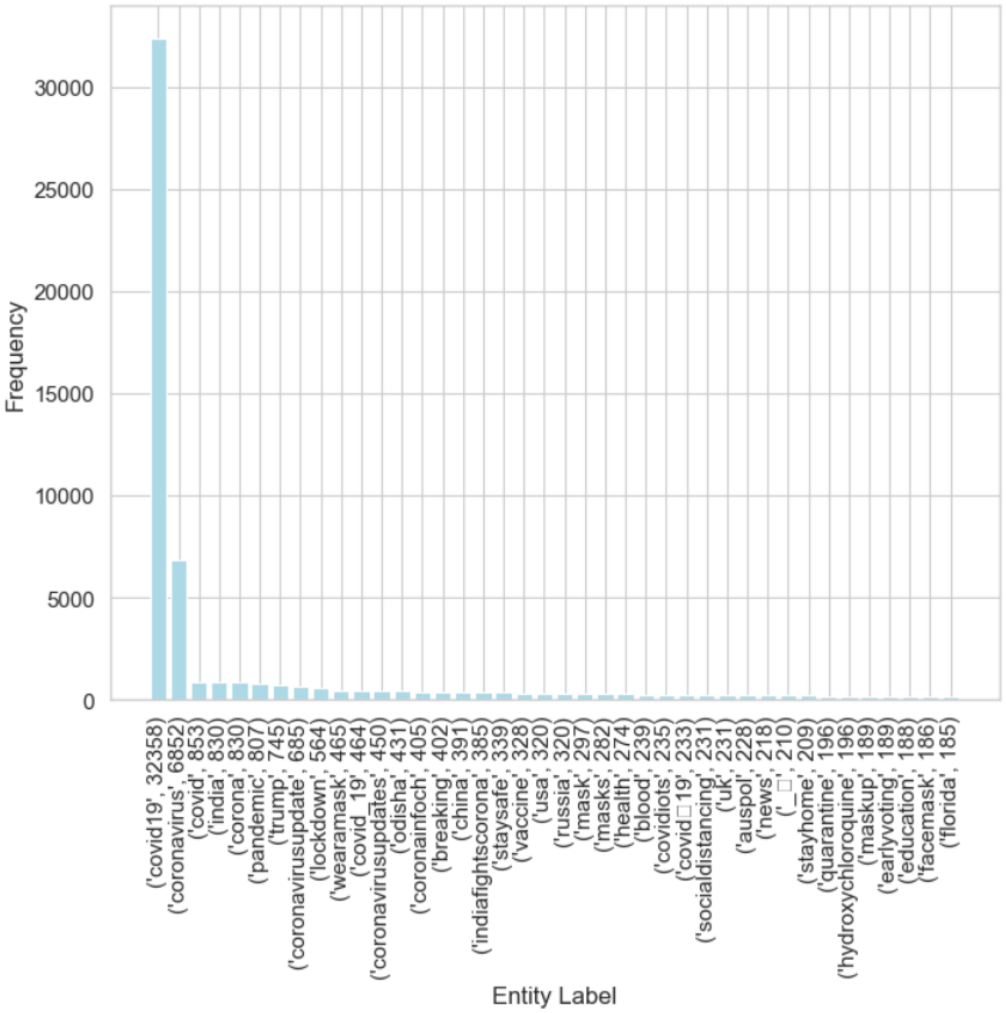


Figure 2: top 40 common hashtags of all tweets

Given our visualization of co-hashtags shows most commonly mentioned co-hashtags for tweets involving covid, I used hashtags in the whole data set for a better visualization of commonly mentioned hashtags. Figure 2 plots top 40 most common hashtags with counts of each. One can see interesting hashtags like ‘trump’, ‘mask’, ‘lockdown’, ‘vaccine’ are among the most common hashtags. This figure allows better visualization of hashtags than figure 1, and also raises a general idea for what are some frequently brought up topics, indicated by hashtags, when twitter users talk about covid 19.

