Big Data Analytics On People’s perception of the Covid-19 vaccine through Twitter Sentiment analysis

University Name

Course Name

Student Name

# Introduction

Sentiment analysis is a type of natural language processing that classifies the tone of text as positive, negative, or neutral. Through order to track customer feelings about a product or brand in reviews and get a feel for what those customers want in the moment, many companies turn to sentiment analysis of textual data. The primary goal of this research is to use sentiment analysis to gauge public opinion of the COVID-19 vaccination. According to (Jang, et al., 2022), the induction of Covid-19 by different research labs in the world was meant to prevent the spread of covid-19 virus and by preventing the infection the rate of covid19 deaths reduced. However, the induction of people brought doubts. Despite the fact that the vaccine has given people new reason for optimism, it has also sparked a number of anti-vaccination campaigns. As the pandemic continues, I plan to use recent Twitter data to study public sentiment about the vaccine.

Data collection

After topic selection, data collection is the next step. I collected data related to my topic “Peoples perception on Covid-19 vaccine”. Using Twitter’s stream API, we have collected 20,000 tweets using the following search terms ”vaccination AND covidvaccine OR covid AND pro OR antivax OR antivaxxer. Since there was an issue with the tweet rate I have also used an external data set from [Kaggle](https://www.kaggle.com/kaushiksuresh147/covidvaccine-tweets) which uses the term Covidvaccine, this contains data from Dec 2020 to March 2021. After combining the datasets, I have checked for duplicates, the total data size is about 50k tweets.

# Why is Our Big Data in-terms of 5V model?

The term "big data" refers to the huge, often unwieldy amounts of data (both organized and unstructured) that flood modern enterprises on a daily basis. What really matters is what businesses do with the data they collect, not just the data themselves. Insightful, self-assured business choices are possible thanks to the analysis of big data (Skuza & Romanowski, 2015). The primary goals of Big Data are to enhance a company's or system's responsiveness to a vast amount of collected data, boost productivity, and refine knowledge of customer behavior to deliver targeted marketing and develop novel markets.

Accessing and storing massive amounts of data for analytical purposes is not novel. Still, industry analyst Doug Laney defined the now-standard definition of big data as the five V's in the early 2000s (Bhuta, Doshi, Doshi, & Narvekar, 2014). The explained the five V’s of big data as follows;

Volume. Businesses gather information from a wide range of channels, such as monetary transactions, Internet of Things (IoT) gadgets, manufacturing machinery, and even consumer-facing media like movies, photos, and audio recordings. The expense of storing all that data used to be prohibitive, but now it's easier than ever thanks to data lakes, Hadoop, and the cloud.

Velocity. As the number of connected devices increases, organizations will be inundated with data at a rate never seen before. The need to cope with these deluges of data in near-real-time is being driven by RFID tags, sensors, and smart meters.

Variety. Text documents, emails, films, audios, stock ticker data, and financial transactions are just some of the many different sorts of data out there.

Additionally, data flows are unpredictable, changing frequently and fluctuating substantially, adding to the growing speeds and types of data. Although it may be difficult, businesses nowadays need to be aware of social media trends and be able to handle the increased data volume that comes with things like holidays and unexpected events.

Quality of information, or "veracity," is what is meant by this term. The disparate origins of data make it challenging to unify information through aggregation, matching, cleansing, and transformation across systems. Relationships, hierarchies, and numerous data links need to be connected and correlated for businesses to function effectively. In the absence of proper safeguards, their data might easily become unmanageable.

# Research Question

What do people think about the vaccination developed to stop the spread of the Covid-19 virus?

# Research Hypothesis

1. United states location produces the highest number of tweets on covid-19 vaccine perception.
2. Majority of the tweets sent via twitter contains negative opinions about covid-19 vaccine.
3. China location contains the highest number of positive tweets.

# Text Preprocessing

During preprocessing, offensive tweets are removed. Some examples of cleaning up text are changing all capitalization to lowercase, renaming the link to "URL," renaming the user name to "AT USER," getting rid of extra spaces, getting rid of hashtags, and getting rid of slashes and quotes. In my case, the following steps were followed to preprocess our text dataset and make it ready for analysis and modelling.

* Remove punctuation, hashtags, and URLs
* Tokenization is the process of breaking down a statement into its component words.
* Remove stopwords, Lemmatization: the process of reducing a word to its simplest form
* Plot Unigram, bigram, and trigram charts of often used words

## Importing Necessary Libraries

To load the downloaded data set and begin processing it, it is necessary to import many libraries. Our project implementation will mostly make use of the following python data preparation libraries;

Stop words – to handle stop words in our text

Tweepy – to scrape twitter tweets based on different variables.

