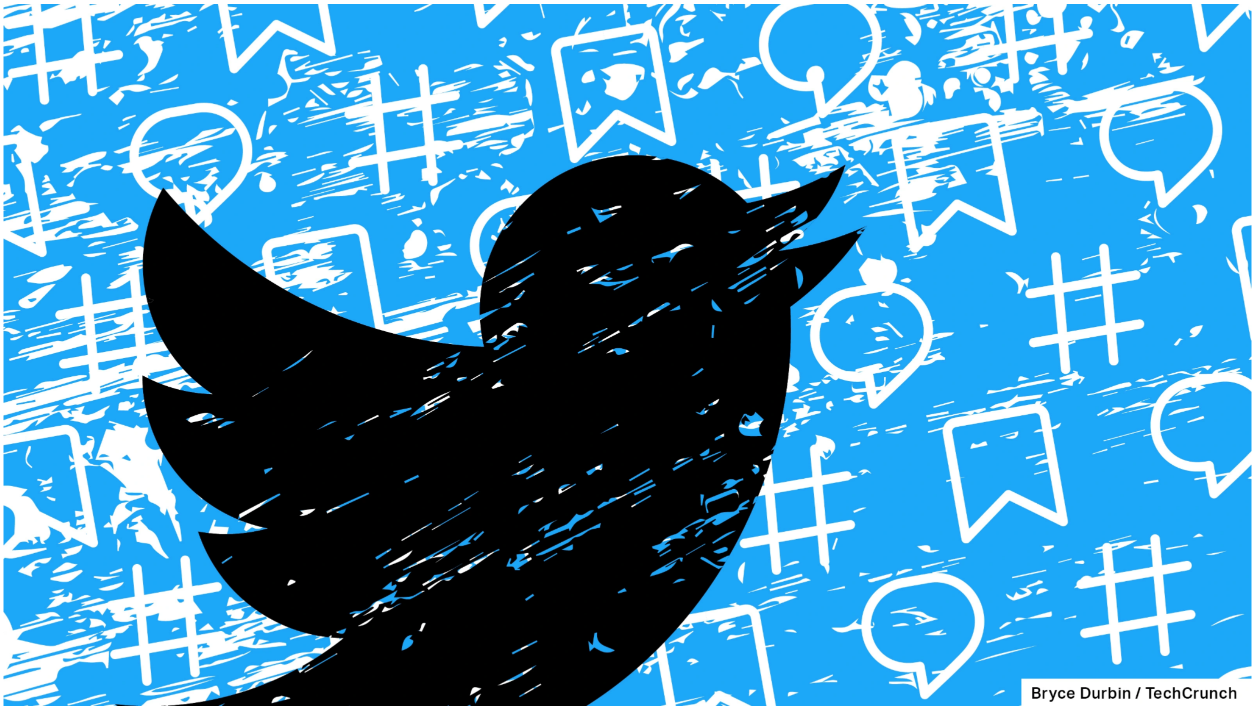
IBM-312 Data Mining

Group Project



Group 1

19113072- Krishna Agarwal
19113099- Parul Chaudhary
19113138- Shubhank
19113156- Vidhi Mittal

Objective

To analyze the common sentiments/views on effects & progress of COVID vaccination drive globally expressed on twitter via sentiment analysis using NLTK, VADER, Textblob API and ensemble method.

Motivation

After considering our syllabus for the course, we were looking for some practically relevant problem statements where we could apply the learnings as well as extract some valuable result with whatever we had learnt.
COVID was (and is), still a part of our lives and hence the pandemic period dominated our thinking and led us to finding something there. Also, the recent maneuver of Elon Musk with twitter is hot and hence the off-spring: Twitter's functioning in COVID.

