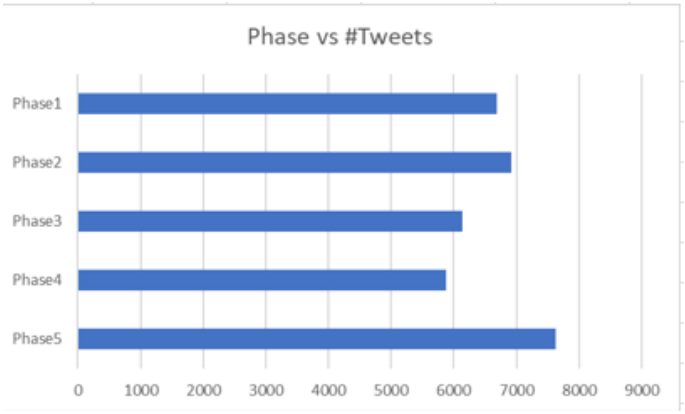
Introduction :
Data Extraction :

TWINT

Twint is an advanced Twitter scraping tool written in Python that allows for scraping Tweets from Twitter profiles without using Twitter's API. Twint utilizes Twitter's search operators to scrape Tweets from specific users, scrape Tweets relating to certain topics, hashtags & trends, or sort out sensitive information from Tweets like email and phone numbers. Twint also makes special queries to Twitter allowing it to scrape a Twitter user's followers, tweets a user has liked, and who they follow without any authentication, API, Selenium, or browser emulation.

P- No	#Tweets
Phase5	7623
Phase4	5877
Phase3	6142
Phase2	6918
Phase1	6696



v Dataset Collection:

Using Twint, initially we only had around 1150 tweets extracted for the January – February 2020 phase of COVID 19 based on the filter "Covid" included along with the 75 keywords provided along with the project requirement.

However, any kind of data processing, analysis or modelling is not possible with such a small dataset. In order to increase the size of the dataset, we broke down the Covid timeline into 5 phases, as follows:

- Jul'19 to Dec'19 - pre-COVID
- Jan'20 to June'20 - COVID Block I

· July’20 to Nov’20 - COVID Block II

· Dec’20 to May’21 - COVID Block III

· June’21 to present - COVID Block IV

Even after extracting tweets phase-wise, we had around 6000 tweets which was inadequate.

We furthermore incorporated the inputs from our Dr. Xia and Prof. Baird. We had to focus on health-care instead of just Covid in the keyword used for tweet extraction. After removing "Covid" from the primary keyword and extracting only based on the 75 keywords provided we successfully increased the number of tweets in our dataset.

We observed that Twint config () was extracting a greater number of tweets for the end dates in the duration provided as the start and end dates. Hence, finally, we extracted for each and every month of the phases defined above and prepared a dataset of around 33,000 tweets.

v Data Pre-Processing:

All the different components of a tweet extracted were discarded and a dataframe was created using the tweet column only. The tweets were further thoroughly pre-processed to prepare them for analysis, modelling and for visualizing. The tweets were first tokenized using the NLTK library.

The tweets were further refined by removing stopwords, removing tweets in languages other than English, removing punctuations, hashtags and mentions.

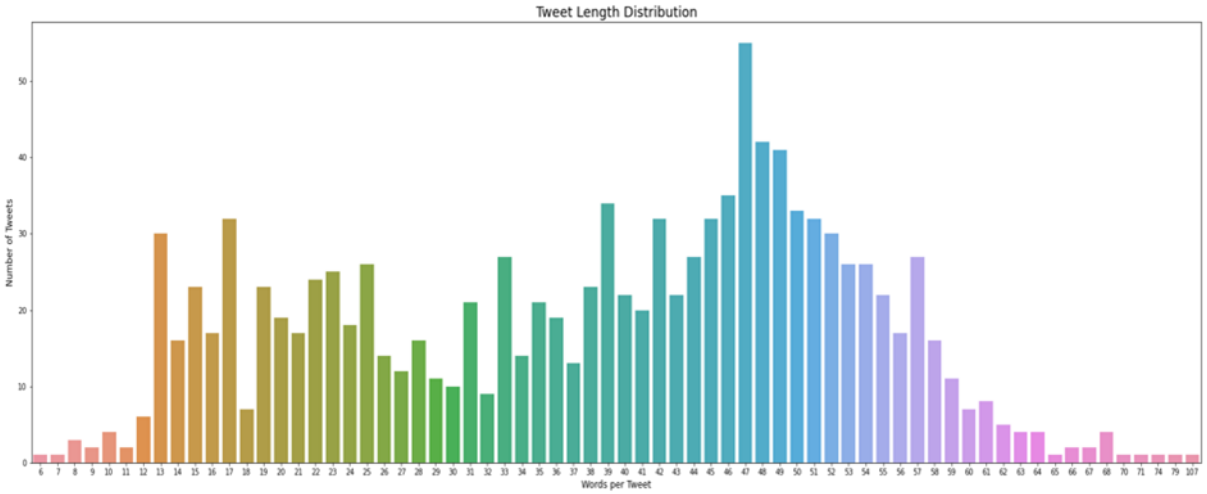


Fig 1: Tweet Length Distribution

Fig 2: Most Common Words – Word Cloud

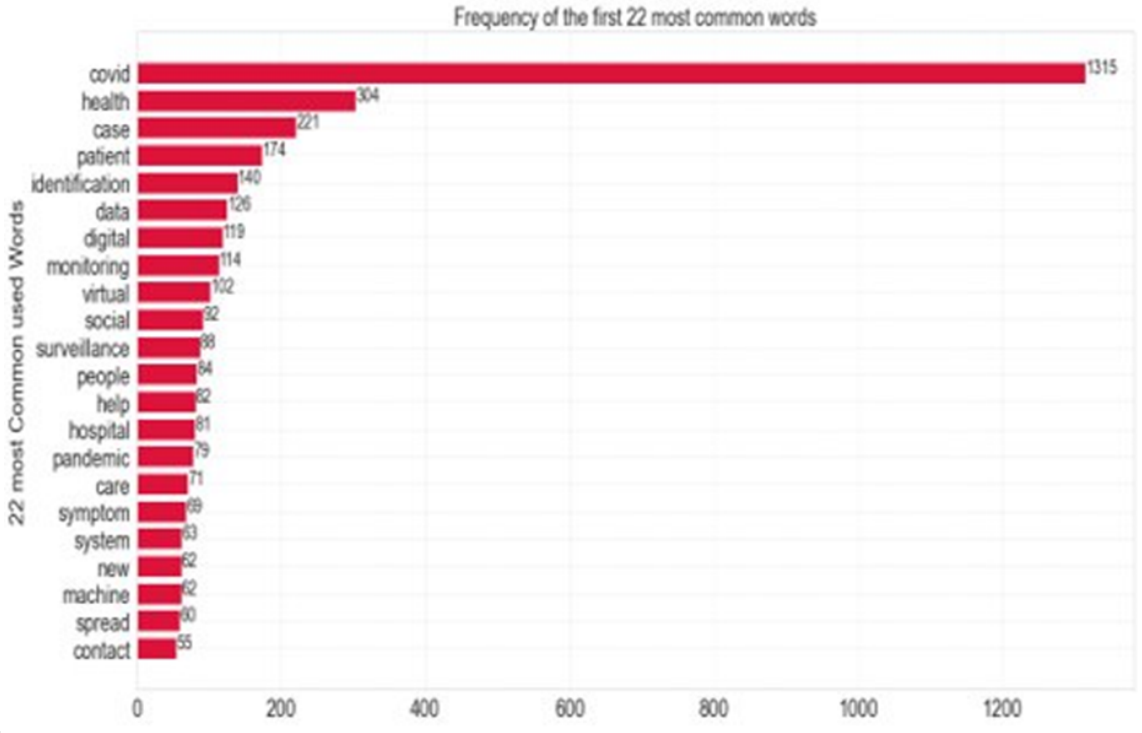


Fig 3: Frequency of the top used words

Latent Dirichlet Allocation(LDA):

LDA is a form of unsupervised algorithm that means it draws the pattern from the documents as a bag of words. The way a document is generated is it just picks up the topic and then for each topic it picks up a set of words.The number of topics to choose will be dependent on the coherence score. Lets now discuss the insights that we observed from our project while using LDA.

