Analyzing Impact of Covid19 and Vaccine

Akram Shaik

Abstract:

Text is the most common way to communicate on any platform. Among many Twitter is one of the top social media platforms where users express their views or share information in the form of text. There are ~500M tweets sent per day and this creates a large corpus to be analyzed. In this research, a dataset of ~10M tweets related to Covid19 over the period of 8 months is used to analyze sentiments of people and their concerns related to Covid Vaccine. The dataset is separated per month to extract geolocation or coordinates from the location of the tweets and analyze sentiments before and after the creation of vaccine and visualize the results on the world map. On analyzing the dataset, it was found that there was a positive change in sentiments of people after the vaccine was developed. Also, most of the tweets specially in United States were from major cities. Concerns of people over the vaccine was mostly regarding the social distancing, safety, government, number of cases, origin of virus and helping people in the pandemic also wearing mask was one of the top discussed or used words in the tweets. In general, most of the tweets had neutral sentiments but there were a greater number of negative tweets than positive.

Introduction:

The popularity of Social Media was not good in the initial years until Facebook was created, where people can share images, videos or messages of their important events or simply daily routine [3]. In 2010, the number of social media users were ~0.97 billion and today in 2021 it has increased to more than ~3.09 billion. Being connected to the people you know or finding new people to interact is one of the many reasons for this growth. This was led by advertisements on these media platforms, it was also the growth where people using the platforms were able to earn from it as influencers or with brand advertisements. Different platforms are created for specific type of communications like YouTube was share videos of different lengths, Instagram was created to share images and Twitter for messages with a limited length as tweets. This led to increase in the amount of data and created opportunity for analyzing it to understand any pattern to help improve marketing strategies in the fields like fashion, health and more. Of all the platforms Facebook has the highest number of users followed by YouTube and Instagram [2]. Twitter is ranked as the 6'th most popular platform. Twitter is very popular for political discussions and it promoted the hashtags feature. Being a limited type of media communication platform, its popularity couldn't reach to a large number of population, but this didn't stop it from becoming one of the top microblogging site and great marketing platform.

Twitter is not just a platform to share one's personal life events, but it has cemented itself as the platform for asking questions to authorities of different levels, branding, as a newsroom, expressing opinions in limited words. Gifs are popularly shared with the text to express different emotions more easily. It also started a revolution of hashtag ('#') where tweets can be shared on any specific subject [4]. This revolution caused any subject or topic to go viral and reach different parts of world where people agreeing or disagreeing on it can share their opinion. This led many complaints or questions to be noticed and solved by the respective authorities. It has been more than 10 years from the time platform was created and it has more than 330 million active users where more than 500 million tweets are posted per day [1]. The men and women user distribution is almost the same unlike Instagram and Pinterest where they have a lot more number of female users than male. A tweet can be a video, photo or links, a text tweet has a limit of 280 characters. User can not only follow another user but also any hashtag to be updated on its topic.
Users can create a thread of their own tweet to convey any larger message and can retweet any one's tweet as a reply or an opinion, but also has an option of directly message. Having said the positives, one of the

many problems with such platform is that the rise of Fake News where links or statements made on the platform can be blindly accepted, followed, or shared by the other users. This creates a large area of concern to be solved not just on twitter but all social media platforms.

Covid-19 has caused the global pandemic disrupting the continuity of most of the events in everyone's day to day life and this change was brought because of fast spreading of virus causing large number of deaths all over the world. Due to which most of the countries closed their borders and imposed lockdowns in major affected areas. All workplaces, educational institutions and other places were closed except the few important stores. This sudden change in everyone's daily life for months was affecting people's mental state and everyone wanted a social interaction and there was no better place to communicate with people without meeting personally than social media. This caused people being idle and use internet to learn about the world and share their status. All Health organizations including World Health Organization were using social media to spread awareness on how on could possibly avoid themselves from getting in contact with virus.
Many educational videos, links, text messages were shared to help people understand the problem and learn about the virus. Twitter before the pandemic in 2018 had ~320 million tweets per day but after the pandemic it got increased to ~500 million tweets per day [5]. Also, majority of the tweets during the pandemic were of shorter length. With this large amount of data researchers of different level had the opportunity to extract patterns from the data collected and analyze it.

A sudden large amount of information was available on twitter where people have expressed their emotions, status, news, questions, and many more important things mostly in the form of text. This project has attempted to make an analysis on tweets related to hashtag Covid-19 from all around the world to analyze people's sentiments and concerns over its vaccine, this was step closer to analyze the purity or validation of any information shared in the form of news. It will help in understanding the effect of virus at the same time in different parts of the world.

Here these tweets extracted just on hashtag Covid-19 over 8 months of time period have not been used before to analyze and answer few questions such as:

1) What are the sentiments of people tweeting from every part of the world? Do we see any difference in it over the time?
2) What is people's concern over the vaccine being developed for the virus?
3) What are people's sentiments on vaccine?

To answer these questions, it was not feasible to work on 10 million tweets on a normal laptop.
Due to which sample of 80K tweets in English language were extracted separately from each month. To analyze difference of sentiments in different locations, it was important to first extract the coordinates of the location from where the tweet was sent. This analysis can help understand people's mindset and their emotions, this way it can help the required officials to take necessary steps to avoid any chaos for any such situations in future and make people aware of steps on handling any such situations.

The next chapters include Related Work, Data, Methodology, Results with discussion, and Future Work.

Related Work:

Twitter has millions of users every day sharing their views, attitudes and opinions [16]. Sidi Yang and Haiyi Zhang used a probabilistic Latent Dirichlet Allocation (LDA) for topic modelling to analyze and find popular topics from large set of tweets. In this paper they used 2 approaches LDA and sentiment analysis by examining tweets from twitter in English. Few limitations of this paper were like not using any emojis, videos and links in their analysis and because of low computing power small number of tweets were used for analysis.

A research done by Bing Xiang and Liang Zhou was focused on improving the approaches of sentiment analysis using mixture model and a semi supervised training framework [17]. LDA was used to generate topics and in their experiment they were able to improve the sentiments of tweets using their model over sem-eval 2013 models and the conclusion was made using average F scores. It was possible because of their state of art mixture model.

To resolve text locations from tweets into a correct physical location an experiment was done by Wei Zhang and Judith Gelernter [18]. In this paper they discussed about the complexity of extracting location from a tweet when it is not indented as length of tweets are limited also many locations in the world are referred by similar or same names. They also brought an interesting point when doing geocoding the possibility of different source conflicts in results. They used a supervised machine learning model that uses different fields of any twitter message and few world gazetteer features to create a model which will select the right field of text from tweet for extracting geolocation. They were able to provide better results than the state of art models by evaluating their model using F1 score.

From past few years as the popularity of twitter increased, this platform is being used for observing human behaviors specially in disastrous events [19]. Lie Zou, Nina S. M. Lam and few other researchers have worked on this experiment. They made a point on how these analysis could be huge failures when there is uneven use of social media in different regions. Their research was focused on 2017's Hurricane Harvey in Texas and Louisiana and not all states of the country where they tried to answer if social-geographical disparities existed in twitter during that emergency period. They used different data mining techniques and regression analysis to find that communities which were less effected by hurricane was seen having most of the twitter usage from these states than other communities.

As web-based data of patient opinions and experiences of health care is available on twitter Kara C Sewalk, Gaurav Tuli, Yulin Hswen and few other researches did a study on it to provide sentiments of patient experience across all states of United States over a 4 year period [20]. Out of 27.1 million tweets of patient experience only 31% of those tweets were identified using a geolocation classifier and they produced some interesting analysis like 27.83% had positive experience, 36.22 % had neutral and 35.95% had negative patient experience. Also most of the tweets from metropolitan areas were negative.