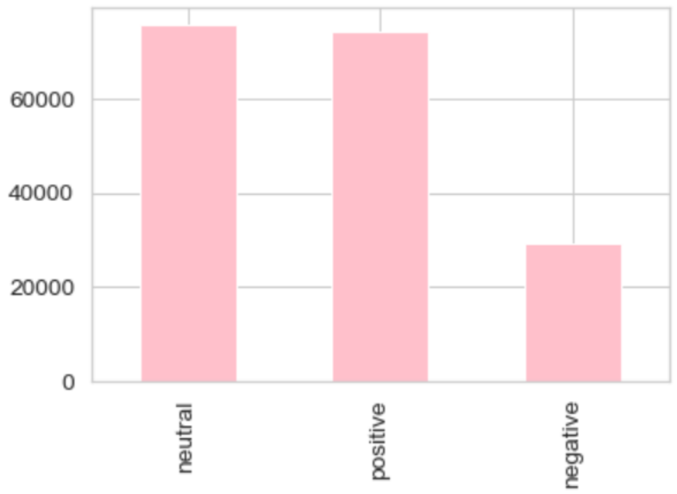


Figure 3: sentiment distribution of all tweets

Figure 3 shows a general distribution of users' sentiments when talking about covid19. Surprisingly, the most common sentiments are neutral and positive. This means that even though the pandemic in general is disastrous, most twitter users posting about covid19 during July and August feel positive or neutral when talking about issues related to covid19. Next, to further look at the word frequencies, I chose to plot out the most commonly used words in these tweets. To do so, I lemmatize and tokenize all texts in tweets.