Word cloud – to visualize mostly used random words

Pandas – loading different file formats, data manipulation, and analysis.

Matplotlib – for 2D plotting of figures

Seaborn - alternative 2D plotting library to matplotlib.

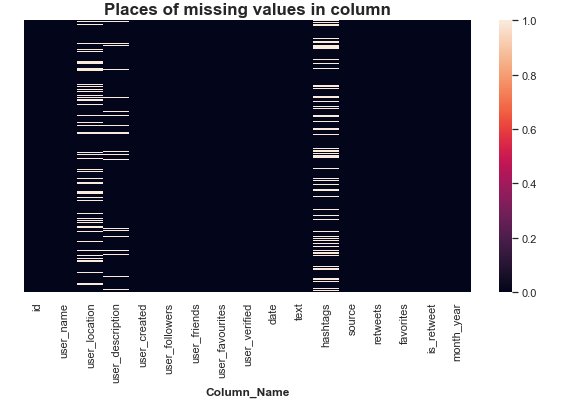
## Import the Datasets

In this step, the dataset must be loaded into the working environment to allow data exploration and preparation. Therefore, we first set the current directory as the working directory, then using panda’s library “the read\_csv()” function, import the CSV dataset files. Keep in mind that the tweets we scraped from twitter, we exported them as CSV file formats. Thus, upon loading Pandas function that is read\_csv() data is loaded in your pc.

## Identifying and Handling Missing Values

It is pivotal to identify and handle the missing values correctly. Failure to work on these results in inaccurate and faulty conclusions and inferences from the data used. Therefore, to get accuracy in our classification techniques employed, we had to find out the missing values and work on them. (Jia & Zhang, 2021). In this project, we handled the missing values by calculating the mean for the numeric data variables like latitudes, years, and temperature and replacing the missing values with the mean calculated. On the other hand, for categorical variables like weather conditions, the missing values were replaced by finding the mode in the whole column and using the mode to replace the missing values. This practice added variance to our dataset; hence it yields better results as compared to dropping the missing values rows, which would have led to data loss (Chu, Ilyas, Krishnan, & Wang, 2016).

Below is a diagram showing missing values per features in our dataset. Most of the dataset features were clean(contains no null values) except for a few like hashtags, user description and user location.



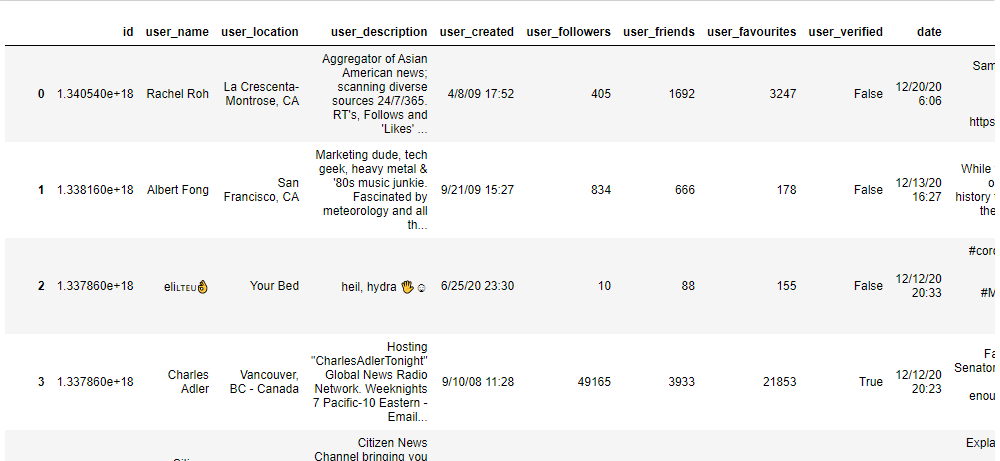
## Splitting the Dataset

In data preparation, data splitting occurs after data cleansing is finished. Every machine learning technique requires dividing the dataset into a training set and a test set. The term "training set" is commonly used to refer to the data used to teach the model. In this case, we are already familiar with the result. Instead, the trained model is put through its paces using data taken from a separate set designated as the test data. Predictions are made using this data set as a test bed for the model. The most common data splits are 70 to 30 and 80 to 20, with the former suggesting that 70% or 80% of the data would be used for training and the latter for testing. The ratio we used was 70/30. It's important to remember, though, that the partitioning procedure always adapts to the specifics of each dataset. Sci-kit lean’s train test split() method is responsible for this.