As people were contributing their opinions on services, products and events on twitter and other social media platforms [21]. Liqiao Zhang, Hui Yuan and Raymond Y. K. Lau did a research to understand consumer intelligence to bring change in business functions and marketing strategies by building a novel framework on top of Apache Spark. In their analysis they found out that their gibbs sampled algorithms outperforms existing algorithms in predicting opinions of consumers.

Tweets are used to express opinions on a wide range of topics [12]. This project by Lei Huang, Alec Go, and Richa Bhayani's goal is to develop an algorithm that can correctly identify tweets as positive or negative in relation to a query term. When it comes to classifying sentiment in tweets, machine learning

techniques work admirably well. However, in order to further enhance the accuracy we can use Semantics, Part of Speech (POS) tagger, Domain-specific tweets etc.

The paper by Aliza Sarlan, Chayanit Nadam, and Shuib Basri describes the development of a sentiment analysis that extracts a large number of tweets [13]. This project makes use of prototyping. The results divide customers' opinions into positive and negative categories, which are displayed in a pie chart and html tab, due to Django's limitations in terms of running on a Linux server or LAMP, this prototyping approach will have to be used in the future.

There are a multitude of tools with various outputs that can be used to quickly analyze data, which complicates the method [14]. This paper by Andry Alamsyah, Farhan Renaldi and few more researchers uses Nave Bayes Sentiment Analysis based on time-series, specifically on a regular basis, and topic modeling based on Latent Dirichlet Allocation to assess the sentiment of the subject as well as the model of the topics addressed (LDA). Real-time data has become a critical component in assessing consumer sentiment. The paper uses Uber as a case study because it is one of the most common modes of transportation in most parts of the world.

The study by Dr. Rajesh Prabhakar and Dr. Krishna Prasad focuses on the information flow on Twitter during the Corona virus outbreak [15]. The #coronavirus tweets were investigated using sentiment analysis and topic modeling with Latent Dirichlet Allocation post-processing. The research verifies prevalence of negative sentiments like fear as well as positive sentiments like trust. Governments and healthcare authorities and organizations have successfully used Twitter to disseminate accurate and reliable information.

Data:

The data was collected by Ashiqur Rahman from twitter in a json file and converted it into a csv file. All the tweets were collected on hashtag (#) Covid-19, it was extracted for a project to identify fake news on social media. The idea was to extract information spreading on social media during the pandemic and analyze what miss-information is. But this project was specifically on just tweets to understand more about data and make some interesting analysis on it and there are many hashtag's people use to talk about same topic but '#Covid-19' is the most popular one with more tweets. The tweets are extracted from the Jan 2020 to August 2020. Following table shows some more stats about the tweets.

Total No. of tweets	No. of random tweets from each month	No. of tweets after extracting geolocation
10 Million (8 months)	80,000 per month	60,000 per month

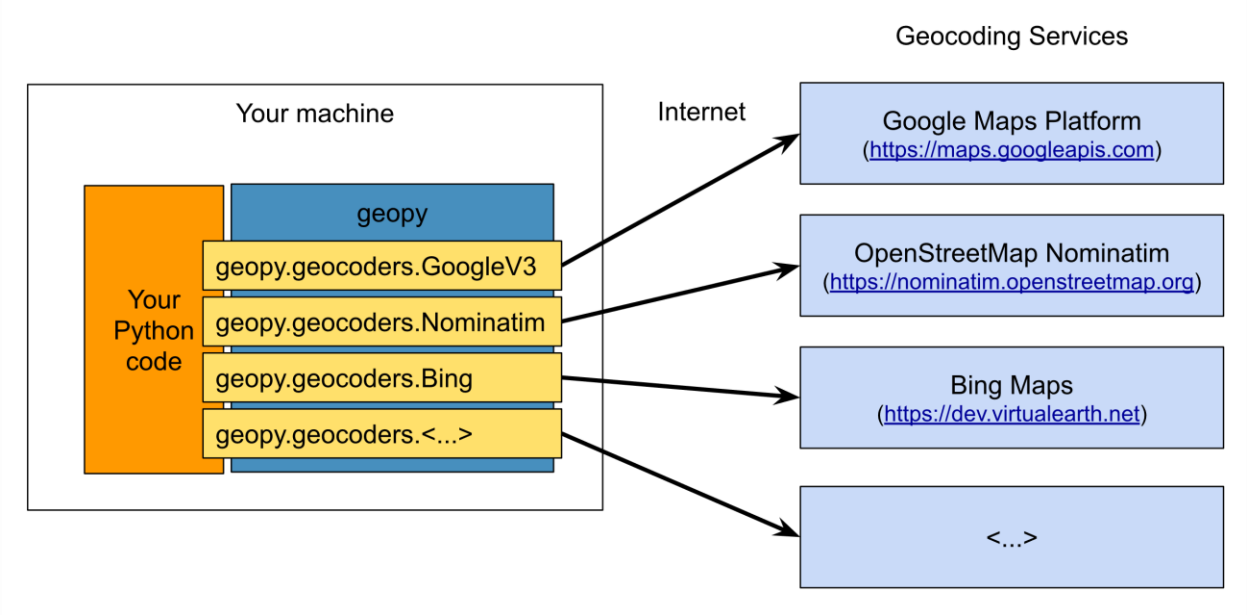
The main problem with data was, it did not have any coordinates of the places where tweets were sent from. The project's aim is to display the analysis on the map, coordinates were required and Geopy library was used to extract them from the locations users provided. This took most of the time for coordinates to be extracted. Eventually random 80k tweets were selected to simplify this process and location coordinates were extracted. But not many locations provided were enough to extract coordinates and number of tweets on average from 8 months is mentioned in the table above.

Methodology:

All the 10 million tweets were extracted simply on just a hashtag of Covid-19 because of which there were tweets which were not in English language, firstly they were filtered by the language. After that tweets were sorted by date and separated for each month. After which many preprocessing steps were applied then Geopy library is used to extract the coordinates of locations. After we get coordinates data is merged from every month into a single data set. Sentiment Analysis using dash library is applied and predicted the sentiments for every tweet. Data is sorted by data to analyze the difference in sentiments over 8 months of time. Later Topic Modelling was applied using LDA to understand what people were talking about. For the final part tweets were filtered on all English language tweets and tweets with coordinated separately on few keywords which identify themselves as Vaccine or any other word related to vaccine. Sentiment analysis was applied on these tweets to analyze how people feel about vaccine and use LDA to analyze what people's concerns are. Following are the steps:

1. Filter tweets on English language and separate dataset for each month.
2. Coordinates were extracted using Geopy on all or randomly selected tweets.
3. Apply Sentiment Analysis and LDA on those tweets to analyze difference in sentiments over the time and understand topics people are talking about, here tf-idf is also used.
4. Similar steps are applied to all tweets in English language irrespective of coordinates to find any general pattern over the keywords related to vaccine.

Extraction of coordinates using Geopy library which is not a geocoding web service but is just a library to be used in such services for locating coordinates of addresses or places with the help of third-party geocoders like Google Maps API's, OpenStreetMap Nominatim API's and more [6].



Since most of the tweets are not geolocated but it had large role to make an analysis in this project and therefore this library was used to convert any address like (Ex: "175 5th Avenue NYC" is converted into geocoordinates like (40.7410861, -73.9896297241625) as latitude and longitude and this process can also be reversed where coordinates given it return the address or place name [11].

All Geocoding services API's have limits on how many request can be sent per day from a particular IP address. Most of these are fast but paid services, so Nominatim API is used in this project as it is free but allows low request limits. The dataset has millions of records and sending these many requests will return errors. To handle these many requests there is class defined in this library as rate limiter which adds delays on geocoding requests and can also retry failed request. Extracting coordinates was the most time consuming part of the project which took me almost 2 weeks of time just to extract them for ~640K sampled tweets. This process was initially applied on whole dataset but since the dataset was large it was taking a lot of time and for extracting coordinates for whole dataset I waited for more than 3 weeks but this caused my browsers to crash and corrupted my file because of which I lost all the data I collected. Having just a 12GB RAM on my laptop did not help in making a better analysis but this was the reason behind sampling the dataset and extracting coordinates just for them.