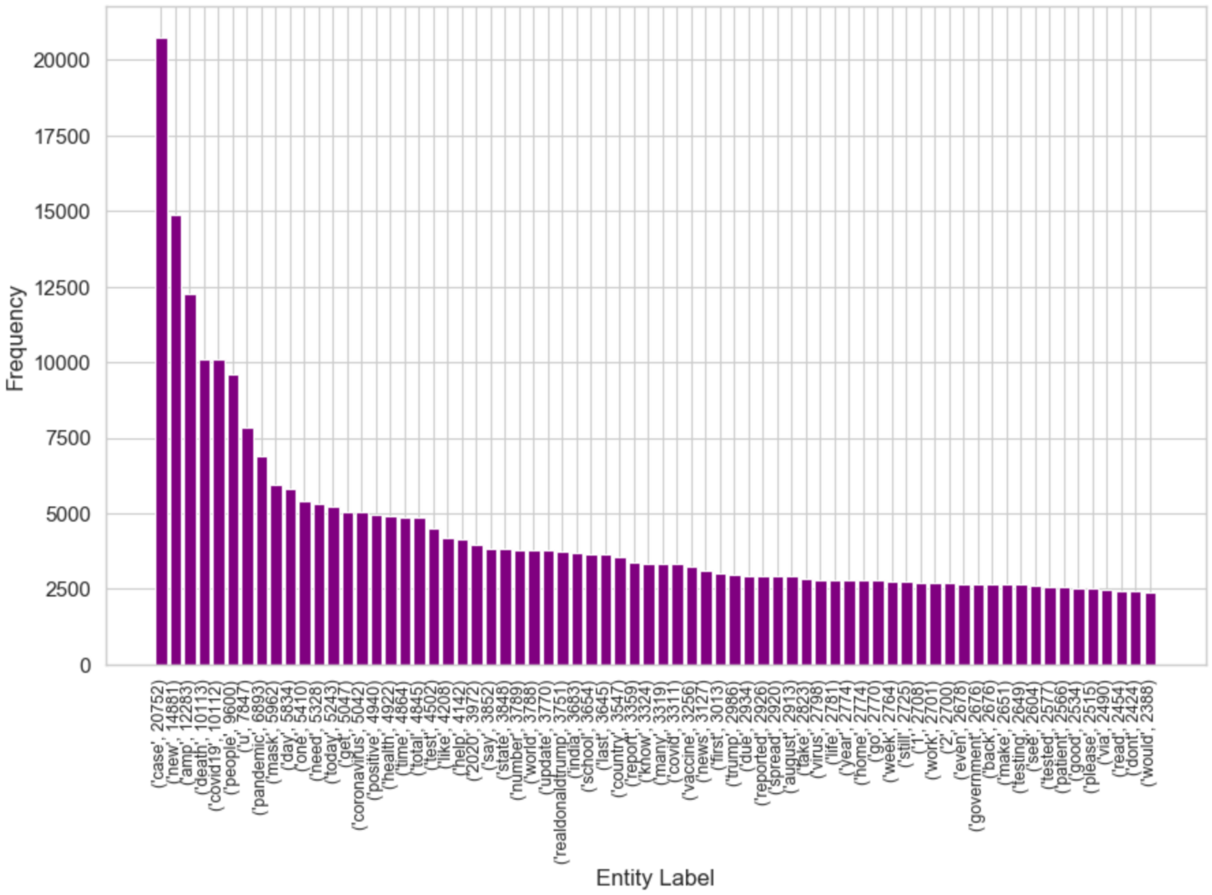


Figure 4: top 40 most common words in all tweets

Figure 4 shows the top 40 most common words in the whole tweet dataset. The word counts gives us more general background regarding the commonly mentioned topics in these tweets. There are in total 6737 word counts for ‘trump’(count of words ‘realdonaldtrump’ that is related with president trump’s twitter account and ‘trump’) and 5962 counts of ‘mask’. This graph also provides a general idea of what are some possible key words mentioned that are shared by these two groups of people. As ‘positive’ and ‘test’ have a high word count, both groups of tweets could be mentioning the covid19 tests in a certain frequency. This finding has made word frequency of keywords related to covid19 tests in trump tweets and mask tweets a reasonable choice for studying.

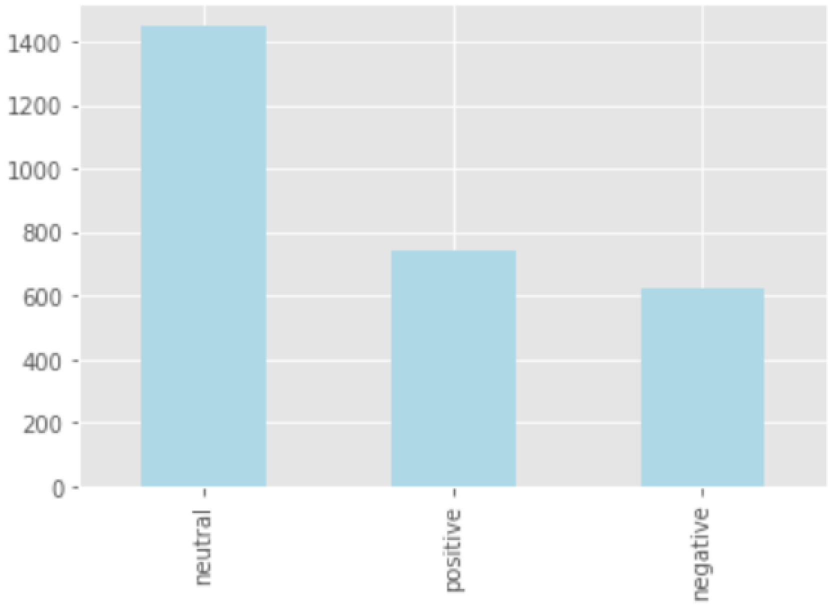


Figure 5: sentiment distribution of trump tweets

Following my prediction that tweets with the hashtag 'trump' and 'mask' are resembling in sentiment distribution, denoted as the proportion of tweets that are negative, positive and neutral, I subset the data based on these hashtags and got two separate dataset. The tweets that contained 'trump' and 'mask' as co-hashtags were taken out of both groups of tweets for hypothesis testing. The group of tweets that uses hashtag 'trump' will be abbreviated as trump tweets, and that of hashtag 'mask' will be abbreviated as mask tweets in this report. Figure 5 shows the sentiment distribution of the trump tweets data set. Surprisingly, a majority of the tweets expressed neutral sentiments. Amount of positive tweets also exceeds the amount of negative tweets.