PHASE-1:

From the figure below the number of topics is found out to be 5 because it has got the highest coherence score.

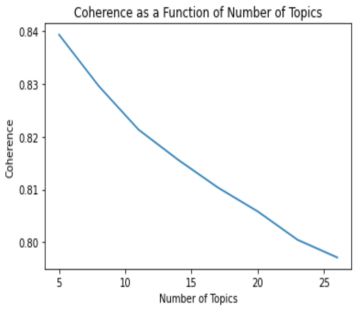


Figure-1(a)

Please refer to Figure-1(a).There are five topics and each topic has 10 words. The topic-1 speaks about the Online Information shared. For example the words like internet,technology,information,visit,e.t.c.The topic -2 speaks about the Covid cases and how technology is used to control or spread the awareness about the covid-19. The topic 3,4 and 5 speaks about the Telehealth services provided by the organisations to people who were affected by the covid-19. From the figure-1(b) says that from the topic 2 and 5 there is some overlapping of the words.

	topic_1	topic_2	topic_3	topic_4	topic_5
word_1	information	surveillance	technology	health	time
word_2	Internet	people	patients	//tco	telemedicine
word_3	technologies	years	COVID-	care	risk
word_4	solution	capacity	Technology	data	coronavirus
word_5	management	digitalhealth	healthcare	post	services
word_6	Things	pleaseretweet	patient	wearabletechnology	monitoring
word_7	visit	virus	mhealth	revenue	support
word_8	help	vision	system	devices	media
word_9	contact	hospital	life	database	telehealth
word_10	triage	record	app	world	sensor

Figure-1(b)

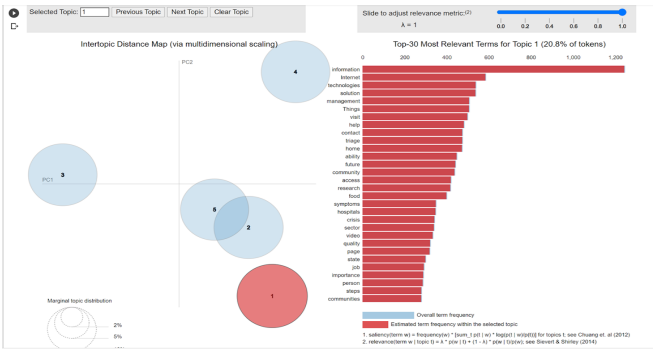


Figure-1(c)

From the below image(Figure-1(d) the number of topics is found out to be 45. Let's see what we get now. The topics 3,4,2,13 e.t.c had unique words in them and most of the other topics are overlapping which says they have common words among them and mostly these topics discuss about the treatment,healthcare of the patients,patients records e.t.c

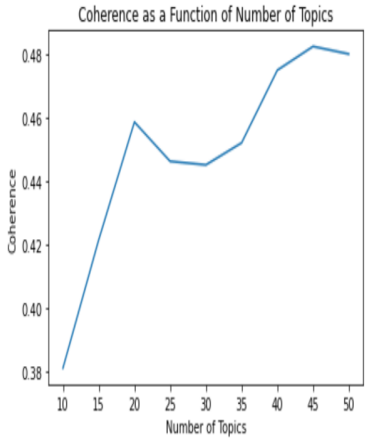


Figure-1(d)

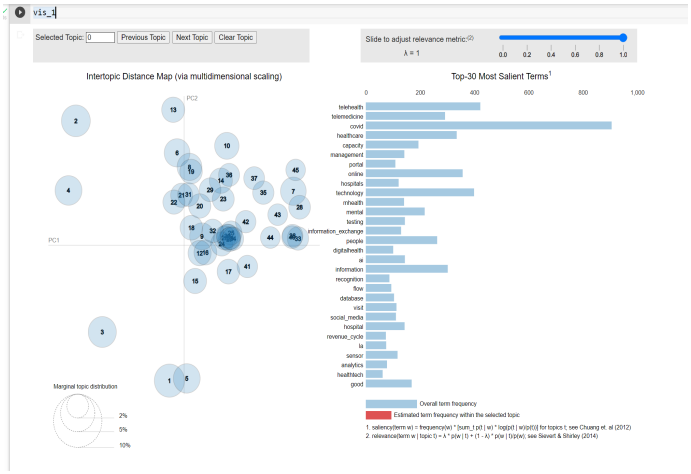


Figure-1(e)

PHASE-2:

From the Figure-2(b),Topic-1 and 5 speaks about accessing the patients' records in the organisations . We can access the data through many ways like databases, web, app and record.Topic-2 speaks about the health condition of the patient admitted in the hospital.Topic-3 speaks about how to detect and recognize the covid-19 using the technology and testing.Topic-4 provide the information of the people who were affected on each day and also providing the stats on the websites to provide the information about the cases.These topics briefly explains on the patients health,record and detection of the covid-19 and provide the stats across the world regarding the confirmed cases per day,per month,e.t.c.From the Figure-2(c) we can observe that topic 1,4, and 5 have words common because even though they explain us the different contexts the words are the same.

	topic_1	topic_2	topic_3	topic_4	topic_5
word_1	technology	health	data	//tco	contact
word_2	time	care	people	patients	services
word_3	patient	monitoring	work	information	way
word_4	interoperability	healthcare	exchange	COVID-	school
word_5	app	access	tracing	day	telemedicine
word_6	web	system	recognition	world	Internet
word_7	today	imaging	detection	year	research
word_8	database	management	week	digitalhealth	Web
word_9	record	help	surveillance	tech	service
word_10	cybersecurity	tool	triage	providers	person

Figure-2(a)

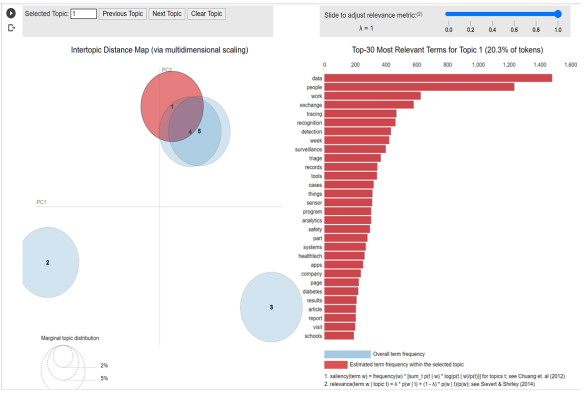


Figure-2(b)