After extracting coordinates, some filtering of tweets was required since not all tweets had valid or any location mentioned and API retries for these entries for at least 3 times and if it couldn't extract it then it returns a NaN or Null value. Final dataset after removing records with Null values in coordinates the dataset was reduced to ~480k tweets.
Finally, sentiment analysis was applied using vaderSentiment library of nltk package [7]. Vader is abbreviated as 'Valence Aware Dictionary and sentiment Reasoner', it was specially created for analyzing sentiments on social media and it is lexicon and rule-based tool [8]. Now, I had the sentiment polarity scores for all tweets and they were later separated into individual columns for better analysis and understanding. To get a general view or understanding WordCloud was used to visualize the words in those tweets. How WordCloud works is it first creates a bag of words and applies inverse document frequency on to have some linearity in results and visualizes most common words with different sizes and color based on the frequency of words.

Color names were assigned to all tweets where Red was represented as Negative sentiment, Green as Positive sentiment and Blue as Neutral sentiment. Now using the scatter_mapbox module of plotly library coordinates i.e., latitude and longitude were used to plot all the tweets on the world map [10].

On the same dataset Topic Modelling was applied using LatentDirichletAllocation (LDA) from skLearn library. To apply LDA many preprocessing steps were applied to clean the texts and TF-IDF vectorizer was also applied from skLearn library. Number of topics was selected as 10 and WordCloud was applied on each topic for visualizing the words.

Later, original dataset was filtered for tweets having words similar to Vaccine were extracted to analyze sentiments of people on the Covid-19 Vaccine. This filtered had ~275K records and same Vader sentiment analysis and LDA topic modelling was applied on it.

Results:

Because of the request limits on extracting coordinates for tweets location, the filtered dataset was separated for each month and requests were made separately.

	created_at	id	lang	text	user_location	city
0	2020-01-23 22:36:49	1220475499366862849	en	RT @WHO: BREAKING: "I am not declaring a publi...	Waikiki	Honolulu
1	2020-01-26 19:35:19	1221516989115531264	en	Thinking of those affected by the Wuhan/corona...	New York	New York
2	2020-01-24 01:56:26	1220525734839701505	en	RT @cnni: The first person diagnosed with the ...	Cloud N.09	Limoges
3	2020-01-26 06:39:18	1221321697263001603	en	RT @Unkle_K: This is how the coronavirus started	Maryland, USA	Maryland
4	2020-01-27 18:43:30	1221866335979941889	en	Coronavirus, Yesterday's tragedies all those f...	NYC	New York

Fig 1: January Tweets Before Extracting Coordinates

	created_at	id	lang	text	user_location	city	country	latitude	longitude
0	2020-01-23 22:36:49	1220475499366862849	en	RT @WHO: BREAKING: "I am not declaring a publi...	Waikiki	Honolulu	United States of America	21.304547	-157.855676
1	2020-01-26 19:35:19	1221516989115531264	en	Thinking of those affected by the Wuhan/corona...	New York	New York	United States of America	40.712728	-74.006015
2	2020-01-24 01:56:26	1220525734839701505	en	RT @cnni: The first person diagnosed with the ...	Cloud N.09	Limoges	France	45.835424	1.264485
3	2020-01-26 06:39:18	1221321697263001603	en	RT @Unkle_K: This is how the coronavirus started	Maryland, USA	Maryland	United States of America	39.516223	-76.938207
4	2020-01-27 18:43:30	1221866335979941889	en	Coronavirus, Yesterday's tragedies all those f...	NYC	New York	United States of America	40.712728	-74.006015

Fig 2: January Tweets After Extracting Coordinates

	created_at	id	lang	text	user_location	city
0	2020-07-16 07:02:18	1283658200361861120	en	Covid-19 App In Xamarin Forms https://t.co/FfC...	RJ - City of Love	London
1	2020-07-20 23:03:56	1285349754482786305	en	People noticing each other dying prompts a res...	Sydney, New South Wales	Sydney
2	2020-07-16 14:43:19	1283774219365625856	en	RT @andrewbostom: (Today's) Evidence The Atlan...	Land of Oz	Greifswald
3	2020-07-26 22:27:19	1287514866748727299	en	RT @GovKaduna: Covid-19 Update, 26 July 2020: ...	Kaduna	Kaduna
4	2020-07-15 00:37:55	1283199079258955778	en	@IWashington @Acosta RT PLEASE \nHeres the SCI...	Chicago,Illinois	Chicago

Fig 3: July Tweets Before Extracting Coordinates

	created_at	id	lang	text	user_location	city	country	latitude	longitude
0	2020-07-16 07:02:18	1283658200361861120	en	Covid-19 App In Xamarin Forms https://t.co/FfC...	RJ - City of Love	London	United Kingdom	51.507322	-0.127647
1	2020-07-20 23:03:56	1285349754482786305	en	People noticing each other dying prompts a res...	Sydney, New South Wales	Sydney	Australia	-33.854816	151.216454
2	2020-07-16 14:43:19	1283774219365625856	en	RT @andrewbostom: (Today's) Evidence The Atlan...	Land of Oz	Greifswald	Deutschland	54.095791	13.381524
3	2020-07-26 22:27:19	1287514866748727299	en	RT @GovKaduna: Covid-19 Update, 26 July 2020: ...	Kaduna	Kaduna	Nigeria	10.382532	7.853323
4	2020-07-15 00:37:55	1283199079258955778	en	@IWashington @Acosta RT PLEASE \nHeres the SCI...	Chicago,Illinois	Chicago	United States of America	41.875562	-87.624421

Fig 4: July Tweets After Extracting Coordinates

The above images are screenshots of sample data of tweets of January and July month, similarly data was extracted for all 8 months from January to August. In fig 2 and 4 we can see based on user location coordinates were extracted. When applied sentiment analysis using Vader the resulted column had dictionary of sentiments with their polarity scores. Below image shows the sample data after extracting the sentiments.

	created_at	id	lang	text	user_location	city	country	latitude	longitude	Tweets	sentiments
0	2020-01-23 22:36:49	1220475499366862849	en	RT @WHO: BREAKING: "I am not declaring a publi...	Waikiki	Honolulu	United States of America	21.304547	-157.855676	[breaking, declaring, public, health, emergenc...	{'neg': 0.245, 'neu': 0.755, 'pos': 0.0, 'comp...
1	2020-01-26 19:35:19	1221516989115531264	en	Thinking of those affected by the Wuhan/corona...	New York	New York	United States of America	40.712728	-74.006015	[thinking, affected, virus]	{'neg': 0.444, 'neu': 0.556, 'pos': 0.0, 'comp...
2	2020-01-24 01:56:26	1220525734839701505	en	RT @cnni: The first person diagnosed with the ...	Cloud N.09	Limoges	France	45.835424	1.264485	[first, person, diagnosed, wuhan, treated, med...	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
3	2020-01-26 06:39:18	1221321697263001603	en	RT @Unkle_K: This is how the coronavirus started	Maryland, USA	Maryland	United States of America	39.516223	-76.938207	[started]	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
4	2020-01-27 18:43:30	1221866335979941889	en	Coronavirus, Yesterday's tragedies all those f...	NYC	New York	United States of America	40.712728	-74.006015	[yesterdays, tragedies, family, members, lost,...	{'neg': 0.366, 'neu': 0.634, 'pos': 0.0, 'comp...

