Using Sentiment Analysis to Predict General Feelings Toward COVID-19 Vaccines

An Analysis of the General Opinions Toward COVID-19 Vaccines and the Relationship Between Vaccine Producers, Region, and Temporal Data as Found From COVID-19 Related Tweets

Madelyn Weber
CSCI 4502 - Data Mining
University of Colorado Boulder
Boulder, CO USA
mawe0876@colorado.edu

ABSTRACT

During the later stages of the COVID-19 pandemic, after vaccines began showing promise in clinical trials and later became available for public distribution, many took to social media to share their opinions on the vaccines available. Through the use of a Sentiment Analysis model, I will be examining text data collected from Twitter on Tweets related to COVID-19 vaccines. By looking at both the Tweet itself, as well as the hashtags provided alongside it, we can predict the classification of sentiment — the inclination toward positive or negative connotation — that each Tweet holds. The dataset used for this analysis holds Twitter data that is related to COVID-19 vaccines in general, and also includes Tweets that specifically mention at least one of the following major vaccines: Pfizer/BioNTech, Sinopharm, Sinovac, Moderna, Oxford/AstraZeneca, Covaxin, and Sputnik V. Through employing Natural Language Processing (NLP) techniques to the text data of each Tweet, sentiment classifications can be extracted, which can provide us with insight as to how users *generally* feel about a particular vaccine. Such sentiment classifications will fall into one of three categories: positive, negative, or neutral. Through data visualization, we can observe interesting patters by plotting histograms for the sentiment classifications provided for each vaccine. Given the results of the assigned classifications from the model, we can compare generalizations made to readily available information on the approximate number of individuals who have received doses of each vaccine, to analyze whether or not there is any correlation between the number of people who have received any of the vaccines mentioned and the sentiment classification assigned to the vaccine. The motivation for this project is to combine NLP and data mining techniques together in an attempt to propose a solution which may be put toward solving the current, real-world problem of increasing the numbers of the vaccinated population. Perhaps if we can learn more about how people generally feel toward particular vaccines, we can apply more strategic and widespread methods of distribution.

1 Introduction

To gain insight as to how people generally feel about the available COVID-19 vaccines, as well as if there is any correlation between the overall sentiment given to a specific vaccine in connection to the number of doses given amongst the vaccines examined, I have analyzed the *COVID-19 All Vaccines Tweets* dataset [1]. Through the help of a Sentiment Analysis model, sentiment classifications were assigned to each Tweet. The classifications, which were assigned as either positive, negative, or neutral, were saved to a specified dictionary corresponding to the given classification, along with each of the corresponding vaccine(s) that the were mentioned within the Tweet. This information was then visualized on a plot to further collect data as to how people generally feel about the vaccines available, how this corresponds to the current approximate number of doses distributed of each vaccine, and whether this information can further be used to improve current vaccination rates. This information can be used as an application in both the business and public health sectors, as it can provide us with the information necessary as to how we can develop an understanding on the public's general view of COVID-19 vaccines, as well as providing us with better insight as to a new distribution method to more effectively increase the numbers of the total vaccinated population.

2 Related Work

Plenty of research has been done in regards to NLP tasks using Sentiment Analysis to gather the general public's opinion on a wide variety of topics. Throughout the COVID-19 pandemic, Sentiment Analysis has been employed to various online platforms to gather the public's feelings on the situation. One study titled *COVID-19 Vaccination Awareness and Aftermath: Public Sentiment Analysis on Twitter Data and Vaccinated Population Predictions in the USA* [2] researched the sentiment classifications associated with COVID-19 vaccines from data collected on Twitter, where it was revealed that within Tweets referencing COVID-19 over a five week timespan, the general US population's feelings toward the vaccines available at the time were largely positive. This particular study employed two Sentiment Analysis tools, *Vader* and *TextBlob*, to classify their data, and then generated word clouds to visualize the most frequently occurring key-words associated with each vaccine for the negative, positive, and neutral classifications. This study analyzed the general sentiment associated with the vaccines, as well as other COVID-19 topics, which included pandemic-related keywords such as "social distancing" or "mask wearing". From this, another group of word cloud visualizations were created. The end result of this study provided a prediction as to the proportion of the US population that was projected to have one or both of the COVID-19 vaccine doses by the end of July 2021.

In a similar fashion, another study titled *Sentiment Analysis of COVID-19 Vaccine Tweets* [3] made use of *TextBlob* to gather sentiment classifications of Twitter data, which also made use of word clouds to visualize the most frequently used words associated with positively, negatively, and neutrally classified keywords. For my analysis of Twitter data related to the topic of COVID-19 vaccines, I have written my own Sentiment Analysis model to classify the sentiment given to COVID-19 vaccines into either a positive, negative, or neutral classification. I have also made use of the *TextBlob* [4] Python library as a comparison tool for checking the accuracy of my model. Rather than using word clouds to visually display words and sentiment classifications associated with the different COVID-19 vaccines, I have plotted the sentiment classifications for each vaccine for the positive, negative, and neutral classifications, which are compared graphically against plots of readily available COVID-19 vaccine distribution data found online, to examine whether any patterns can be found.

3 The Data

In order to properly train the Sentiment Analysis model, I selected the *Amazon Cells Labelled* dataset, the *IMDB Labelled* dataset, and the *Yelp Labelled* dataset, which can be found for available use from the *UCI Machine Learning Repository* [5]. Each dataset consists of 1000 lines of data, which includes a text review and a sentiment classification of either positive or negative. By combining all three datasets into one larger training set, the model was able to train itself off of 3000 instances of data. Each of these three datasets contained 500 instances of positive classifications, and 500 instances of negative classifications; this allowed for the model to have a balanced set of positive and negative classifications for it to use in its classification process, which helped ensure that the model wouldn't result in classifications skewed too far in one direction.

