April 9, 2021

Francisco Barba

Final Report:

Covid-19 Vaccine Sentiment Analysis

I am pleased to provide you with this report of my findings in conjunction to my analysis that utilizes Twitter data from approximately between March 1st, 2020 and January 1st, 2021. Hence, the recommendations and/or findings herein are purely objective and derived from the data collected.

# **1. Executive Summary**

My analysis indicates that the sentiment towards the covid-19 vaccine on the Twitter platform has increased in volume as time progressed within the interval of time in which the data was collected. The items shown below are discussed in greater detail within the body of the report.

* The tweets collected were filtered in order to accurately conduct the analysis.
* The frequency of tweets regarding the vaccine has increased by approximately 300 to 350 percent at the end of 2020.
* At the end of 2020, majority of the sentiments from the tweets were positive and supported the idea of the vaccine.
* Applying Bernoulli Naïve Bayes Classifier, I was able to achieve an accuracy score of 81.9 percent on correctly predicting if the data is positive, negative, or neutral on ‘unseen’ or future tweets.
* Applying the Random Forest Classifier, I manage to achieve an accuracy score of 82.6 percent on correctly predicting if the data is positive, negative, or neutral on ‘unseen’ or future tweets.

# **2. Project Understanding**

I understand that the purpose of this analysis is to gauge on the sentiment of the public with regards to the covid-19 vaccine and whether they will support it or not. Further, this information can be utilized to carefully plan the timing of when to rollout the vaccine to the public as well as when to plan the full reopening of public locations such as restaurants, schools, cafes, etc. This analysis is meant to provide insight to assist in preparations for a smoother transition from ‘lockdown’ levels back to ‘normal’ level, and to avoid either economic ‘shock’ or resurgence of an outbreak.

Initially, the data I receive consisted of approximately 38,450 tweets with 13 features such as “user\_name”, “user\_location”, “text”, “hashtags”, etc. Through process of elimination and filtering, I’ve managed to reduce the amount of tweets to approximately 24,500. Further, I’ve transformed each of the tweets to their most basic, fundamental forms which will then be the basis for evaluation during the modelling phase of this analysis.