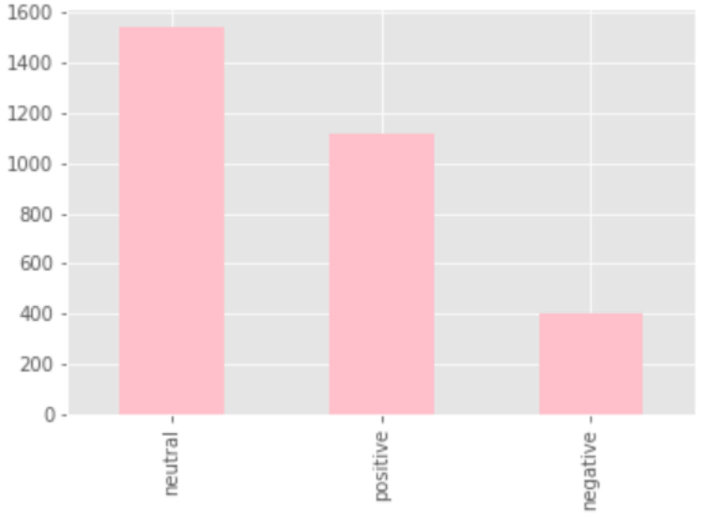


Figure 6: sentiment distribution of mask tweets

Figure 6 is the sentiment distribution of the mask tweets. The sentiment distribution among these two different groups resembles each other. Further statistical analysis could be conducted on them to test for the difference in proportion of each sentiment level.