For the initial accuracy test of the model, I used an 80-20 ratio split in the training dataset. This means that the model was trained on 80%, or 2400 instances, of the dataset that were randomly selected from the entire training set. To test the accuracy of each classification, 20%, or 600 instances, of the dataset were also randomly selected from the initial training set and were set aside, thus ensuring the model did not see these instances during its training phase. This splitting process allows for a more defined measure of accuracy by reducing the possibility of the model overfitting to the data. Next, each text data instance was run through the model and assigned a classification. Measures for true positive, true negative, false positive, and false negative instances were collected and used to measure the accuracy of the model. These measures were collected by comparing the output classification from the model to the actual classification given within the dataset. The results of this analysis are given in more detail in section 5 below.

When it comes to selecting a dataset to train a Sentiment Analysis model on, it is important to ensure that the data is classified with the same classifications that we wish to examine. For this analysis, the Twitter dataset was to be classified for positive and negative sentiment, based on the Tweet author's personal sentiment toward a topic, so it is important that the training data reflects this. While all Sentiment Analysis models do classify for personal sentiment toward a topic, some training sets may classify their positive and negative labels differently, such as whether or not a Tweet holds hate speech rather than whether or not it is simply positively or negatively inclined. Thus, it is important to throughly examine the dataset to be used for the training stage. Finally, it is important to have enough data to train the model. Sentiment Analysis models are a supervised learning method, meaning that their accuracy heavily relies on what has been seen previously within the training phase. This means that as the model receives more data not previously seen during training, the classification outputs have the potential to become less accurate. This is the reasoning behind combining all three datasets named above into the larger single set.

For conducting the analysis of sentiment classifications on Twitter text data, the *COVID-19 All Vaccines Tweets* dataset was run through the Sentiment Analysis model. This dataset consists of 16 columns of data for each Tweet, with information for the following traits: ID, username, user location, user description, user account creation date, number of user followers, number of user friends, user favorites, whether the user is verified, date of Tweet, the Tweet text, hashtags, source, number of retweets, favorites, and if it is a retweet. For the purposes of my analysis, I only made use of the Tweet text data and the hashtag data, both of which are text data, allowing them to be combined into one long string of text. This was done because hashtags have the potential to contribute greatly to the overall sentiment of a given piece of text, even when the text itself may be ambiguous at times. Combining these together works because, for a Bayesian approach to Sentiment Analysis, the order of the words in the sentence does not matter; therefore there is no worry that appending individual hashtags onto the end of a sentence will throw off the model's results. After removing all empty text fields from the dataset, the model had a total of 152,229 text instances for analysis.

4 The Model

The Sentiment Analysis model created for this particular analysis project uses a Naïve Bayesian approach to classification, which takes on a bag-of-words approach. This simply means that the position of the words within a particular piece of text does not matter when it comes to classification. Instead, the count of each word's occurrence within each positive and negative class is what allows the algorithm to make predictions as to the likelihood of a classification for a given piece of text. This is done by collecting counts of frequency for each word, seen during training, in dictionaries corresponding to negative and positive classifications.

Due to this being an NLP task, preprocessing the text data is extremely important for acquiring accurate results. The first step in preprocessing any linguistics text for sentiment classification is to covert everything into lower case, as well as removing any non-alphabetical characters, including numbers, punctuation, and non-English characters in the case of an all English data classification. Without this, two word such as "example" and "Example" would be counted as two separate words by the model, due to the difference in capitalization for the letter 'e'. Tokenization of the text is another important step, which simply refers to splitting sentences up into individual tokens, or words. For each token, contracted forms must be expanded. This means that words such as "don't" will become "do not". This typically happens in the step between converting all letters to lowercase, but before removing non-alphabetical characters. Next is to lemmatize all forms of a word to retrieve their lemma — the underlying, or base form, of a word. This means that words such as "cats" or "eating" are converted to their base forms, "cat" or "eat", respectively. These steps are necessary when collecting occurrence counts of a word to be used in a Naïve Bayesian classification, as without them, words such as "Cats" and "cat" or "don't" and "doesn't" would be counted as two distinct words, even though their meanings are ultimately the same. Once all words have been converted to lowercase, stripped of all non-alphabetical characters, and converted into their underlying lemmas, stop words must be removed from the data. Stop words are any words within a language that hold no real sentimental purpose or semantic meaning within a sentence. For instance, words such as "the", "I", "and", etc. are all considered stop words in English, as they don't contribute much to the overall meaning of a sentence. For the Twitter analysis task at hand, I created my own list of stop words, which includes internet and Twitter expressions such as "lol", "user" or "retweet". While many stop word lists can be found online, it is important to keep in mind that we do not want to use just any list of stop words when conducting Sentiment Analysis, as words such as "not" or "never" are considered stop words in English, but will greatly impact a Sentiment Analysis model's classifications when removed. For example, in the given sentence with negative sentiment, "I do not like it", if we remove the stop word "not", we get a sentence with positive sentiment, "I do like it." Thus, for Sentiment Analysis tasks, it is usually better to maintain a smaller list of stop words rather than using a pre-made list of all English stop words. For these NLP pre-processing steps, I made use of the *NLTK* library in Python [6], which is an existing tool commonly used for conducing NLP processes.

