**Problem**

Many people and places around the world are still seeing the effects of COVID-19 after over a year of dealing with the virus. There are many ways to find information about COVID-19 such as websites, social media, word of mouth, and news. There is a mix of true and false pieces of information all over the place about this virus and vaccines that will help prevent people from getting it.

The Centers for Disease Control and Prevention (CDC) defines COVID-19 as “a respiratory disease caused by SARS-CoV-2, a new coronavirus discovered in 2019. The virus is thought to spread mainly from person to person through respiratory droplets produced when an infected person coughs, sneezes, or talks. Some people who are infected may not have symptoms. For people who have symptoms, illness can range from mild to severe. Adults 65 years and older and people of any age with underlying medical conditions are at higher risk for severe illness” (CDC, 2021).

COVID-19 vaccines have been a hot topic at the end of 2020 and into 2021. Finding ways to communicate the correct information with people all over the world is tough. There is so much information that it is hard to differentiate what is true or not. One way to identify and communicate with people is through social media websites where companies and individuals post, share, and reply to posts. For instance, companies can use Twitter to Identify what other companies, people, and news organizations are tweeting and retweeting about and target them with specific information to tweet. This could include where to get vaccinations, facts about vaccines, vaccines that are available, and other COVID-19 discussions.

**Hypothesis**

There will be a lot of COVID vaccine information people and companies will tweet and retweet about on Twitter. There will be common words that should appear in tweets such as COVID, vaccines, COVID-19, and other words. There will also be a lot of positive and negative connotations for words associated with COVID-19. There will be more positive words in tweets than there will be negative words. There will be a lot of screen names that tweet about COVID-19 such as companies that want to get vaccine and testing information out to consumers, bots, and individuals.

**Data Used**

The dataset I used is Twitter tweets that are not included in retweets. Twitter only allows a person to download up to 18,000 records at a time every 15 minutes. Without a Twitter subscription, the data only goes back 8 days. There are a lot of fields that Twitter stores information for. The fields I mainly used to review the tweets are screen\_name, text (tweet), source (what device was used to tweet), location, followers\_count, friends\_count, and account\_created\_at.

**Methods**

I used various methods to evaluate the Twitter tweet information. There was text mining completed by downloading through a Twitter API and cleaning the data to remove various web page links, non-characters, hashtags, and common words. I evaluated the word counts for screen\_name, source (device used to tweet), and tweet text.

Another method I used was word networks. Word networks are used to identify words that were used in tweets that appeared in several posts. The below chart shows high common words that appeared in the tweets. For instance, the word “world” is correlated with “health” and “demand” is correlated with “declines”. The thick black line is where there were words with high word counts that appeared in the various tweets.

Diagram

Description automatically generated

Word cloud was also used to evaluate words that were used a lot in the tweets. In the image below, vaccine and covid were used a lot.

Text

Description automatically generated

Lastly, I used natural processing language (NLP) which is defined as “a field of Artificial Intelligence that gives the machines the ability to read, understand and derive meaning from human languages” (Lopenz Yse, 2019). Sentimental analysis is one example of NLP. The method identifies, based on words, whether the data is positive, negative, or neutral. “Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs” (MonkeyLearn, 2021).

In the sentimental chart below, it shows there is a lot of negative connotative words within the cleaned words. There is also a lot of positive words. While reviewing the words behind the counts of each of the sentiments, I noticed there were several variations of certain words. The R library that was used seemed to assign the words correctly.

Chart

Description automatically generated

**Conclusion**

Overall, there are many ways to analyze Twitter tweets. Cleaning the data in Twitter tweets can take time as there can be so many different variations of words. Throughout the last four weeks, it has been interesting to see how the various tweet word utilizations have changed when analyzing the keywords “COVID Vaccine”. After reviewing the sentimental analysis, I found that there were more negative words than positive words. Also, there were various screen names that tweeted about COVID within the last 8 days which included consumers, bots, and individuals.

**Ways to Improve**

There are several ways I can improve utilizing Twitter to analyze tweets and retweets. There could potentially be more data cleaning that could be done such as removing more non-characters and different word variations. Also, I could review the other fields more closely such as creating maps of where the tweets were from. Lastly, I could identify other ways to evaluate the data that utilize machine learning and text mining.

**Appendix**

Questions to answer in Presentation:

* How many Twitter tweets were used in the dataset?
* What is text mining?
* How was each tweet analyzed?
* Were there any similarities between the fields that you used?
* How did you clean the data?
* What words were commonly used?
* How did you identify the keywords you wanted to use to pull the tweets to evaluate?
* What did you find interesting about analyzing a social media dataset?
* What words were the most similar to each other?
* Were the tweets accurately predicting the positive, negative, and neutral words?

**References**

Centers for Disease Control and Prevention. 2021. Coronavirus Disease 2019 (COVID-19). Retrieved from <https://www.cdc.gov/dotw/covid-19/index.html>

Twitter. 2020. Tweets retrieved from [www.twitter.com](http://www.twitter.com)

Lopez Yse, Diego. 2019, January 15. Your Guide to Natural Language Processing (NLP). Retrieved from <https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1>

MonkeyLearn. 2021. Sentiment Analysis: A Definitive Guide. Retrieved from <https://monkeylearn.com/sentiment-analysis/>

Wasser, L and Farmer, C. 2021. Earth Data Analytics Online Certificate. Retrieved from <https://www.earthdatascience.org/courses/earth-analytics/get-data-using-apis/text-mining-twitter-data-intro-r/>

Van den Rul, Celine. 2019, September 20. A Guide to Mining and Analysing Tweets with R. Retrieved from <https://towardsdatascience.com/a-guide-to-mining-and-analysing-tweets-with-r-2f56818fdd16>