Fig 5: DataFrame After Extracting Sentiments

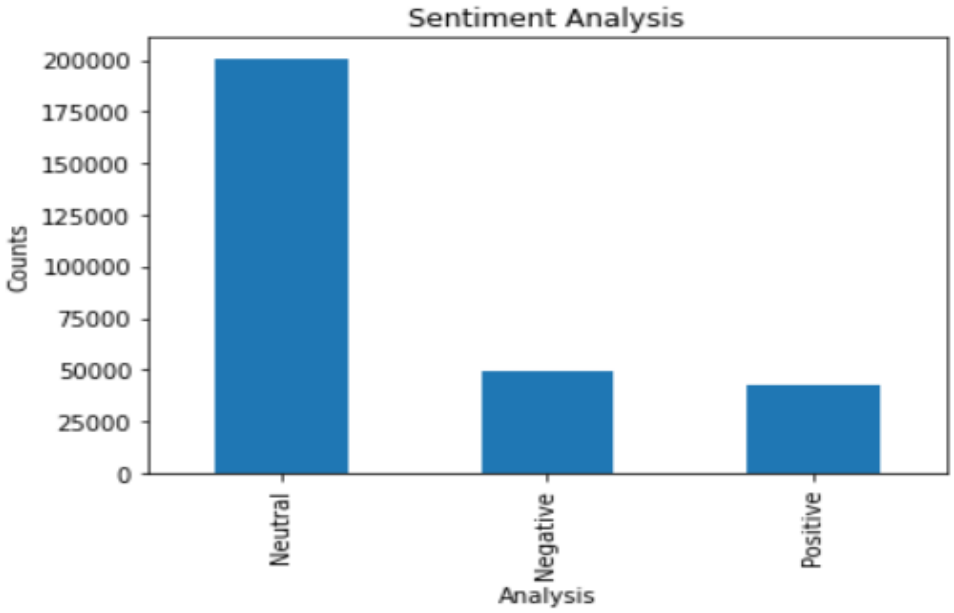


Fig 6: Sentiments Bar Plot

The Fig 6 bar plot shows the distribution of sentiments where majority are neutral, but we also notice the number of negative tweets is more than positive. There is not a big difference but of few thousands. Below Fig 7 bar plot shows distribution of sentiments per month.

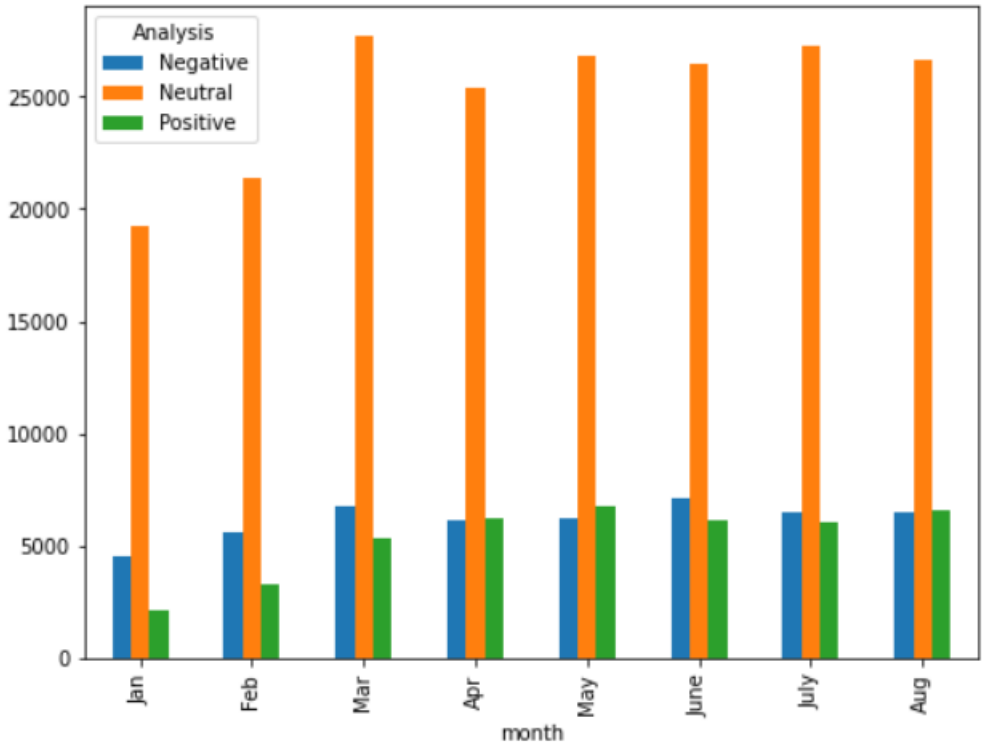


Fig 7: Sentiments Distribution Bar Plot for Each Month

From Fig 7 it can also be seen how number of positive tweets have increased and this is a big difference since the time vaccine testing was announced was during the month of May. But these results could change based on the random sample tweets taken from the data.

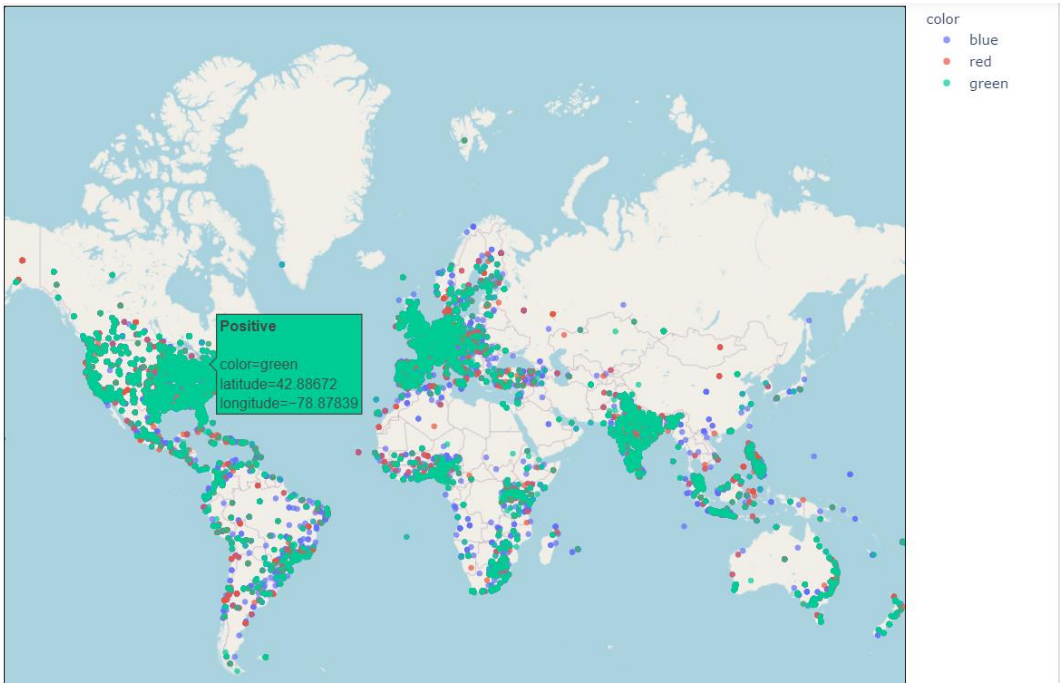


Fig 8: Scatter Plot of All Sentiments on World Map

An interactive scatter plot visualization was plotted using Plotly library, here different sentiments were represented in different colors and each point on hover shows more detail about the particular tweet. On the right there are 3 small clickable tabs where on selecting individual tab we will be able to see tweets of that tab or type only. It also has an option to focus in and out of the map and this helps a lot in understanding the pattern in any specific country or any region. In US, the major cities like New York, Chicago, California, Washington had most of the tweets from. We can also set opacity in visualizing the overlapping sentiments on map and this was the most challenging part of the project to find the right tool for it. After testing on many tools like Folium and Bokeh, I was not able to visualize them but the plotly tool also had some problems where even after setting the opacity on overlapping points of sentiments on the map. The result were not very clear to interpret. But having the option to select any individual sentiments it was easy to understand the distribution of sentiments. Since this visualization was made on all tweets of 8 months. I split the dataset into two equal parts based on dates, first dataset had tweets from month January to April and second dataset had tweets from months May to August.