From the below image the number of topics is found out to be 45. Let's see what we get now. The topics 1,3,23,39 e.t.c had unique words in them and most of the other topics are overlapping which says they have common words among them and mostly these topics discuss about the detection,treatment of the patients using the technology e.t.c

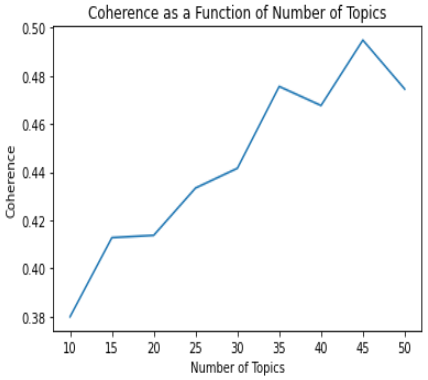


Figure-2(c)

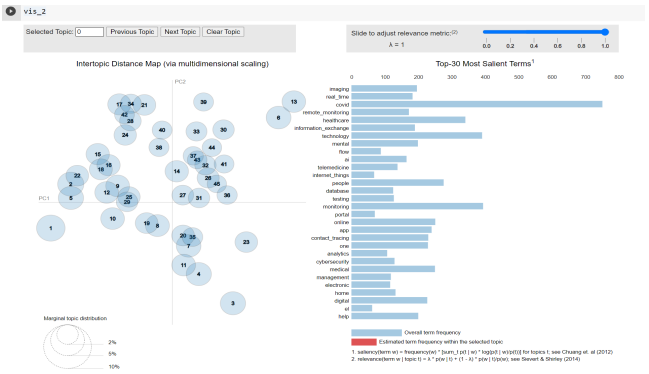


Figure-2(d)

PHASE-3:

From the Figure-3(a), Topic-1 speaks more about the websites,links where people could get more information about the covid-19 and tells about the security of the date.How important the data is.Topic-2 and 5 speak about the accessing the tele-health information and revenue management. Topic-3 speaks about how to prevent the covid-19, they are many blogs written on the pandemic.Topic-4 speaks about the monitoring the health of the patients and stats of the positive cases over the months.From the figure-3(b) we can infer that the words in topic 2 and 5 overlap because they explain the same context.

	topic_1	topic_2	topic_3	topic_4	topic_5
word_1	web	health	prevention	time	information
word_2	technology	data	tools	monitoring	people
word_3	//tco	sensor	part	patients	Internet
word_4	healthcare	access	intelligence	system	telehealth
word_5	care	video	security	today	number
word_6	visit	work	recognition	blockchain	imaging
word_7	patient	management	digitalhealth	day	info
word_8	sensors	revenuecyclemanagement	blog	COVID-	month
word_9	cybersecurity	year	way	surveillance	revenue
word_10	risk	years	devices	app	solutions

Figure-3(a)

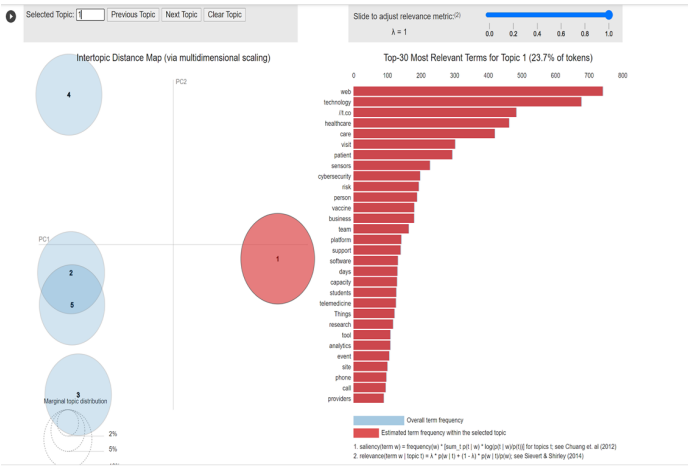


Figure-3(b)

From the below image Figure-3(c) the number of topics is found out to be 45. Let's see what we get now. The topics 1,3,23,39 e.t.c had unique words in them and most of the other topics are overlapping which says they have common words among them and mostly these topics discuss about the advancement of the web applications and technology

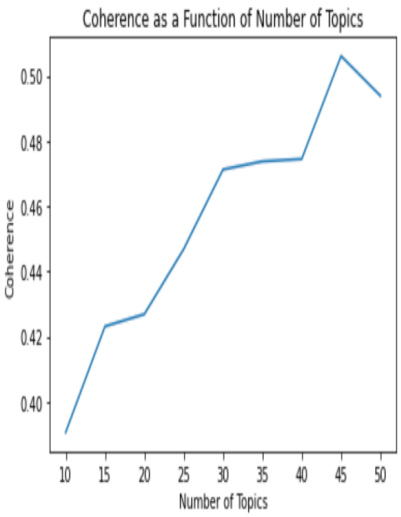


Figure-3(c)

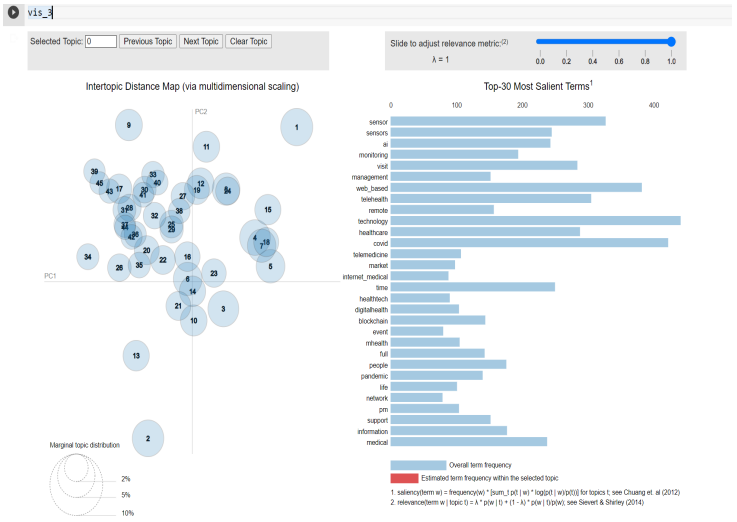


Figure-3(d)

PHASE-4:

From the Figure-4(a),Topic-1 speaks about the Patients health monitoring and the risk of being affected by covid-19,information/awareness of the disease.Topic-2 and 3 speaks about the web technology on how it displays the importance of the requirement of the mask to avoid the covid.Topic-4 speaks about the vaccines and prevention of the covid-19 and Topic-5 speaks on the patient treatment and how media is helping to take care of ourselves Analysis on how many are getting affected,how many deaths,e.t.c.How internet played a crucial role during these tough times.

	topic_1	topic_2	topic_3	topic_4	topic_5
word_1	health	technology	care	//tco	patient
word_2	patients	web	data	time	healthcare
word_3	people	flow	hospital	application	media
word_4	information	work	Things	business	Internet
word_5	monitoring	database	recognition	hospitals	analytics
word_6	triage	cases	COVID-	app	exchange
word_7	contact	testing	record	vaccine	treatment
word_8	risk	homework	applications	prevention	capacity
word_9	year		interoperability	way	home
word_10	paper	today	system	world	issues