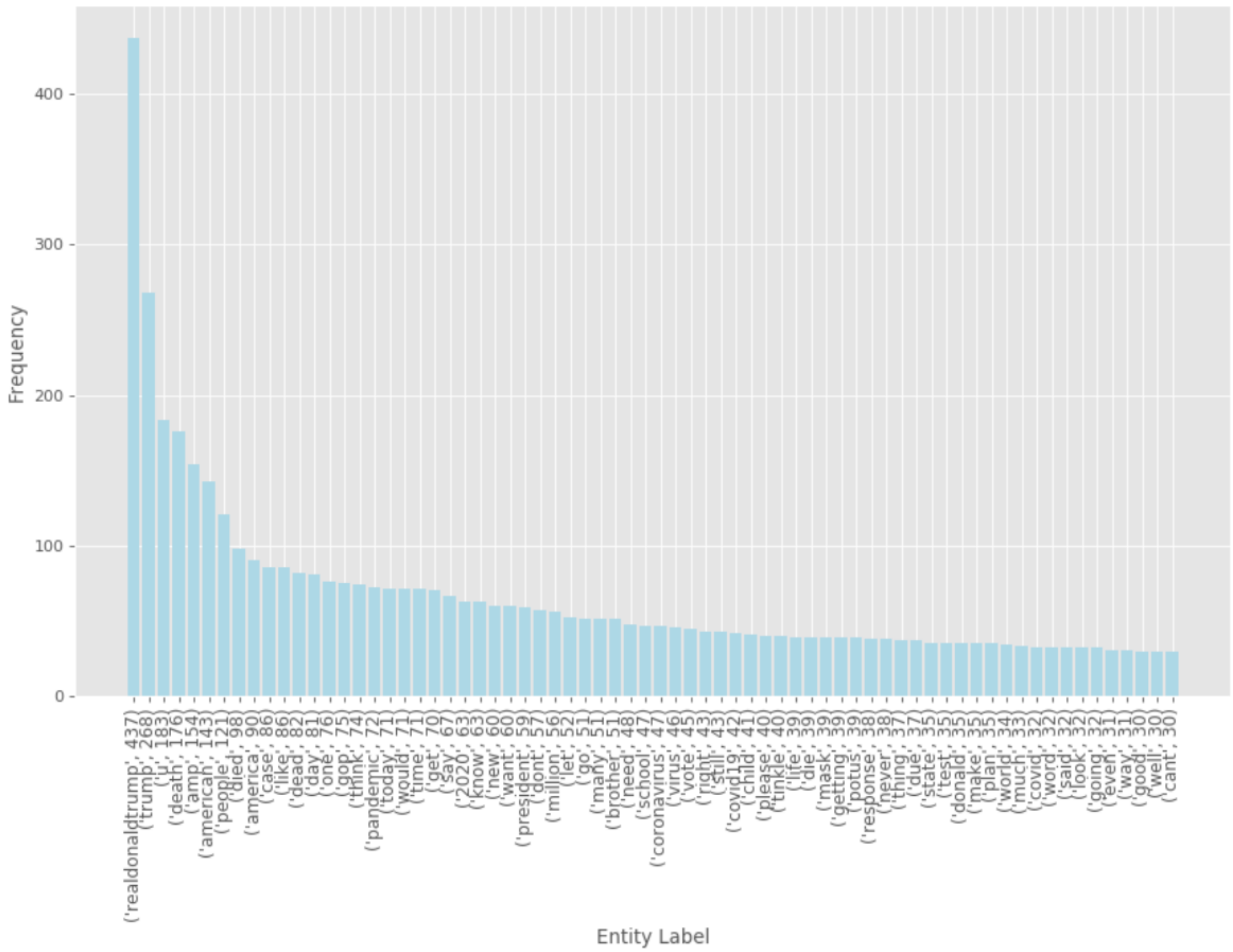


Figure 7: top 40 most common words in all tweets

This figure 7 shows the top 40 common words in trump tweets. 'Tested' and 'positive' both have a high count in the tweets. This confirms that our interested keywords prevail in the tweets.