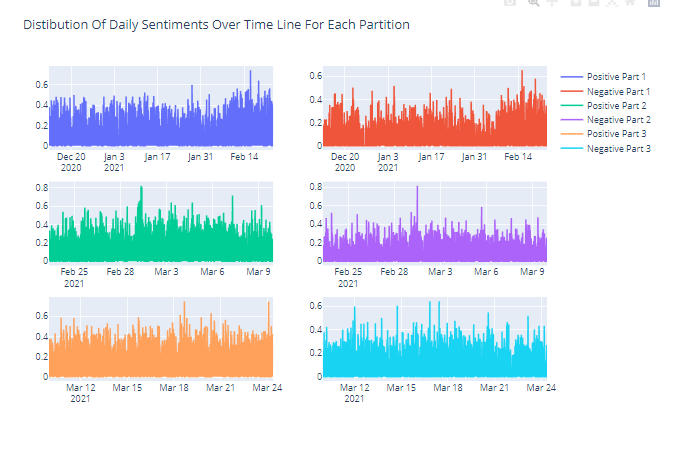
# Data Visualization

In this section, several graphs, plots are presented. The graphs and plots are produced as a result of doing the dataset exploration while finding out to understand more about our data.

Tweet data collected view in a data.frame after loading in the environment.

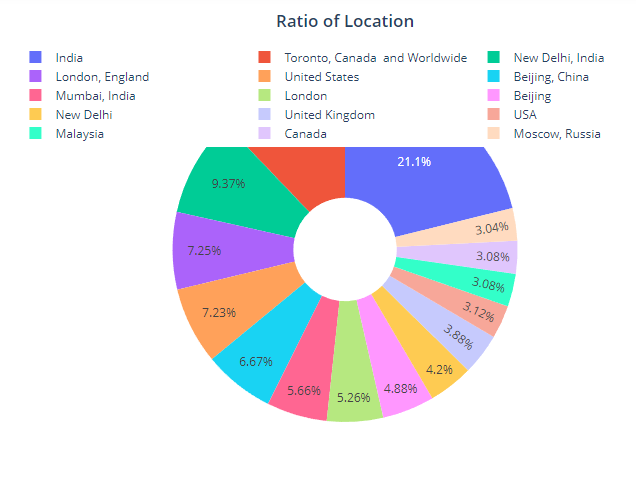


Below is a grid plot which shows the distribution of daily Sentiments over time for each partition(here, we partitioned our dataset into 3parts and find the mean and standard deviation of each partition. The aim of partitioning was to at least try to get access to all data parts to understand its distribution ). From the plot, its clear that there is high density of positive tweets between the month of Dec 2020 to March 2021. We also observed that, there seems to be less density tweets between the month of March and April.



Next, there’s a pie chart diagram below. The pie chart was made to easily show the percentage of tweets made based on location. From the pie chart, we can easily conclude that, people from India had more opinions on covid-19 vaccine as seen by the number of tweets at around 21.1% while People from Moscow, Russia participated less in their contribution about tweeting their opinion on covid-19 vaccine. This is the country with less tweets by 3.04%

The next visual shows the location which made more tweets on covid-19 vaccine. From the pie chart its clear that, India is the leading location with highest number of covid-19 vaccine tweets perception at 21.1%



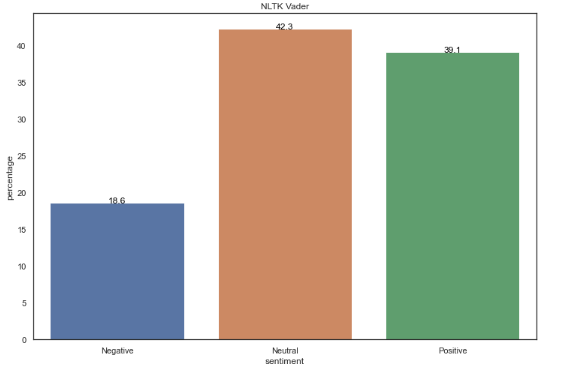
The common words among the most positive and negative tweets are shown by the word cloud visualization below.

# 

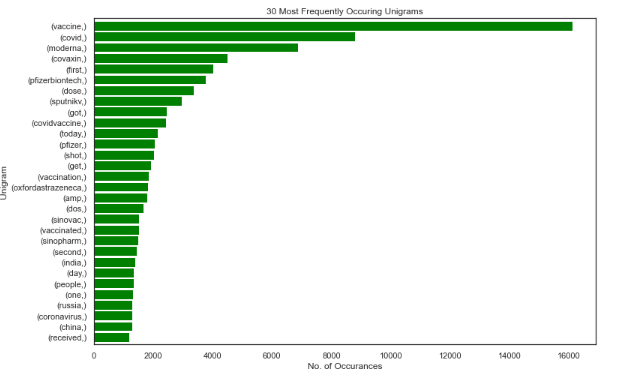
The word cloud below shows the top ten positive and negative tweets