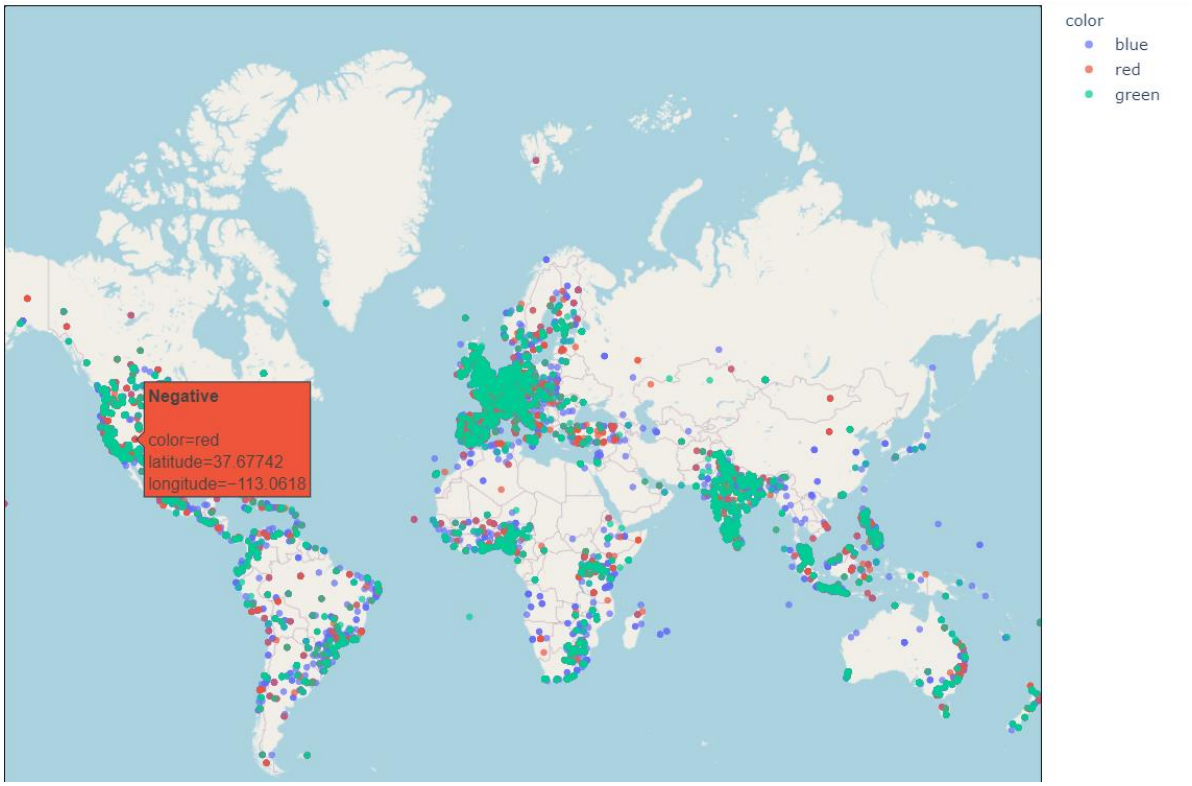


Fig 9: Sentiments of Tweets From Months January to April

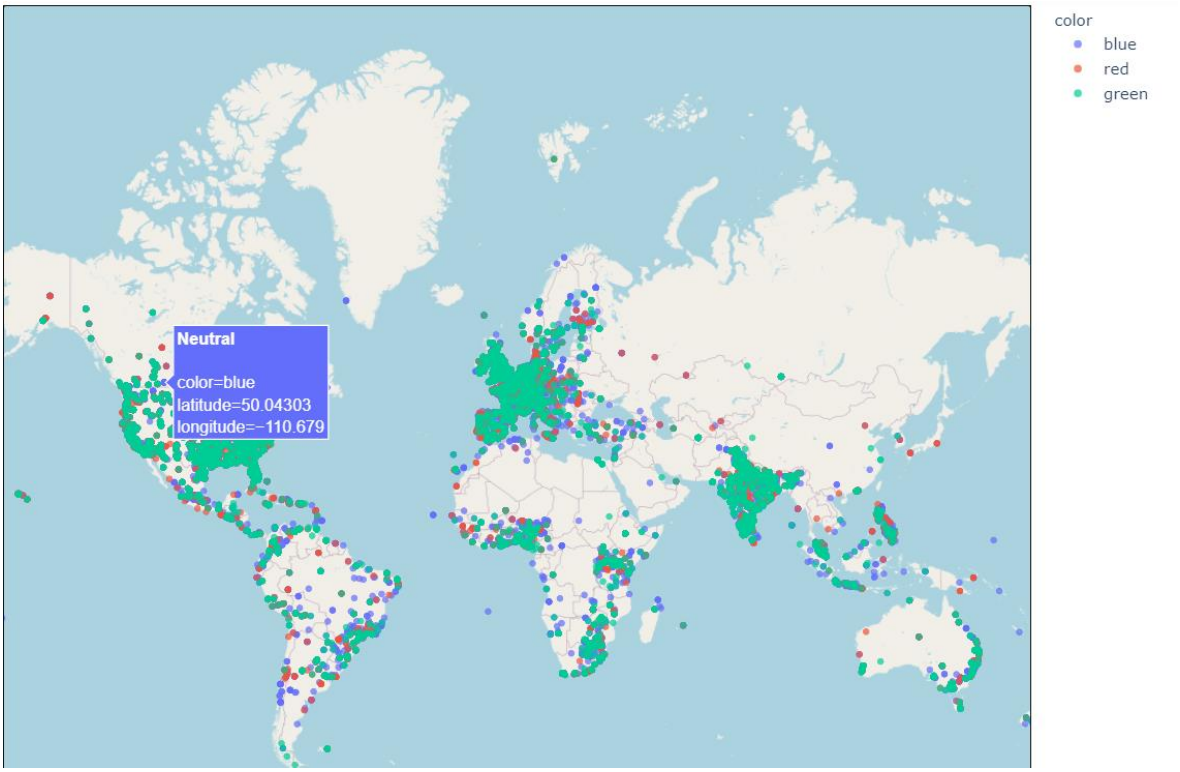


Fig 10: Sentiments of Tweets From Month May to August

On focusing over Fig 9 and 10 visualizations it was noticed number of positive tweets have increased over the time. There are many possible reasons for this change like testing of vaccine being was first brought into news in May, people were having hope and giving support to the families affected through their tweets or could be any other reason.