Now that the text data has been preprocessed, we can consider the Naïve Bayesian model. Naïve Bayes is a supervised probabilisitic classification model, meaning that it uses pre-labeled training data to learn probabilities which it will use in its approach to classification. The first step using this approach is to go through the pre-labeled training data and collect frequency counts for each word that occurs for any corresponding negative or positive classification label. Once these are collected, a count of the total vocabulary — the unique and individual words that the model has seen — is collected, as well as the frequency counts for words belonging to both the positive and negative classes, as assigned from the training dataset. Once a frequency count has been created for both classes, dictionaries containing all the individual words seen are compared against each other. When instances of a word belonging to one class's dictionary, but not the other, are encountered, that word will be added to the dictionary in which it is missing from with a frequency count of zero; this is done to ensure that both classes hold the same items. Once this training is complete, we can move onto feeding the model our text data for classification. As mentioned previously, this is a Naïve Bayesian model, which uses probabilities for classification. Thus, during this classification phase, if any words with a probabilisitic outcome of zero are encountered, a smoothing technique will be employed to ensure that the probability results are small, but not a zero-value. This is done because having a zero probability within a calculation is not ideal, as it throws off the probability measure and reduces the accuracy of the model. This is an important step, as there is never a case when a word's probability will be truly zero — in the field on linguistics, natural language is seen as being infinite. Therefore, while the probability of any such word being used within a sentence may be very low, we can not say that it's impossible for that word to ever be used. To avoid inaccuracies within the probability value, we employ Laplace Smoothing, which is seen in the following equation, where V is the size of the vocabulary, *Size(C)* is the size of our class *C*, and the numerator is 1, due to the fact that this method is only employed when *Count*(w_i, *C*) is 0.

$$P(w_i|C) = \frac{Count(w_i, C) + 1}{Size(C)\ +\ V}$$

Now that we've addressed the problem of zero-probabilities for any word within the classes examined, we can move onto gathering the probabilities used for classifying any given piece of text. This is done by employing the formula below, where the calculation on the left-hand side is the prior probability distribution, which is added to the log likelihood, as seen on the right-hand side.

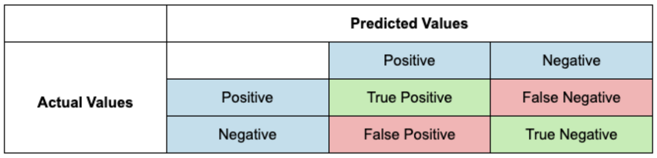
$$log\frac{P(pos)}{P(neg)} + \sum_{i=1}^{n} log\frac{P(w_i|pos)}{P(w_i|neg)}$$

The prior probability distribution is the ratio of the probability of a piece of text being positively classified over that of it being negatively classified, as found from the training data. This value is simply the probability of classification *before* factors from the text being examined are taken into consideration. The log likelihood is the calculated summation for every word in a sentence composed of n-words. This will give us the value for an assumed probability distribution for classification. For each i^{th} word, (w_i), the probability of that word belonging to the positive and negative class are calculated, and the ratio between them is found. The final summation is the combined probability of a sentence belonging to either class. We take the log of both sides because the final value has the potential to be too small for evaluation. From this calculation, the classification of any sentence can be determined. If the result is equal to zero, then it is classified as a neutral statement. If the result is greater than zero, it is classified as a positive statement, and if the result is less than zero, it is classified as a negative statement.

After the model has assigned a sentiment classification to each piece of text from the dataset, the text is then screened for mentions of each of the following major vaccines: Pfizer/BioNTech, Sinopharm, Sinovac, Moderna, Oxford/AstraZeneca, Covaxin, and Sputnik V. For each mention of one of these vaccines within the text, a frequency count within the dictionary corresponding to each classification will be incremented. There are three dictionaries corresponding to positive, negative, and neutral classifications. If multiple vaccines are mentioned within the same piece of text, then the count will be incremented for the corresponding vaccine in the corresponding dictionary. To exemplify this, if one piece of text mentions two difference vaccines, the count will increase by one for each vaccine within the corresponding dictionary. However, if the same vaccine is mentioned multiple times within the same text, the frequency is only incremented once for that text. If none of the vaccines being screened for are mentioned, then a count for the label "N/A" is incremented — here, "N/A" is used as a tag to refer to the fact that the Tweet is related to COVID-19 vaccines in general, but there is not a specific vaccine mentioned by name. These dictionaries are used for plotting, which will be displayed and analyzed later in section 6.

5 Analysis for Accuracy

Accuracy in a Sentiment Analysis model can be measured by collecting performance measures of accuracy, precision, and recall, which are then used to calculate an F1-score. All these measures come from instances of true positive, true negative, false positive, and false negative occurrences, which are depicted in the contingency table below.



		Predicted Values	
Actual Values		Positive	Negative
	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

These instances are collected by running the testing set through the model and comparing the assigned classifications from the model to the actual classifications provided within the dataset. Each value for the precision, recall, and F1-score are calculated from the following formulas:

$$Precision = \frac{True\ Positive}{True\ Positive\ +\ False\ Positive}$$

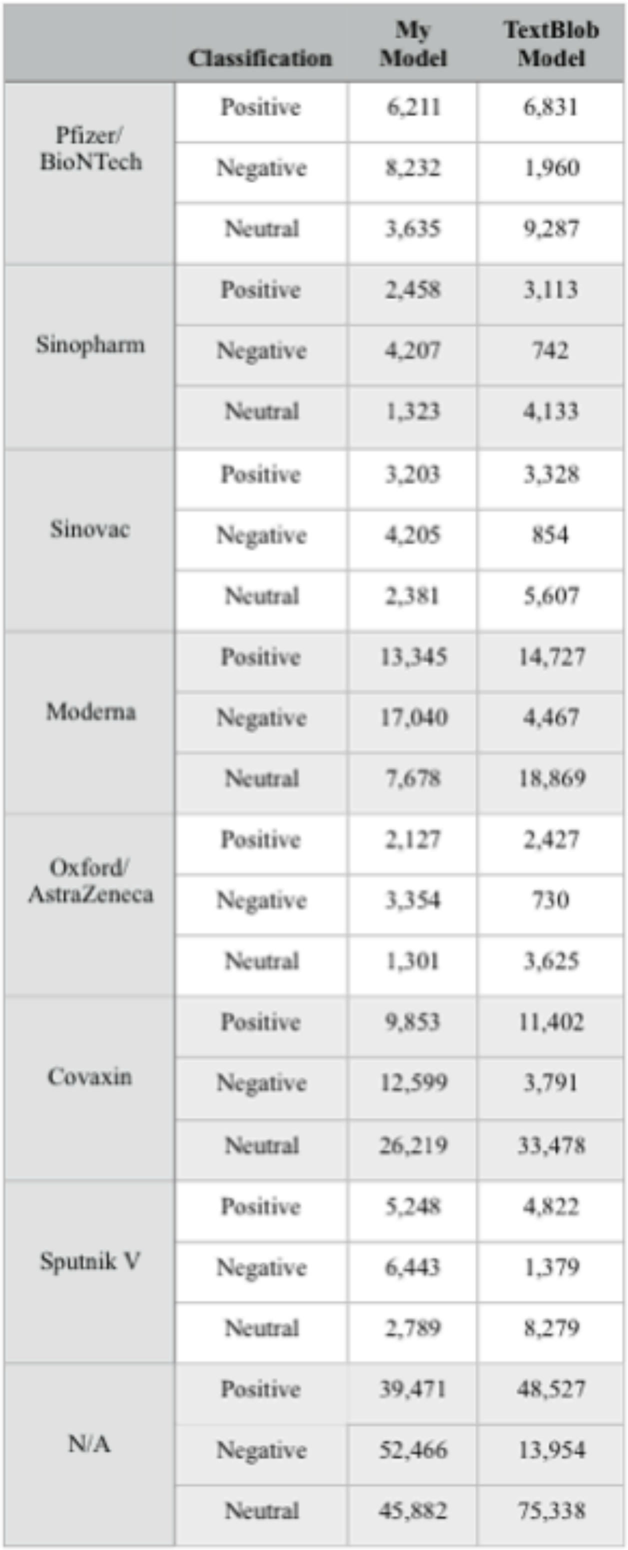
$$Recall = \frac{True\ Positive}{True\ Positive\ +\ False\ Negative}$$

$$F1 = \frac{2(Recall * Precision)}{Recall\ +\ Precision}$$

The F1-score is a better measure of accuracy than an actual accuracy score measure, due to the fact that the F1-score takes into account both the false positive and false negative cases, as we can see from the formulas above. This means that, while an accuracy measure is only calculating its score for occurrences in which the model was correct, the F1-score also takes into account the occurrences in which the model was incorrect. The average F1-score from three separate run-throughs of the model was calculated to be about 0.9622. As a perfect F1-score is a score of 1, we can conclude that this model is reasonably accurate.

As a secondary means of analyzing my model's classification accuracy, the classification results for the *COVID-19 All Vaccines Tweets* dataset were compared against the results from a *TextBlob* model. *TextBlob* is a publicly available Python library used for conducting various NLP tasks, including Sentiment Analysis. Due to TextBlob's accuracy, it is a convent tool to use toward ensuring accuracy within my model. A comparison of the total number of vaccines mentioned in text assigned to a classification of either positive, negative, or neutral can be seen in the table below. This was done with the intent of checking for major discrepancies between the two models, to get a better sense of where the model is doing well, and where it may be falling short. The results can be seen in the following table:

Figure 1 - Comparison between model classifications



	Classification	My Model	TextBlob Model
Pfizer/ BioNTech	Positive	6,211	6,831
	Negative	8,232	1,960
	Neutral	3,635	9,287
Sinopharm	Positive	2,458	3,113
	Negative	4,207	742
	Neutral	1,323	4,133
Sinovac	Positive	3,203	3,328
	Negative	4,205	854
	Neutral	2,381	5,607
Moderna	Positive	13,345	14,727
	Negative	17,040	4,467
	Neutral	7,678	18,869
Oxford/ AstraZeneca	Positive	2,127	2,427
	Negative	3,354	730
	Neutral	1,301	3,625
Covaxin	Positive	9,853	11,402
	Negative	12,599	3,791
	Neutral	26,219	33,478
Sputnik V	Positive	5,248	4,822
	Negative	6,443	1,379
	Neutral	2,789	8,279
N/A	Positive	39,471	48,527
	Negative	52,466	13,954
	Neutral	45,882	75,338

As can be seen from the table, the number of classifications assigned to the positive class is similar between the two models. However, there seem to be many discrepancies between what the two models consider to be negative or neutral classifications. This could be explained by a number of reasons. For one, although the training set for my model had a close 50-50 ratio between negative and positive classifications, this high discrepancy between negative and neutral classifications from the TextBlob model could mean that the model requires more instances of negative data to train itself on. It is also very likely that the algorithm in which TextBlob bases its calculations from is different from the Naïve Bayesian approach to classification implemented within my own model. As a result, this is a problem that should be addressed for future work using this model.

6 Evaluation of Results

In order to analyze correlations between the classification results from the Sentiment Analysis model and real-world data, I took the data from two sources which describe the current numbers of vaccine doses and quantity of distribution, and converted the information into visual graphs. The first source of data came from *Nikkei Asia*, a site which provides news from various perspectives around the Asian content. The visual bar graph below displays the number of doses distributed by country and/or region per vaccine producer; the y-axis is a measure for the number of hundred million doses in which contracts were signed for per vaccine [7].