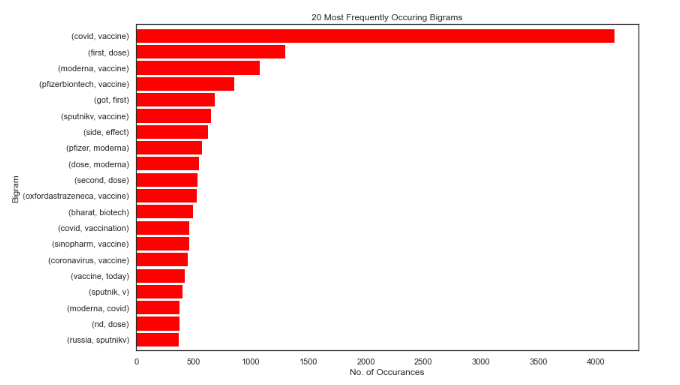
Bar plot showing the percentage of negative, neutral and positive tweets from the people around the world between the year 2020 and 2021.



Top 30 mostly used / occurring one word in users’ tweets. Vaccine takes the lead followed by covid as the most two word used by users while tweeting.



Top 20 mostly used / occurring two word in users’ tweets. Covid and vaccine are the most two word used by users while tweeting.



# Conclusion

In this research paper, we find out people’s opinion on covid-19 vaccine. These opinions were classified as either neutral, negative or positive. In addition, we made several hypotheses about our collected tweets.

The hypothesis states that:

**Null hypothesis 1**: United states location produces the highest number of tweets on covid-19 vaccine perception.

**Alternate hypothesis:** India location produces the highest number of tweets on covid-19 vaccine perception.

**Null hypothesis 2:** Majority of the tweets sent via twitter contains negative opinions about covid-19 vaccine.

**Alternate hypothesis:** Majority of the tweets sent via twitter contains neutral opinions about covid-19 vaccine.

**Null hypothesis 3:** China location contains the highest number of positive tweets.

**Alternate hypothesis:** London location contains the highest number of positive tweets.

In summary, it is evident from our analysis that the dominant sentiment of covid vaccine-related tweets are positive. The tweets tend to be more positive than negative on average as indicated in the visualizations presented to us upon analyzing data. Also, there was a trend of decreasing negative sentiment strength from the 18th of December to the 27th, which was when they started vaccination. I also observed that there are recurring words both in the extremely negative and extremely positive tweets. On the other hand, after data preprocessing, I embarked on the three distinct classification techniques. In this paper, I worked on decision trees, logistic regression and Naïve Bayes classification techniques. From the results found, decision trees outperform both logistic and Naïve Bayes with an accuracy of 99.99%, Naïve Bayes comes the second with 98.57% and lastly logistic regression with 93.07%. Thus, we choose decision tree classification technique as the best technique to work on this dataset.

In future work, I highly recommend on conducting the sentiment analysis on a larger number of tweets. Ensure that you make comparison between text analysis results with other packages e.g., Text blob. Improve the model to design a suitable cut off for positive and negative tweets, PyMC3 is a Python package for Bayesian statistical modeling and Probabilistic Machine Learning which focuses on advanced Markov chain Monte Carlo and variational fitting algorithms.

# References

Bhuta, S., Doshi, A., Doshi, U., & Narvekar, M. (2014). A review of techniques for sentiment analysis of twitter data. . *In 2014 International conference on issues and challenges in intelligent computing techniques* , 583-591.

Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016). Data cleaning: Overview and emerging challenges. . *In Proceedings of the 2016 international conference on management of data*, 2201-2206.

Jang, H., Rempel, E., Roe, I., Adu, P., Carenini, G., & Janjua, N. Z. (2022). Tracking Public Attitudes Toward COVID-19 Vaccination on Tweets in Canada: Using Aspect-Based Sentiment Analysis. . *Journal of Medical Internet Research*, 24.

Jia, B. B., & Zhang, M. L. (2021). *Multi-Dimensional Classification via Decomposed Label Encoding. .* IEEE Transactions on Knowledge and Data Engineering.

Shah, D., Xue, Z. Y., & Aamodt, T. (2021). *Label Encoding for Regression Networks.* In International Conference on Learning Representations.

Skuza, M., & Romanowski, A. (2015). Sentiment analysis of Twitter data within big data distributed environment for stock prediction. *In 2015 Federated Conference on Computer Science and Information Systems* , 1349-1354.

# Appendix

“Sentiment Analysis.ipynb” Jupyter file that contains project code

“Sentiment Analysis.pptx” file contains the project presentation.

“vaccination\_all\_tweets.xlsx” that contains the collected data.