# **3. Findings, Trends and Insight**

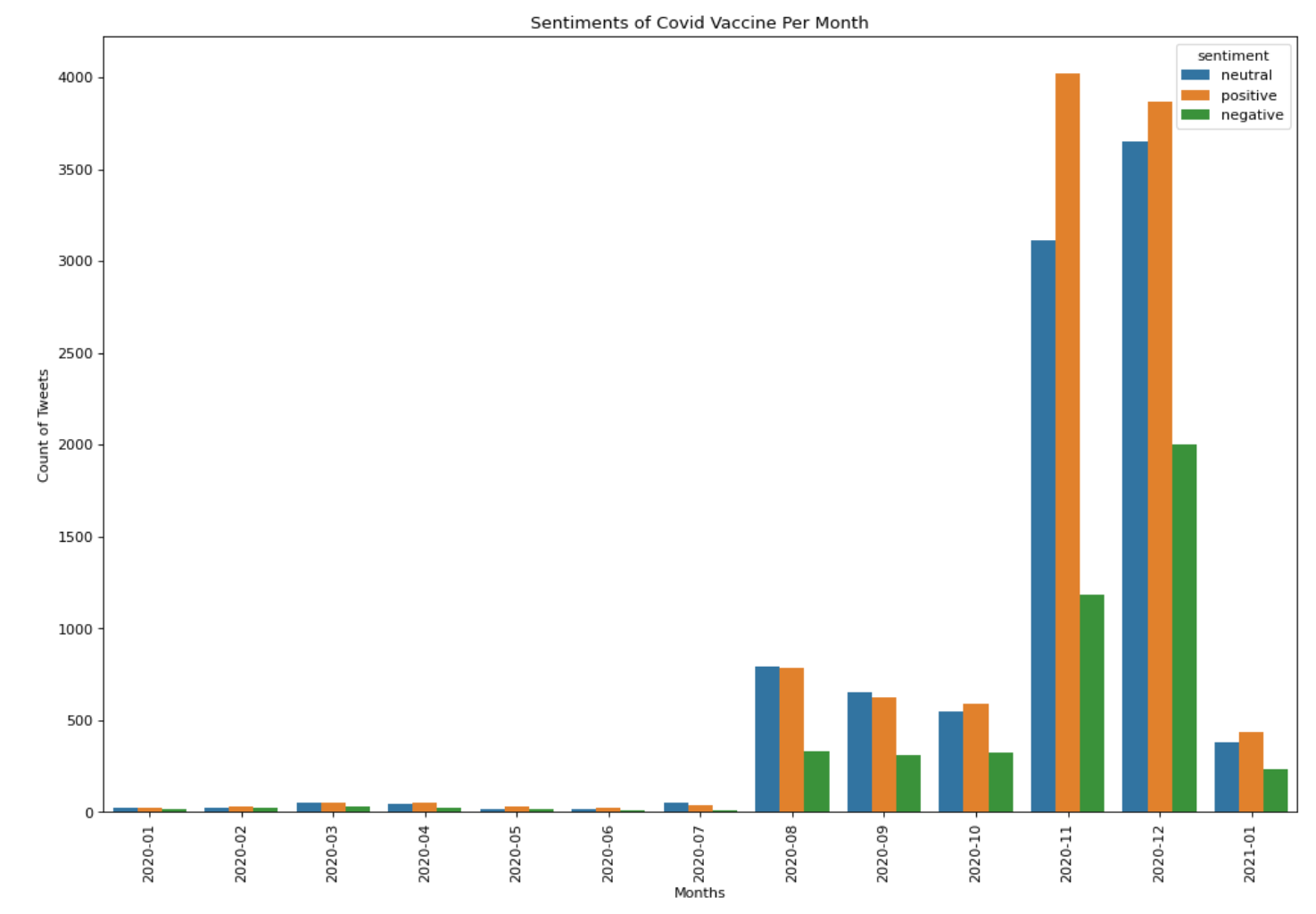
During the exploratory phase of the analysis, trends with respect to time were made more prevalent. The data showed that, prior to the month of November, there was relatively minimal amounts of tweets mentioning anything along the subject of covid-19 vaccines. Based on my research, this was not because the vaccine was not already under development.

*”A Phase 1 clinical trial evaluating an investigational vaccine designed to protect against coronavirus disease 2019 (COVID-19) has begun at Kaiser Permanente Washington Health Research Institute (KPWHRI) in Seattle.” (1)*

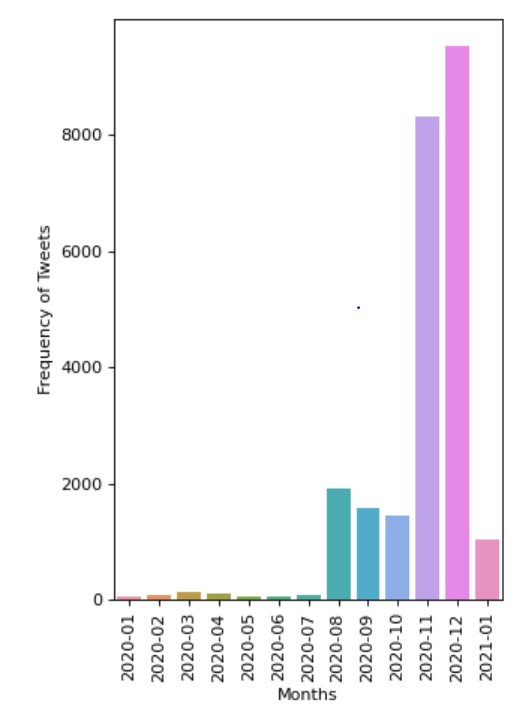
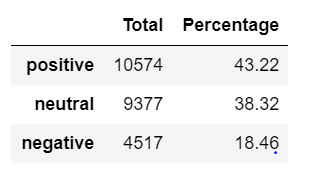
The above statement was from an article within the National Institute of Health website and is dated March 16th, 2020.

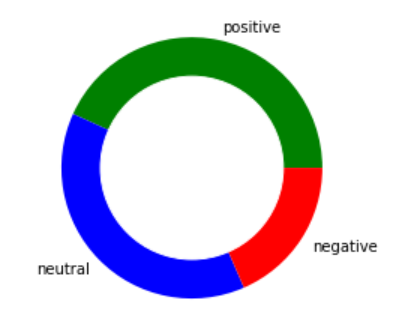
We can speculate that due to the holidays that lie within the months of November and December, that there was an increase of safety precautions the CDC, and government officials, released with regards to avoiding celebrating those holidays with family and loved ones. Hence, this stirred up popularity about the vaccines that led to the rise of more people supporting it – as shown on the following page in **Figure 1**. When observing the chart, please keep in mind that ‘positive’ means that the user supports the vaccine, ‘negative’ means the user is against it, and ‘neutral’ means that the tweet from the user was more likely an informative tweet (i.e. updates on the trial dates) and therefore has no subjectivity.

1 “NIH Clinical Trial of Investigational Vaccine for COVID-19 Begins.” *National Institutes of Health*, U.S. Department of Health and Human Services, 16 Mar. 2020, www.nih.gov/news-events/news-releases/nih-clinical-trial-investigational-vaccine-covid-19-begins.

**Figure 1: Sentiments of Covid Vaccine Per Month**

Furthermore, the volume of tweets with regards to covid-19 has proportionally increased as well. From between the months of October and November, the volume has increased by approximately 300 percent, as shown on the following page as **Figure 2**. The increase volume in conjunction to the sentiment shown in **Figure 1** depicts that there is far more support towards the vaccine, and that individuals are willing to get vaccinated. Overall, 43.22 percent of the tweets within the span of our study were positive, as opposed to the 18.46 percent that were negative. These percentages are depicted in the following page as **Figure 3**.



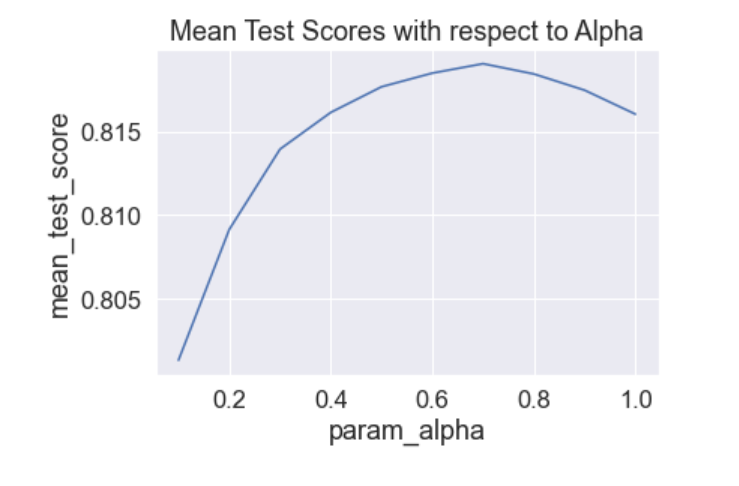


**Figure 2: Frequency of Tweets per Month Figure 3: Sentiment Percentage of Tweets**

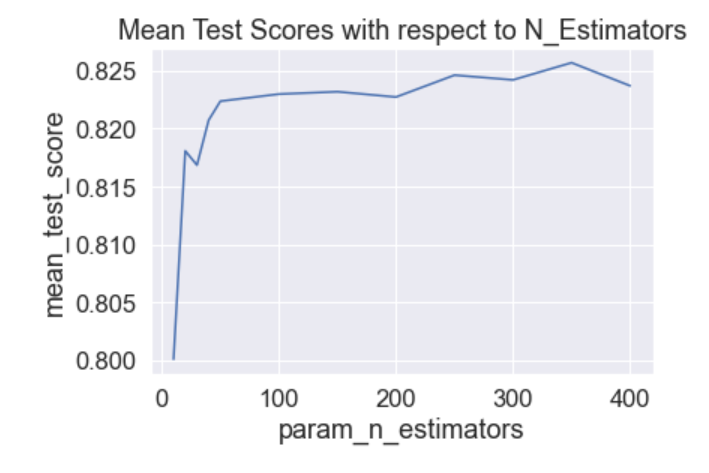
# 4. Model Selection and Metrics

For the modeling phase of the analysist, I chose a semi-supervised approach. I encoded the 3 sentiments into discrete values to then be tested by 2 machine learning classification models: Bernoulli Naïve Bayes and Random Forest Classifier.

The main function of the model is to feed it tweets to then evaluate whether each individual tweet is “positive”, “negative” or “neutral” without the need for myself to tamper or adjust the model. To accomplish this, I used the Tfidf vectorize with the “cleaned” tweets so their contents can be properly ‘weighted’ and have primed the models with 80 percent of the ‘weighted’ tweets and encoded sentiments.

* **Bernoulli Naïve Bayes Classifier:** What is great about using this classifier is that it treats all the labels independently, is fast, and can make real-time predictions. Therefore, since each tweet is not influenced by the other -for as far as we know-, it makes perfect sense to use this classifier. Using GridSearchCV to tune the hyperparameter “alpha”, and implementing stratified K-folds to split the tweets and its labels into 5 equally proportioned folds, I fitted the training set to the model and used the model to predict the labels on the test set. The model performed best when the “alpha” was set to 0.7 – resulting in an accuracy score of 81.9 percent. Below I have provided a line plot that shows the mean accuracy score of the model with respect to alpha, which we will refer to as **Figure 4**.

**Figure 4: Mean Test Scores with respect to Alpha**

* **Random Forest Classifier:** This ensemble method is similar to Bernoulli Naïve Bayes approach with regards to each tree being independent from one another. I chose it due to the idea that once the decision trees are fully drawn out, they will combine together to make a cohesive system of decision trees that I felt was ideal for this type of analysis. The process of tuning the hyperparameters of this model was similar to the Bernoullli Naïve Bayes except that the hyperparameter “alpha” was replaced with “n\_estimators” (or number of decision trees in the forest). Initially, I tried to tune the hyperparameters “n\_estimators” and “max\_depth” but found better results when I removed the “max\_depth” parameter completely. Be that as it may, the model performed best when the “n\_estimators” was equal to 350, providing an accuracy score of 82.6 percent – approximately 0.7 percent higher than Bernoulli Naïve Bayes model. Below I have provided a line plot that shows the mean accuracy score of the model with respect to n\_estimators, which will be referred to as **Figure 5**.

**Figure 5: Mean Test Scores with respect to N\_Estimators**

During the modeling process, the metric of focus was accuracy. The reason being is because I felt that precision and recall had equal weights with regards to this specific analysis and I wanted to maximize the general accuracy of its predictive prowess rather than maximize the focus of what to be more accurate on. Hence, making the model less bias on precision and recall, and more fixated on make the right prediction overall.

# 5. Limitations and Constraints

The analysis comes with its own flaws that should be brought to light and taken into account. Of all the flaws the biggest one is that the data is constricted by time. Since I originally started this analysis in late December to early January, the data collected begins in March 1st, 2020 and stems out to January 1st, 2021. At the time of this writing it is April 9th, 2021, and therefore, there is approximately a 3 month gap of missing information that could further improve the analysis and provide better clarity to the meaning of the results as well. Hence, the results of this analysis should not be viewed as an “exact” representation of the Twitter community’s opinion on the matter, but more so as a “roadmap” on where it may be possibly heading.

# 6. Conclusion

Based on the charts, there is undeniably a growing interest in supporting the covid-19 vaccinations. However, even though these sentiments are specifically from Twitter users and not the general public, it still does hold some truth to public opinion. At at the time of this writing, the vaccinations have been rolled out to the public, and society is steadily returning to ‘normal’.

Further, based on its accuracy score, Random Forest Classifier is the clear winner in predictability of sentiments. The goal of this analysis was to understand public opinion about the covid-19 vaccine via Twitter in order to gauge on the approach of how/when would be the optimal approach to efficiently release the vaccine to the public to avoid economic setbacks, wasting resources, prepare medical personnel, and to propagate a smoother and healthier transition to ‘normal’ societal functions. The Random Forest Classifier will be able to take in new tweets and return a result with 82.6 percent accuracy, in which can be further used to appropriately gauge on when would be an ideal time to begin the vaccine rollouts as we approach the dawn of the post covid-19 pandemic era.