Methodology

Our main objective is to analyze the Tweets regarding Covid 19 vaccines and classify them as positive, neutral and negative. We have used two methods namely Sentiment analysis using Textblob and Sentiment analysis using NLTK Vader. Then we combined the two methods and did the composite sentiment analysis with the ensemble method. Our methods for the analysis of Covid-19 vaccine tweets involve four significant steps :-

1. Data Collection

The following work employs a Kaggle dataset called "All COVID-19 Vaccines Tweets". The data was collected using a Python package called Tweepy, which enables a user to access the Twitter API if they have successfully created a Twitter Developer account and obtained access credentials.
Below is the link to the original dataset.

https://www.kaggle.com/datasets/gpreda/all-covid19-vaccines-tweets?resource=download

	id	user_name	user_location	user_description	user_created	user_follo	user_frien	user_favo	user_verif	date	text	hashtags	source	retweets	favorites	is_retweet
2	1.34054E+18	Rachel Roh	La Crescenta-Mont	Aggregator of Asian	08-04-2009 17:52	405	1692	3247	FALSE	20-12-2020 06:06	Same folks said d	['PfizerBioNTech	Twitter for Android	0	0	FALSE
3	1.33816E+18	Albert Fong	San Francisco, CA	Marketing dude, tec	21-09-2009 15:27	834	666	178	FALSE	13-12-2020 16:27	While the world has been on the v		Twitter Web App	1	1	FALSE
4	1.33786E+18	eliðŸ‡±ðŸ‡¹ðŸ‡ªðŸ‡ºðŸ'	Your Bed	heil, hydra ðŸ–â˜º	25-06-2020 23:30	10	88	155	FALSE	12-12-2020 20:33	#coronavirus #Sp	['coronavirus', 'S	Twitter for Android	0	0	FALSE
5	1.33786E+18	Charles Adler	Vancouver, BC - Ca	Hosting "CharlesAdl	10-09-2008 11:28	49165	3933	21853	TRUE	12-12-2020 20:23	Facts are immutable, Senator, eve		Twitter Web App	446	2129	FALSE
6	1.33785E+18	Citizen News Channel		Citizen News Chann	23-04-2020 17:58	152	580	1473	FALSE	12-12-2020 20:17	Explain to me aga	['whereareallthe	Twitter for iPhone	0	0	FALSE
7	1.33785E+18	Dee	Birmingham, Engla	Gastroenterology tr	26-01-2020 21:43	105	108	106	FALSE	12-12-2020 20:11	Does anyone have any useful advi		Twitter for iPhone	0	0	FALSE
8	1.33785E+18	Gunther Fehlinger	Austria, Ukraine an	End North Stream 2	10-06-2013 17:49	2731	5001	69344	FALSE	12-12-2020 20:06	it is a bit sad to cl	['vaccination']	Twitter Web App	0	4	FALSE
9	1.33785E+18	Dr.Krutika Kuppalli		ID, Global Health, VI	25-03-2019 04:14	21924	593	7815	TRUE	12-12-2020 20:04	There have not	['BidenHarris', 'E	Twitter for iPhone	2	22	FALSE
10	1.33785E+18	Erin Despas		Designing&selling or	30-10-2009 17:53	887	1515	9639	FALSE	12-12-2020 20:01	Covid vaccine;	['CovidVaccine',	Twitter Web App	2	1	FALSE
11	1.33784E+18	Ch.Amjad Ali	Islamabad	#ProudPakistani	12-11-2012 04:18	671	2368	20469	FALSE	12-12-2020 19:30	#CovidVaccine	['CovidVaccine',	Twitter Web App	0	0	FALSE
12	1.33784E+18	Tamer Yazar	Turkey-Israel	Im Market Analyst,	17-09-2009 16:45	1302	78	339	FALSE	12-12-2020 19:29	while deaths are	['PfizerBioNTech	Twitter Web App	0	0	FALSE