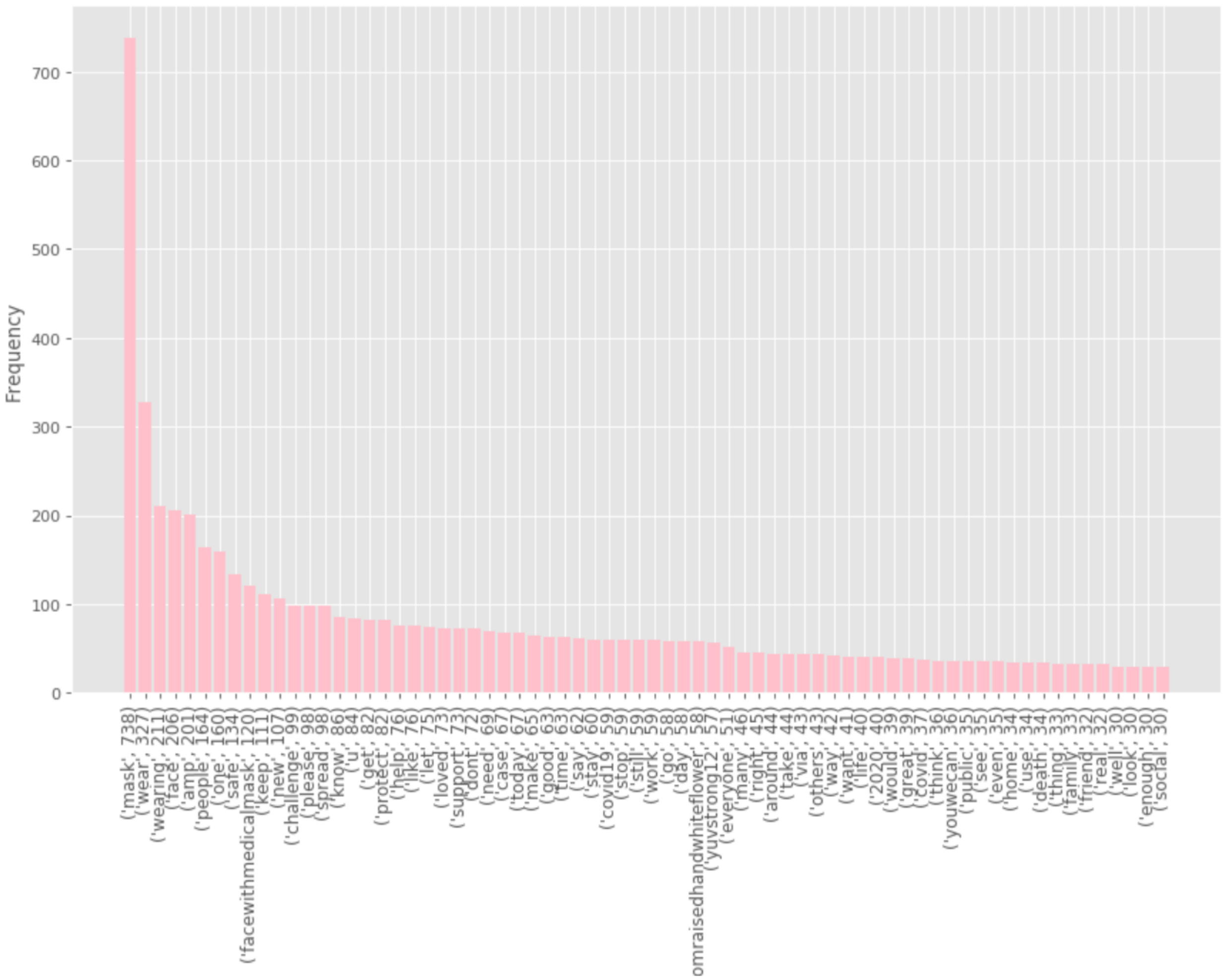


Figure 8: top 40 most common words in mask tweets

This figure 8 shows the top 40 common words in the group of trump tweets. Together with figure 7, both plots provide strong support that confirms that our interested keywords associated with covid19 tests prevail in the tweets.

Question of interest restated

Based upon the EDA, I choose to split the dataset with two hashtags interesting to look at: ‘trump’ and ‘mask’. The reason why these hashtags are chosen for study is because both hashtags apparently talk about the most commonly mentioned issues among during this pandemic, and the number of tweets containing these hashtags are similar. Performing sentiment analysis on these groups of tweets should provide insights on what the general sentiment of twitter users holds when they use these hashtags. Investigating in word frequency of interested topics also elicit insights into some commonly discussed issues, specifically mentioning of covid19 test during this special period of time.

For the following section of hypothesis testing, I seek to solve these two questions of interest:

1. Do tweets with hashtags related to ‘Trump’ and those with hashtags related to ‘mask’ exhibit the same distribution of positivity vs. negativity vs. neutrality?
2. Do tweets with hashtags related to ‘Trump’ and those with hashtags related to ‘mask’ mentions covid 19 test in a similar frequency?

Hypothesis test

<u>Part 1: Sentiment Analysis</u>

Chi-square Test and Variables

In this hypothesis test, I seek to investigate whether the proportion of positive, negative and neutral between trump tweets and mask tweets are different. The chosen dependent variable is the frequency of sentiment level in trump tweets and mask tweets. Our population of interest is all covid 19 tweets with the hashtags ‘trump’ and ‘mask’. The sample sizes of groups are different(2814 rows in trump tweets, NUM rows in mask tweets). Our population mean and variance is unknown. The population distribution is also unknown. Given that the data sets are independent and tweets that mention both hashtags are being removed, it fits into the assumptions of chi-square test for group-categorical comparison. As the chi-square test has assumptions that fit into these given conditions, I chose the chi-square test for hypothesis testing.

Contingency Table

Group	Positive counts	Negative counts	Neutral counts	Total counts
‘mask’ tweets	1120	403	1542	3065
‘trump’ tweets	741	623	1450	2814
Total tweets	74161	29395	75551	179108

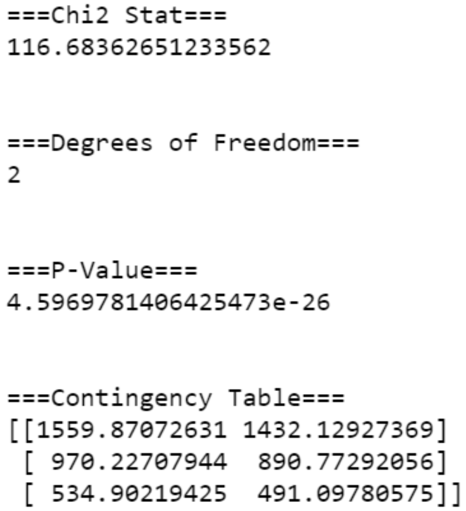
Group	Positive frequency	Negative frequency	Neutral frequency
‘mask’ tweets	0.365416	0.131485	0.503100
‘trump’ tweets	0.221393	0.263326	0.515281
Total tweets	0.414057	0.164119	0.421818

Hypothesis

Null hypothesis for this statistical analysis The null hypothesis would be that there is no difference between three proportions of positive posts, negative posts, neutral posts. The alternative hypothesis is that at least one of the proportions is different between two groups.

