Create Sentiment Analysis Model of English-Language Twitter Tweets Concerning COVID-19 During the Course of April 2020

Problem Statement:

This capstone project aims to create a sentiment analysis model from the English-language, coronavirus disease 2019 (COVID-19) Tweets to predict sentiment towards the circumstances of the pandemic during the course of April 2020. The sentiment in this project will be categorized as positive, neutral, and negative. The primary clients this project targets are those that work in or deal with mental health and social analysis. Particularly, these clients would want to see how the social circumstances surrounding the COVID-19 pandemic and its lockdowns affect the social-welfare and mental health of the populace. COVID-19 is an infectious disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that primarily affects the lungs. However, it was discovered that it can also negatively affect other organs like the heart, kidneys, and brain, causing a multitude of other health complications that can leave a lasting impact even after recovery from the initial infection. It was first identified in Wuhan, China on December 2019 and has since spread worldwide, causing long-term lockdowns in many parts of the world as early as March 9, 2020. Besides taking a toll on medical health, it is indicated that the COVID-19 pandemic is also taking a toll on mental and social health, especially due to the stay-at-home orders. Hence, this project could give mental healthcare workers and sociologists insights on how the COVID-19 pandemic situation is affecting the mental health of the populace who are heavily limited in their movements and social interactions to prevent the spread of the virus. Also, the resulting model of this project could give indications on how resources of psychological care could be distributed and utilized to mitigate any negative effects long-term pandemics and lockdowns may have on emotional health.

The datasets that were used for this project were created by Shane Smith (https://www.kaggle.com/smid80) and posted on Kaggle.com under the CC0: Public Domain. They are available for download via the following links:

- Main Page: https://www.kaggle.com/smid80/coronavirus-covid19-tweets
- Early April: https://www.kaggle.com/smid80/coronavirus-covid19-tweets-early-april
- Late April: https://www.kaggle.com/smid80/coronavirus-covid19-tweets-late-april

Dataset Description:

It was decided that a supervised, classification prediction model using NLP will be created to predict whether the sentiment of a Tweet will be determined to be positive, negative, or neutral. However, as the original data does not have the sentiment labels, the sentiment analysis model will be trained with the pre-labeled corpus of Twitter samples from the Natural Language Toolkit (NLTK) Python library to predict and create sentiment labels for the original data. To create and clean the df_tweets.csv file for data exploration and for the main analysis of this project, multiple csv files pertaining to each day from March 29, 2020 to April 30, 2020 were concatenated together. This section describes the data wrangling and cleaning steps and methods used to create the df_tweets.csv file.

Using Python 3 in IPython Notebook, the multiple csv files containing tweets pertaining to the COVID-19 pandemic during the month of April 2020 were concatenated together via the glob module into one pandas dataframe denoted as df_tweets which contained 22 columns and

14,607,013 rows. Rows in which the Tweet text were not in the English language were dropped via a Boolean array. Also, columns in the df_tweets dataframe were removed by a Boolean array based on a 60% missing (null) percentage criteria as well as whether or not the columns would contribute to the data exploration and the main purpose of this project. It was decided that status_id (the ID of the actual Tweet) and user_id (The ID of the user account that Tweeted) would be dropped while the other variables were kept. This left the df_tweets dataframe with 13 columns and 8,133,785 rows. Due to the limited computational power of the hardware used to analyze the data for this project, 90% of the rows were randomly dropped from the df_tweets dataframe. This left the dataframe with 13 columns and 813,378 rows.

Text cleaning was done on the text column to prepare the df_tweets dataframe for the NLP analysis of the project. First, a function called, emoticon_text, was created to convert emoticons into their respective texts via a custom dictionary containing commonly used emoticons with their respective textual meanings (ex. :) : happy / smile, xD : laugh, :(: frown / sad / pouting, etc.) before it was applied on the text column with the new column called, cleaned_text, which will contain the cleaned text of the Tweets. Next, a function called, replace_contractions, was written to convert contractions in the text of the cleaned_text column into their basic words (ex. don't : do not, hadn't : had not, hadn't've : had not have, etc.). Then, a function called, tweet_cleaner, was made to remove url links and Twitter handles (@user), convert emojis (ex. 😁, 😆, 😍, etc.) into their respective text, keep only alphabetical words, remove whitespaces, and convert the text to lowercase under the cleaned_text column. Additionally, functions called, lemmatize_text and stem_text, were written to help to lemmatize and stem the words of the text under the cleaned_text column. Finally, the remove_stopwords function was created to remove the stop words of the text under the cleaned_text column. After applying the text cleaning funtions, the rows containing null values of the cleaned_text column were removed from the df_final dataframe. The final df_tweets dataframe has 14 columns and 813,378 rows.