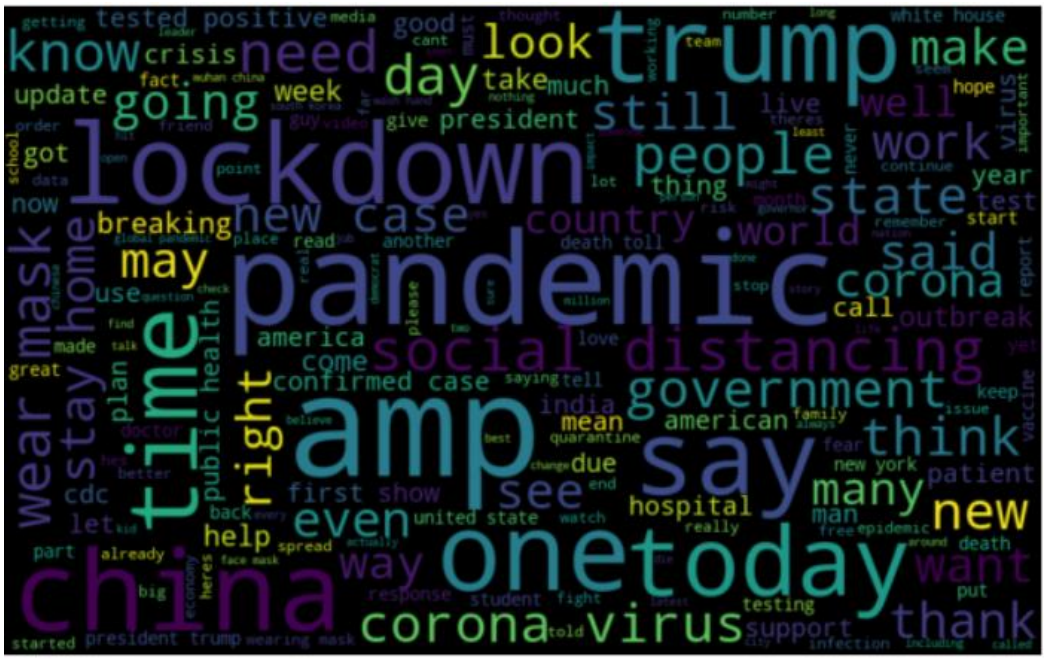


Fig 11: WordCloud of All the Tweets

Fig 11 displays the most repeated words from all tweets in the form of a WordCloud where higher the frequency of the word the bigger the font size would be. Here we see that Pandemic, Lockdown, China, Trump, Social Distancing, Time and Wear Mask are few most repeated words that people were talking about during the pandemic of Covid-19.

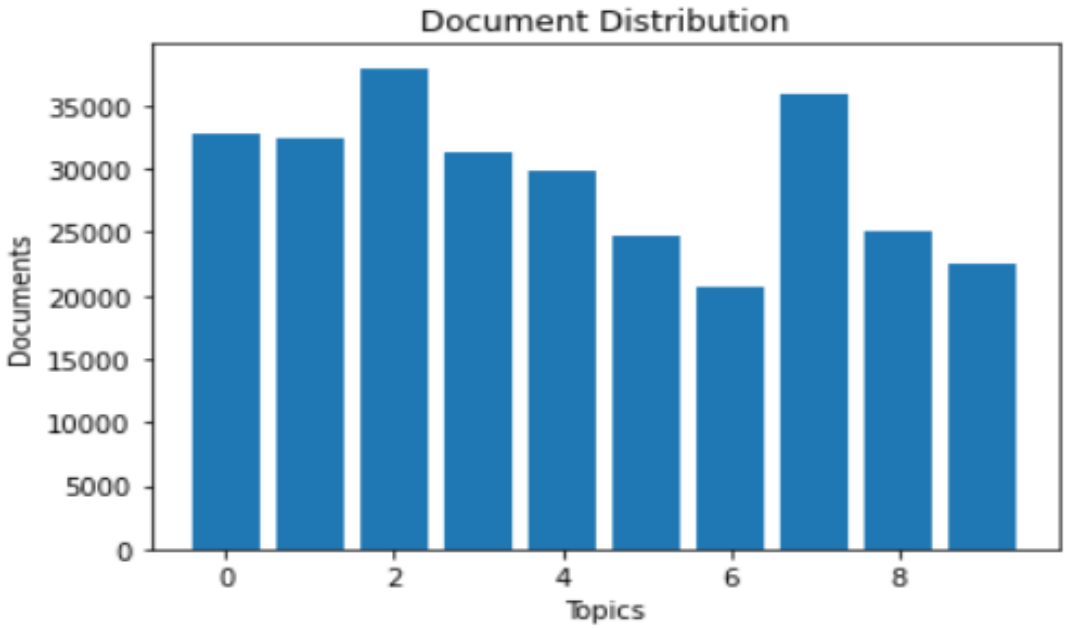


Fig 12: Document Distribution in Each Topic

Using LDA, tweets were distributed in 10 topics and Fig 12 shows the number of tweets or documents in each topic where Topic 2 had most number of tweets and Topic 6 had least. For better understanding pandas has a library named pyLDAvis which is a interactive visualization tool by extracting data or topics from fitted topic model [9].

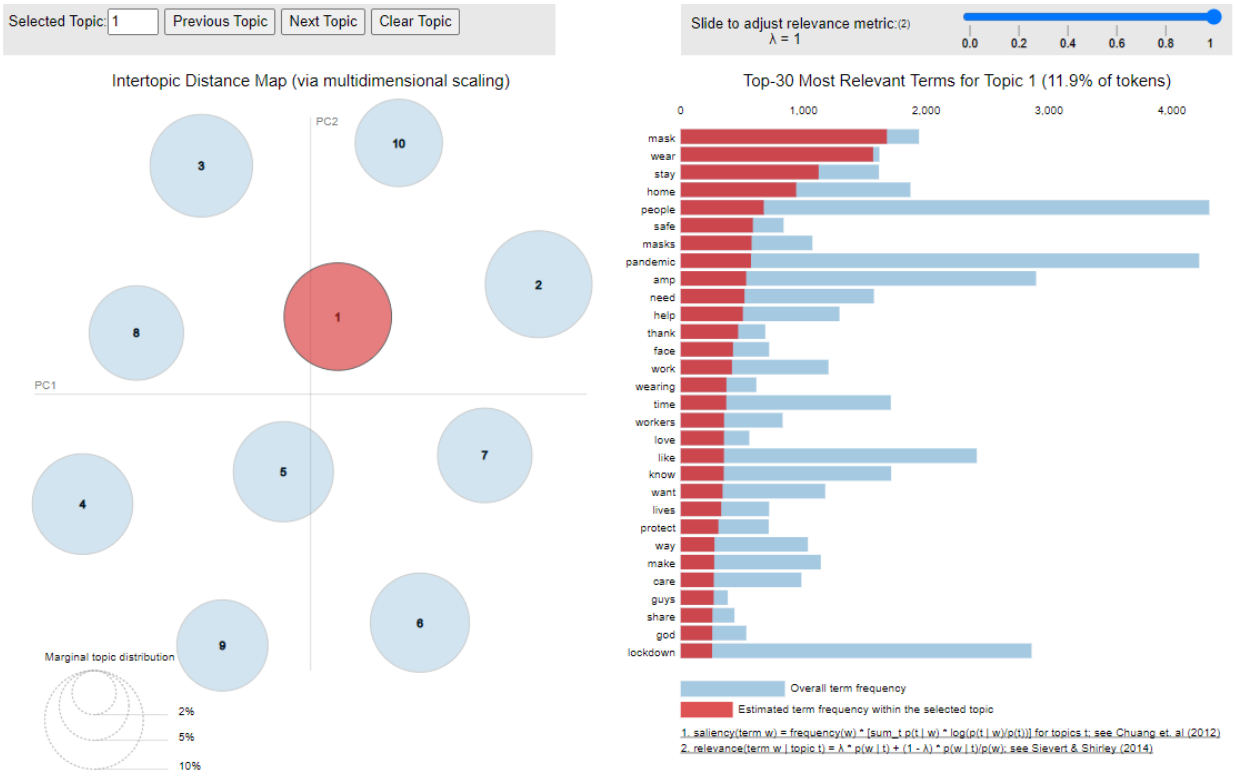


Fig 13: pyLDAvis of All Topics

In Fig 13 interactive visualization when focused on topic 1, on the right side it shows the number of most repeated words from its documents by showing the frequency in red color. This is great tool for visualization as it also allows changing the lambda value for the inverse document frequency and comparing the results between all topics. Also from the graph it can be analyzed how different the topics are by distance between their blobs. Topics 1 and 5 are very similar since distance between their blobs is least compared to other blobs similarly, Topics 10 and 9 blobs are very far from each other stating they are not related.