Results

Using the package of scipy.stats.chisquare in python, the expected observations and true observations were implemented to the chi-square model. The following output is shown below in figure 9.



```
===Chi2 Stat===
116.68362651233562


===Degrees of Freedom===
2


===P-Value===
4.5969781406425473e-26


===Contingency Table===
[[1559.87072631 1432.12927369]
 [ 970.22707944  890.77292056]
 [ 534.90219425  491.09780575]]
```

Figure 9: output of chi-square test for sentiment analysis

The shown output indicates that given the sample size and counts of sentimental levels in each data, there is extremely strong proof that the sentiment distribution among these two groups of tweets are different. The p-value of the test is close to 0, significantly smaller than the significance level.

<u>Part 2: word frequency analysis</u>

Chi-square test and variables

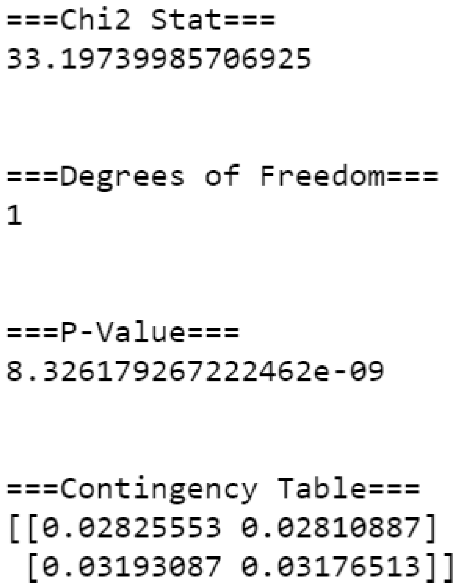
In this hypothesis test, I seek to investigate whether the word frequency of words related with covid19 tests between trump tweets and mask tweets are different from the word frequency of the whole data set. The chosen dependent variable is the count of presence of keywords: ‘positive’, ‘negative’, ‘test’, ‘tested’, in trump tweets and mask tweets. In statistical analysis, I will fit the observed frequencies and expected word frequencies into this model. The observed word frequencies are calculated from combining the counts of key words and counts of total tweets in both groups. Our population of interest is all covid 19 tweets with the hashtags ‘trump’ and ‘mask’. The population mean and variance are unknown. The population distribution is also unknown. Given that the data sets are independent and tweets that mention both hashtags are being removed, it fits into the assumptions of chi-square test, and thus chi-square is chosen for test of hypothesis.

Contingency Table for word frequency

Group	Keywords count	Total count	Word frequency
‘mask’ tweets	81	3065	0.026427
‘trump’ tweets	95	2814	0.033759
‘mask’ and ‘trump’ tweets	176	5879	0.029937
Total tweets	14621	179108	0.081632

Results

Using the package of scipy.stats.chisquare in python, the expected frequencies(0.030) and observed frequencies (0.026,0.033) were implemented to the chi-square model. The following output is shown below in figure 10.



```
===Chi2 Stat===
33.19739985706925


===Degrees of Freedom===
1


===P-Value===
8.326179267222462e-09


===Contingency Table===
[[0.02825553 0.02810887]
 [0.03193087 0.03176513]]
```

Figure 10: output of chi-square test for word frequency analysis

Consistent with what the contingency table is showing, either or both word frequencies in mask tweets and trump tweets are different from the expected frequency (average of word frequencies for both groups). The p-value of the test is close to 0, significantly smaller than the significance level. Thus we can reject the null hypothesis that observed frequencies are the same as expected frequencies.