13	1.33784E+18	VoiceM		campaigner & optin	31-08-2020 10:38	2	25	20	FALSE	12-12-2020 19:22	@cnnbrk #COVID	['COVID19', 'Cov	Twitter Web App	0	0	FALSE
14	1.33782E+18	WION	India	#WION: World Is Or	21-03-2016 03:44	292510	91	7531	TRUE	12-12-2020 17:45	The agency also released new info		TweetDeck	0	18	FALSE
15	1.33781E+18	Dr.Krutika Kuppalli		ID, Global Health, VI	25-03-2019 04:14	21924	593	7815	TRUE	12-12-2020 17:19	For all the wome	['PfizerBioNTech	Twitter for iPhone	48	82	FALSE
16	1.33781E+18	Opoyi		High-quality trusted	13-01-2019 18:33	10332	49	16	FALSE	12-12-2020 17:10	"Expect 145 sites across all the sta		TweetDeck	0	0	FALSE
17	1.33779E+18	City A.M.	London, England	London's business n	09-06-2009 13:53	66224	603	771	TRUE	12-12-2020 16:00	Trump	['vaccine']	Twitter for iPhone	0	1	FALSE
18	1.33779E+18	STOPCOMMONPASS.OR	Global	'Trust' is not carte-b	25-10-2020 20:33	406	176	479	FALSE	12-12-2020 15:59	UPDATED:	['YellowFever', '	Twitter Web App	2	2	FALSE
19	1.33778E+18	ILKHA	TÃ¼rkiye	Official Twitter acco	22-05-2015 08:31	4056	6	3	TRUE	12-12-2020 15:38	Coronavirus: Iran	['Iran', 'coronavi	TweetDeck	3	5	FALSE
20	1.33778E+18	Braderz73ðŸŒ¹#GTTO ð	Bristol, UK	One of those lefty	24-07-2012 08:18	6430	6292	45007	FALSE	12-12-2020 15:27	.@Pfizer will rake	['CovidVaccine']	Twitter for Android	3	3	FALSE
21	1.33778E+18	Alex Vie	Los Angeles, CA	Marine vet. Yogi. Kr	24-01-2010 04:43	125	442	5401	FALSE	12-12-2020 15:10	The trump admin	['COVIDIOTS', 'c	Twitter for iPhone	0	0	FALSE
22	1.33777E+18	Mani			10-10-2019 13:41	26	33	2515	FALSE	12-12-2020 15:00	How much did th	['fda', 'vaccine']	Twitter Web App	0	0	FALSE
23	1.33777E+18	Richard Dunne, MD	Rochester, NY	Husband, Girl Dad, G	23-04-2012 12:18	1982	608	9110	FALSE	12-12-2020 14:59	Anyone wonderin	['PfizerBioNTech	Twitter for iPhone	0	2	FALSE
24	1.33777E+18	City A.M.	London, England	London's business n	09-06-2009 13:53	66224	603	771	TRUE	12-12-2020 14:59	Trump	['vaccine']	Buffer	1	0	FALSE
25	1.33777E+18	BOOM Live	Mumbai, India	IFCN certified fact-c	16-03-2014 03:52	64185	1183	1794	TRUE	12-12-2020 14:58	The US Food and Drug Administrat		Twitter Web App	1	5	FALSE
26	1.33777E+18	DOCNOS Official		An Innovative Healt	10-11-2020 16:53	7	0	5	FALSE	12-12-2020 14:46	Presenting you	['docnosofficial'	Twitter Web App	0	3	FALSE
27	1.33777E+18	Devan Surendran	Nottingham, Engla	NHS Doctor | Singer	16-01-2010 23:59	116	86	268	FALSE	12-12-2020 14:43	No.1 of 2 done.	['ThankYouNHS'	Twitter Web App	1	10	FALSE
28	1.33777E+18	Tamer Yazar	Turkey-Israel	Im Market Analyst,	17-09-2009 16:45	1302	78	339	FALSE	12-12-2020 14:42	Wear a mask, wa	['stayhome', 'Sta	Twitter Web App	0	0	FALSE
29	1.33777E+18	Michael Finn	Crete, Greece	Hellenic National He	17-10-2011 19:03	151	235	838	FALSE	12-12-2020 14:41	â¦@AvgerinosM	['PfizerBioNTech	Twitter for iPad	0	0	FALSE
30	1.33777E+18	Emiliano Pacelli	Rome, Italy	#IBMCloud and #AI/	29-07-2019 20:03	75	291	685	FALSE	12-12-2020 14:26	Interesting and ve	['supplychain', 's	Twitter for iPad	0	0	FALSE

