BACKGROUND

COVID 19 disease has been crippling the world since 2019, bringing the world to a standstill, crashing global economies, families stranded and unable to meet during holidays, occasions, students attending schools and universities virtually for the last two years. Besides that, around the world 5 million[1] have died due to this deadly disease and in the USA alone around 804,758 [1] have died due to COVID-19. So far there has been no sign of the pandemic truly ending and we are completely going back to our "normal lives", the only respite we have had so far is the introduction of COVID-19 vaccination. Though introduction of the vaccine has proven itself to be the best way to limit the spread of the supplemented by social distancing and wearing facemask in public spaces, being fully vaccinated doesn't mean you will 100% contract the virus, you might still fall sick due to COVID-19 but it wouldn't be as severe and life threatening if not fully or not vaccinated at all. To say the least the introduction of the vaccine has not been necessarily positive across the world for multiple reasons. This coupled with people's apprehension on how fast the vaccine was developed and released for public usage has not instilled much confidence either. In current day with expansive reach of the internet, there surely has been a lot of mis-information or myths about the vaccine like

"A COVID-19 vaccine can make me sick with COVID-19."

"COVID-19 vaccines contain microchips."

"The natural immunity I get from being sick with COVID-19 is better than the immunity I get from COVID-19 vaccination."

"COVID-19 vaccines authorized for use in the United States shed or release their components."

These being a few them, we are here analysis tweets on twitter to understand people sentiments and emotional reaction to COVID19 vaccine. It's important for governments to understand the potential drivers that affect public's attitudes towards COVID-19 vaccines based on social media, which generates abundant user-based data. We will be focusing our study in USA, which we believe in correlation with stats on fully vaccinated people around all the USA states would be rather helpful for

health professionals and policy makers to understand the sentimental awareness of states that are lacking or not as highly vaccinated as other states.
This project contains four parts for analysis:
(1) sentiment and emotional analysis,
(2) keyword modeling and cloud mapping
(3) predict the sentiment from tweets with machine learning.
(4) Spatial analysis of the tweets - positive and negative.

DATA COLLECTION

To begin with we will be using twitter's developer app to extract data from 1st January 2021 to 14th December 2021 by filtering the location with USA or United States. We research for the most used hashtags in social media pertaining to COVID 19 vaccine and we narrow it to #covidvaccine #Covid19vaccine, and to further fine tune and be precise with the analysis we also extract tweets with #pfizervaccine and #Modernavaccine as these are the two trending vaccine providers in the US.

Once we are done extracting tweets we clean the data by removing punctuations, URLs amongst others, further we homogenise the tweets, location and data for easy handling and consistency, converting all tweets to lowercase, etc. Initially we run the sentimental and emotional analysis on all the 50 states through Tweepy and TextBlob to calculate the polarity and subjectivity, then label the sentiment score depending on the positively, negatively and neutrality of the tweets. Then we perform emotional analysis with NLTK on primary emotion types, such as fear, anger, anticipation, trust, surprise, positive, negative, neutral, sadness, disgust and joy. We then explore the temporal pattern based on the sentimental score each of these rank for each state. Based on the trend of sentimental score as well as the number of geo tweets, we will identify some key points. To investigate their potential drivers, we will model some key words. For these keywords for the duration of our study we will create word cloud mapping in positive or negative sentiment. Finally, this study devises a sentimental analyzer to identify Tweets test sentiment, whether positive or not. Several models, including logistic regression, TesnsorFlow-based model, SVM, LSTM, etc., were built, trained and compared. The one with the highest accuracy will be chosen as the most optimal one.

TEXT MINING AND TOPIC MODELING

Text mining is the process of extracting high-quality information from text. [2] Topic modeling is also a form of text mining which employs statistical machine learning techniques to identify patterns in large amounts of text. It can take your huge collection of documents and group the words into clusters of words, identify topics, by a using process of similarity.[3] In our project we use text mining and topic modeling, to identify the most recurring words from our dataset (all the tweets we have extracted between 1st Jan 2021 to 14th December 2021, and while excluding the most common words sets like 'a' 'the' 'if' etc.,). Based on this we are able to decipher – positive, negative and neutral words and mindsets associated with the vaccine. We further visualise this into 3 different word colour based on this interpretation. To better understand which are the most used words in each month, we have visualised this in the form of a bar graph with a widget which can be used to change the month and the corresponding bar graph with 15 most used words of that month are revisualised.

SENTIMENT and EMOTIONAL ANALYSIS

Let's start off with understanding the terms polarity and subjectivity in sentiment analysis. Polarity refers to the strength of an opinion, it could be either positive or negative. It varies between -1 and +1 and the strength of the polarity depends on various factors and situations under which these tweets are posted. Subjectivity refers to the degree to which a person is personally involved in a topic. What matters the most here are personal connections and individual experiences pertaining to the topic. From our line graph of polarity and subjectively along with total number of tweets, we can without doubt tell that, most tweets are more subjective, which we think is a good indicator as this means that people are basing their opinions towards vaccine depending on their personal experiences and not being sweets away by median content circulating against the vaccine. This is also a good indicator as this would mean, we governments communicating with these populations who might not be very keen on vaccinating themselves will be able to create awareness and educate them and they don't necessarily have to be worried about public opinion being swayed by misinformation on vaccines.

