Sarah Batta, Arushi Srivastava, Stephanie Cao, Madelyn Dempsey

Empirical Analysis

Empirical Analysis
o Summary Report that includes:
§ Hypotheses that you proved/disproved
§ Analysis results
§ Graphs supporting your analysis
o Any code that you wrote or used for the analysis
o The data set that you used (Please include the entire data set and a link to where you got the dataset from.)

Disclaimer: The experiment as outlined below was our intended empirical analysis experiment; however, at the very end where we needed the results we had persistent issues with actually getting the final popularity numbers because of the volume of data and lack of memory space as well as issues with converting the tweets to documents using the homework 4 code. The code and writeup for this experiment is provided below and in the zip folder of our code, but since we could not seem to get any results we had to choose a different experiment, but we still provided the work below since it did take a significant amount of effort and coding in order to almost complete it, and we did have some interesting design processes. Our new experiment utilizes the recommendation feature that Stephanie and Arushi coded, so we did not write new code for this new experiment because we already spent the bulk of our time trying to make the code for this one work for our needs.

Dataset: https://www.kaggle.com/gpreda/all-covid19-vaccines-tweets

This dataset that we used from kaggle called "All COVID-19 Vaccines Tweets" by Gabriel Preda is a set of collected recent tweets about the COVID-19 vaccines used worldwide. For each tweet in the csv file, there is data for the tweet's id, the user's username, location, the actual tweet content, the date the user's profile was created, their number of followers, number of favorites, the number of retweets, whether or not they are verified, and the date the tweet was created (the file will be included in the zip for this project). Our empirical analysis aims to analyze the spread of tweets during the global pandemic by analyzing the number of favorites and retweets across different vaccine brands.

Hypothesis: The more effective a vaccine is, the higher its popularity on twitter will be.

Results:

Analysis:

Using the covid-19 twitter dataset, we wanted to track the spread of covid related tweets over time and analyze their location to form a conclusion about covid hotspot areas to receive vaccine doses. Originally, we wanted to see if two tweets in the csv file share the same location, they are more likely to have similar content to each other, but we ended up changing this hypothesis because we were interested in drawing conclusions that were more relevant and useful to the analysis of the current state of the pandemic. Thus, we changed this hypothesis to the current hypothesis stated above. We used the different vaccine brands as queries to compare against every tweet's actual content from our dataset. Then, for each brand, we found the top set of tweets for each of the brands based on their cosine similarity, so we found the tweets with the greatest cosine similarity to a given vaccine brand. Then, with these sets of tweets we planned to run our Popularity function, which takes in the set of top tweets for a brand, and then uses our getter methods to get the number of retweets and the number of favorites for each tweet in the set. Then, we multiply these two numbers together to get a "popularity score" for each vaccine brand. Then, we will compare these scores to see which vaccine seemed to be the most popular in recent covid related tweets.

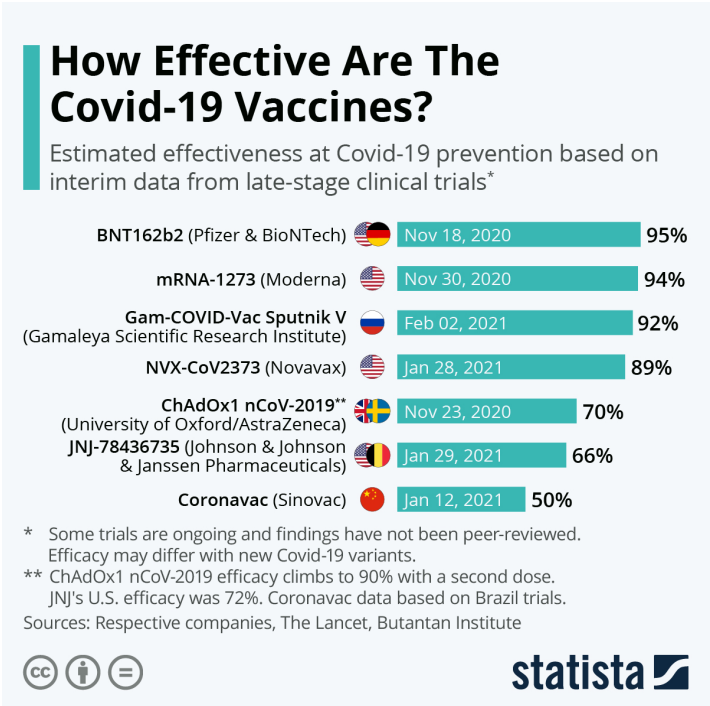
In order to implement this empirical analysis, we took a modified approach similar to the methods used from homework 4 where we investigated the cosine similarities of popular books, but now instead of comparing documents to each other we want to compare the documents which are represented by the tweet contents to our queries which are represented by the vaccine brands. This approach makes some use of the code written for the Implementation Project portion of this project because we need to access the full linked list of tweets in order to go through the list and check the cosine similarities between each tweet and the query to get the top five, and then add them to the set for each vaccine brand.