Figure-4(a)

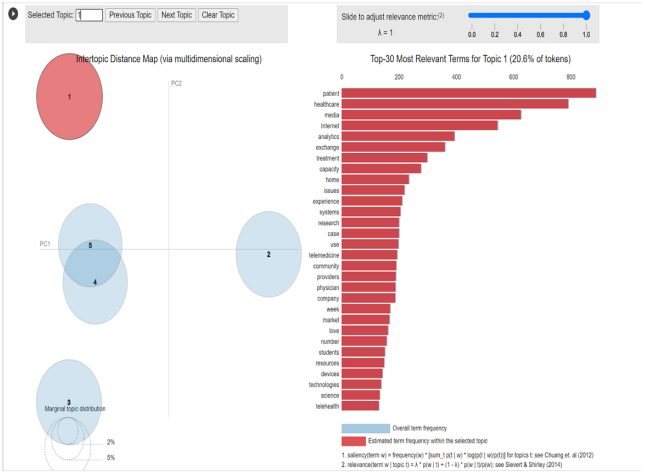


Figure-4(b)

From the below image Figure-4(c) the number of topics is found out to be 50. Let's see what we get now. The topics 3,4,20,33 e.t.c had unique words in them and most of the other topics are overlapping which says they have common words among them and mostly these topics discuss about the covid,technology and monitoring of the patients health.

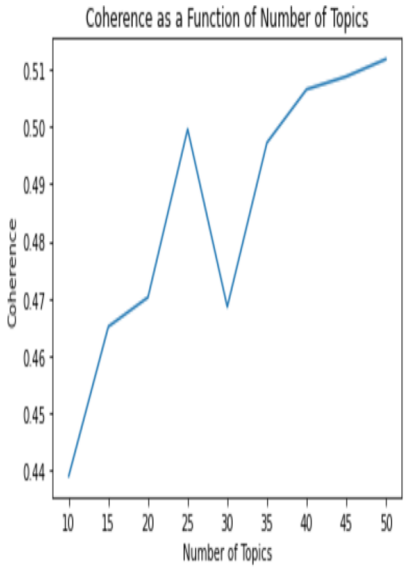


Figure-4(c)

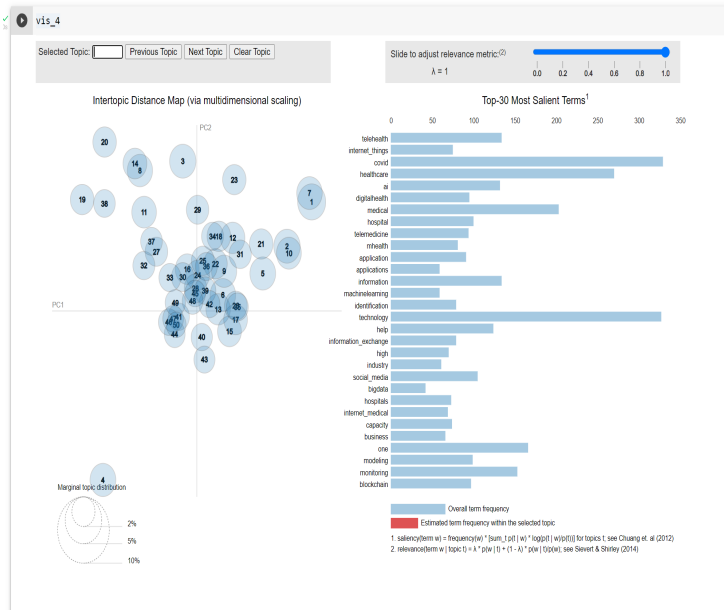


Figure-4(d)

PHASE-5:

From the below image (figure-5(a)) the number of the topics are 45.This phase of topics is mostly on vaccines,prevention and how to control the disease. Analytics had also played a crucial role in this phase, for example, how many received the vaccines today or in a month and how many are yet to receive a vaccine. Technology is being updated daily in our day to day life.

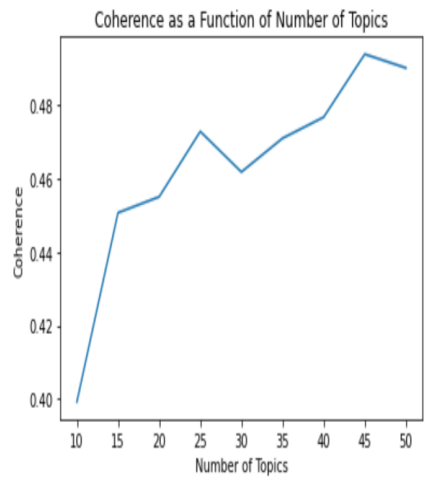


Figure-5(a)

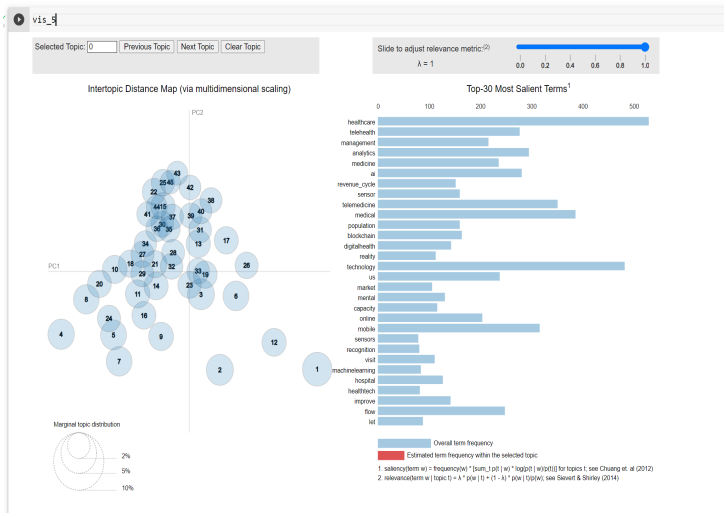
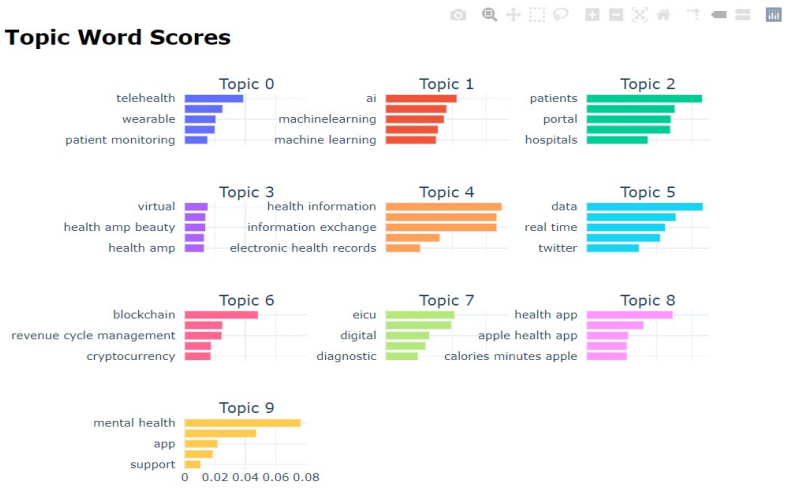
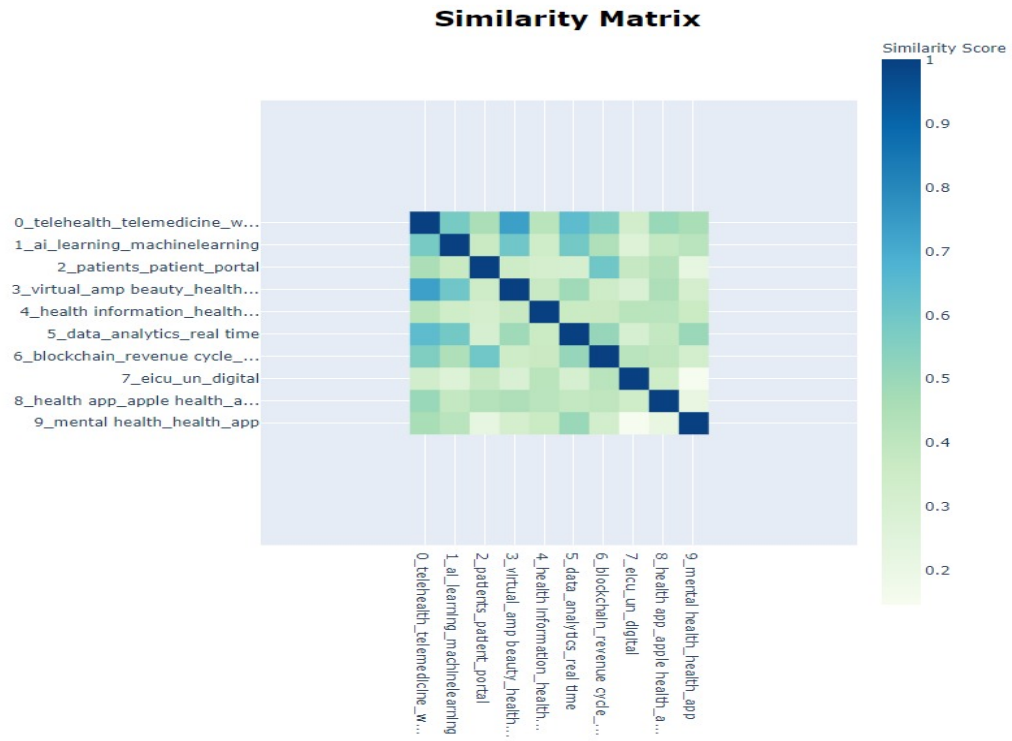