Fig 14 shows the WordCloud visualization of all these 10 topics. Topic 0 is more about precautionary or safety steps to be followed by all like social distancing and lockdown. In Topic 1 words interpret the effect of pandemic on schools and people also how the US President Trump response is to the situation. Topic 2 also focusses on safety measures like staying home, wearing mask and helping others. Topic 3 focusses on saving lives of people effected by pandemic and lockdown. Topic 4 focusses on the umber of covid cases in US and relief packages provided by the government. Topic 5 is about the effect of pandemic in other countries like Iran and China. Topic 6 is about the virus outbreak and how it started, how it is being spread between people. Topic 7 talks about new positive tested cases and deaths of people. Topic 8 Wuhan and China from where the virus was spread from and people travelling from these places. Topic 9 talks about cure of the virus and symptoms of virus in people like flu.

Fig 14: WordCloud of All 10 Topics

Next, I tried to make similar analysis on tweets about Vaccine of Covid-19. Since the datasets was generally on tweets of Covid-19, I tried to extract tweets taking about vaccine. There were like ~275K records in the filtered dataset. Same preprocessing steps were applied to clean the tweets and finally sentiment analysis was applied using the Vader library.

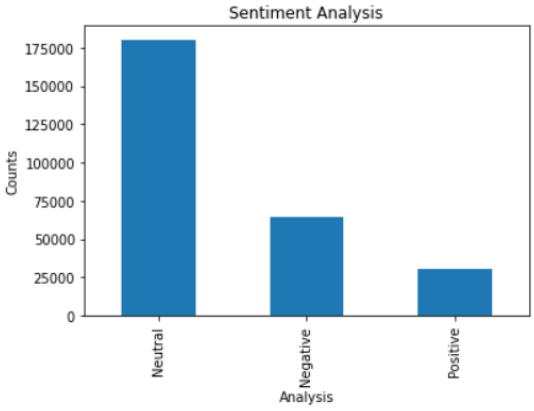


Fig 15: Vaccine Sentiment Distribution of Tweets

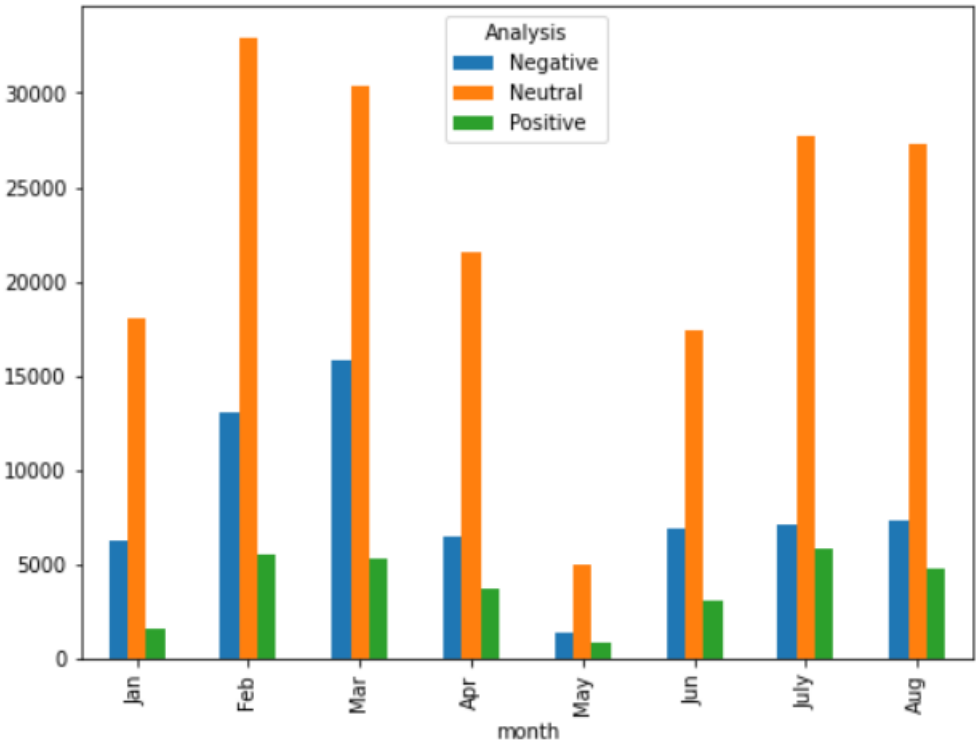


Fig 16: Vaccine Sentiments Distribution per Month

Fig 15 and 16 shows the distribution of sentiments of people's tweets on Covid Vaccine. Although the number of positive tweets were less than negative in every month but there is drastic decrease in number of negative tweets after March. In general, based on Fig 15 stats number of positive tweets were very less. Although this analysis or results could be very different or similar when dataset is a lot bigger. This analysis was made on significantly small dataset of 275K records of data compared to original ~10 Million tweets. But extracting tweets on just words similar to vaccine resulted in this small dataset compared to general Covid analysis on ~640K sampled tweets dataset.

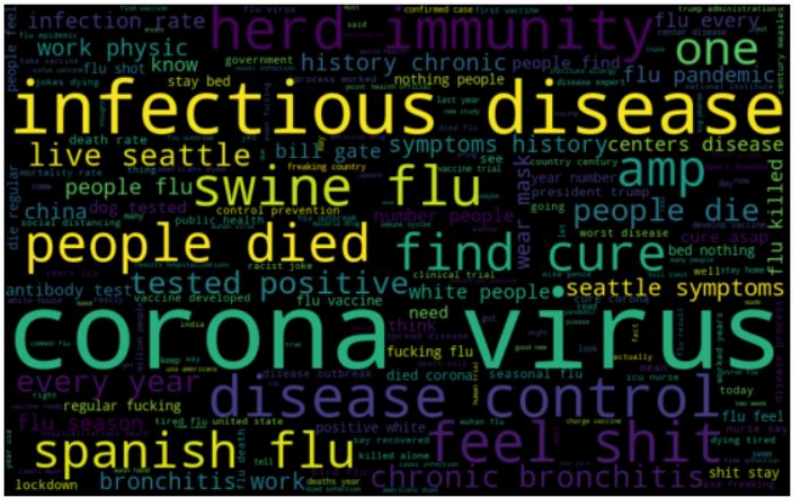


Fig 17: WordCloud of Vaccine Related Tweets

In Fig 17 the WordCloud shows words with more focus on Corona Virus, Infectious, disease, Swine Flu, Spanish Flu, disease control, deaths and immunity. It's not wise to interpret all these words to a meaningful sentence since these words are from total dataset and not any specific tweets. But it is understandable that people were talking about the previous major virus outbreaks their symptoms, how immune system is important and death of people because of how not enough measures were for controlling the disease.

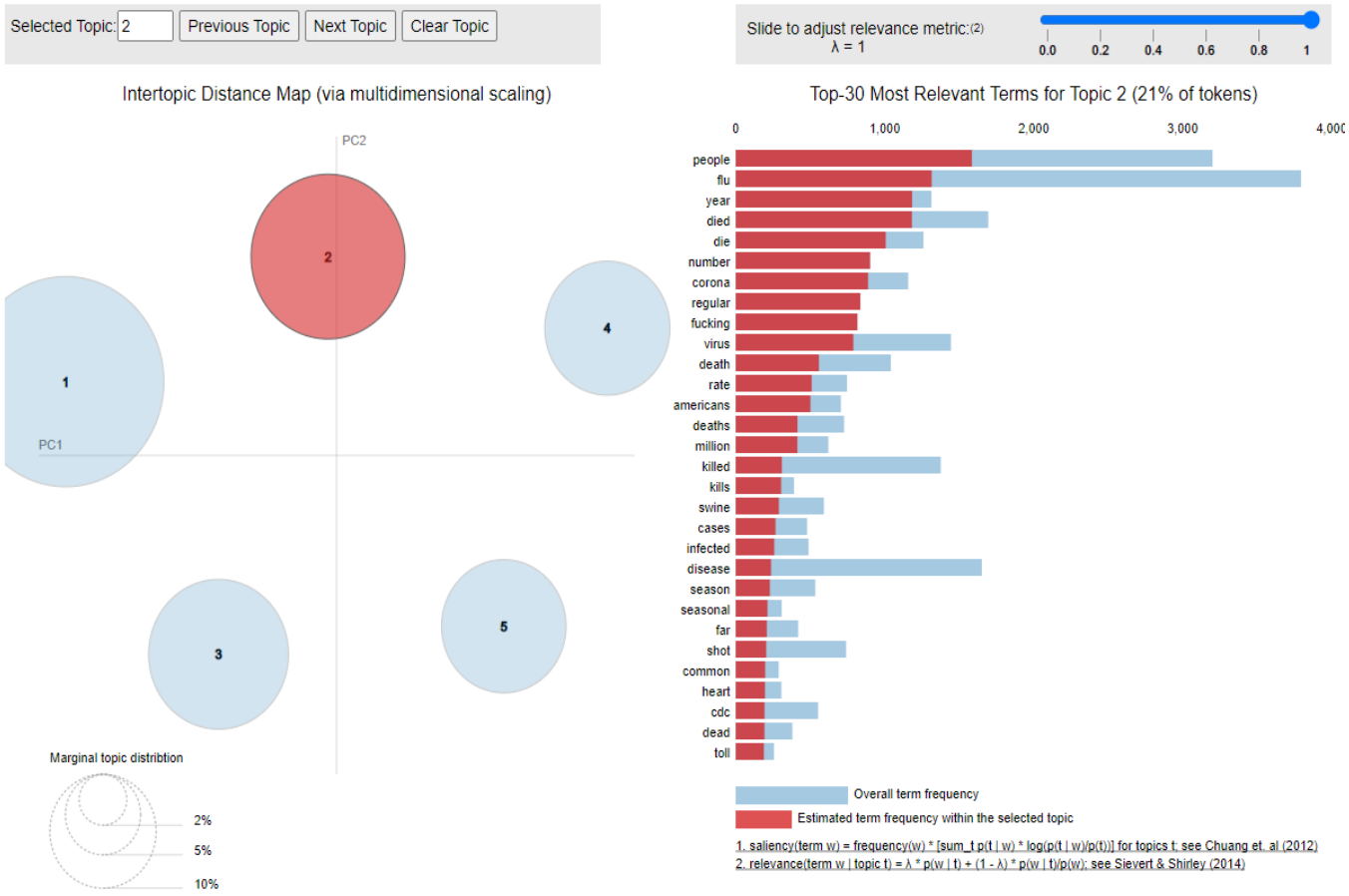


Fig 18: pyLDAvis for Vaccine Tweets

In Fig 18 again pyLDAvis interactive visualization tool was used for understanding the relation between topics and get more information about them. The topics blobs distance between them defines the similarity between them as topic 1 and 2 from the graph have minimal distance but topics 1 and 4 have highest distance so they can be considered as different from each other. On the right side of the visualization Blue bar shows frequency of word from all topics and the bar in red is aligned with the blob selected which is topic 2. Words appearing on the bar graph will be changed based on the topic selected. Finally, WordCloud is used to understand each topic generated by LDA. The number of topics was selected as 5 since the dataset size of previous analysis was twice the current size or shape of dataset.

Fig 19: WordCloud of Vaccine Related LDA Topics

Here Topic 0 is more about situation of nurses and doctors in this pandemic and the process being followed by the administrations. Topic 1 is about heath and infection being spread among people and how America or world is worried. Topic 2 is about racism being spread during pandemic how people are dying because of it. Topic 3 talks about cure or vaccine of virus and its getting worse. Topic 5 talks about rate of deaths of people due to corona virus.

Future Work:

There are few setbacks in this project which I would like to improve in future like ~10M tweets of dataset was filtered initially on English language and this way a lot of tweets is not taken into analysis, So working with tweets of all languages would be an important step to include in future. Even after filtering tweets with English language this project did not include all those tweets since extracting coordinates for such large data requires a large size of RAM memory. Next I would also try using different paid services of Geopy library as it would work fast. Finally working on laptop with high end specifications to manage large and time-consuming computations.

References:

1. 60 incredible and Interesting Twitter stats and statistics. (n.d.). Retrieved April 08, 2021, from https://www.brandwatch.com/blog/twitter-stats-and-statistics/#:~:text=Twitter%20usage%20statistics,That's%206%2C000%20tweets%20every%20second
2. Walton, J. (2021, January 26). Twitter vs. Facebook VS. Instagram: What's the difference? Retrieved April 08, 2021, from https://www.investopedia.com/articles/markets/100215/twitter-vs-facebook-vs-instagram-who-target-audience.asp
3. Ortiz-Ospina, E. (n.d.). The rise of social media. Retrieved April 08, 2021, from https://ourworldindata.org/rise-of-social-media
4. Brown, D. (2019, December 30). Remember Vine? These social networking sites defined the past decade. Retrieved April 08, 2021, from https://www.usatoday.com/story/tech/2019/12/19/end-decade-heres-how-social-media-has-evolved-over-10-years/4227619002/
5. Twitter revenue and usage Statistics (2020). (2021, March 08). Retrieved April 08, 2021, from https://www.businessofapps.com/data/twitter-statistics/#1
6. Welcome to GeoPy's documentation!¶. (n.d.). Retrieved April 08, 2021, from https://geopy.readthedocs.io/en/stable/#module-geopy.extra.rate_limiter
7. Vadersentiment. (n.d.). Retrieved April 08, 2021, from https://pypi.org/project/vaderSentiment/
8. Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.
9. Pyldavis. (n.d.). Retrieved April 08, 2021, from https://pypi.org/project/pyLDAvis/
10. Scatter plots on Mapbox. (n.d.). Retrieved April 08, 2021, from https://plotly.com/python/scattermapbox/
11. Albon, C. (2017, December 20). Geocoding and reverse geocoding. Retrieved April 08, 2021, from https://chrisalbon.com/python/data_wrangling/geocoding_and_reverse_geocoding/
12. Go, A., Huang, L., & Bhayani, R. (2009). Twitter sentiment analysis. Entropy, 17, 252.
13. Sarlan, A., Nadam, C., & Basri, S. (2014, November). Twitter sentiment analysis. In Proceedings of the 6th International conference on Information Technology and Multimedia (pp. 212-216). IEEE.
14. Alamsyah, A., Rizkika, W., Nugroho, D. D. A., Renaldi, F., & Saadah, S. (2018, May). Dynamic large scale data on Twitter using sentiment analysis and topic modeling. In 2018 6th International Conference on Information and Communication Technology (ICoICT) (pp. 254-258). IEEE.
15. Prabhakar Kaila, D., & Prasad, D. A. (2020). Informational flow on Twitter–Corona virus outbreak–topic modelling approach. International Journal of Advanced Research in Engineering and Technology (IJARET), 11(3).
16. Yang, S., & Zhang, H. (2018). Text mining of Twitter data using a latent Dirichlet allocation topic model and sentiment analysis. Int. J. Comput. Inf. Eng, 12(7), 525-529.
17. Xiang, B., & Zhou, L. (2014, June). Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 434-439).
18. Zhang, W., & Gelernter, J. (2014). Geocoding location expressions in Twitter messages: A preference learning method. Journal of Spatial Information Science, 2014(9), 37-70.
19. Zou, L., Lam, N. S., Shams, S., Cai, H., Meyer, M. A., Yang, S., ... & Reams, M. A. (2019). Social and geographical disparities in Twitter use during Hurricane Harvey. International Journal of Digital Earth, 12(11), 1300-1318.

20. Sewalk, K. C., Tuli, G., Hswen, Y., Brownstein, J. S., & Hawkins, J. B. (2018). Using Twitter to examine Web-based patient experience sentiments in the United States: Longitudinal study. Journal of medical Internet research, 20(10), e10043.
21. Zhang, L., Yuan, H., & Lau, R. Y. (2016, November). Predicting and visualizing consumer sentiments in online social media. In 2016 IEEE 13th International Conference on e-Business Engineering (ICEBE) (pp. 92-99). IEEE.