Discussion

The results from hypothesis tests show that twitter users tagging trump in tweets do share a different sentiment distribution than those of tweets that tag 'mask'. Additionally, both groups of tweets are having a statistically different keyword frequency regarding covid19 tests. Referring from the contingency tables, one can also see that there is a trend of more negative tweets for trump tweets, and higher covid test keyword frequencies that in trump tweets, when compared to mask tweets. This finding is consistent with one of the claims of Trump made throughout the pandemic, that emphasized specifically on the United States' ample resource for covid 19 tests and its corresponding effect on the pandemic statistics.

Comparing the mask tweets' and trump tweets' sentiment ratio and word frequencies to those of total tweets from the whole dataset, one can see that both groups have different sentiment distributions and different keyword frequencies from the total tweets. For trump tweets, negative tweets occupy a higher ratio of tweets, while positive tweets occupy a lower ratio of tweets. This is possibly due to a general negative sentiment towards Trump among twitter users. For mask tweets, both negative tweets and positive tweets occupy a lower ratio of tweets, compared to the overall ratio in all tweets. A possible illustration for this finding is that most of the mask tweets, as shown from the frequent word 'please' and frequent hashtag 'wearamask', are talking about the need to wear a mask during a pandemic. Thus there is reasonably an increase in positive posts and neutral posts. While there are certainly negative posts that accuse people for not wearing masks, possible explanations are: these tweets occupy a small proportion of all tweets, or those negative tweets are more commonly associated with other negative hashtags like 'covidiots'.

Conclusion

In this observational study, I aim to answer the following questions: Do tweets with hashtags related to 'trump' and those with hashtags related to 'mask' exhibit the same distribution of positivity vs. negativity vs. neutrality? Do tweets with hashtags related to 'Trump' and those with hashtags related to 'mask' mentions covid 19 test in a similar frequency? I used a chi-square test to test the statistical significance of my hypothesis. Based on the output, it was concluded that trump tweets and mask tweets show different sentiment distributions and different covid19 test word frequencies.

By answering the question of interest, this study has provided insights into two commonly discussed issues among covid19 related posts. We can tell that although these are all covid19 tweets, a general pattern of sentiment distribution under each associated hashtags are different.

Some of these distributions reflect the general sentiments among twitter users, towards the discussed topics, like trump in covid19 tweets. The word frequency study reflected a dominant type of contents that are included in the selected tweets. By linking the effect of Trump's claim and its corresponding effect of tweets, one can also infer the power of twitter data for reflection of public sentiment as well as public influence of political figures. Both sentiment analysis and word frequency analysis illustrate great potential for future research studies.

Future studies

Because the used dataset is limited in both its size and time covered, it is not sufficient enough to answer my question of interest. In prospect of future studies, it will be optimal if more tweets can be included in, while covering a longer period of time, so that the dataset will not be sensitive to any bias caused by possible special incidents happening during this collected time. Future studies could also filter out hashtags 'covid19' or other related hashtags like 'coronavirus', because this dataset is already pulled from twitter dataset using API that specificies for these hashtags. Thus we already know that the whole dataset is related to covid19. When studying other related hashtags, a pull or filter out of the covid19 related hashtags could save more spaces and allow more hashtags to be visualized.

A challenge that arose was the finding the optimal hashtags that relates to common discussed topics amongst covid tweets. Certain hashtags do not fully represent the topics tweets text are representative of, thus an introduction of additional models into the analysis could be helpful. For example, adding in Latent Dirichlet Allocation models for topic modeling for trump tweets and mask tweets can provide a better insight into the most commonly discussed topics.

Resources for citation:

1. Hazra A, Gogtay N. Biostatistics Series Module 4: Comparing Groups - Categorical Variables. *Indian J Dermatol*. 2016;61(4):385-392. doi:10.4103/0019-5154.185700
2. https://towardsdatascience.com/running-chi-square-tests-in-python-with-die-roll-data-b9903817c51b
3. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chisquare.html
4. https://www.kaggle.com/gpreda/covid19-tweets