Figure-5(b)

Topic Modeling and similarity matrix using BERT :

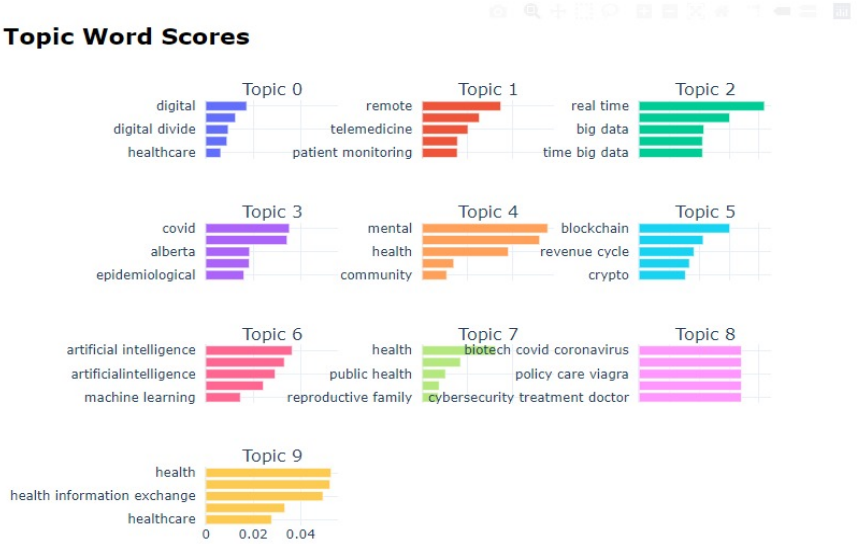
Phase 1:

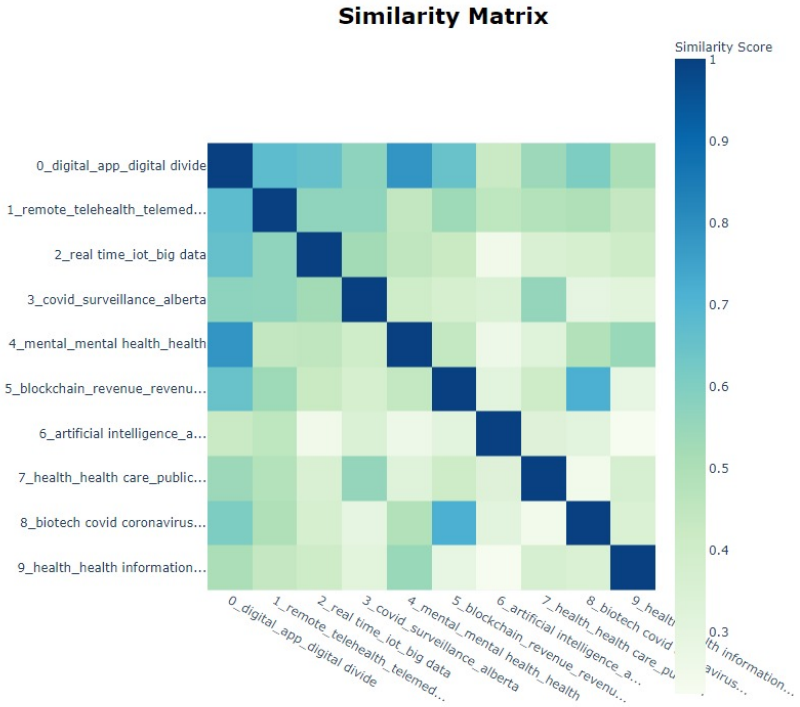




In phase 1 , from the similarity matrix we can see that topic 0(telehealth) and virtual health amp are similar followed by data analytics and telehealth are similar topics discussed in the twitter. Also in Phase 1 we havent seen any topic related to covid/corona or healthcare topics because this is pre covid phase

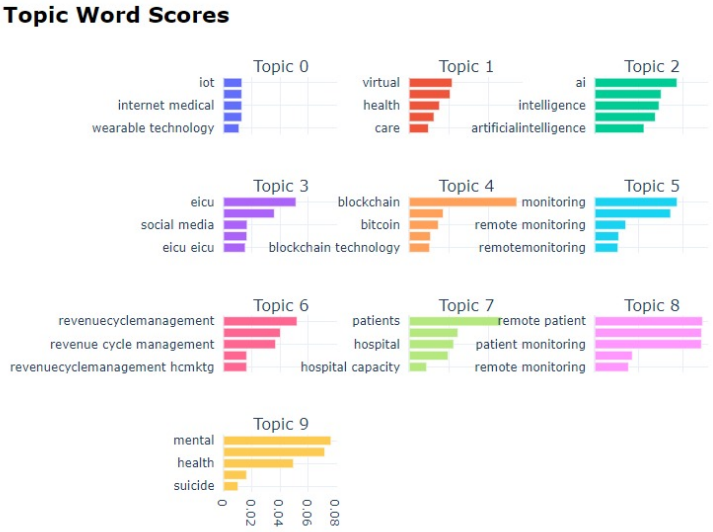
Phase 2 :