Understanding the emotional analysis, from our emotional tree. 46% of the tweets are expressed positively towards the vaccine, that is almost half the tweets. Following positive emotion, is fear and trust being at 19% and 15% respectively. Our analysis is not very clearly indicating the nature of the trust if its positive or trust against the vaccine and hence we are not able to fully comment on it. 19% fear the vaccine is something of concern and would be helpful to have some educational and awareness programs to help this segment. From our Emotional line graph, we can tell the reaction had been very positive overall till June 2021 and October 2021, the overall emotion of the tweets during the period seems to be very confusing as we don't see any emotion being very prevalent.

For plotting our sentimental analysis and Global Moran's I, we used cartro and Pysal packages respectively.

MACHINE LEARNING

Since the input of machine learning model should be numeric, we start off by transforming the sentiment type (y-variable) and tweet text (x_variable) into numeric types. We set 'negative' to 0, neutral to '1' while positive to '5'. Then for tweet text, CountVectorizer is applied to transform the text to array format. By comparing the accuracy score as well as the classification report, random forest is tested to be the most optimal model to predict the positive sentiment of tweets. To further evaluate its performance, a random sentence that contains obvious positive sentiment words is chosen, and the results shows that its positive probability is 86.88%, while 13.11% for negative. We believe the accuracy of model can be further enhanced in the future.

SPATIAL ANALYSIS

Finally we conducted spatial analysis of positive and negative emotions by calculating their spatial autocorrelations with Pysal for tweets across US. Spatial weight matrix is first built by finding the 5 nearest neighbours (here k=5).

- Local Moran's I
  Moran's I is a correlation coefficient that measures the overall spatial autocorrelation of your data set. In other words, it measures how one object is similar to others surrounding it. If objects are attracted (or repelled) by each other, it means that the observations are not independent. This violates a basic assumption of statistics — independence of data. In other words, the presence of autocorrelation renders most statistical tests invalid, so its important to test for it. Moran's I is one way to test for autocorrelation.
  Spatial autocorrelation is multi-directional and multi-dimensional, making it useful for finding patterns in complicated data sets. It is similar to correlation coefficients, it has a value from -1 to 1. However, while other coefficients measure perfect correlation to no correlation, Moran's is slightly different (due to the more complex, spatial calculations):
    - -1 is perfect clustering of dissimilar values (you can also think of this as perfect dispersion).
    - 0 is no autocorrelation (perfect randomness.)
    - +1 indicates perfect clustering of similar values (it's the opposite of dispersion).

  To plot the Local Moran's Maps in Spatial Analysis of Positive and Negative tweets, we use the Pysal package. Both positive and negative tweets behave strong spatial autocorrelation in the east coast, while in the west this relationship is week.

- the Global Moran’s I

  The Spatial Autocorrelation (Global Moran's I) tool measures spatial autocorrelation based on both feature locations and feature values simultaneously. Given a set of features and an associated attribute, it evaluates whether the pattern expressed is clustered, dispersed, or random. The tool calculates the Moran's I Index value and both a a z-score and p-value to evaluate the significance of that Index. P-values are numerical approximations of the area under the curve for a known distribution, limited by the test statistic.[4]

The Global Moran’s I for both positive and negative tweets are around 0.2, with P-value smaller than 0.05. This suggests that the spatial autocorrelation of both positive and negative tweets is week at global scale.

LIMITATIONS

The biggest limitations of our study is the lack of inclusion of external factors, like calendar events, seasonal changes etc., we understand if they have any role to play in changing the sentiments and emotions of the tweet patterns. That limitation can to a certain extent is reflected in our lack of understand of the bar graph representing sentimental classification of each month, where the sentimental value of the tweets is the highest during the month of December 2021. And the second limitation of the study, in hindsight we didn't not realise that the emotion – Trust could be either ways, either they trust the vaccine or they don’t trust the vaccine, unlike all the other emotions would be generalised.

CONCLUSION

Overall, from our sentimental analysis we can tell that the US's attitude towards the vaccine is not very clearly leaning towards a positive or negative mindset towards the vaccine. There are states that very strongly incline towards a positive attitude, there are definite states that are still in the border line and leaning strongly towards negative sentiments towards the vaccine, leaving room for potentially significant populations not coming forward to being fully vaccinated. This is something that needs to be of at most priority. Currently we are going through our 4-5 variants which keep getting stronger as the new one comes. Currently with Omnicore WHO is suggests booster shot which requires prior 2 shots of vaccination. Having given our analysis we would like to say that we recognise that our analysis has 2 strong limitations. 1. Being that we dint analyse external factors like political environment in the country, seasons, travel restrictions being few important ones, which we think could have affected the emotion of the tweets. Secondly, in retrospect we dint not realise that the emotion – Trust could be either ways, either they trust the vaccine or they don’t trust the vaccine, unlike all the other emotions would be generalised. Finally, for reference of the public we have uploaded all our code in the readme file in our github repository, so please feel free to use it for your reference.