The final df_tweets dataframe was exported as a CSV file named, df_tweets.csv.

Initial Statistical Findings:

According to Figure 1 and 2 below, the word frequency plots of both the text and cleaned_text columns of the df_tweets dataframe show how well the text cleaning methods worked in removing url links, Twitter handles, special characters, and stop words of the text as well as lemmatizing and stemming them. Also, Figure 2 shows how many more meaningful words appear in the word frequency count of the text from the text column after the text cleaning methods were applied to it.

Even though this project is an NLP analysis, additional exploratory data analysis was done to see how other features of the df_tweets dataset might affect the sentiment of the COVID-19 Tweets once the sentiment labels were made for the Tweets later in the project. Verified Twitter accounts are accounts that received the blue verified badge on Twitter, letting people know that an account of public interest is authentic. According to Figure 3 below, there are far less verified Twitter accounts than unverified Twitter accounts in the df_tweets dataset. However, according to Figure 4 and 5 below, verified Twitter accounts have more followers and friends than unverified Twitter accounts, especially for number of followers. Thus, there is a possibility that sentiment might follow the verified Twitter accounts as they contain more

Figure 1: 100 Most Frequent Words of “text” in df_tweets

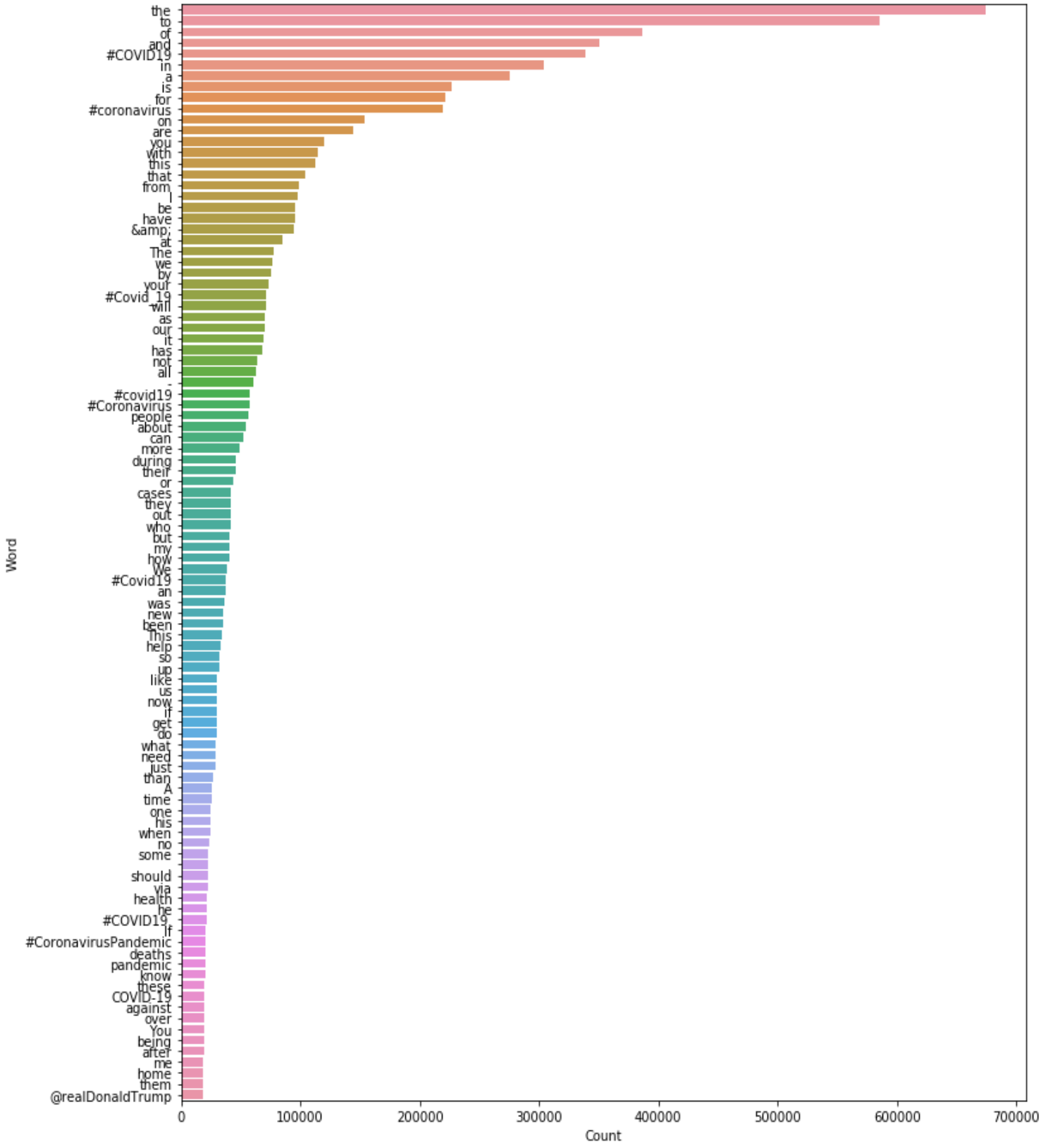


Figure 2: 100 Most Frequent Words of “cleaned_text” in df_tweets

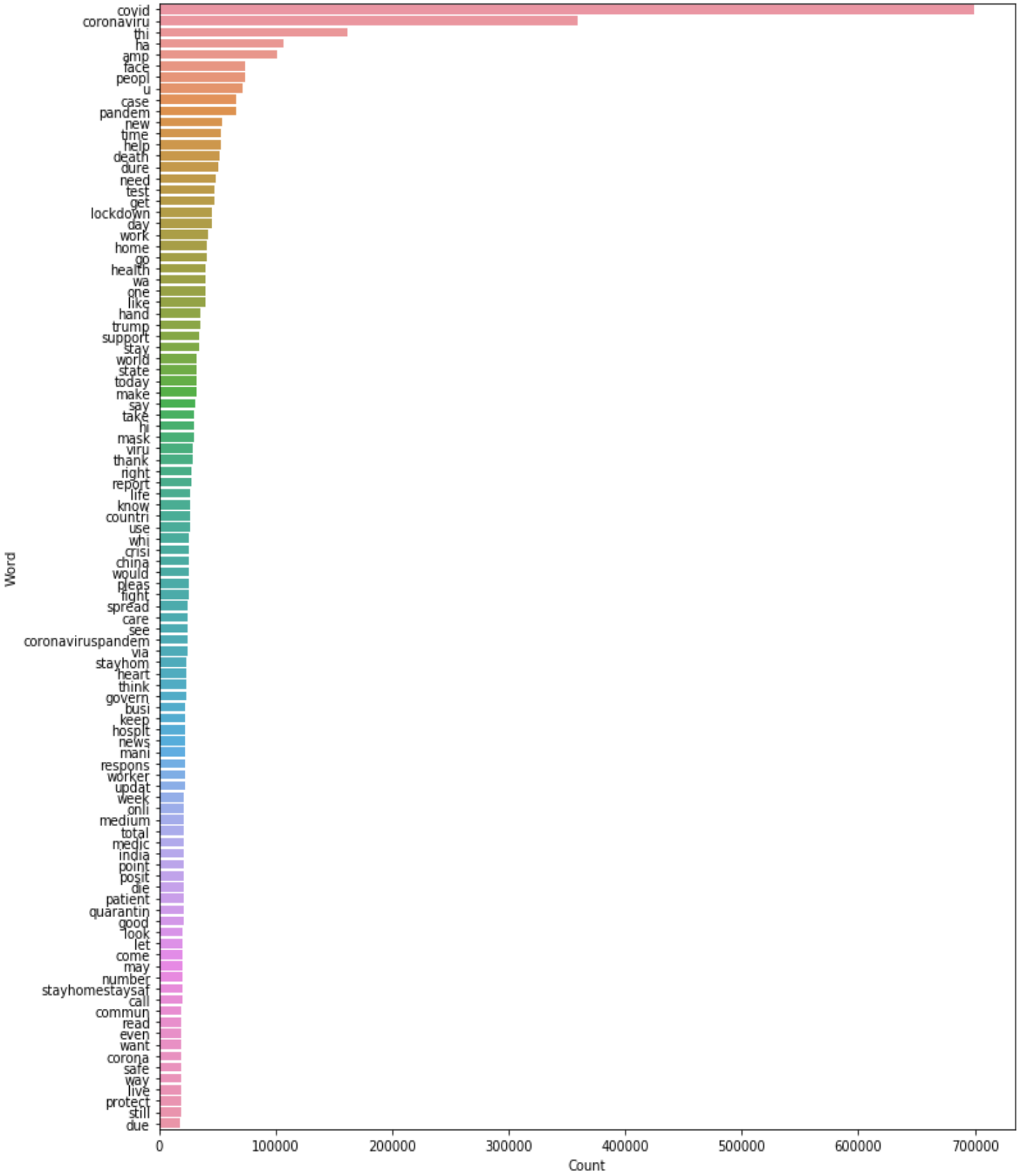


Figure 3:

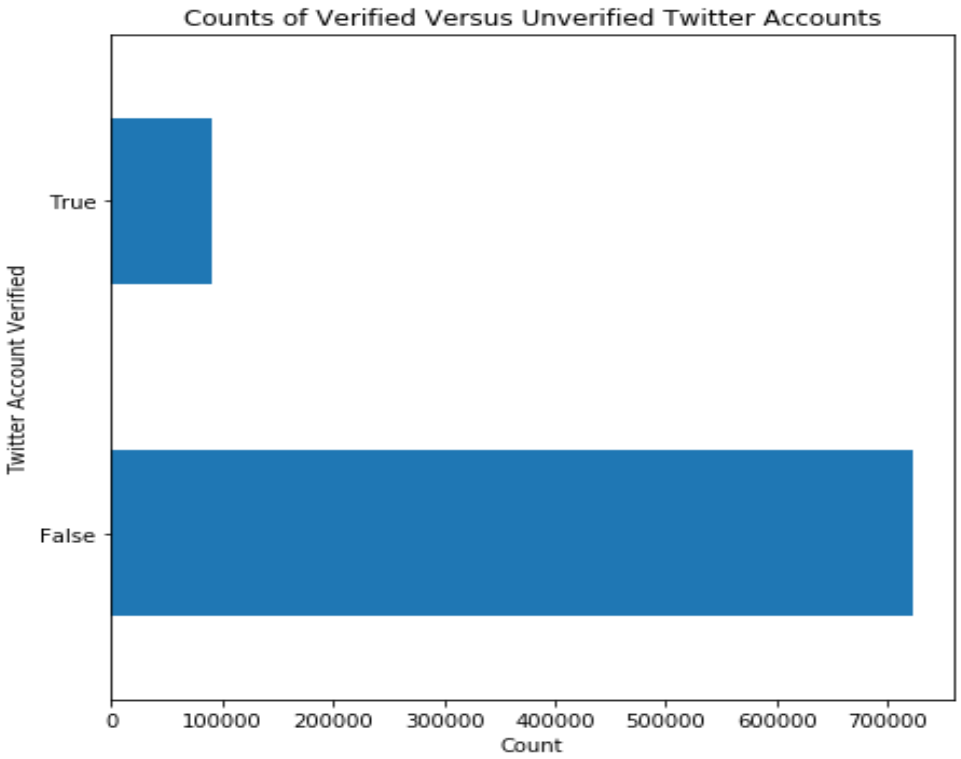


Figure 4:

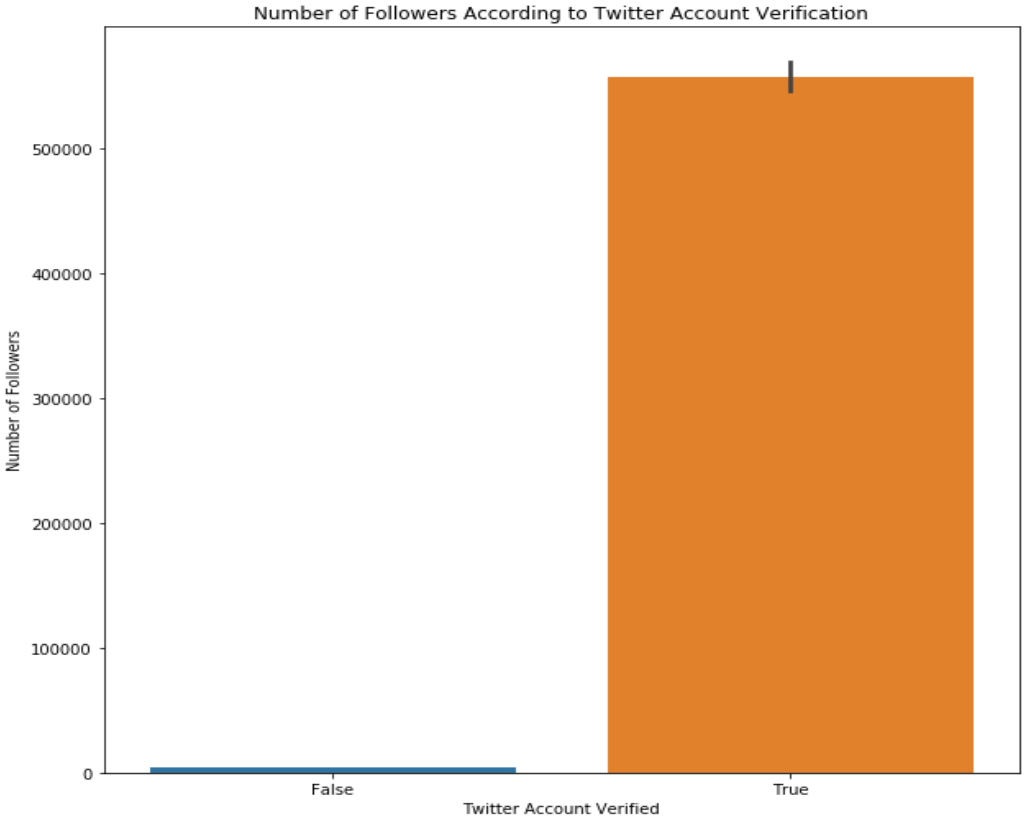
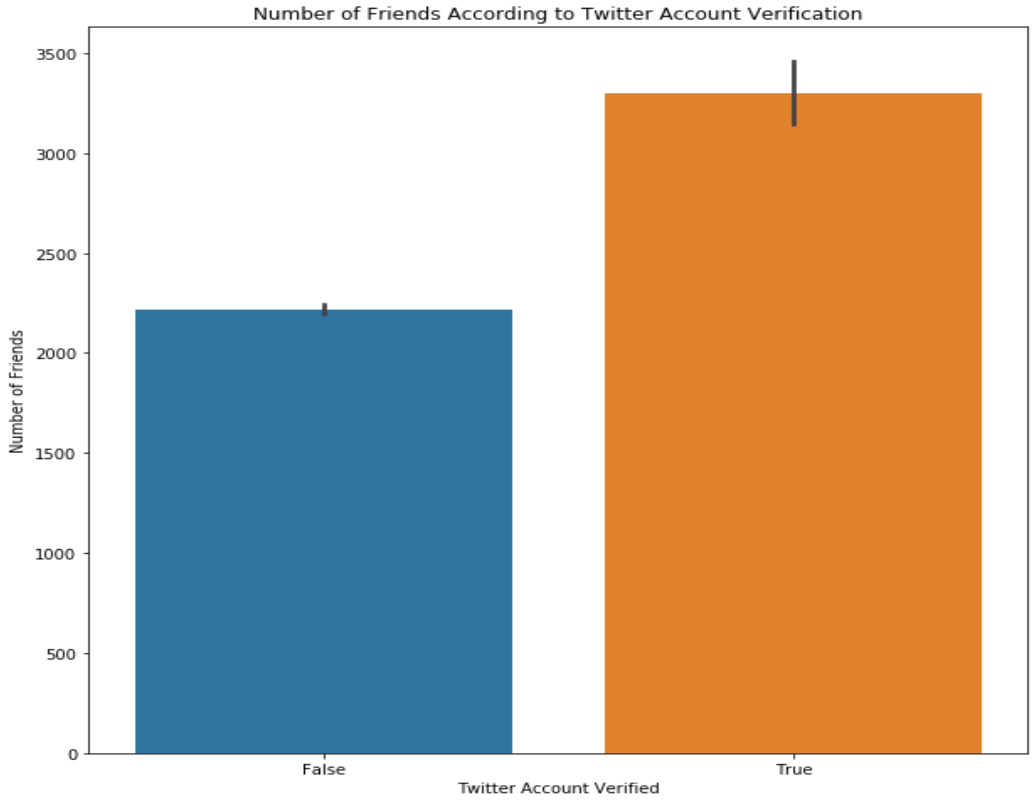