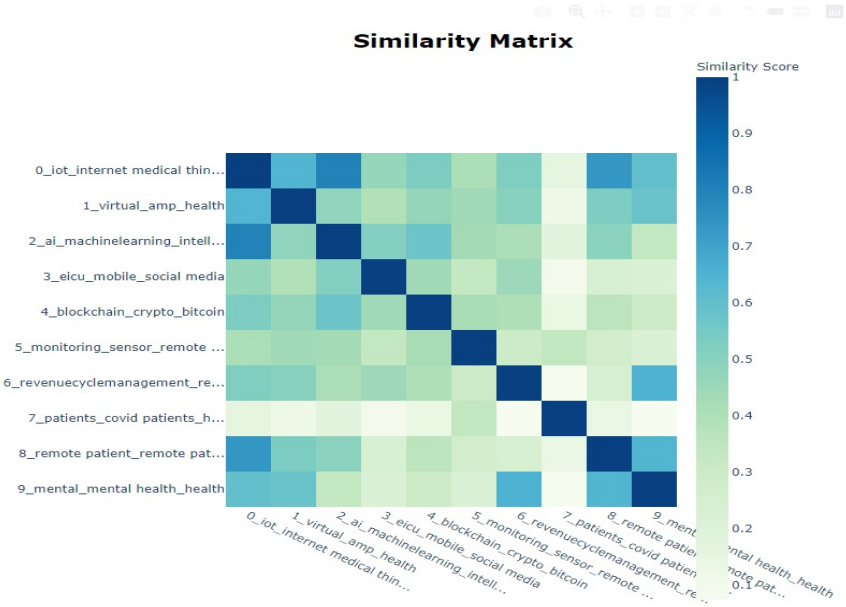




From the phase 2 topics and similarity matrix we can see that mental, Health and digital_healthcare share the most similar topics followed by Covid_corona_biotechnology and block chain revenue. Also we can see new topics of corona virus has appeared in this phase which is starting phase of covid19

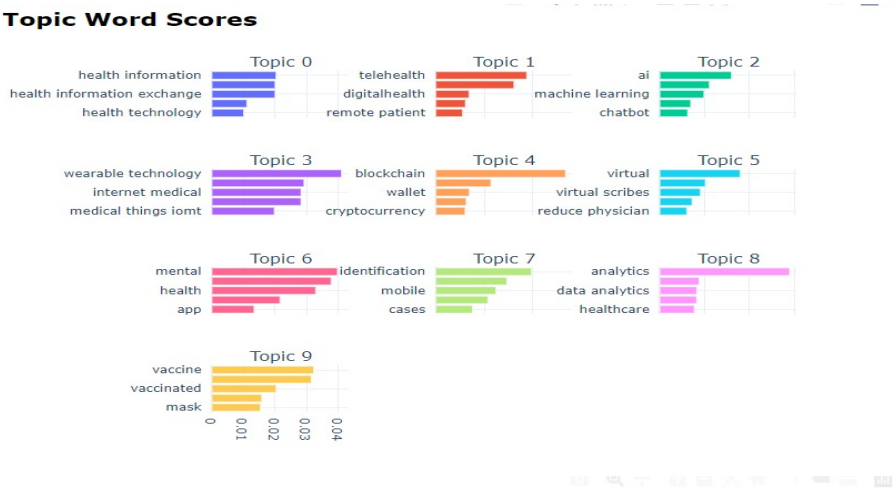
Phase 3:

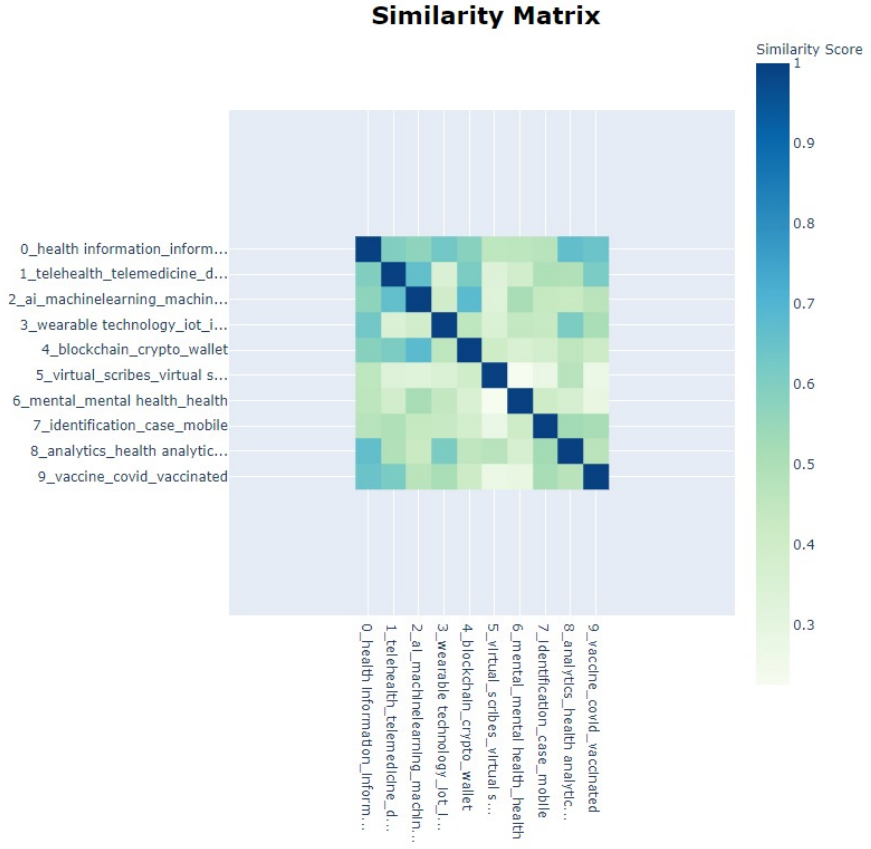




From the above phase 3 topic words we can see that Patients, blockchain and remote patients are the frequent topics words from the overall topics. As this was the crucial phase where lockdows started and covid pandemic took many lives, hence patients' words came into picture in most of the topics. Also, observing the topic similarity matric, there is ai_machinelearning and iot_internet_medical healthcare share the similarity topics followed by Remote_patients vs mental health which shows healthcare development started and sectors came into remote considerations.

