SENTIMENTAL ANALYSIS OF COVID 19 VACCINATION ON TWITTER DATA

Abstract

Delving into the Twitterverse, where millions of voices converge in the succinct language of tweets, this study harnesses the power of sentiment analysis and natural language processing to unveil the public's stance on COVID-19 vaccinations.

With a treasure trove of over ten thousand tweets spanning from January to April 2021, our research journey begins. By employing cutting-edge techniques like counting and text summarization, we uncover the prevailing themes within this sea of tweets and transform them into digestible narratives.

But what makes this study truly groundbreaking is its real-time applicability. This analytical pipeline, custom-crafted for Twitter's fast-paced world, offers healthcare experts a unique window into public reactions to health interventions as they unfold.

To paint this sentiment-driven landscape, we kick-start with Twitter's API and authentication tokens, sifting through raw tweets with NLP wizardry. Then, the magic of sentiment analysis steps in, categorizing tweets as positive, negative, or neutral, enabling us to gauge the collective emotional pulse.

From these, we unearth the key themes, summarizing them for a bird's-eye view of the overarching conversations. It's like painting a masterpiece, one stroke at a time, except our canvas is a million tweets, and our brush is an algorithm.

In a nutshell, this study is a deep dive into the world of tweets, translating them into insights, and offering a real-time radar for the health community to navigate the turbulent waters of public opinion.

Keywords: COVID-19, Tweets, Dataset, NLP , Vaccine, Sentiment Analysis, Public Health , India, The United States of America.

Introduction

Under the umbrella of computational linguistics and data mining, the field of sentiment analysis is referred to as opinion mining.

Its fundamental goal is to infer from text documents the person's attitude, behaviour, and opinion. As social networking sites have been more widely used, sentiment analysis approaches have begun to leverage the open data on these sites to conduct sentiment analysis research in a variety of sociological fields, including politics, sociology, the economy, and finance.

This kind of unstructured data makes for around 80% of all data worldwide. As a result, it is challenging to evaluate and make insightful decisions using such data. The key method for identifying people's thoughts in social media data is sentiment analysis, often known as opinion mining.

Following the stimulation of social media activity reported to accompany disease outbreak occurrences, a body of work has arisen over the last decade that particularly examines how trends in online activity and discourse might assist inform epidemiological models. In addition, a set of computer frameworks and

models based on Twitter data have been developed to answer particular research questions about viral phenomena and their societal implications.

To that purpose, we create a unique pipeline that uses cutting-edge natural language processing (NLP) tools to thoroughly assess Twitter debate about and public views regarding vaccinations throughout the Covid-19 period.

In machine learning, ensemble methods aggregate the effect of numerous machine learning algorithms on a particular problem set to produce a stronger prediction capacity than its constituent algorithms alone. NLP was used to preprocess the raw data. The semantic structure of the document was determined using topic modelling.

The analysed text data is transformed to polarity and subjectivity for classification.

Literature survey

Amir Hussain, Ahsen Tahir, Surveys conducted in the UK and US revealed varying sentiments towards COVID-19 vaccinations. The UK had 58% positive sentiment, while the US had 56%. The study suggests that AI-enabled social-media analysis could help assess public confidence and trust in vaccinations, addressing vaccine-sceptic concerns. [1]

Healthcare professionals in the UK and US are experiencing increased emotional display during the COVID-19 pandemic, highlighting the importance of understanding their experiences and emotions on Twitter to improve public health responses during emergencies.[2]

Twitter has been utilized by leaders in the global health response to COVID-19, with 88.9% of G7 leaders having verified accounts. These leaders communicated public health information, with a preference for tweets containing official government-based sources.[3]

WATA analysis reveals important discussions on vaccination programs in public health, revealing differences in the importance of non-government scientific experts and university presidents in vaccination discussions across eight countries. [4]

Twitter, a popular social media platform, has been instrumental in promoting global vaccines, particularly those from homegrown sources. A study by Zahra Bokaee Nezhad analyzed Persian tweets to compare Iranian views on homegrown and imported vaccines, finding a subtle difference in positive sentiments. The World Health Organization suggests promoting positive messaging on Twitter to counter opposing views.[5]

Twitter, a global social media platform, plays a crucial role in spreading news and discussing ideas. This paper explores topic identification and sentiment analysis in Brazil and the USA, comparing English and Portuguese tweets to understand public reactions and consensus building during the COVID-19 pandemic.[6]

Twitter data reveals public perceptions and attitudes towards nonpharmaceutical interventions (NPIs) in six countries: Australia, Canada, New Zealand, Ireland, the United Kingdom, and the US. The study reveals that New Zealand and the US have different levels of attention to NPIs, with less restrictive regimes receiving more support. Understanding these perceptions can help inform government decision-making and communication strategies according to study of Caitlin Doogan.[7]

Twitter, a global platform, has been used to analyze the evolution of emotions related to COVID-19 vaccinations in countries like India, the United States, Brazil, the United Kingdom, and Australia. The

study found that hesitancy towards vaccines was the most prevalent, with negative emotions like rage and sorrow being the most significant.[8]
Twitter has played a crucial role in expressing emotions during the COVID-19 pandemic, revealing the devastating effect of the virus on people's minds. This research used deep learning techniques to understand sentiments about vaccines, revealing a diverse range of responses, highlighting the importance of understanding public opinions on vaccines.[9]
Twitter data from the US, UK, and India revealed proximal and distal defenses during the COVID-19 pandemic. Latent Dirichlet Allocation analysis revealed cultural differences in defenses, with implications for public health practitioners and social media platform managers.[10]
Using a vast analysis of Twitter messages collected during the COVID-19 pandemic in Europe, researchers found that lockdown announcements correlated with a deterioration in moods across all countries.[11]
Twitter users in the United Kingdom have expressed mixed attitudes towards remote care delivery during the COVID-19 pandemic. The study analyzed Twitter content from January 2018 to October 2020, revealing a surge in discussions about remote care delivery. The study highlights the importance of patients having a choice over consultation types and ensuring technology does not become a barrier. The mixed attitudes highlight the need for continued examination of people's preferences in remote care delivery.[12]
Twitter's widespread use of COVID-19 tweets has led to the spread of uncontrolled conspiracy theories and propaganda, according to the WHO. This infodemic has caused psychological panic, misleading medical advice, and economic disruption. The study highlights the importance of using credible users for effective social media management during crises.[13]
The U.K. and U.S. have strong social media presence, with political and health sources dominating posts. The U.K.'s Twitter accounts show a strong elite orientation, while the U.S.'s social media accounts show a diverse range of sources. [14]

The COVID-19 pandemic has significantly impacted public sentiment in the UK, USA, and India. Using data science methods, the study found that governments and policymakers remained the primary focus of public discussion on Twitter, with the USA showing the most neutral sentiment. [15]
The research work analyzed sentiment in tweets from twelve countries related to COVID-19, revealing that while most people are positive, fear, sadness, and disgust are prevalent. Four countries, France, Switzerland, Netherlands, and the USA, displayed greater distrust and anger.[16]
People's sentiment towards COVID-19 vaccines was analyzed using VADER on Twitter API in the UK and US. Results showed celebrities influence opinion shifts, with 40% of the population having negative attitudes. Pfizer vaccine was most popular, but side effects and safety concerns were also noted.[17]

Transfer learning was utilized to analyze over 2 million tweets during the COVID-19 pandemic, providing insights into emotional wellbeing. The study utilized the Robustly Optimized BERT Pretraining Approach (RoBERTa) and a Reddit-based standard Emotion Dataset. The results improved existing AI-based emotion classification methods, offering valuable insights for effective pandemic management strategies and predictive strategies for future shocks.[18]
The United Kingdom's social media sentiments towards COVID-19 vaccines were analyzed using artificial intelligence. The study found that 58% of UK citizens had positive sentiments, while 22% had

negative sentiments. The findings suggest that AI-enabled social media analysis could help assess public confidence in vaccines, address vaccine skeptics, and develop effective policies.[19]
The global epidemiological research is utilizing a vast data source to understand the social dynamics of the COVID-19 pandemic, enabling researchers to conduct diverse studies on emotional responses, misinformation sources, and sentiment measurement in real-time.[20]

System design

The suggested system collects data from Twitter and uses NLP techniques to frame the feature vectors. The training model is then fonned using ensemble methods applied to the training data. Following training, the extracted feature vectors are categorised using the training model, and the results are reported.

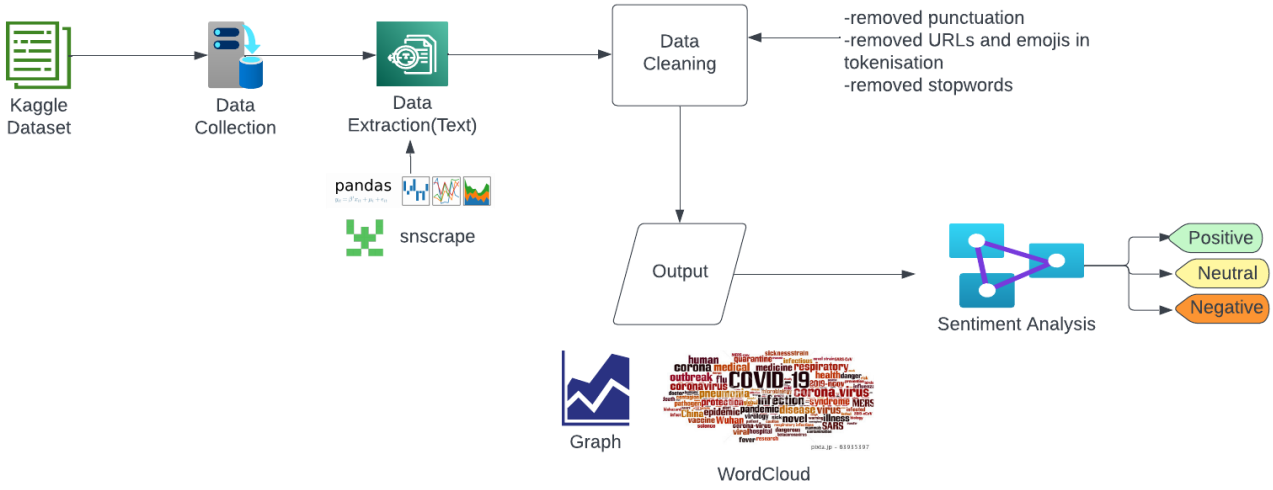

Fig 1. Architectural diagram

NLP and VADER

Sentiment analysis is a critical area of natural language processing (NLP) that categorises opinions conveyed in text based on their polarity (e.g., positive, negative, or neutral).

A. Lexicon

A lexicon is a man's, dialect, or a department of expertise's vocabulary that contains the lexemes in that linguistics. Polarity lexicons are collections of words with varying degrees of polarity. It is one of the most important resources for computerised sentiment and opinion analysis in texts. Building polarity lexicons can be accomplished in three ways: interpreting existing lexicons from other languages, extracting polarity lexicons from corpora, and annotating sentiments Lexical Knowledge Base. There are well-known hand generated lexicons for major languages, such as General Inquirer, OpinionFinder, SO-CAL, and others. The authors of [18] and [19] investigated the methodology of translating English resources into Romanian and Spanish, respectively. There are several English polarity lexicons available online such as SentiWordNet1 , VADER etc.

B. English Grammar

The English language is rich with important sentiment analysis works, such as VADER. Researchers used a combination of qualitative and quantitative methodologies to create a gold-standard sentiment lexicon, which was then empirically validated against especially receptive microblog scenarios. VADER integrates essential lexical elements derived from five generalised rules that embody human speech grammatical and syntactical patterns. It also preserves the benefits of traditional sentiment lexicons like LIWC. The author investigated three methods for developing polarity lexicons: interpreting existing lexicons from other languages, explaining sentiments from lexical knowledge bases, and extracting polarity lexicons from corpora.Different levels of human effort are needed for each model. The Elhuyar SpanishBasque2 dictionary, which offers five interpretations for each Spanish word, was used to translate the vocabulary in this case.

C. VADER

VADER produces a gold standard sentiment lexicon by combining qualitative and quantitative methods.Basic sentiment (or opinion) lexicons are heavily used by many sentiment analysis techniques.A sentiment lexicon is a listing of lexical capabilities (e.g. words) which can be commonly categorised with their semantic orientation as either positive or negative . After creating and testing a lexicon, they assessed the sentiment of texts using the architecture depicted in Figure.The architecture (Fig 2) mostly uses the procedures outlined below to determine the polarity of the sentences:

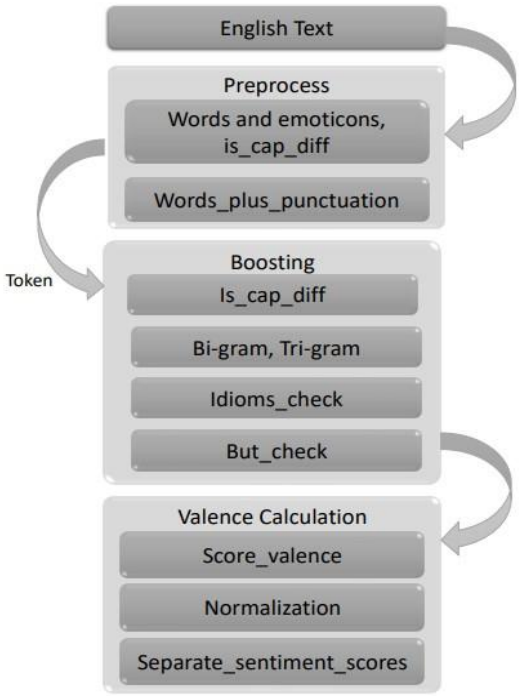

Fig 2. Implementation architecture of English VADER

1) Preprocessing: Tokenization of the English text input occurs during preprocessing. To begin with, all words with capital letters, emojis, and words are tokenized. Words and punctuation are tokenized later. Some punctuation changes a word's valence, which is retained with the term.

2) Boosting: All tokens are examined for valence boosting purposes after tokenization is complete. For boosting, VADER employs the bigram and trigram models. When a boosting word like "extremely, very, great" is identified, the word's valence is boosted. The valence is boosted further by taking into account all uppercase letters. The text is further examined for idioms and phrases; if any are discovered, the valence gets boosted once again. When the word "but" is detected, the sentence is split into two pieces and the valence is calculated for each of them. The overall valence of such a statement is then determined.

3) Valence Calculation: A sentence's valence, which ranges from -4 to +4, is calculated in this stage. It is further normalised to have a value between -1 and +1. Every statement is given its appropriate polarity in this way.

Mathematical Formulas used in VADER and its implementation

The VADER sentiment analysis algorithm assigns sentiment scores to individual words and phrases based on a predefined sentiment lexicon (Fig 3.2). Each word and phrase receive a polarity (positive, negative, or neutral) score and a magnitude (intensity) score.
The overall polarity score (compound score) is calculated as follows:
Polarity Score = sum of (word_polarity_score * word_magnitude_score) / sum of word_magnitude_scores
In this formula:
word_polarity_score is the polarity score of an individual word or phrase.
word_magnitude_score is the magnitude (intensity) score of the same word or phrase.
The resulting polarity score can range from -1 to 1:
A score near 1 indicates a highly positive sentiment.
A score near -1 indicates a highly negative sentiment.
A score near 0 indicates a neutral sentiment.

Example-1: "Received first dose covid-19 vaccine, highly appreciated efforts"
here are the word-level sentiment scores:

"Received": 0.3182 (positive)
"first": 0.0 (neutral)
"dose": 0.0 (neutral)
"Covid-19": -0.296 (negative)
"vaccine": 0.2263 (positive)
"Highly": 0.4215 (positive)
"appreciated": 0.6486 (positive)
"efforts": 0.4404 (positive)
To calculate the compound score, you sum all these scores and then normalize the result. The formula for the compound score is:

Compound Score = (Sum of individual word scores) / (number of words)

So, in this case:

Compound Score = (0.3182 + 0.0 + 0.0 - 0.296 + 0.2263 + 0.4215 + 0.6486 + 0.4404) / 8

Compound Score = 2.159 / 8

Compound Score ≈ 0.2699

So, the corrected compound sentiment score for the text "Received first dose covid-19 vaccine, highly appreciated efforts" using VADER is approximately 0.2699, indicating a slightly positive sentiment.

Example-2: India’s Covid-19 cases rising which is alarming

Here are the word-by-word sentiment scores breakdown for this text:

"Indias’s": 0.0 (neutral)

"Covid-19": -0.296 (negative)

"cases": -0.3182 (negative)

"rising": -0.2263 (negative)

"which": 0.0 (neutral)

"is": 0.0 (neutral)

"alarming": -0.5859 (negative)

To calculate the compound score, we sum these individual word scores and normalize the result using the formula:

Compound Score = (Sum of individual word scores) / (number of words)

Implementing this in the example

Compound Score = (0.0 - 0.296 - 0.3182 - 0.2263 + 0.0 + 0.0 - 0.5859) / 7

Compound Score = (-1.4264) / 7

Compound Score ≈ -0.2038

The VADER compound sentiment score for the text "India’s Covid-19 cases rising which is alarming" is approximately -0.2038, which proves that it is a negative sentiment. This sentiment reflects concerns about the rising Covid-19 cases in India and the alarming situation.

Algorithms and Functions

A. VADER (Valence Aware Dictionary and sEtiment Reasoner), part of the NLTK function library is a rule-based sentiment analysis tool it is used to perform sentiment analysis on tweets by assigning a sentiment score (positive, negative, or neutral) to each tweet designed to understand the sentiment of text (e.g. a sentence or tweet) by evaluating the heavy use of words in the text. It uses a dictionary of words and their associated scores to determine the overall meaning of the text.

Key Terms :

i. Dictionary Definition: VADER uses a word dictionary where each word is scored from -4 to +4. For example, "good " will have more happiness, while " bad" will have more unhappiness.

ii.

iii. Emotional Intensity Calculation: The emotional intensity of a word is determined by taking into account various factors such as capitalization (for example, the emotionally of "BIG " still means " good"), exclamation marks, degree, etc. modifiers (like "very good ") and negators (like "bad").

iv. Valence score of a sentence: The valence score of a sentence is calculated by summing the scores of individual words, including the intensity and polarity of each word.

v. Final Sentiment Classification: The final sentiment of the text (positive, negative or neutral) is determined by the valence score and some biases.(Fig 4)

vi. Implementation in the given code: In the given code, VADER is used to evaluate the views of tweets. It calculates the sentiment score for each tweet and classifies them as positive, negative or neutral based on this score.(Fig 3.1)

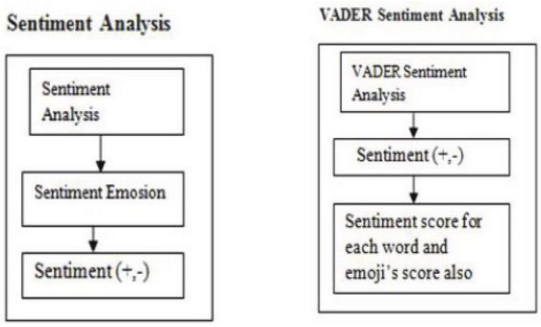

Fig 3.1 difference between sentiment analysis and vader's analysis[24]

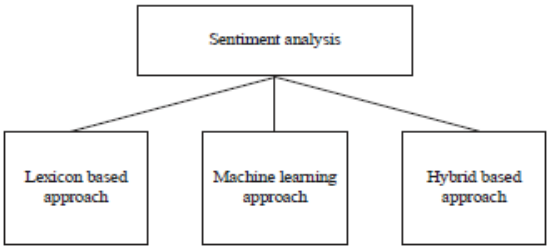

Fig 3.2 types of approaches towards sentiment analysis[25]

A. The Counter algorithm in Python is a part of the collections module and is used to count the occurrences of elements in a collection, commonly used to count word frequencies in a text corpus. This analysis identifies the most frequently used words in tweets

Key Terms:

i. Key Characteristics: The Counter algorithm utilizes a hash table to efficiently count the occurrences of each element (words, in this case) in a collection. It provides a convenient and efficient way to calculate the frequency of items in a dataset.

ii. Usage and Application : In the provided code, the Counter algorithm is used to count the frequency of words in the processed tweet text. This frequency analysis helps in identifying the most common words used in the tweets.

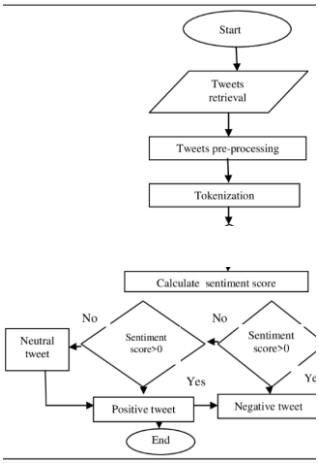

Fig 4 process of sentiment analysis[21]

STEPS INVOLVED/METHODOLOGY

1. Problem Definition and Objective

Scope and objective:

Analyse sentiment in tweets related to a specific topic (e.g., COVID-19) to understand public opinion. Identify the most common words used in the tweets for insights into the discussion.

2. Data Collection

Collect the necessary data for analysis:

Obtain a dataset containing tweets related to the chosen topic (e.g., COVID-19 tweets) from reliable sources(kaggle[22]) or APIs.

We utilised a comprehensive database/dataset obtained from the Kaggle platform. Kaggle, renowned for its extensive collection of datasets, offers a wide array of data for users to utilise. From Kaggle, we acquired datasets pertaining to both the United States and India, enabling a comparative analysis of these distinct yet rapidly advancing nations.

3. Data Preprocessing

Clean and prepare the data for analysis:

Load the dataset into a structured format (e.g., a DataFrame in Python).

Extract relevant columns (e.g., 'user_id', 'username', 'date', 'tweet')(refer table 1.1 and 1.2).

We've employed cat codes for the purpose of encrypting Twitter user usernames to ensure their privacy.(Table 2.1 and 2.2)

Once the entire dataset of 39 columns and approximately 4000 entries has been loaded, we extract the relevant columns that will be used for subsequent processing.We perform feature selection on the loaded dataset, extracting the relevant subset of columns that will be used for downstream machine learning tasks.

Sn0.	user_id	username	date	tweet
0	1386063854	SaifUllah_Dogar	2021-03-16 18:09:39	Today visited Expo Center with my mother for h...
1	2160285660	DrRupani	2021-03-16 15:08:32	@anjanaomkashyap is it possible to get data n...
2	117822932	V_with_RG	2021-03-16 13:42:40	@StayingReal0511 More alarmingly, over 12,000 ...
3	117822932	V_with_RG	2021-03-16 13:42:30	@StayingReal0511 More alarmingly, over 12,000 ...
4	7639837012438548 49	IraAtreya	2021-03-16 12:37:52	My mother’s b.p shot up to 180/90 for 25 days ...

Table 1.1:Extracted columns (In India)

S.No.	User_name	date	tweet
0	□□☻ɻ₡₮	2020-07-25 12:27:21	If I smelled the scent of hand sanitizers toda...
1	Tom Basile □□	2020-07-25 12:27:17	Hey @Yankees @YankeesPR and @MLB - wouldn't it...
2	Time4fisticuffs	2020-07-25 12:27:14	@diane3443 @wdunlap @realDonaldTrump Trump nev...
3	ethel mertz	2020-07-25 12:27:10	@brookbanktv The one gift #COVID19 has give me...
4	DIPR-J&K	2020-07-25 12:27:08	25 July : Media Bulletin on Novel #CoronaVirus...

Table 1.2:Extracted columns (In USA)

Sno	user_iD	username	date	tweet
0	1386063854	617	2021-03-16	Today visited Expo Center with my mother for h...
1	2160285660	226	2021-03-16	@anjanaomkashyap is it possible to get data n...
2	117822932	747	2021-03-16	@StayingReal0511 More alarmingly, over 12,000 ...
3	117822932	747	2021-03-16	@StayingReal0511 More alarmingly, over 12,000 ...
4	763983701243854849	318	2021-03-16	My mother’s b.p shot up to 180/90 for 25 days ...

Table 2.1 :Depicts usernames in cat codes (In India)

S.No.	User_name	date	text
0	89755	2020-07-25	If I smelled the scent of hand sanitizers toda...
1	76403	2020-07-25	Hey @Yankees @YankeesPR and @MLB - wouldn't it...
2	76147	2020-07-25	@diane3443 @wdunlap @realDonaldTrump Trump nev...
3	84572	2020-07-25	@brookbanktv The one gift #COVID19 has give me...

4	**18398**	**2020-07-25**	**25 July : Media Bulletin on Novel #CoronaVirus...**

Table 2.2:Depicts usernames in cat codes (In USA)

4. Text Preprocessing

Process tweet text to facilitate analysis:

I. Convert text to lowercase for standardisation (Table 3.1 and 3.2).
Converting the tweet text to lowercase is a common preprocessing step for sentiment analysis using VADER. This is because VADER uses a lexicon of words that are associated with positive, negative, and neutral sentiment. Some of these words have different meanings depending on their capitalization (e.g., "LOVE" and "love"). By converting all words to lowercase, we can ensure that VADER is using the same meaning for each word, regardless of its capitalization.

0	**today visited expo center mother first dose va...**
1	**anjanaomkashyap possible get data n effectiven...**
2	**stayingreal0511 alarmingly 12000 people receiv...**
3	**mother’s bp shot 18090 25 days first dose covi...**
4	**vicustomercare vodaideanews check screenshot d.**

Table 3.1:Converting text to lowercase (In India)

0	**if i smelled the scent of hand sanitizers toda..**
1	**hey @yankees @yankeespr and @mlb - wouldn't it...**
2	**@diane3443 @wdunlap @realdonaldtrump trump nev...**
3	**@brookbanktv the one gift #covid19 has give me...**
4	**25 july : media bulletin on novel #coronavirus...**

Table 3.2:Converting text to lowercase (In USA)

II. Remove URLs(refer Table 4.1 ,4.2) and punctuation (Table 4.3 and 4.4).
This is because URLs and punctuations do not typically contain sentiment-related information. In fact, they can sometimes add noise to the data and make it more difficult for VADER to accurately assess the sentiment of the text.

Sno	Tweets (removed urls)
0	**Today visited Expo Center with my mother for h..**
1	**@anjanaomkashyap is it possible to get data n..**
2	**@StayingReal0511 More alarmingly, over 12,000 ...**
3	**My mother’s b.p shot up to 180/90 for 25 days ... ...**

4	**@ViCustomerCare @VodaIdea_NEWS Check this scre... 3027**
5	**Vehicles should be heavily fined for stoppi...**

Table 4.1:urls removal (In India)

Sno	Tweets (removed urls)
0	**If I smelled the scent of hand sanitizers toda...**
1	**Hey @Yankees @YankeesPR and @MLB - wouldn't it..**
2	**@diane3443 @wdunlap @realDonaldTrump Trump nev..**
3	**@brookbanktv The one gift #COVID19 has give me.**
4	**25 July : Media Bulletin on Novel #CoronaVirus...**

Table 4.2:urls removal (In USA)

Sno	Tweet(punctuation removal)
0	**today visited expo center with my mother for h..**
1	**anjanaomkashyap is it possible to get data n ..**
2	**stayingreal0511 more alarmingly over 12000 peo..**
3	**my mother’s bp shot up to 18090 for 25 days af...**
4	**vicustomercare vodaideanews check this screens...**

Table 4.3 :punctuation removal (In India)

Sno	Tweet(punctuation removal)
0	**if i smelled the scent of hand sanitizers toda...**
1	**hey yankees yankeespr and mlb wouldnt it have...**
2	**diane3443 wdunlap realdonaldtrump trump never ...**
3	**brookbanktv the one gift covid19 has give me i..**
4	**25 july media bulletin on novel coronavirusup...**

Table 4.4 :punctuation removal (In USA)

III. Handle stopwords and perform word tokenization.

Stop words are common words that do not add much meaning to a sentence, such as "the", "is", and "of". Removing stop words can help to reduce the noise in the data and make it easier for VADER to focus on the most important words. Tokenization is the process of splitting a text string into individual words or tokens. This is necessary for VADER to be able to process the text and identify the sentiment of each word(table 5.1.1,5.1.2 and 5.2.1 and 5.2.2)

0	**today visited expo center mother first dose va...**
1	**anjanaomkashyap possible get data n effectiven...**
2	**stayingreal0511 alarmingly 12000 people receiv...**
3	**stayingreal0511 alarmingly 12000 people receiv... 4 mother's bp shot 18090 25 days first dose cov..**
i... ...	**... ...**
3024	**ashwinravi99 bcci pls request chennai crowd we...**
3025	**□□□□□□□□□□□□□ today date13022021 testing power...**

Table 5.1.1: stopwords removal(In India)

['today','visited','expo','center','mother','first','dose','vaccination','extremely','helpful','staff','superb','arrangements','kudos','captainusman','dclahore','proving','public','sector','efficient','deliver','intent','right','people','lead','initiative','□□','anjanaomkashyap','possible','get','data','n','effectiveness','vaccine','1','many','developed'....

Table 5.1.2: tokenized words (In India)

0	**smelled scent hand sanitizers today someone pa...**
1	**hey yankees yankeespr mlb wouldnt made sense p...**
2	**diane3443 wdunlap realdonaldtrump trump never ..**
3	**brookbanktv one gift covid19 give appreciation..**

... ...	... ...
109235	**achieve long terms goals health wellness women...**
109236	**sarscov2 type coronavirus group viruses corona..**

Fig 5.2.1: stopwords removal(In USA)

['smelled', 'scent', 'hand', 'sanitizers', 'today', 'someone', 'past', 'would', 'think',
'intoxicated', 'that…', 'hey', 'yankees', 'yankeespr', 'mlb', 'wouldnt', 'made', 'sense', 'players', 'pay',
'respects', 'a…',
'diane3443', 'wdunlap', 'realdonaldtrump', 'trump', 'never', 'claimed', 'covid19', 'hoax', 'claim', 'effort',
'to…', 'brookbanktv', 'one', 'gift', 'covid19', 'give'....

Fig 5.2.2: tokenized words(In USA)

5. Sentiment Analysis with VADER

I. Analyse sentiment in the preprocessed tweet text using VADER
II. Utilise the VADER lexicon and rules to assign sentiment scores to each tweet.

VADER is a Python library that uses a predefined lexicon of words and scores to assign a sentiment score to each tweet. This lexicon, called the AFINN list, contains millions of words and their associated sentiment scores. When VADER analyzes a tweet, it first tokenizes the text and then looks up each word in the AFINN list. If a word is found in the lexicon, VADER assigns the corresponding sentiment score to the word. VADER then calculates the overall sentiment score for the tweet by taking an average of the sentiment scores of all the words in the tweet.
Categorize tweets into positive, negative, or neutral based on the sentiment scores (Table 6.1 and 6.2).

VADER gives a sentiment score to each tweet, which is a number between -1 and 1, where -1 is the most negative score and 1 is the most positive score. A score of 0 indicates a neutral sentiment.

To label a tweet as positive, negative, or neutral based on the VADER sentiment score, we can use the following criteria:

- **Positive:** Score > 0
- **Negative:** Score < 0
- **Neutral:** Score = 0

Sno.	neg	neu	pos	compound	label
0	**0.000**	**0.629**	**0.371**	**0.9268**	**Positive**
1	**0.000**	**1.000**	**0.000**	**0.0000**	**Neutral**
2	**0.222**	**0.556**	**0.222**	**0.0000**	**Neutral**

3	**0.222**	**0.556**	**0.222**	**0.0000**	**Neutral**
4	**0.000**	**0.831**	**0.169**	**0.6705**	**Positive**
…	**...**	**…**	**…**	**…**	**…**
12	**0.000**	**1.000**	**0.000**	**0.0000**	**Neutral**
13	**0.000**	**0.518**	**0.482**	**0.8519**	**Positive**
14	**0.118**	**0.798**	**0.084**	**0.2263**	**Negative**

Table 6.1: Categorizing tweets (In India)

S.no.	neg	neu	pos	compound	label
0	**0.0**	**0.758**	**0.242**	**0.4939**	**positive**
1	**0.11**	**0.709**	**0.181**	**0.2263**	**positive**
2	**0.0**	**0.846**	**0.154**	**0.2057**	**positive**
3	**0.0**	**0.592**	**0.408**	**0.7351**	**positive**
4	**0.0**	**0.813**	**0.187**	**0.3182**	**positive**
…	**….**	**…**	**……**	**……**	**…..**
12	**0.0**	**0.429**	**0.571**	**0.7906**	**positive**
13	**0.0**	**0.776**	**0.224**	**0.3818**	**positive**
14	**0.0**	**1.0**	**0.0**	**0.0**	**neutral**

Table 6.2: Categorizing tweets (In USA)

6. Word Frequency Analysis

I. Analyze word frequency in the preprocessed tweet text:

II. Use the Counter algorithm to count the frequency of each word(Fig 5.1.2,5.2.2).

Identify the most common words and their occurrences in the tweets(Fig 5.1.1,5.2.1).

We are applying the most_common() function of the counter object to get the most common words, along with their frequencies.We can employ the counter algorithm to check the frequence of words in a corpus of tweets, with a focus on the most constantly repeated words. These words can also be imaged in a variety of ways, similar as through bar maps or word shadows

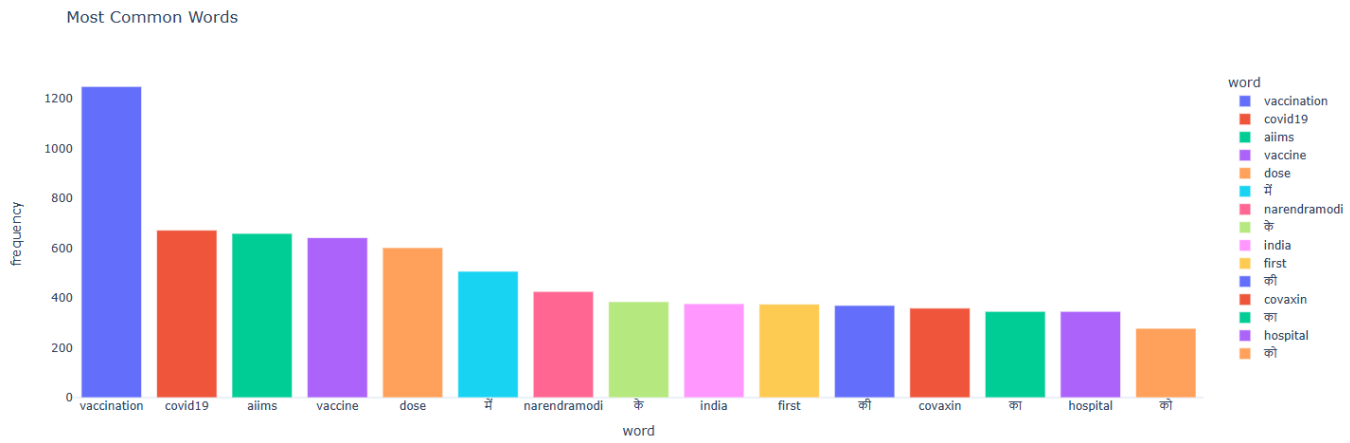

Fig 5.1.1 Most common words used (In India)

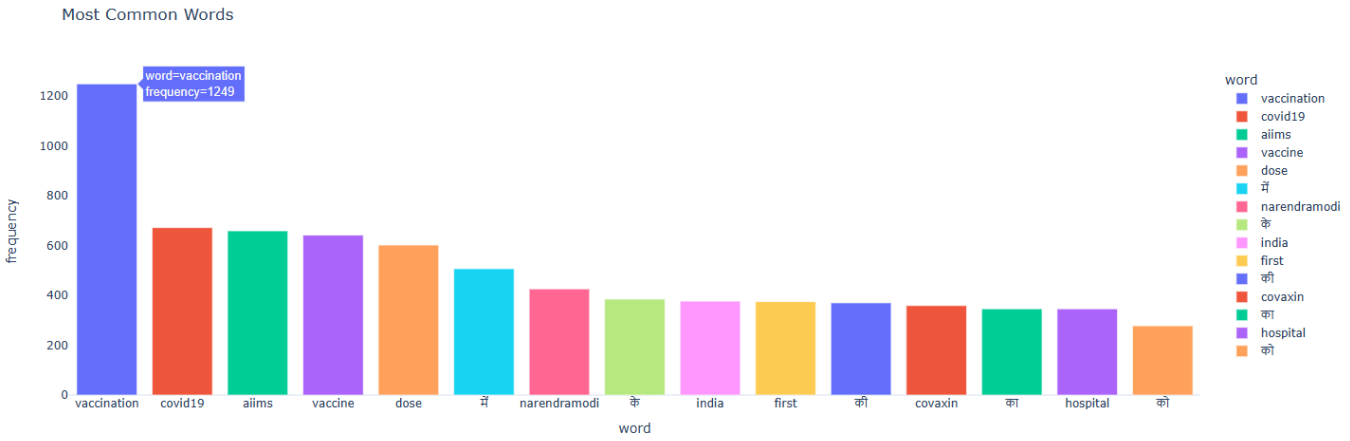

Fig 5.1.2 Most common words used with pointer (In India)

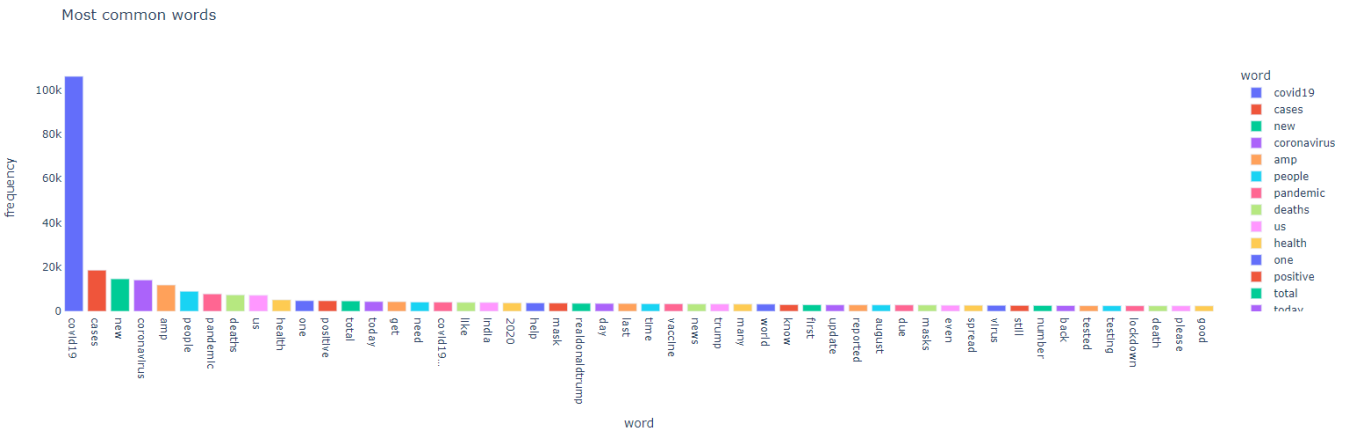

Fig5.2.1 Most common words used (In USA)

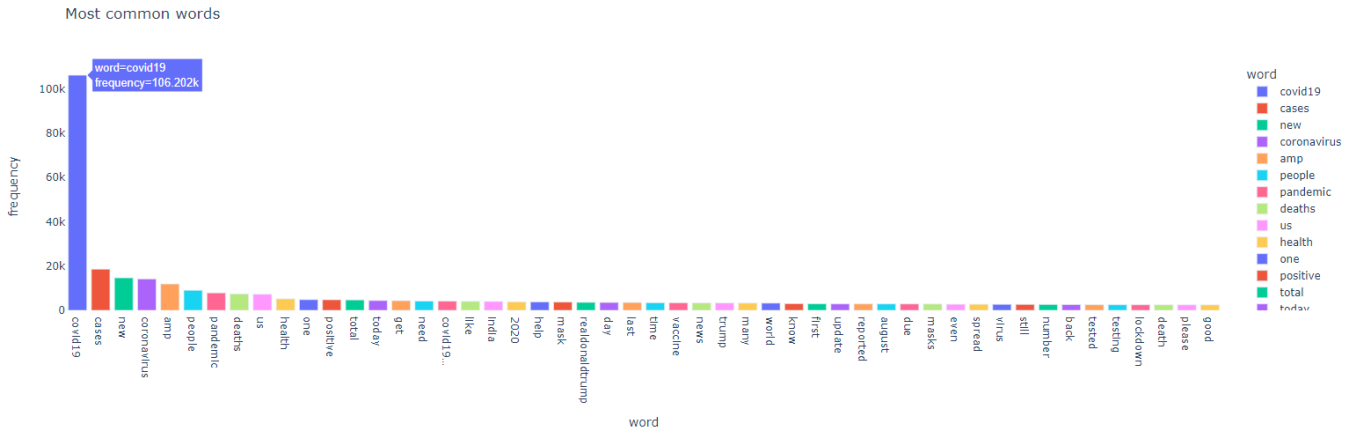

Fig 5.2.2 Most common words used with pointer(In USA)

7. Data Visualization

Present the analysis results through visualizations:

Create plots (e.g., bar chart, line chart) to display word frequencies and sentiment distribution.

We have created interactive graphical representations of the most used words, the number of positive, negative, and neutral tweets in a wordcloud image, bar graph, and the comparison over a month of these tweet sentiments via a line graph using a variety of tools in python respectively(Fig 6.1 and 6.2)

8. Interpretation and Insights

Interpret the analysis results to draw meaningful insights:

Summarize the sentiment distribution and major themes from the common words identified.

Derive conclusions based on the sentiment analysis and word frequency findings.

According to the word cloud image, the size of a word in a word cloud corresponds to its frequency of use, with the most used words appearing in the largest font. In the specific word cloud for India, "vaccination" is the most prominent word, indicating its high frequency of use. In the word cloud for the United States, "covid-19" stands out as the most frequently used word(Fig 8.1 and 8.2).

9. Conclusion and Recommendations

Conclude the analysis and provide recommendations:

Summarize the project's findings and insights.

Suggest potential actions or strategies based on the sentiment and word frequency analysis.

As can be seen from the graph, there is a significant difference in negative opinion between the USA and India.This suggests that Indians were more accepting of the vaccine in contrast to people in the US. This trend is also reflected in the word cloud in India, especially with the maximum frequency of the word “vaccination”, showing that people are focused on solving problems. On the other hand, the fact that the

keyword in the word cloud in the USA is "covid-19" shows that people are still very worried about this problem(Fig 7.1.1,7.2.1)

10. Documentation and Reporting

Prepare a detailed project report:
Document the entire process, including code, data sources, preprocessing steps, and analysis outcomes.
Include visualisations and interpretations for easy understanding.
Provide recommendations and future scope for the project.

Result Discussion

In India

The sentiment distribution among the tweets was visualised using a bar graph(Fig 6.1), clearly presenting the proportion of positive, negative, and neutral sentiments. The graph showcased that the majority of the tweets were neutral, indicating a generally balanced outlook regarding COVID-19. Additionally, a notable portion of tweets fell within the positive category, suggesting an optimistic viewpoint, while a smaller fraction conveyed negative sentiments.

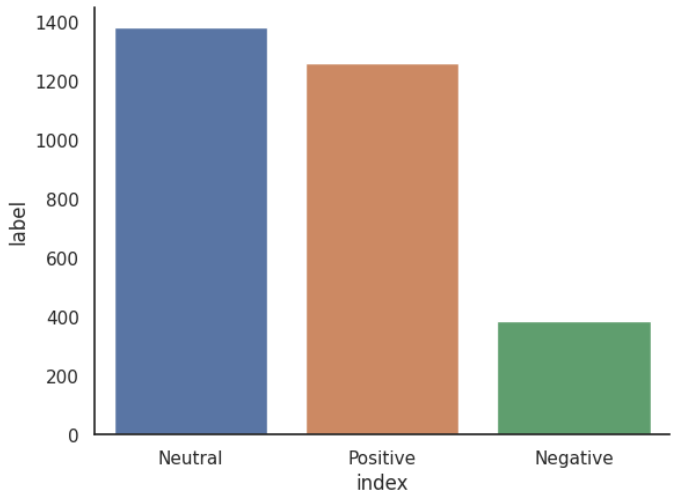

Fig 6.1 proportion of sentiment (In India)

In USA

We utilized a bar graph(Fig6.2) to visually represent the sentiment distribution of positive, negative, and neutral expressions in tweets The graph provides a clear and concise overview, indicating a predominant

presence of high positive sentiments. Additionally, there is a notable but smaller proportion of neutral sentiments, and a slightly lesser representation of negative sentiments. This visual representation underscores the observation that a substantial number of individuals expressed negative opinions regarding vaccination.

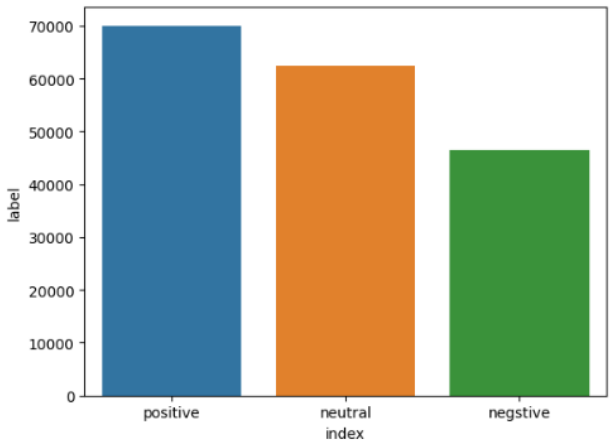

Fig 6.2 proportion of sentiment (In USA)

In India

To comprehend the overall sentiment trends over time, a line chart(Fig 7.1.1) is generated. This chart portrayed how positive, negative, and neutral sentiments fluctuated throughout the timeline(Fig 7.1.2) of the analyzed tweets. It was evident that the sentiment pattern evolved, showcasing variations in public perception regarding COVID-19 over the analyzed period. Notably, the positivity and neutrality in tweets experienced a surge at the end of February , potentially linked to significant developments in the Healthcare system and the Vaccination drives.

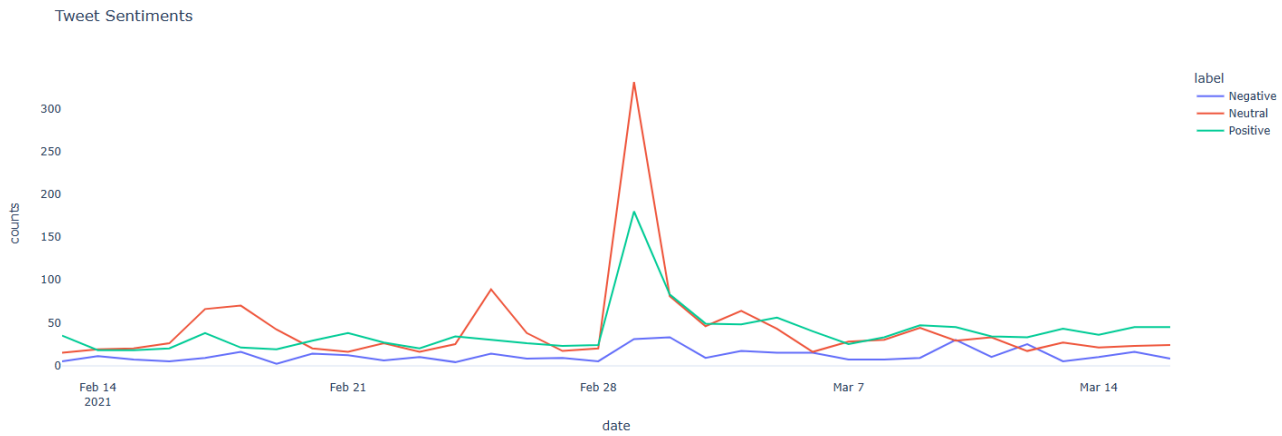

Fig 7.1.1 Depicts sentiment trends

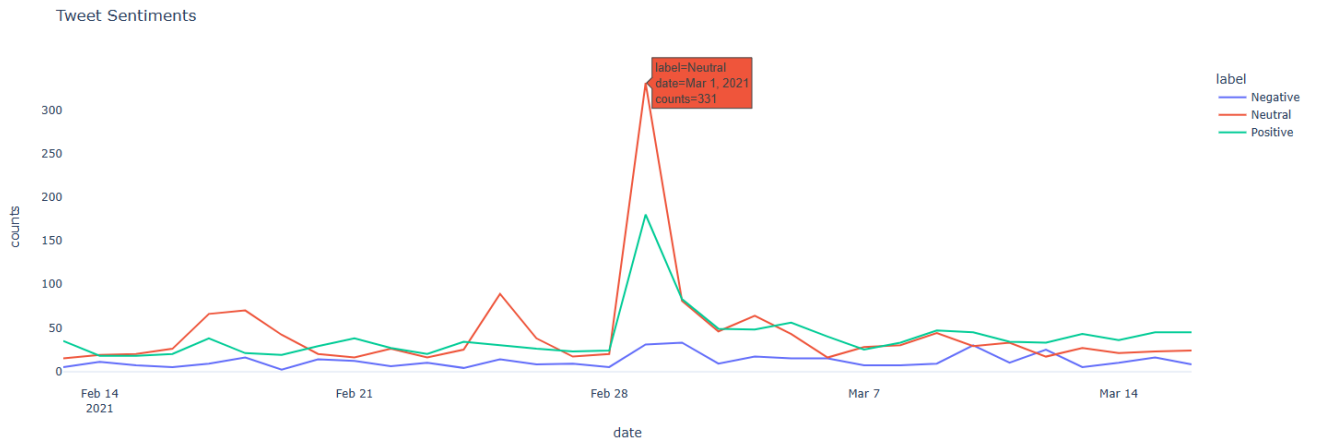

Fig 7.1.2 Depicts sentiment trends with pointer

In USA

We compared tweets over a specific period, generating a line chart(Fig 7.2.1) to visualize fluctuations in public sentiments. The chart effectively highlights(Fig 7.2.2) variations in positive, negative, and neutral comments. Notably, there was a surge in highly positive comments observed in July, providing valuable insights into the evolving sentiments of the online community during that period.

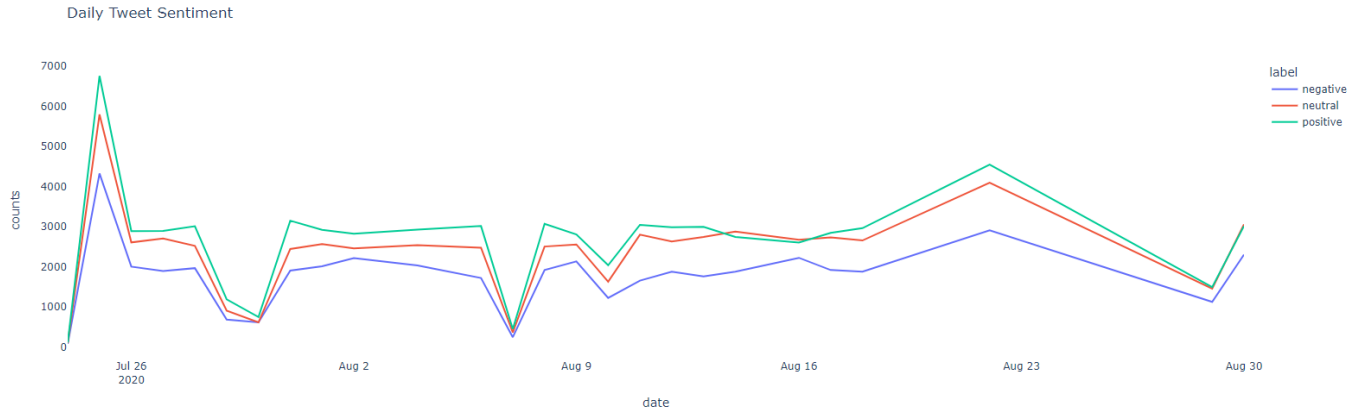

Fig 7.2.1 Depicts sentiment trends

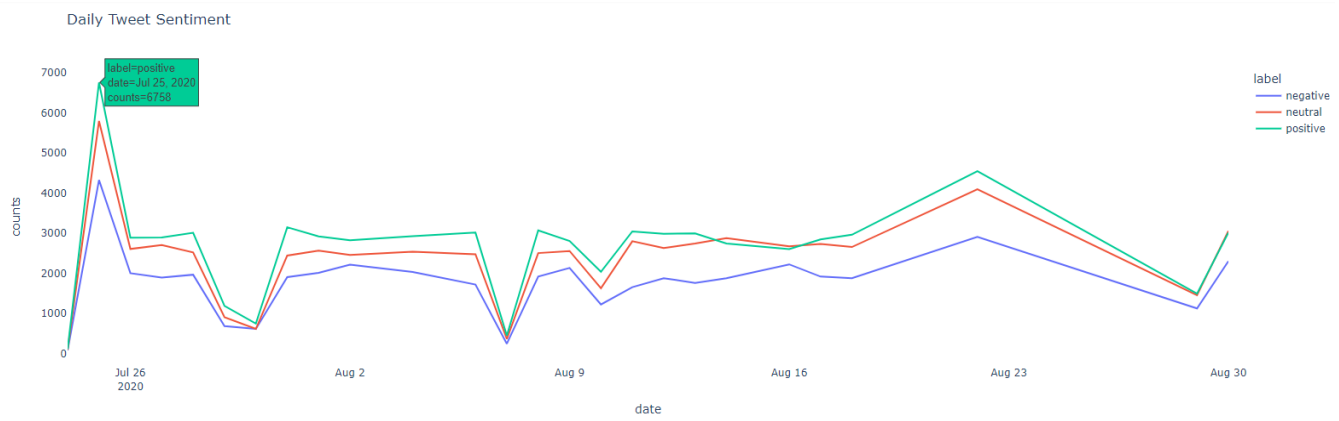

Fig 7.2.2 Depicts sentiment trends with pointer

In India

Word clouds(Fig 8.1) were generated to visualize the most frequently used words in the tweets. The word clouds provided a graphical representation of the prominent terms, giving valuable insights into the prevailing topics and discussions. Common terms such as "vaccination," "first dose," and "hospital" were prominent in the clouds, indicating their significance in the COVID-19 discourse.

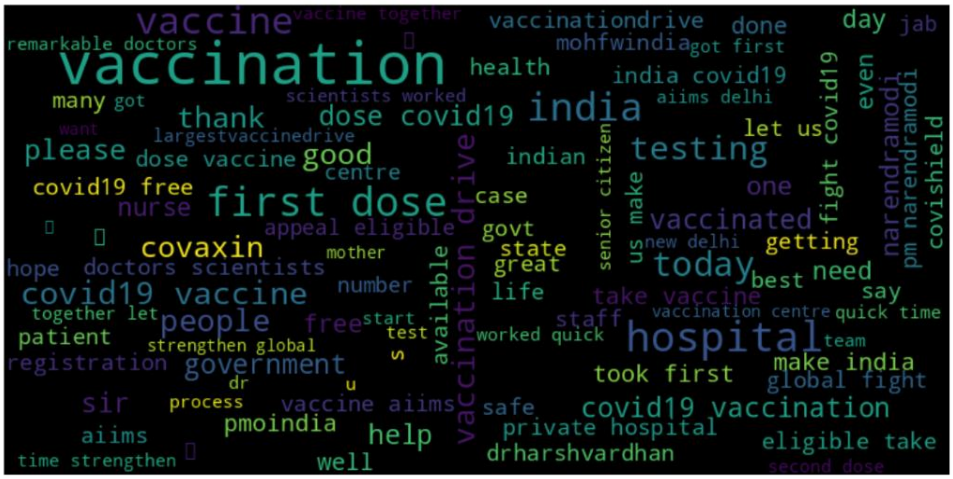

Fig 8.1 word density representation on wordcloud (In India)

In USA

A word cloud(Fig 8.2) is a visual representation of text data, where words are displayed in different sizes based on their frequency in the given text. The more frequently a word appears in the tweets, the larger

and more prominent it appears in the word cloud. Here we found some common terms such as "covid19", "risk case", "health" are prominent in this word cloud.

Fig 8.2 word density representation on wordcloud (In USA)

Conclusion

Given the escalation of the pandemic into a global crisis, as well as the extent to which the virus's implications have changed in the public eye over time, semantic analyses like the one presented here are becoming increasingly relevant as sources of information to the medical research community for a variety of health-related considerations.

Collection of data from the US(representing the world) and Delhi (representing India) helps in gaining valuable data and insights into the contrast and similarities between the two. Incorporating geolocation technology has proven instrumental in the acquisition of data for sentiment analysis related to the COVID-19 pandemic. By harnessing geospatial information, we can systematically gather and analyze sentiments expressed across diverse geographic regions. This approach not only enables a comprehensive understanding of local and global sentiments but also aids in identifying regional variations in public sentiment towards the pandemic, which can be pivotal for targeted public health interventions and effective crisis management.

As we can see, analyzing Twitter data enables the quick summarization of opinions about empirically proven illness prevention measures.

Moreover, Twitter sentiment analysis benefits the WHO and UN by enabling real-time monitoring of public sentiment on global health issues. It gives input on policy, assists in crisis management, and acts as an early warning system for impending health crisis.This analysis helps tailor communication strategies, identify stakeholders, and inform resource allocation decisions. Ultimately, sentiment analysis improves data-driven decision-making, enabling these organizations to respond faster and more successfully to challenges in global health.

We intend to develop this pipeline into a valuable tool that will assist health professionals and policymakers in assessing public response to health interventions in the ongoing global health crisis in

the future. This could include identifying categories that are underserved by current campaigns, as well as predictive modelling of public health messaging responses to help health organisations create and optimise outreach initiatives.

References

1.View ORCID Profile Amir Hussain, Ahsen Tahir, View ORCID Profil Zain Hussain, Zakariya Sheikh, View ORCID ProfileMandar Gogate, View ORCID ProfileKia Dashtipour, Azhar Ali,View ORCID Profile Aziz Sheikhdoi: https://doi.org/10.1101/2020.12.08.20246231

2. Ortal Slobodin 1,Ilia Plochotnikov 2,3,Idan-Chaim Cohen 4,Aviad Elyashar 3,5,Odeya Cohen 6,* and Rami Puzis 2,3,* doi: https://doi.org/10.3390/ijerph19116895

3.Sohaib R Rufai, Catey Bunce World leaders' usage of Twitter in response to the COVID-19 pandemic: a content analysis *Journal of Public Health*, Volume 42, Issue 3, September 2020, Pages 510–516, doi:https://doi.org/10.1093/pubmed/fdaa049

4.Mike Thelwall, Statistical Cybermetrics Research Group, University of Wolverhampton, UK. Orcid: 0000-0001-6065-205X Can Twitter Give Insights into International Differences in Covid-19 Vaccination? Eight countries' English tweets to 21 March 2021 doi:https://doi.org/10.48550/arXiv.2103.14125

5.Zahra Bokaee Nezhad, Mohammad Ali Deihimi Department of Computer Science, Shiraz University, Shiraz, Iran doi:https://doi.org/10.1016/j.dsx.2021.102367

6.Klaifer Garcia, Lilian Berton
Institute of Science and Technology, Federal University of Sao Paulo, São José dos Campos, São Paulo, 12247-014, Brazil. doi:https://doi.org/10.1016/j.asoc.2020.107057

7.Doogan C, Buntine W, Linger H, Brunt S
Public Perceptions and Attitudes Toward COVID-19 Nonpharmaceutical Interventions Across Six Countries: A Topic Modeling Analysis of Twitter Data
J Med Internet Res 2020;22(9):e21419
doi: 10.2196/21419

8.Harshita Chopra, Aniket Vashishtha, Ridam Pal, Ashima, Ananya Tyagi, Tavpritesh Sethi Computation and Language (cs.CL); Social and Information Networks (cs.SI) .doi: https://doi.org/10.48550/arXiv.2104.01131

9.Kazi Nabiul Alam, Md Shakib Khan, Abdur Rab Dhruba, Mohammad Monirujjaman Khan, Jehad F. Al-Amri, Mehedi Masud, Majdi Rawashdeh, "Deep Learning-Based Sentiment Analysis of COVID-19 Vaccination Responses from Twitter Data", *Computational and Mathematical Methods in Medicine*, vol. 2021, Article ID 4321131, 15 pages, 2021.doi: https://doi.org/10.1155/2021/4321131

10.Soyeon Kwon: Conceptualization, Validation, Formal analysis, Writing – original draft, Project administration. Albert Park: Conceptualization, Software, Formal analysis, Data curation, Writing – review & editing.
doi:https://doi.org/10.1016/j.chb.2021.107087

11.Anna Kruspe, Matthias Häberle, Iona Kuhn, Xiao Xiang Zhu Social and Information Networks (cs.SI); Machine Learning (stat.ML) doi: https://doi.org/10.48550/arXiv.2008.12172

12.Ainley E, Witwicki C, Tallett A, Graham C Using Twitter Comments to Understand People's Experiences of UK Health Care During the COVID-19 Pandemic: Thematic and Sentiment Analysis
J Med Internet Res 2021;23(10):e31101
doi: 10.2196/31101

13.A. Mourad, A. Srour, H. Harmanani, C. Jenainati and M. Arafeh, "Critical Impact of Social Networks Infodemic on Defeating Coronavirus COVID-19 Pandemic: Twitter-Based Study and Research Directions," in *IEEE Transactions on Network and Service Management*, vol. 17, no. 4, pp. 2145-2155, Dec. 2020, doi: 10.1109/TNSM.2020.3031034.

14.Claudia Mellado, Daniel Hallin, Luis Cárcamo, Rodrigo Alfaro, Daniel Jackson, María Luisa Humanes, Mireya Márquez-Ramírez, Jacques Mick, Cornelia Mothes, Christi I-Hsuan LIN, Misook Lee, Amaranta Alfaro, Jose Isbej & Andrés Ramos (2021) Sourcing Pandemic News: A Cross-National Computational Analysis of Mainstream Media Coverage of COVID-19 on Facebook, Twitter, and Instagram, Digital Journalism, 9:9, 1261-1285, DOI: 10.1080/21670811.2021.1942114

15.Ilyas, H., Anwar, A., Yaqub, U., Alzamil, Z. and Appelbaum, D. (2022), "Analysis and visualization of COVID-19 discourse on Twitter using data science: a case study of the USA, the UK and India", *Global Knowledge, Memory and Communication*, Vol. 71 No. 3, pp. 140-154.
https://doi.org/10.1108/GKMC-01-2021-0006

16.Dubey, Akash Dutt, Twitter Sentiment Analysis during COVID-19 Outbreak (April 9, 2020). Available at SSRN: https://ssrn.com/abstract=3572023 or http://dx.doi.org/10.2139/ssrn.3572023

17.Tao Na, Wei Cheng, Dongming Li, Wanyu Lu, Hongjiang Li Computation and Language (cs.CL); Social and Information Networks (cs.SI) doi:https://doi.org/10.48550/arXiv.2106.04081

18.Choudrie, J., Patil, S., Kotecha, K. *et al.* Applying and Understanding an Advanced, Novel Deep Learning Approach: A Covid 19, Text Based, Emotions Analysis Study. *Inf Syst Front* 23, 1431–1465 (2021).doi: https://doi.org/10.1007/s10796-021-10152-6

19.Hussain A, Tahir A, Hussain Z, Sheikh Z, Gogate M, Dashtipour K, Ali A, Sheikh A
Artificial Intelligence–Enabled Analysis of Public Attitudes on Facebook and Twitter Toward COVID-19 Vaccines in the United Kingdom and the United States: Observational Study
J Med Internet Res 2021;23(4):e26627

doi: 10.2196/26627

20.Banda, J.M.; Tekumalla, R.; Wang, G.; Yu, J.; Liu, T.; Ding, Y.; Artemova, E.; Tutubalina, E.; Chowell, G. A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research—An International Collaboration. *Epidemiologia* 2021, *2*, 315-324. https://doi.org/10.3390/epidemiologia2030024.

21.Fig4
https://www.researchgate.net/figure/Flow-chart-for-sentiment-analysis-IV-RESULTS-AND-DISCUSSIONS-I-There-have-been-four_fig1_330136547

22 dataset delhi
Twitter sentiment analysis (kaggle.com)

23 dataset USA
https://www.kaggle.com/datasets/gpreda/covid19-tweets

24. Fig3.1
https://www.researchgate.net/figure/Sentiment-Analysis-vs-VADER-Sentiment-Analysis_fig5_325896826

25. Fig3.2
https://scialert.net/fulltext/?doi=sjsres.2019.45.51