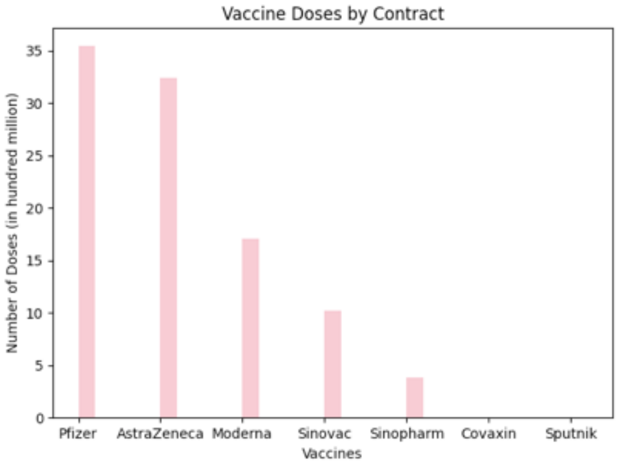


Figure 2 - Number of vaccine doses distributed by contract

As we can see from this graph, Pfizer and AstraZeneca are by far the two top administered vaccines. It is also worth noting that both Covaxin and Sputnik have counts of zero, as there was no data provided for these two vaccines from the source. From this graph alone, it can be predicted that Pfizer, AstraZeneca, and perhaps Moderna can be expected to see the most counts for classification from the Sentiment Analysis model, as they are all highly administered vaccines, meaning that more individuals have likely had exposure to them, and thus are more likely to hold opinions relating to them.

The second means of comparison is to look at the data collected in which the global outreach of COVID-19 vaccines is displayed. This data was found from *BBC News* [8], and is displayed in the graph below. It shows the total number of countries and/or territories which have officially reported the use of any of the vaccines listed.

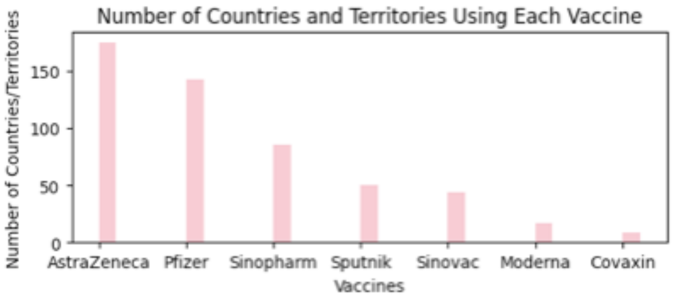


Figure 3 - Vaccine outreach by official country/territory

As we can see from this graph, AstraZeneca and Pfizer are yet again the top two distributed vaccines. This further backs up the prediction in which we will likely see many classifications for these two vaccines, as they are among the listed vaccines which have distributed the greatest number of doses, as well as those with the highest global outreach.

We can now examine the classification results from the Sentiment Analysis model for the *COVID-19 All Vaccines Tweets* dataset. The first graph to be examined is that of the classification results of my Sentiment Analysis model. The second displays the classifications assigned to the same dataset, but from the *TextBlob* classification model.

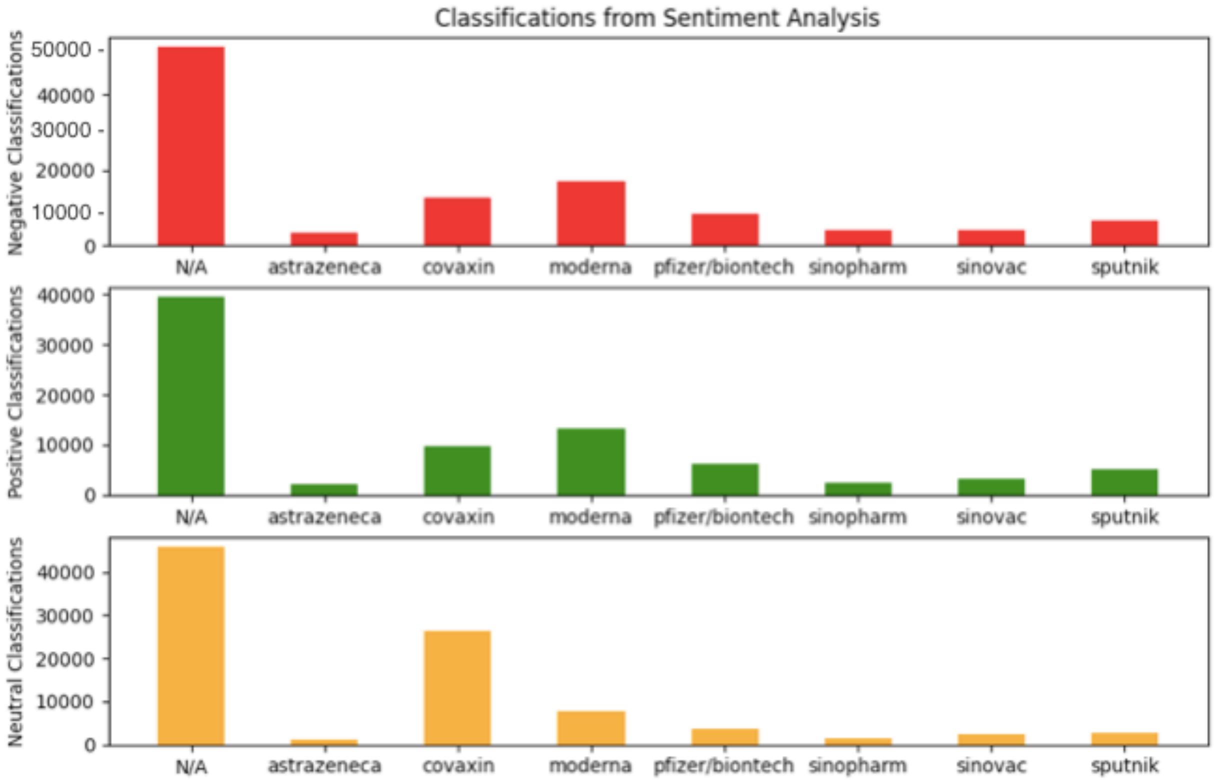


Figure 4 - Classification results from Sentiment Analysis model created for this analysis