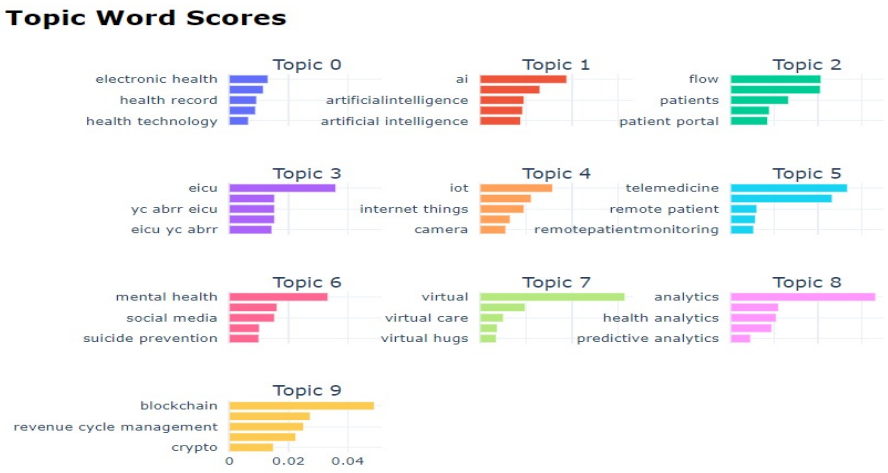
Phase 4:

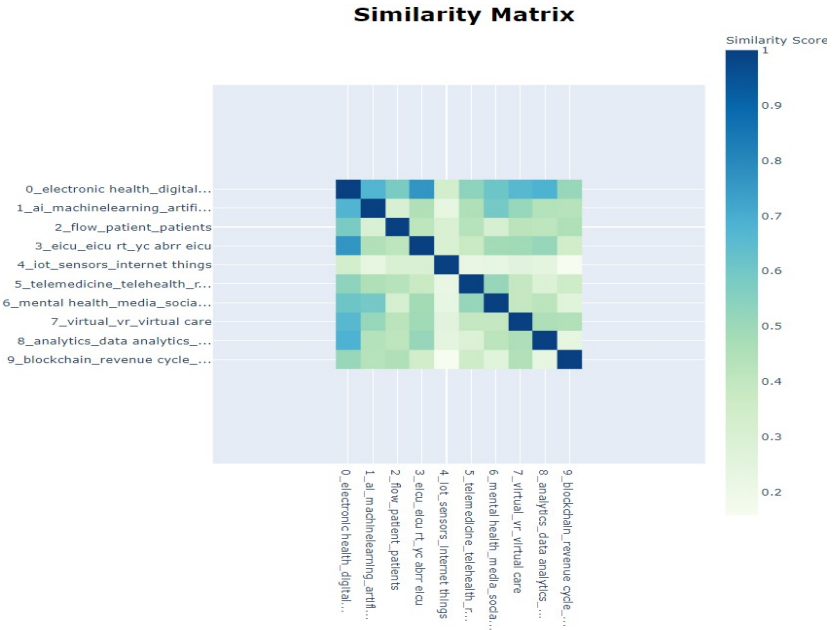




From the above topic word phase 4 figure we can see that there is new topic 9 with words vaccine , vaccinated and mask came into the picture which shows that this is the phase where people and government healthcare started talking about different vaccines and importance of vaccination to shield the corona effect which was not seen in the frequent words of other phases. Also, from the similarity matrix, it is inferred that health information is similar share the similar topics with the topics from anlytics_health_analytics. Followed by blockchain crypto wallet with ai and ML.

Phase 5





In the Phase 5 topic modelling, we can see that virtual care , patients, eicu,healrecord came into picture which means there is stress on care , healthcare development where its importance is talked about. From the similarity matrix we can say that ICU and electronic health digital have high similarity scores followed by electronic_health with analytics_revenue cycle.

Sentiment Analysis:

Introduction:

Sentiment analysis, in the field of data analytics is the process to quantify subjectivity in the data available.
For Team 3, the data is simply Tweets, and the labels are '**Positive**', '**Negative**' and '**Neutral**'.

Labelling and Preprocessing:

The team collectively labelled about 3100 tweets and cross labelled to ensure quality of labels among the team. Furthermore, it has also been cross checked with Vader score, which ranges between [-1,1], -1 being extremely negative, 1 being extremely positive and 0 being neutral.
The distribution of the above mentioned label is shown below:

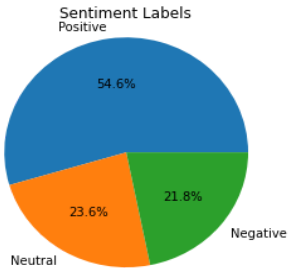


Figure-6(a)

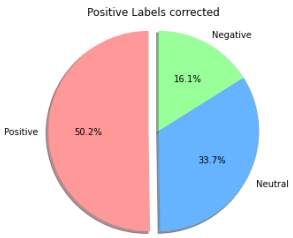


Figure-6(b)

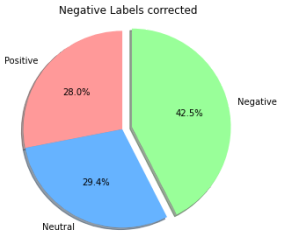


Figure-6(c)

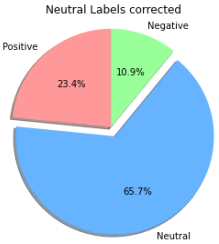


Figure-6(d)

Figure - 6(b) shows the corrected labels to positive, Figure - 6(c) shows corrected labels to negative, Figure - 6(d) shows corrected labels to neutral.

Preprocessing for Sentiment Analysis include:

- Vader Compound Scores Extracted ranging between [-1,1]
- Remove duplicate tweets (Total 128)
- Target label categorized
- Tokenization, Digit & Punctuation Removal
- Conservative Stopword removal approach
- Lemmatization using Wordnet

Sentiment Word Clouds for Positive vs Negative are shown below:

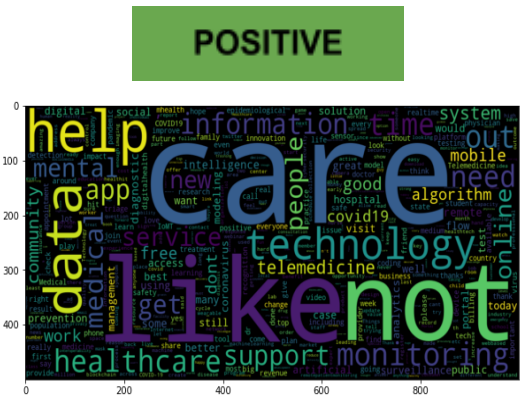


Figure-7(a)

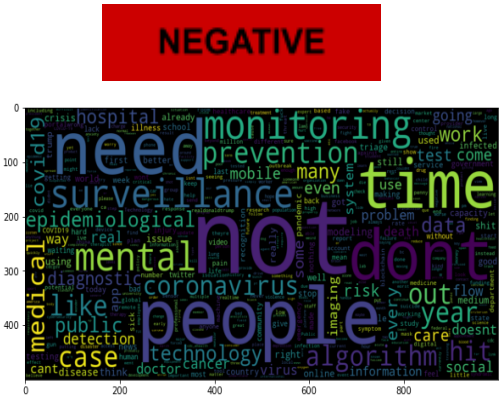


Figure-7(b)

Modelling:

- Vectors: **TF-IDF**
- Split: **80:20**
- **Stratification during train-test split -** to ensure equal class-distribution
- Multi-class classification problem:
  - ○Logistic Regression
  - ○Random Forest
  - ○Linear SVC
  - ○**DistilBERT (**Best Results**)**

Results and Inferences:

	BERT	Random Forest	Logistic regression	SVM
Precision	0.759143	0.679758	0.667730	0.680058
Recall	0.750861	0.618726	0.605450	0.650786
Accuracy	0.758446	0.663851	0.653716	0.677365
F1 score	0.752146	0.626976	0.618293	0.660845
MCC	0.623540	0.466787	0.450835	0.489917

Figure-8

DistilBERT outperforms the rest of the models by around **8**%.