vaccination_all_tweets

2. Exploratory Data Analysis

The primary goal of the exploratory data analysis phase of this project was to get acquainted with the columns of the data frame and start brainstorming research questions.

Starting with the step of loading the data using pandas, some basic data frame operations allow us to see that, for each tweet, all of the following information is available:

Information about the user who tweeted

user_name: Twitter handle
user_location: where in the world the person tweets from (NOTE: there is no validation here… "your bed" is technically acceptable)
user_description: user-written biography
user_created: when they created their Twitter account
user_followers: number of followers
user_friends: number of accounts the user is following
user_favourites: number of tweets the user has liked
user_verified: indicates if the user is a well-known figure (boolean)

Information about the tweet itself

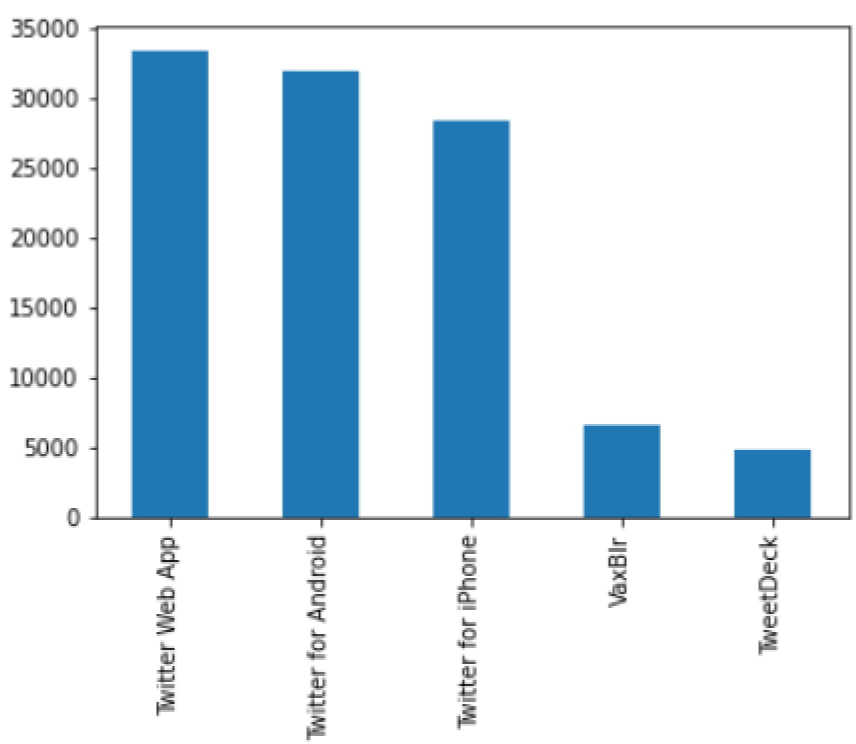
id: indexing value for Twitter API
date: a datetime object in the form of YYYY-MM-DD HH:MM:SS
text: the tweet itself (**MOST IMPORTANT**)
hashtags: list of hashtags used in the tweet (without '#' character)

source: which device was used for the tweet
retweets: number of retweets received at the time the data was collected
favorites: number of likes received at the time the data was collected
is_retweet: indicates if the tweet is original or a retweet (boolean)

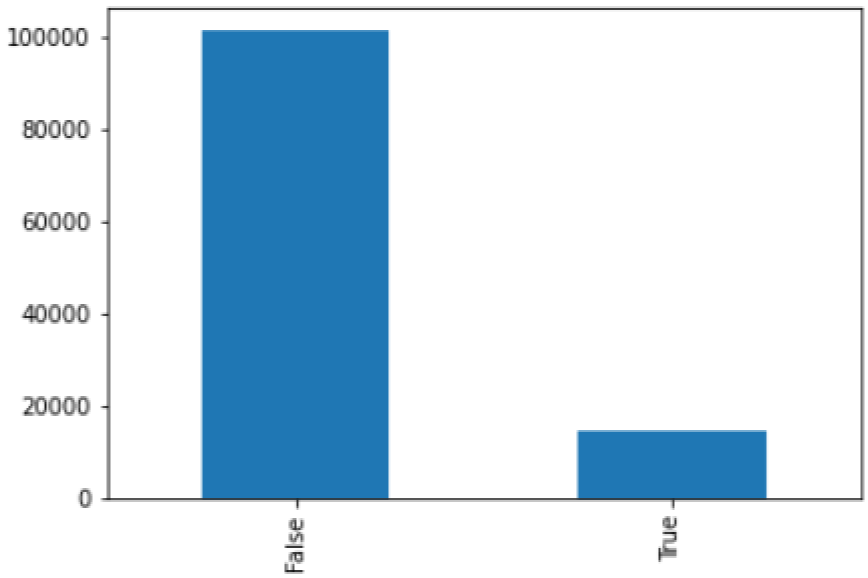
Out of the above columns, text, date, user_name, user_location, hashtags, favorites, and retweets will be most relevant for this analysis. Though the tweets were queried using vaccine-related keywords, the specific vaccine being referred to in the tweet is not explicitly included as a column in the dataset. Therefore, we will need to filter for vaccine references in order to do any comparative analysis.

	user_followers	user_friends	user_favourites	retweets	favorites	polarity	subjectivity	nltk_cmp_score	composite_score
count	1.158490e+05	115849.000000	1.158490e+05	115849.000000	115849.000000	115849.000000	115849.000000	115849.000000	115849.000000
mean	1.594820e+05	1387.750710	1.537447e+04	3.456379	15.505839	0.107977	0.278626	0.125007	0.116492
std	1.115311e+06	6970.271513	4.346030e+04	65.137923	254.309936	0.234791	0.302504	0.349607	0.252114
min	0.000000e+00	0.000000	0.000000e+00	0.000000	0.000000	-1.000000	0.000000	-0.968200	-0.979950
25%	1.140000e+02	126.000000	2.670000e+02	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	6.210000e+02	398.000000	2.185000e+03	0.000000	1.000000	0.000000	0.200000	0.000000	0.025800
75%	3.397000e+03	1149.000000	1.112800e+04	1.000000	3.000000	0.200000	0.500000	0.421500	0.263350
max	1.635305e+07	582461.000000	1.214813e+06	12294.000000	54017.000000	1.000000	1.000000	0.971800	0.973650

Which device are people tweeting about the vaccine from?



User verified?



Top 10 most retweeted tweets

	text	date	user_name	user_location	hashtags	favorites	retweets
221427	This video fits the last almost 2 years into 2 minutes. At #SputnikV we strongly believe that it is only through Va… https://t.co/Ggi7X5qO8x	2021-11-11	Sputnik V	Moscow, Russia	['SputnikV']	54017	12294
68358	RDIF, Laboratorios Richmond launched production of #SputnikV in Argentina, the first country in Latin America to ma… https://t.co/oEMaUwVR92	2021-04-20	Sputnik V	Moscow, Russia	['SputnikV']	25724	11288
46053	Why we need Two Doses of mRNA Vaccine 💉 #vaccines #COVID19 #Pfizer #moderna #VaccinesSaveLives #vaccinated https://t.co/RFRmPAyubD	2021-04-01	hotvickkrishna	Manhattan, NY	['vaccines', 'COVID19', 'Pfizer', 'moderna', 'VaccinesSaveLives', 'vaccinated']	19622	7695
66822	ICMR study shows #COVAXIN neutralises against multiple variants of SARS-CoV-2 and effectively neutralises the doubl… https://t.co/0IYwr0KymJ	2021-04-21	ICMR	New Delhi	['COVAXIN']	11995	4851
76306	#Argentina's actor breaks into a live TV to show his #SputnikV vaccination certificate & express his gratitude. \n\nT… https://t.co/N1NwjkD83y	2021-05-19	Sputnik V	Moscow, Russia	['Argentina', 'SputnikV']	14412	2550
17118	Got my jab. For the curious, it was #Covaxin. \n\nFelt secure, will travel safely. https://t.co/8PL7PZMEsf	2021-03-01	Dr. S. Jaishankar	New Delhi, India	['Covaxin']	22815	2360
53045	I see it's going around with signature cropped....so here is the original:) #covid 19 #vaccine #pfizer #moderna… https://t.co/eoqT74V78A	2021-04-12	dawnymock	Fredericton New Brunswick	['covid', 'vaccine', 'pfizer', 'moderna']	10175	2299
7126	New research published in Microbiology & Infectious Diseases, immunologist J. Bart Classen warns #mRNA technology u… https://t.co/OWUTf5ShHO	2021-02-10	Robert F. Kennedy Jr	Los Angles, California	['mRNA']	3090	2247
24268	#Covaxin ɪɴ , made by Hyderabad-based Bharat Biotech International Limited, has been declared "Safe, Immunogenic wi… https://t.co/FAUOEHJmAw	2021-03-09	Megh Updates 🚨	Turn on Notification 🔔	['Covaxin']	9458	2095
32826	A batch of fake Sputnik V vaccines was confiscated in Mexico. See this comparison of the genuine #SputnikV with a f… https://t.co/J7PxMq2e1M	2021-03-18	Sputnik V	Moscow, Russia	['SputnikV']	3473	1980

3. Preprocessing Data

Data cleaning involves transforming the raw data into a form that is more understandable, useful and efficient. Cleaned data can be easily interpreted by our machine learning algorithm. It is one of the most crucial steps in any machine learning model since it impacts the success and accuracy of our model. This process involves removing the duplicates, the redundant data and the outliers. It improves the quality of our dataset, helps in making accurate predictions and thereby, increases the overall accuracy of our model.
The extracted data in the CSV file is in raw form, so we need to clean and preprocess the data before training our model.

Major cleaning tasks we have performed include:

- We have dropped the ID column as it does not help in our analysis.
- Removing duplicate tweets.
- Strip each tweet of mentions, hashtags, retweet information, and links using regular expressions.

4. Sentiment analysis

- Sentiment Analysis with TextBlob

Pivoting to the sentiment analysis portion of this work, we can take this intuition of some of the tweets being informative and some of the tweets being opinionated to partition the greater discourse into separate sets of tweets with similar quantitative features. These features can be obtained using a Python package called TextBlob, which provides an API for NLP tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.
TextBlob returns polarity and subjectivity of a sentence. Polarity lies between [-1,1], -1 defines a negative sentiment and 1 defines a positive sentiment. Negation words reverse the polarity. TextBlob has semantic labels that help with fine-grained analysis. For example — emoticons, exclamation mark, emojis, etc. Subjectivity lies between [0,1]. Subjectivity quantifies the amount of personal opinion and factual information contained in the text. The higher subjectivity means that the text contains personal opinion rather than factual information. TextBlob has one more parameter — intensity. TextBlob calculates subjectivity by looking at the 'intensity'. Intensity determines if a word modifies the next word. For English, adverbs are used as modifiers ('very good').

TextBlob is not context-aware, so the scores returned from its API should be interpreted loosely, in a way that prompts further analysis.

- <u>Sentiment Analysis with NLTK Vader</u>

Valence Aware Dictionary and sEntiment Reasoner (Vader) model, which is a lexicon and rule-based sentiment analysis tool aimed at sentiment analysis of social media text. It uses a bag of words approach with simple heuristics (e.g. increasing sentiment intensity in presence of certain words like "very").

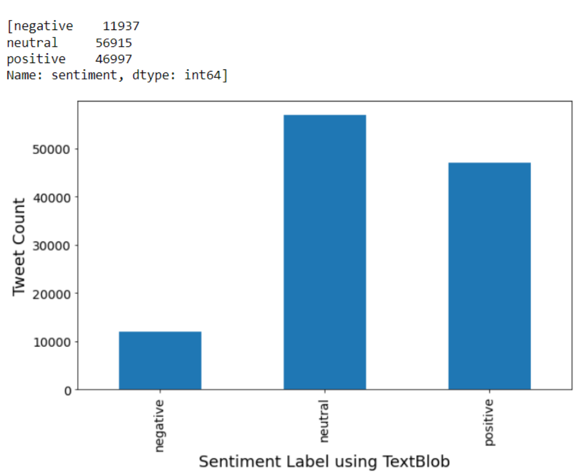
Vader returns compound scores, which are single unidimensional sentiment measures for a given text. The score ranges from -1 (most negative) to +1 (most positive), and the score for neutral sentiment is set arbitrarily between -0.05 and 0.05. We have chosen neutral threshold to be 0.01.

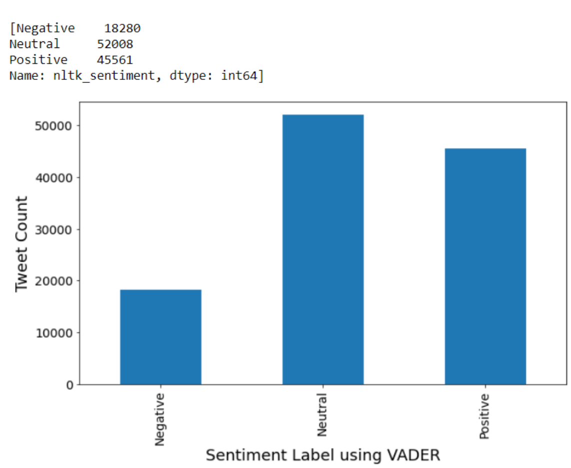
- <u>Composite sentiment with ensemble method -</u>

Ensemble method is used to create a composite score from the average of different sentiment scores. Both NLTK Vader and TextBlob scores were already in the range [-1,1], and were given equal weights when computing the mean.
As with previous steps, score ≥0.05 was defined as Positive, score ≤-0.05 was defined as Negative, and anything in between was set as Neutral.

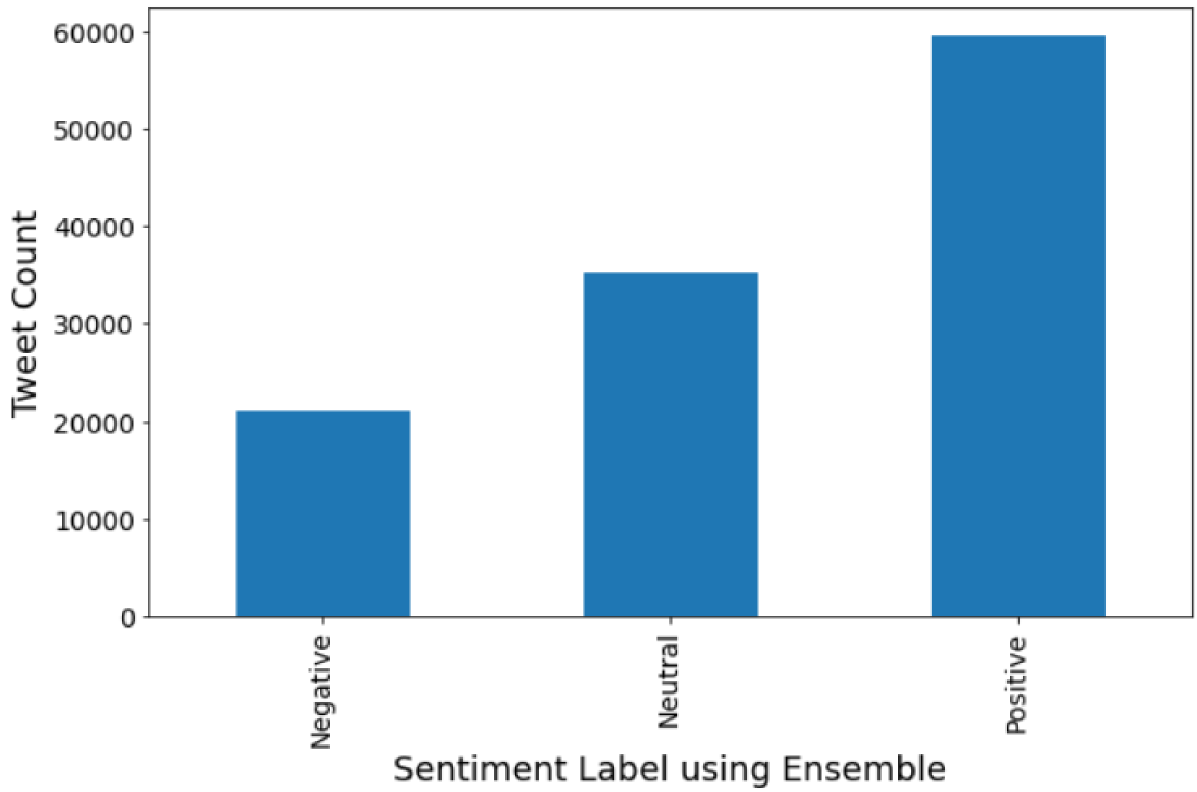
Results

We were looking for some practically relevant output from this project and this is what we found.





```
[Negative     21046
Neutral      35216
Positive     59587
Name: composite_vote_2, dtype: int64]
```



Conclusion/Findings

Our limited analysis indicates that the proportion of positive sentiments (51.4%) is greater that of negative sentiments (18.7%). This suggests that the general sentiment towards COVID-19 vaccine at the point of analysis tends to be on a positive side which indicates that people were willing to get vaccinated.