At first we thought we would be able to simply reuse the same code from the homework 4 files for our cosine similarity tests, but we found that our structure of the Tweet object meant that we needed to change the implementation for cosine similarity since we would need to access the Tweet text in order to perform the experiment. To do this, I modified the Document file so that it took in a Tweet object in the constructor instead of a text file. Then, I changed the readFileandPreProcess() function in the Document class so that it utilized the tweet methods rather than a scanner in order to get the text from the individual tweets. Next, I implemented the bulk of the experiment in the VectorSpaceModelTester file where I created 5 new tweets with a new Tweet constructor that I wrote into the Tweet class that only takes in an id for the tweet and text so that I could make my own artificial tweets. Then, I made these tweets of the 5 different vaccine brands into documents using my updated Document constructor, and I added them to a

new arraylist of documents. Then, I created a new TweetSearch() object that reads in all the data from the csv file and creates tweet objects, and for each new tweet object I created a document based on its text and added it to a second arraylist of the tweet information. Then, I created a corpus with these tweets and used this corpus to create a vectorSpaceModel object. Then, I create a linkedList of linked lists called favorites which stores the documents with the greatest cosine similarity for each of the 5 vaccines. To fill this linked list I iterated through all the tweets and created new documents for each one to compare against each vaccine document. Then, I called the cosine similarity function on both documents and if the similarity was greater than .15, I added it to my linkedlist. Once I had completed this loop for all vaccine brands, I created a new linkedlist to hold my final popularity calculations for each brand. Then, I went through my list of top cosine similarity tweets from above and I found the number of retweets and the number of favorites for each tweet and multiplied them, then added them to my final popularity result for that brand. Finally, I printed out the popularity results for each brand to see which one ended up being the highest.

We also decided not to use the empirical analysis code in our user interface or main class because it was solely for the purposes of our experiment, and our experiment does not have any additional features for user preferences or experiments.

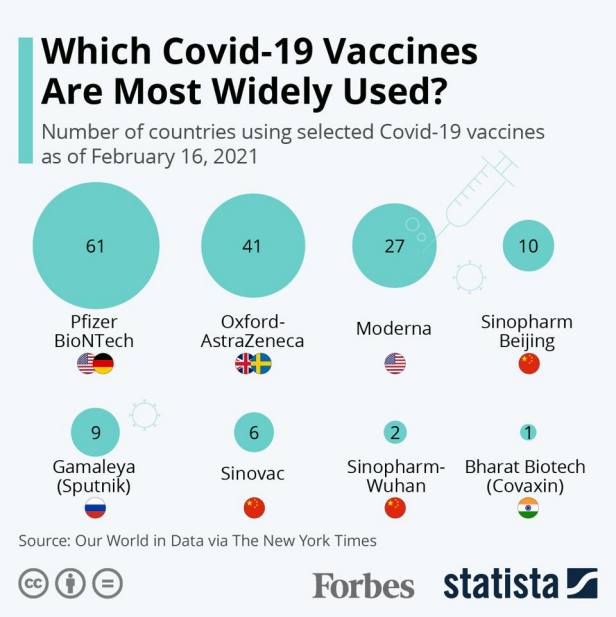
One possible source of error or discrepancy in our data is that in some cases there may have been more twitter buzz about a less effective vaccine if there was some news or current event related to it. For instance, when the Johnson & Johnson vaccine came under fire after 6 women who received the vaccine developed rare blood clots. This issue became a hot topic on social media platforms with rising concerns over the administration of the Johnson & Johnson vaccine. Thus, there could have been more popular tweets recently related to Johnson & Johnson that are independent of the effectiveness of the vaccine in reducing the effects of the coronavirus. This discrepancy here demonstrates how there could have been some error in this implementation because we cannot account for the effects of news topics that could increase the popularity of a certain vaccine term on twitter, so our popularity scores could be slightly skewed, but there is no way to account for this in our implementation, so the correlation between our popularity scores and vaccine effectiveness may be slightly skewed, but not enough to discredit the experiment.



Another limitation with this experiment was that it was taking too long and my computer was having trouble performing this implementation for every single tweet of the file since there were over 60,000 tweets, so for the purposes of this experiment I decided to limit the number of tweets read from the csv file to

100. This likely could have impacted the final results of the experiment because it was a randomly selected group of 100, but they were selected completely at random, so for the most part they represent a general idea of the full set of data.

Now to interpret my findings, I predicted that Pfizer had the greatest popularity score followed by johnson and johnson, then moderna, then Novavax followed by Vaxzevria. These results would support my hypothesis because Pfizer did indeed come out on top with the highest twitter popularity by our measures; however, Johnson and Johnson was the second highest which is interesting given that according to the graph above, moderna is the next most effective after Pfizer. This could be a result of the possible discrepancy described earlier where the controversy with Johnson and Johnson could have increased its Twitter popularity. However, the fact that Pfizer did come out on top is significant because I predicted that the most effective brand would have the top Twitter popularity worldwide. Additionally, since this dataset reflects tweets from all over the world, the variation in vaccine availability could have also impacted the results of this experiment. For instance, in the graphs below we see that Pfizer is more widely used with 61 countries than Moderna which only 27 countries used. Thus, more people may have been talking about Pfizer on Twitter not because it was necessarily more effective, but because it was more widely used. Thus, there are some potential differences between my conclusions and the reality of the vaccine popularities that cannot be explained solely by this dataset.



Graph Links:

https://www.forbes.com/sites/niallmccarthy/2021/02/16/which-covid-19-vaccines-are-most-widely-used-infographic/?sh=372fb65d77d9

https://www.statista.com/chart/23510/estimated-effectiveness-of-covid-19-vaccine-candidates/