	precision	recall	f1-score	support
0	0.74	0.78	0.76	135
1	0.78	0.64	0.70	191
2	0.75	0.83	0.79	266
accuracy			0.76	592
macro avg	0.76	0.75	0.75	592
weighted avg	0.76	0.76	0.76	592

Figure-9(a)

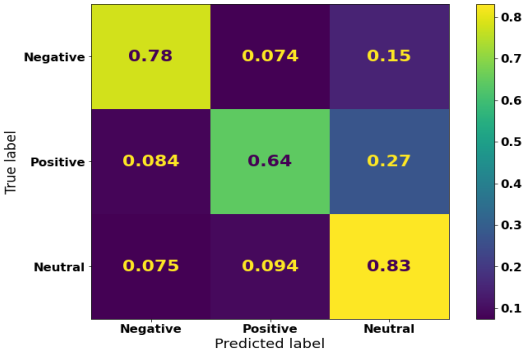


Figure-9(b)

Here, from Figure-9(b), the confusion matrix can be seen to predict well for Neutral tweets. It is also apparent that the model misclassifies Negative and Positive quite often (0.15 and 0.27 respectively), but these results indicate that BERT still performs relatively better.

Sentiment Analysis (Team 3 & 4):

What is Sentiment Analysis?

Sentiment Analysis:

Sentimental Analysis is contextual mining of text to identify the subjectivity in the source. The final goal is to find the subjectivity in the text i.e., to find whether the text data we have is Positive, Negative, Neutral.

Example:

Positive	I am very happy today
Negative	I had a bad day
Neutral	I am not sure if I like that book

Sentiment analysis on combined data - Team 3 & 4:

Combined the tweets data from Team 3 and text data from articles from Team 4. Sentiment analysis is done on combined data. The data for Team 3 is extracted from twitter and the data for Team 4 is extracted from news articles. The data preprocessing is done, and data is cleaned to apply sentiment analysis. The tweets are classified into **Positive** and **Negative** based on the sentiment in the text. Word clouds for both Positive and Negative text data are generated for all five phases.

Sentiment word clouds for Positive and Negative text data:

Phase 1(July'19 to Dec'19):

Positive **Negative**

For phase 1 we can see positive word cloud containing words like news, research, disease, prevention e.t.c and negative word cloud containing words like cancer, news, prevention, condition, therapy e.t.c. As this phase is pre-covid phase the word clouds contain generic terms related to health care in both positive and negative word clouds.

Phase 2(Jan'20 – Jun'20):

Positive **Negative**

Phase 2 is when WHO declares pandemic. We can see in positive word cloud words like company, market, business, research, solution, global which is meaningful because more discussions during phase 2 are related to how covid is impacting the business worldwide and discussions about companies manufacturing vaccines. If we see negative word cloud words like prevention, disease, risk, control, emergency, tracing indicate that people started talking and worrying about the risks, prevention and control of the disease.

Phase 3(Jul'20 – Nov'20):

Positive **Negative**

Phase 3 is when discussions about vaccines and their effectiveness started. Positive word cloud contains words like market, analysis, product, solution, contact, sale which indicates that discussions are about the vaccines and their sales and their impact on the market. Negative word cloud contains words like disease, prevention, risk, severe, contact, tracing, respiratory which clearly shows that people talked about the risks of covid, how it is caused e.t.c

Phase 4(Dec'20 – May'21):

Positive **Negative**

Phase 4 is when vaccines get emergency approvals and vaccine trades started between countries. Also US elections took place during this time. The positive word clouds contains words like market, company, vaccine, million, impact, analysis, revenue which shows discussions about vaccines and their impact in the global market. Negative word cloud containing words like risk, viral, respiratory, severe, prevention, virology, epidemiology show that the discussions are related to complications caused by covid.

Phase 5(Jun'21-present):

Positive **Negative**

Phase 5 is when new strains of covid started appearing and also demand for vaccines increased and people started getting vaccinated. Positive word cloud contains words like solution, analysis, growth, demand, application, control, forecast, impact which shows discussions about market conditions, control of covid, analysis of covid and also it shows that discussions include technical terms like analysis, software, analyst, information which shows that discussions are also about use of technology during covid times to face tough situations. Negative word cloud contains words like risk, syndrome, respiratory, viral, medicine, severe which indicates discussions about the impact of covid.

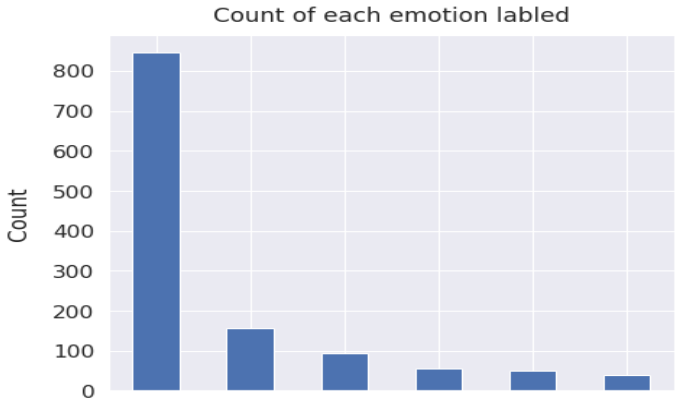
Emotion Analysis

Emotion analysis is a process of identifying and analyzing underlying emotions present in the textual data. For Team 3, the applicable data is Tweets for 5 phases, and respective classes are 'Happy', 'Anger', 'Fear', 'Surprise', 'Sad', 'None'

Labelling and Preprocessing:

The team collectively labeled about 1500 tweets and cross labeled to ensure quality of classes among the team. The majority class here turns out to be 'None', which shows that tweets during these phases were objective in nature more often than not.

The distribution of the classes mentioned label is shown below:



- 'None' class dominates the labelling, majority class
- Happy' slightly higher than

‘sad’, aligns with sentiment

- Minority classes: ‘Anger’, ‘Surprise’, ‘Fear’

Figure-10

Modeling and Inferences:

	BERT
Precision	0.171773
Recall	0.176645
Accuracy	0.680700
F1 score	0.162203
MCC	0.044686

Figure-11(a)

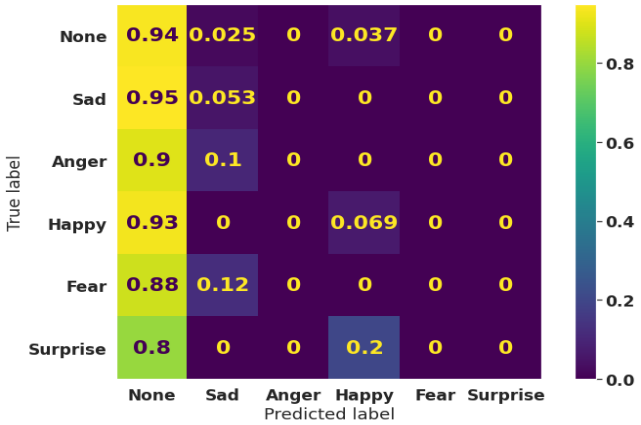


Figure-11(b)

As we can see here, DistilBERT performs poorly on the labeled emotion data.
This could be due various reasons, we suspect the most substantial of the reasons include:

- Data Imbalance in the labelled dataset, as ‘None’ covers more than 60% of the data
- Insufficient data labelled; **a potential fix here would be implementing BERT on a bigger dataset.**