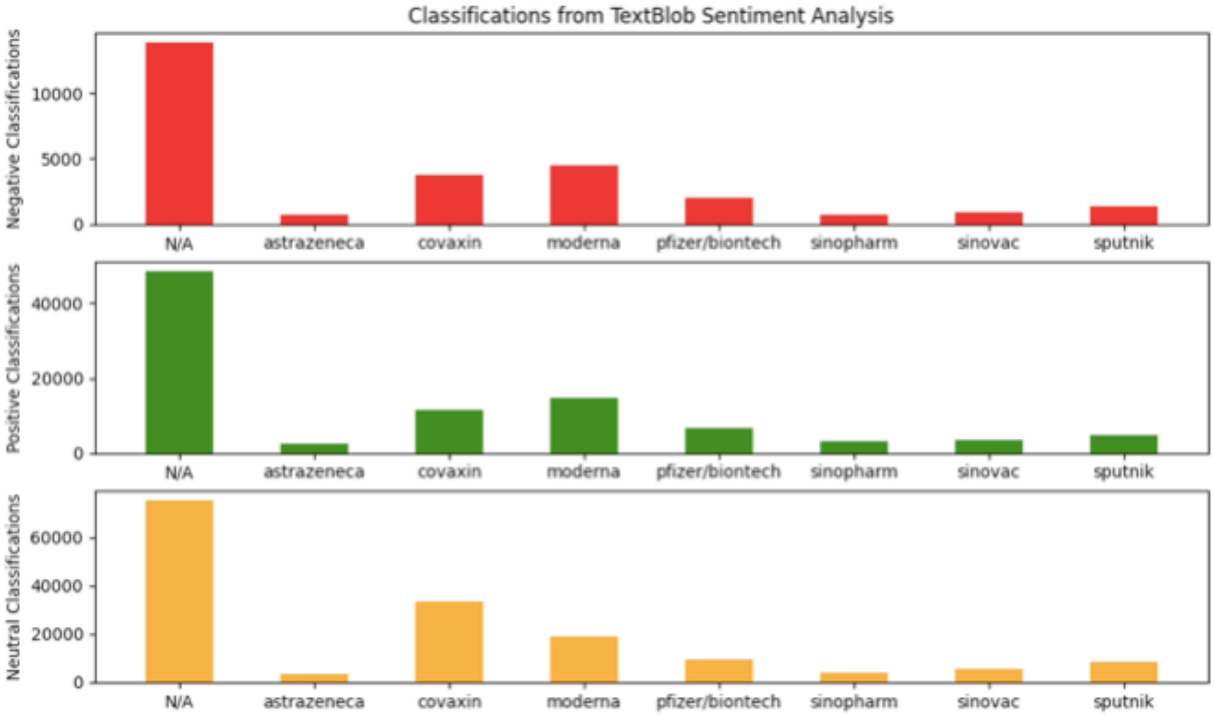
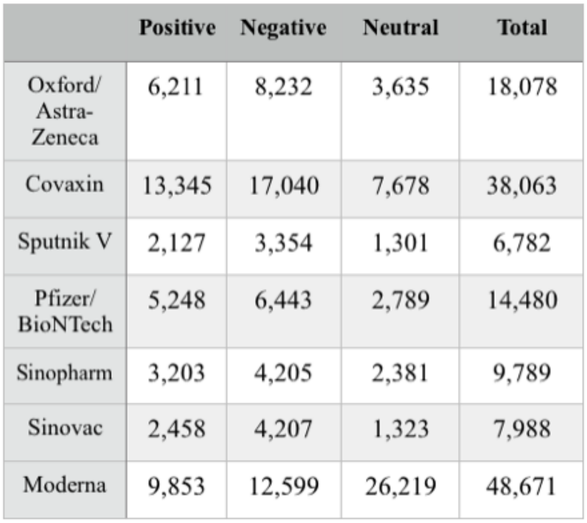


Figure 5 - Classification results from TextBlob Sentiment Analysis model

Although we saw above in section 5 that there were some discrepancies in the number of classifications between my model and the TextBlob model for neutral and negative classifications, we can see from the visual representations above that the relative distribution curves are fairly similar, and thus, the results from my model can confidently be used as a means of overall analysis. Disregarding the counts for "N/A", we can see that Moderna and Covaxin receive the most mentions from the dataset. This is contradictory to the predictions made above from the real-world distribution data, where it was predicted that the dataset would reflect the results of Pfizer and AstraZeneca receiving the most attention. The table below displays the total counts of items belonging to each class for each vaccine, not including the generalized "N/A" category.

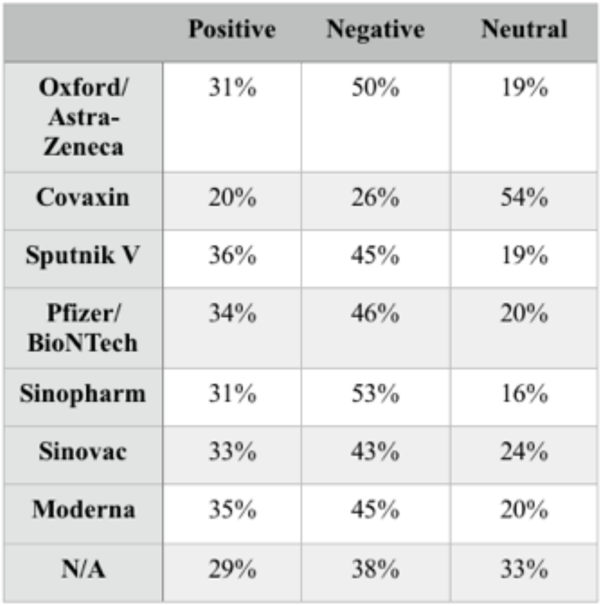


	Positive	Negative	Neutral	Total
Oxford/ Astra-Zeneca	6,211	8,232	3,635	18,078
Covaxin	13,345	17,040	7,678	38,063
Sputnik V	2,127	3,354	1,301	6,782
Pfizer/ BioNTech	5,248	6,443	2,789	14,480
Sinopharm	3,203	4,205	2,381	9,789
Sinovac	2,458	4,207	1,323	7,988
Moderna	9,853	12,599	26,219	48,671

Figure 6 - Counts per classification for each vaccine

While it would appear that the general opinions toward COVID-19 vaccines is largely negative, we can combine the classifications for the positive and neutral classes together to examine two categories of negative versus non-negative classification. This is an acceptable combination for this analysis due to the fact that neutral statements are not the same as negative statements. From a linguistic and psychological point-of-view, when humans make statements of opinion, especially negative statements, we tend to use clear language to express our feelings of negativity. It is very rare that someone who feels strong negative emotions toward a topic would make a comment that is largely neutral. As an example, a Tweet such as, "The Covid vaccines are now available. #COVIDVaccine" is a largely neutral statement, especially when compared to a statement such as, "Unfortunately, the Covid vaccines are now available. #COVIDVaccine." When comporting these two hypothetical Tweets, the likelihood of someone who feels negatively toward the vaccines posting the neutral statement is very unlikely. Thus, due to the higher likelihood of negative sentiment coming through in an individual's language, we can begin to examine the outcome of the sentiment model as being either negative or non-negative, leading us to the conclusion that the feelings toward COVID-19 vaccines, in general, is largely non-negative.

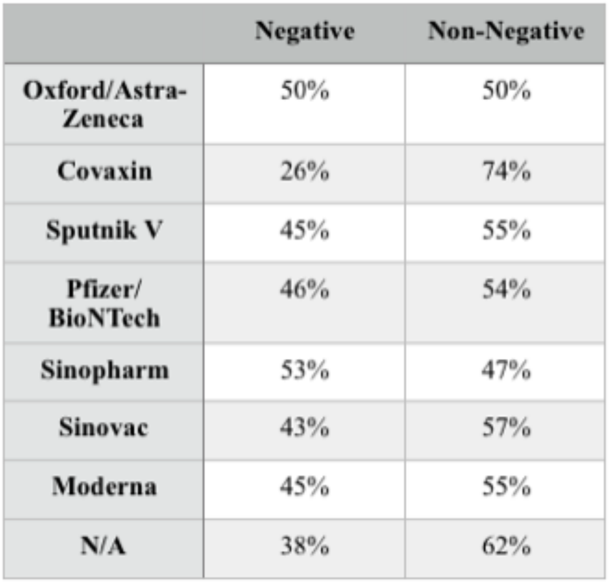
Looking back to figure 4, we cannot say from the graph alone which vaccine appears to have the highest negative and non-negative classifications. From just a visual appearance, one may feel inclined to conclude that Moderna is viewed more negatively than other vaccines, such as Sinopharm, due to its higher frequency of counts within the negative class in the table. This, however, is not a conclusion we can draw from the visualization alone. If we take a look at the total number of classifications displayed in figure 6, we can see that Moderna has far more classifications assigned than other vaccines, such as Sinopharm. Thus, it appears to have more negative sentiment assigned in the model, but this is a direct correlation to the frequency of mentions, and not the overall sentiment. Therefore, due to a lack of equal classifications assigned to each vaccine, we cannot draw conclusions from the graphs or table alone, and must look further into the ratio of classifications to see how each vaccine compares. This is displayed in the table below, where we can examine both the ratios for the positive, negative, and neutral classes, as well as the more condensed negative versus non-negative classes.



	Positive	Negative	Neutral
Oxford/ Astra-Zeneca	31%	50%	19%
Covaxin	20%	26%	54%
Sputnik V	36%	45%	19%
Pfizer/ BioNTech	34%	46%	20%
Sinopharm	31%	53%	16%
Sinovac	33%	43%	24%
Moderna	35%	45%	20%
N/A	29%	38%	33%

Figure 7 - Ratio of classifications per vaccine

Interestingly enough, we can see from the ratios above that the results are fairly similar across the positive and negative classes. Now, if we compare this table to the table of ratios for the negative and non-negative classes, we can draw some interesting conclusions.



	Negative	Non-Negative
Oxford/Astra-Zeneca	50%	50%
Covaxin	26%	74%
Sputnik V	45%	55%
Pfizer/ BioNTech	46%	54%
Sinopharm	53%	47%
Sinovac	43%	57%
Moderna	45%	55%
N/A	38%	62%

Figure 8 - Ratio of classifications per vaccine for negative and non-negative classes

From our ratio measures, we can see that the ratio of classifications are largely non-negative across the board. We can also use the results from figure 8 to draw some conclusions as to the general feelings toward COVID-19 vaccines. Namely, that the Covaxin vaccine has the highest ratio of non-negative sentiment out of all the vaccines mentioned. We can also see that while the majority of vaccines have an overall non-negative ratio, the Sinopharm vaccine is the only one which has a higher ratio of negative sentiment attached to it.

If we look back at figures 2 and 3, we saw that the highest real-word distribution of vaccines was for the AstraZeneca and Pfizer vaccines, whereas the Covaxin and Moderna vaccines were among the lowest in distribution. Sinopharm fell around the mid-range for distribution. Putting this information together, we can draw some ideas for new potential distribution plans. First, although AstraZeneca and Pfizer are amongst the world's top vaccines in terms of distribution, they are closely divided between negative and non-negative classification; AstraZeneca even has an exact 50-50 split in classifications assigned. Covaxin, on the other hand, was amongst the world's lowest distributed vaccines, and yet holds the highest rate of non-negative classifications. From this, it could potentially be useful to increase the number of doses of the Covaxin vaccine available around the world. This may include more countries accepting contracts for doses of this vaccine, or even making it one of their officially recognized vaccines. Similarly, we saw that Sinopharm was mid-range in terms of the world's distribution data, and yet it holds the largest rate of negative classifications amongst all the vaccines examined here. Thus, it could potentially be useful for regions which are distributing this vaccine to increase the availability of other vaccines offered, to give people more options to chose from.

7 Conclusion

As can be seen from the data tables and plots above, the general feelings toward COVID-19 vaccines, as viewed from Twitter, is largely non-negative. We can also see that the vaccines mentioned most frequently from the Twitter data are not necessarily those which are in high distribution when it comes to the number of doses, or even the number of countries/territories, officially distributing each vaccine. For instance, from the real-world distribution data, the Covaxin vaccine had the lowest number of distributions, and yet it had the highest percentage of non-negative sentiment. Similarly, the Sinopharm vaccine fell around the mid-range in terms of distribution rates, yet had the highest negative sentiment, and the AstraZeneca vaccine, which had some of the highest numbers for distribution, was split 50-50 between negative and non-negative sentiment.

Of course, it is worth keeping in mind that the ratios of each classification assigned are directly related to the awareness of each vaccine and the number of people actively talking about them. This means that just because a vaccine, such as Covaxin, shows a higher ratio of non-negative classification, that does not necessarily mean that it is ultimately viewed more favorably over other vaccines, such as Sinovac or Sputnik. This data comes from English-speaking Twitter users, which means that the vaccines largely mentioned here are likely what the general English-speaking population feels toward each vaccine. If more data from Twitter users of other languages were to be examined, we could gather an even more accurate analysis as to how people generally feel about certain vaccines, and further go on to suggest an even more strategic and accurate distribution plan.

We can use this information as an application to the real-world problem of increasing the numbers of the vaccinated public against COVID-19. Through this data, new policies can be introduced. These would include taking measures such as increasing the availability of the Covaxin vaccine, due to the fact that it had such a high percentage of non-negative sentiment collected by the model, and yet it was among the lowest of the vaccines mentioned within the real-world distribution data. Similarly, it could be potentially useful to reduce the availability of some vaccines, such as the Sinopharm vaccine, due to the fact that it landed around the center of skew for the real-world distribution data, but also had the largest ratio of negative classifications assigned.

8 Future Work

As for future implications and improvements to this current system of analysis, it would be more effective to implement a system of regional recognition, to group all user-defined locational data, as given in the Twitter dataset, into a condensed group of locations; cities, states, or provinces could be summarized into one country, continent, or region. This would be useful because not every vaccine is available for distribution worldwide. Therefore, obtaining a more localized idea as to how the people of any given region generally feel toward the vaccines available to them, and the ones that aren't, could help to better understand how we can further develop a more effective system of distribution. Regional data can also help to eliminate bias from the data. As we can see from the table in figure 8 above, the Sinovac vaccine has the highest number of negative classifications. However, these classifications are being assigned to a Chinese-made vaccine from English-speaking Twitter users. From this, a potential causation to these higher negative classifications may come from underlying bias, prejudice, or even racial discrimination against China and the Chinese-made vaccines.

Another potential improvement would be to implement a means of scaling the classifications assigned. Currently, the model simply groups any piece of text into a class of positive, negative, or neutral. However, some statements may have different levels of intensity within their own class. For instance, if we were to put the two sentences, “This place sucks.” and “This place wasn't great.”, into the model, they would likely both come out with a negative classification, even though one is clearly more negatively inclined than the other. Through implementing a scaling system in which each statement is classified on a level of [1,10] (for the positive class) or [-1,-10] (for the negative class), the first example may be assigned a score of: (negative, -8), whereas the second example may be assigned something along the lines of: (negative, -3), to clearly display the difference in intensity. By incorporating such alterations into the analysis system, we could obtain a more accurate idea of just how people feel about each vaccine in a less-general scope.

9 Sources

[1] Preda, Gabriel. “Covid-19 All Vaccines Tweets.” Kaggle, 23 Nov. 2021, https://www.kaggle.com/gpreda/all-covid19-vaccines-tweets.

[2] Sattar, Naw Safrin, and Shaikh Arifuzzaman. “Covid-19 Vaccination Awareness and Aftermath: Public Sentiment Analysis on Twitter Data and Vaccinated Population Prediction in the USA.” MDPI, Multidisciplinary Digital Publishing Institute, 30 June 2021, https://www.mdpi.com/2076-3417/11/13/6128.

[3] Dua, Sejal. “Sentiment Analysis of COVID-19 Vaccine Tweets.” Medium, Towards Data Science, 29 Mar. 2021, https://towardsdatascience.com/sentiment-analysis-of-covid-19-vaccine-tweets-dc6f41a5e1af.

[4] “Simplified Text Processing.” TextBlob, https://textblob.readthedocs.io/en/dev/.

[5] UCI Machine Learning Repository: Sentiment Labelled Sentences Data Set, https://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences#.

[6] NLTK, https://www.nltk.org/.

[7] Takashi Igarashi. “Charting Coronavirus Vaccination around the World.” Nikkei Asia, Nikkei, Inc., 2 Dec. 2021, https://vdata.nikkei.com/en/newsgraphics/coronavirus-vaccine-status/.

[8] The Visual and Data Journalism Team. “Covid Vaccines: How Fast Is Progress around the World?” BBC News, BBC, 22 Nov. 2021, https://www.bbc.com/news/world-56237778.

10 Acknowledgements

CU Boulder Honor Code Pledge:

“*On my honor, as a University of Colorado Boulder student, I have neither given nor received unauthorized assistance.*”