Figure 5:



numbers of followers and friends than the unverified Twitter accounts. Based on the scatterplots of Figure 6 and 7 below, the date in which the Twitter account was created does somewhat have an effect on the number of friends and followers as the higher count points tend to correlate with the older Twitter accounts. However, those are only a small number of data points, and a large number of the data points do not show any correlation with the age of the Twitter account. There is one noticeable outlier in the scatterplots in which a Twitter account was created in the year of 1970. This is supposed to be impossible as Twitter was launched on July 15, 2006. It can be assumed that this outlier was caused by a technical error and thus can be ignored.

To further explore the data, an unsupervised learning algorithm was used to analyze if there are any distinct classes in the text under the cleaned_text column in the dataset of df_tweets. The algorithm that was used was K-Means clustering since it can handle big data well due to its linear time complexity. Due to limitations of the RAM memory space of the computing computer, 97.5% of rows were randomly dropped from the df_tweets dataset which left only 20,334 of its observations to be analyzed under K-Means clustering. Below documents how this inferential statistics technique was used to analyze the data of df_tweets and the inferences made based on the results.

In order to perform K-Means clustering on the cleaned text under the cleaned_text column of the df_tweets dataframe, TF-IDF (term frequency–inverse document frequency) values were computed to vectorize the text of the cleaned_text column in order to create the features and X matrices. Then, the K-Means clustering algorithm was applied to the vectorized

Figure 6:

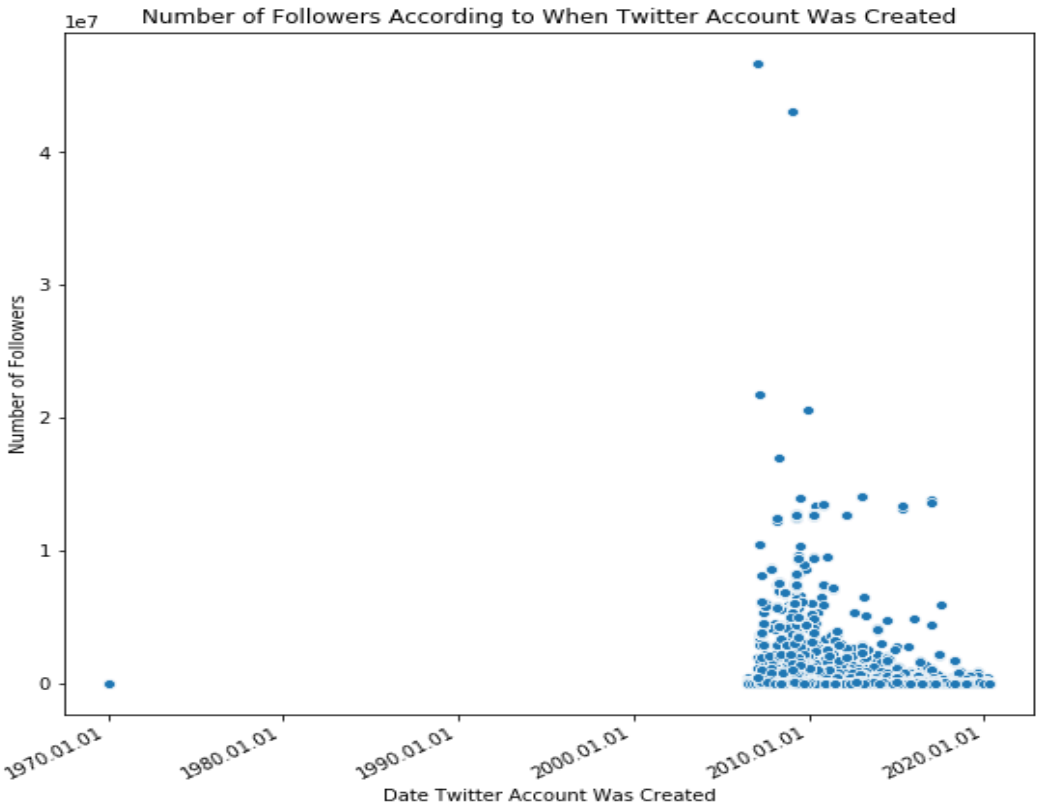
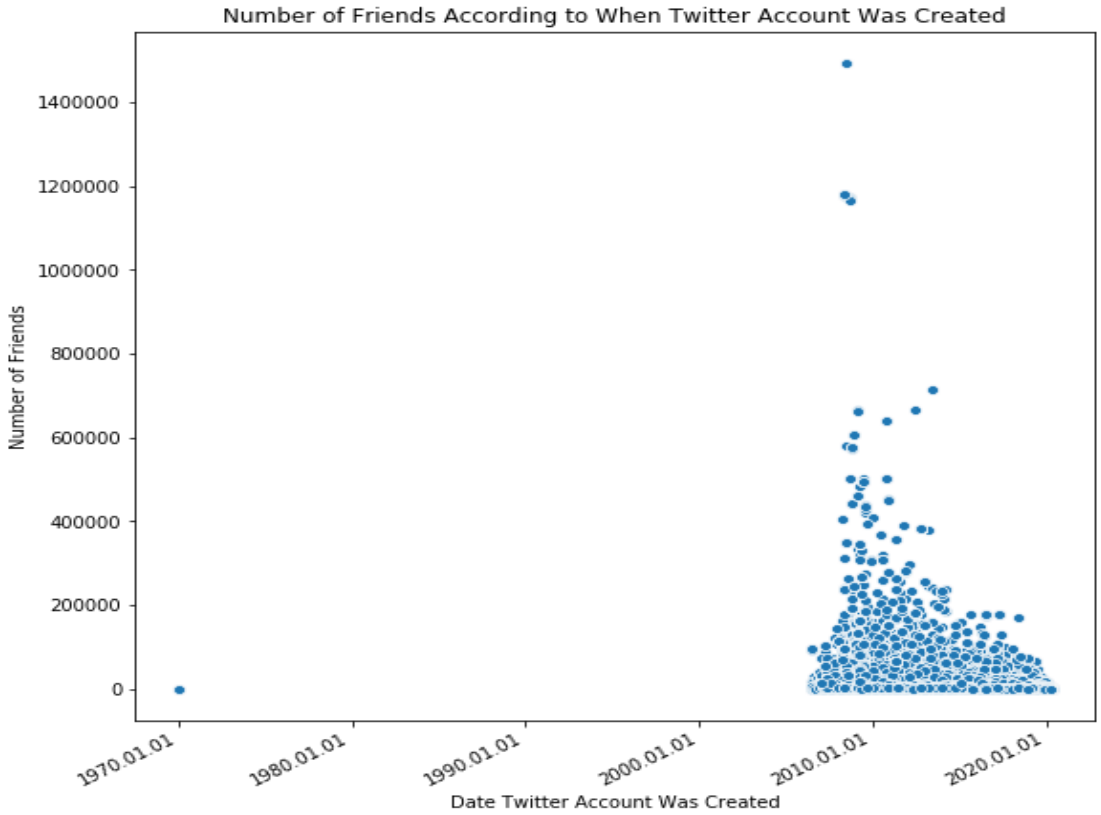


Figure 7:



text. The number of clusters in the algorithm was set to 3 since there are 3 sentiment classes to be analyzed for this project. The centroids and features of the K-Means clustering analysis were also calculated. Table 1 below shows which clusters the centroids, based on the words of the text in the cleaned_text column, belong to.

Table 1: Centroids in the Three K-Means Clusters

Cluster	Centroids
0	covid thi amp ha face peopl help dure time pandem
1	coronaviru covid thi ha pandem peopl trump lockdown test amp
2	case death total new covid report coronaviru confirm deliv number

The K-Means clustering plot in Figure 8 was made by creating an instance of K-Means and then using the PCA function in Python 3 to reduce the calculated features and the cluster centers to 2D. According to Figure 8 below, the scatterplot shows the points overlap instead of forming into three distinct clusters. As the text under the cleaned_text column in the df_tweets dataset is unlabeled for sentiment, the Silhouette Coefficient is used to check the clustering model in Figure 8. The best value of the Silhouette Coefficient is 1 while the worst value is -1, and values

near 0 indicate overlapping clusters. As the Silhouette Coefficient was calculated to be about 0.0017, the clusters very much overlap rather than being distinct from each other. The calculations of K-Means so far indicate that the cleaned_text column in the df_tweets dataset does not have the distinct classes to accurately predict which sentiment label of positive, negative, or neutral would likely occur for a Tweet pertaining COVID-19 in April 2020.

Figure 8: