Exploring the COVID-19 Discourse Through Twitter Data

In the wake of the COVID-19 pandemic, social media platforms, particularly Twitter, have become a vital space for public discourse, information sharing, and sentiment expression regarding the crisis. This blog post aims to delve into the COVID-19 Twitter dataset, a comprehensive collection of tweets related to the pandemic, to understand the global conversation surrounding COVID-19. This analysis could serve public health agencies, researchers, and policymakers in several ways:

- **Informing Public Health Messaging**: By analysing the sentiments and topics prevalent in the COVID-19 discourse on Twitter, health agencies can tailor their communication strategies to address public concerns, misinformation, and sentiment effectively

- **Understanding Public Sentiment and Behavior**: Insights into the public's emotional response and behaviour towards the pandemic can guide interventions and policy-making to better manage the crisis.

- **Tracking Misinformation Spread**: Identifying misinformation trends on Twitter can help in deploying countermeasures to prevent the spread of false information.

- **Source Verification**: The dataset's "source" field can be used to analyse the origins of widespread misinformation, identifying whether it stems from particular accounts, websites, or other media. This can aid in pinpointing misinformation hubs.

Process:

Description:

- The dataset was extracted from Kaggle, and the tweets are collected using Twitter API and a Python script.
- The dataset contains the tweets about Covid posted from July 2020 to august 2020
- The dataset contains **179108 observations** and **13 features.**

Features:

The dataset comprises a variety of features that offer insights into the behaviour and interactions of Twitter users during the pandemic. It includes the following columns:

1. **user_name:** The account name of the Twitter user.
2. **user_location:** The reported location of the user.
3. **user_description:** A description provided by the user for their profile.
4. **user_created:** The date when the user's Twitter account was created.
5. **user_followers:** The number of followers the user has.
6. **user_friends:** The number of friends (accounts the user is following) the user has.
7. **user_favourites:** The number of tweets the user has liked.
8. **user_verified:** A boolean indicating whether the user's account is verified.
9. **date:** The date and time when the tweet was posted.
10. **text:** The content of the tweet.
11. **hashtags:** Any hashtags included in the tweet.
12. **source:** The original source or device used to post the tweet.
13. **is_retweet:** A boolean indicating whether the tweet is a retweet.

Data Observations and Constraints

- The dataset primarily contains tweets from July and August 2020, with the highest number of tweets recorded in these months.
- All of the tweets in the dataset are original content i.e., posted by the users themselves, as indicated by the is_retweet field.
- The dataset has some limitations, such as missing data for certain dates, usernames, and empty and weird data is user_location field.
- Most users in the dataset are not verified, which could affect the spread and perceived credibility of the information.

Model

Feature Engineering

We first did feature engineering on the data set. The approach to clean up the data had the following steps:

- Converted boolean features 'is_retweet' and 'user_verified' from boolean to numeric format (0s and 1s) for consistency and analysis purposes.
- Dropped duplicate rows to ensure data integrity and eliminate redundancy in the dataset.
- Changed the missing values in the User Name and User Location to Unknown, and changed Hashtags columns of Missing / Tweets with no Hashtags to None.
- Removed emojis and unwanted characters from the text data to improve readability and analysis accuracy.
- Separated URLs present in the text field of the dataset and created a separate column for the URLs.
- Extracted components such as date, month, and time from the date column to make it easier to find the relation between tweet volume, and analyse the tweet volume by day

Key Features for Analysis

For our analysis, we will focus on several key features that provide valuable insights into the nature of COVID-19 related discourse on Twitter:

- **User Location:** Understanding the geographical distribution of tweets can shed light on how different regions are discussing COVID-19.
- **User Verified Status:** Analysing the impact of verified accounts on the spread of information and engagement levels.
- **Tweet Date and Time:** Temporal analysis of tweet volumes can reveal patterns and spikes in COVID-19 related conversations.
- **Tweet Content (Text):** Content analysis through sentiment analysis and topic modelling can uncover the prevailing sentiments and topics within the tweets.
- **Hashtags:** Identifying common hashtags can help understand the focus of discussions and how they spread across the network.
- **Engagement Metrics:** Analysing user engagement through followers, friends, and favourites can provide insights into the influence of user profiles on information dissemination.

RESEARCH QUESTIONS

Our four main research questions are:

What are the most commonly used hashtags in COVID-19 related tweets across different geographic areas?

This research question aims to identify the most frequently used hashtags in COVID-19 related tweets across different geographic regions. By analysing the hashtags, we can gain insights into the prevalent topics or themes associated with the pandemic in various locations.

Can sentiment analysis techniques be applied to COVID-19 tweets to infer sentiments or identify key topics based on their content or context?

This research question explores the possibility of analysing the sentiment and identifying key topics in COVID-19 tweets based on their content or context. By employing sentiment analysis techniques and topic modelling algorithms, we can categorise tweets into positive, negative, or neutral sentiments and uncover important themes or discussions surrounding the pandemic.

Is there a correlation between the number of followers/friends of Twitter users and their engagement with COVID-19 related content?

This research question investigates the correlation between the number of followers/friends of Twitter users and their engagement with COVID-19 related content. By examining user engagement metrics such as likes, retweets, and replies, we can assess how social media influence influences the level of engagement with pandemic-related tweets.

How does the volume of COVID-19 tweets change over time, and are there specific periods of higher tweet activity?

This research question focuses on understanding how the volume of COVID-19 tweets changes over time and identifying specific periods of heightened tweet activity. By analysing temporal trends, we can discern patterns in tweet activity related to significant events, announcements, or developments in the course of the pandemic.

DATA CLEANING

The remove_duplicates function is essential for data cleaning, particularly in analysing COVID-19 tweets. It selects a DataFrame and columns to identify duplicates, returning a new DataFrame without these redundancies. This process is vital for ensuring accuracy and quality by guaranteeing that each tweet from a user on a given data is unique. It also makes the dataset smaller and cleaner, ensuring more efficient data processing, which leads to more reliable insights.
After cleaning, the number of rows changed from 179108 rows × 13 columns to 178312 rows × 13 columns.

Before:

Out[2]:

	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date	
0	ᵛ'i●ıɛ†	astroworld	wednesday addams as a disney princess keepin i...	2017-05-26 05:46:42	624	950	18775	False	2020-07-25 12:27:21	If I smelled th s
1	Tom Basile us	New York, NY	Husband, Father, Columnist & Commentator. Auth...	2009-04-16 20:06:23	2253	1677	24	True	2020-07-25 12:27:17	Hey @Yankee and @ML
2	Time4fisticuffs	Pewee Valley, KY	#Christian #Catholic #Conservative #Reagan #Re...	2009-02-28 18:57:41	9275	9525	7254	False	2020-07-25 12:27:14	@diane3 @realDonaldTru
3	ethel mertz	Stuck in the Middle	#Browns #Indians #ClevelandProud #[]_[] #Cavs ...	2019-03-07 01:45:06	197	987	1488	False	2020-07-25 12:27:10	@brookban #COVID1
4	DIPR-J&K	Jammu and Kashmir	🖊Official Twitter handle of Department of Inf...	2017-02-12 06:45:15	101009	168	101	False	2020-07-25 12:27:08	25 July : Media B
...	...	...	...	...	...	...	...	...	...	
179103	AJIMATI AbdulRahman O.	Ilorin, Nigeria	Animal Scientist|| Muslim|| Real Madrid/Chelsea	2013-12-30 18:59:19	412	1609	1062	False	2020-08-29 19:44:21	Thanks nominating me
179104	Jason	Ontario	When your cat has more baking soda than Ninja ...	2011-12-21 04:41:30	150	182	7295	False	2020-08-29 19:44:16	2020! The year #
179105	BEEHEMOTH ⏳	CA Canada	🛠 The Architects of Free Trade 🛠 Really Did ...	2016-07-13 17:21:59	1623	2160	98000	False	2020-08-29 19:44:15	@CTVNews A p by J
179106	Gary DelPonte	New York City	Global UX UI Visual Designer. StoryTeller, Mus...	2009-10-27 17:43:13	1338	1111	0	False	2020-08-29 19:44:14	More than 1,2 po
179107	TUKY II	Aliwal North, South Africa	TOKELO SEKHOPA | TUKY II | LAST BORN | EISH TU...	2018-04-14 17:30:07	97	1697	566	False	2020-08-29 19:44:08	I st Stop\n\n@SABCNe

179108 rows × 13 columns

After:

Out[3]:

	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date	
0	ᐟV°i●ɪꜫ†	astroworld	wednesday addams as a disney princess keepin i...	2017-05-26 05:46:42	624	950	18775	False	2020-07-25 12:27:21	If I smelled th s
1	Tom Basile 🇺🇸	New York, NY	Husband, Father, Columnist & Commentator. Auth...	2009-04-16 20:06:23	2253	1677	24	True	2020-07-25 12:27:17	Hey @Yankee and @ML
2	Time4fisticuffs	Pewee Valley, KY	#Christian #Catholic #Conservative #Reagan #Re...	2009-02-28 18:57:41	9275	9525	7254	False	2020-07-25 12:27:14	@diane3 @realDonaldTrur
3	ethel mertz	Stuck in the Middle	#Browns #Indians #ClevelandProud #[]_[] #Cavs ...	2019-03-07 01:45:06	197	987	1488	False	2020-07-25 12:27:10	@brookban #COVID1!
4	DIPR-J&K	Jammu and Kashmir	✏️Official Twitter handle of Department of Inf...	2017-02-12 06:45:15	101009	168	101	False	2020-07-25 12:27:08	25 July : Media B
...	...	...	...	...	...	...	...	...	...	
179103	AJIMATI AbdulRahman O.	Ilorin, Nigeria	Animal Scientist|| Muslim|| Real Madrid/Chelsea	2013-12-30 18:59:19	412	1609	1062	False	2020-08-29 19:44:21	Thanks nominating me
179104	Jason	Ontario	When your cat has more baking soda than Ninja ...	2011-12-21 04:41:30	150	182	7295	False	2020-08-29 19:44:16	2020! The year #
179105	BEEHEMOTH ⏳	🇨🇦 Canada	⚒️ The Architects of Free Trade ⚒️ Really Did ...	2016-07-13 17:21:59	1623	2160	98000	False	2020-08-29 19:44:15	@CTVNews A p by J
179106	Gary DelPonte	New York City	Global UX UI Visual Designer. StoryTeller, Mus...	2009-10-27 17:43:13	1338	1111	0	False	2020-08-29 19:44:14	More than 1,2 po
179107	TUKY II	Aliwal North, South Africa	TOKELO SEKHOPA | TUKY II | LAST BORN | EISH TU...	2018-04-14 17:30:07	97	1697	566	False	2020-08-29 19:44:08	I st Stop\n\n@SABCNe

178312 rows × 13 columns

```
covid19_tweets['is_retweet'] = covid19_tweets['is_retweet'].replace({True: 1, False: 0})
covid19_tweets['is_retweet'] = covid19_tweets['is_retweet'].replace({True: 1, False: 0})

covid19_tweets['user_verified'] = covid19_tweets['user_verified'].replace({True: 1, False: 0})
covid19_tweets['user_verified'] = covid19_tweets['user_verified'].replace({True: 1, False: 0})

def remove_duplicates(df, columns):
    df_unique = df.drop_duplicates(subset=columns)
    return df_unique

covid19_tweets = remove_duplicates(covid19_tweets, ['user_name', 'date']).reset_index()

covid19_tweets.shape
covid19_tweets.drop(columns=['index'], inplace = True)
covid19_tweets['hashtags'] = covid19_tweets['hashtags'].fillna('None')
```

To accomplish the second part, we transformed boolean values in the column "is_retweet" to 0 or 1 using the function map_bool_to_int. This conversion is essential for numerical analysis and standardising data formats across datasets.

Now, many features in our dataset like user name, user location, text, etc had emojis, trademark symbols, and other special characters, so we removed them from text data. Additionally, we standardise the text format by replacing unwanted symbols with whitespace and handling cases of null or empty values, and setting them to “Unknown”. This comprehensive text cleaning process ensures that the data is ready for analysis by removing noise and standardising the text format.

```python
def remove_emojis(text):
    emoji_pattern = re.compile("["
                               u"\U0001F600-\U0001F64F"
                               u"\U0001F300-\U0001F5FF"
                               u"\U0001F680-\U0001F6FF"
                               u"\U0001F700-\U0001F77F"
                               u"\U0001F780-\U0001F7FF"
                               u"\U0001F800-\U0001F8FF"
                               u"\U0001F900-\U0001F9FF"
                               u"\U0001FA00-\U0001FA6F"
                               u"\U0001FA70-\U0001FAFF"
                               u"\U00002702-\U000027B0"
                               u"\U000024C2-\U0001F251"
                               u"\U000023F3"
                               "]+", flags=re.UNICODE)

    return emoji_pattern.sub(r'', text)

def remove_trademark_symbols(text):
    trademark_pattern = re.compile(r'\b(\w+)\s*[™®©℗]\b', flags=re.IGNORECASE)
    cleaned_text = re.sub(trademark_pattern, r'\1', text)

    return cleaned_text


def replace_special_chars(text):
    special_char_pattern = re.compile("[.@\|!#$%^&*()<>?/\-_|+}{~:,]")
    return special_char_pattern.sub(' ', text)

def clean_value(value):
    if pd.isna(value) or value.strip() == "" or not value:
        return "Unknown"
    cleaned_value = re.sub(r'[^\w\s]', '', value)
    cleaned_value = cleaned_value.strip('. ')
    cleaned_value = re.sub(r'[˙•°]', '', cleaned_value)
    cleaned_value = re.sub(r'®|™', '', cleaned_value)
    cleaned_value = re.sub(r'[0-9□?]', '', cleaned_value)
    cleaned_value = ' '.join(cleaned_value.split())
    return cleaned_value
```

After Data Cleaning:

user_name	user_location
Ѵ'i¿↑	astroworld
Tom Basile	New York, NY
Timefisticuffs	Pewee Valley, KY
ethel mertz	Stuck in the Middle
DIPR J K	Jammu and Kashmir
Franz Schubert	Новоро́ссия
hr bartender	Gainesville, FL
Derbyshire LPC	Unknown
Prathamesh Bendre	Unknown
Member of Christ	location at link below
Voice Of CBSE Students	Unknown
Creativegms	Dhaka,Bangladesh

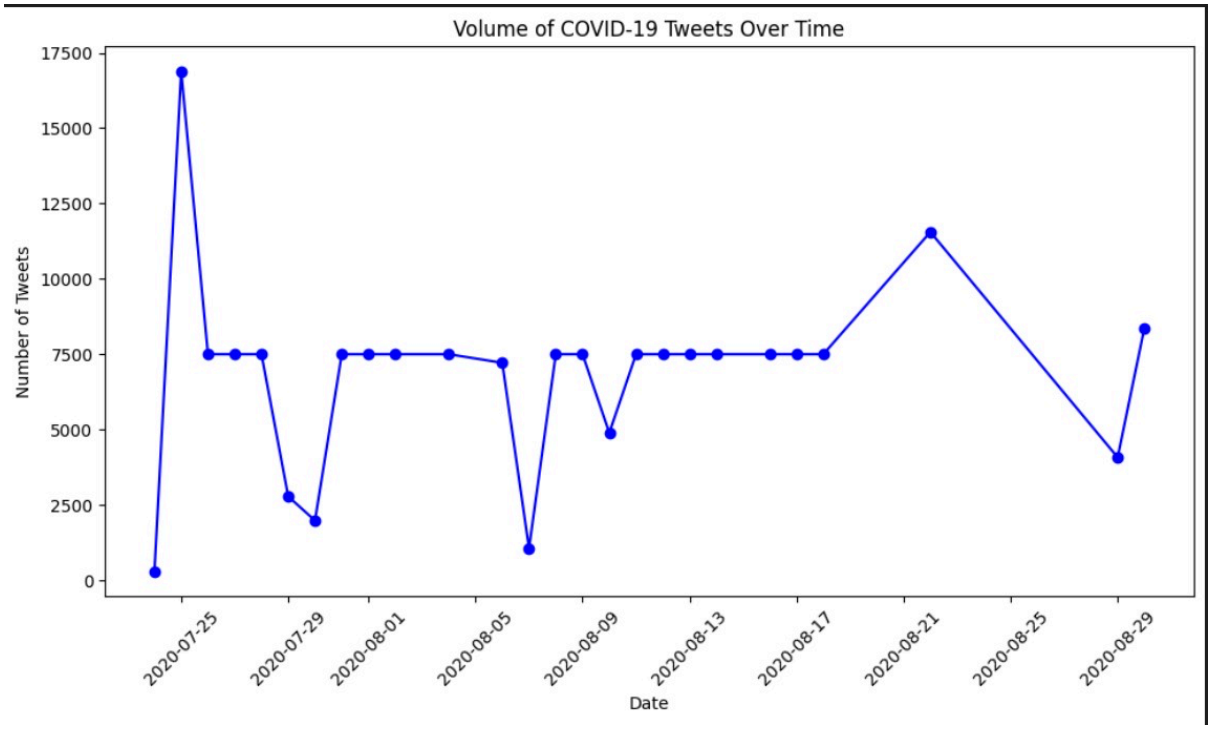
In this part of the data cleaning process, we focused on transforming the 'date' column into more manageable components for analysis. Initially, this column contained dates and times in a mixed format, including some invalid entries like 'FALSE'. To address this, we first converted this column to a datetime format using the pd.to_datetime() function. This step allowed us to standardise the date and time data across the dataset. The 'coerce' parameter was used to handle any conversion errors, ensuring that even entries with unexpected or invalid formats would not cause the conversion process to fail entirely. Next, to ensure data integrity, all rows with invalid dates using the dropna() function were dropped, which helped to clean the dataset by removing irrelevant or corrupted entries.

```
df = pd.DataFrame("covid19_tweets.csv")
df['date'] = pd.to_datetime(df['date'], errors='coerce')
df = df.dropna(subset=['date'])
```

Furthermore, we split the datetime data into three separate columns: 'date', 'month', and 'time'. The 'date' column now contains only the day component of the original datetime values, which can be useful for exploring temporal patterns. The 'month' column represents the month component as string names rather than numerical values, making it more intuitive for interpretation. Lastly, the 'time' column extracts the time component in the format 'HH:MM:SS', enabling analyses that focus specifically on time-related trends. Since the year was the same (2020) for all tweets, we did not make a separate column for that. Doing all of this enhances the overall data quality and usability for further exploration.

```
df['month'] = df['date'].dt.month_name()
df['date'] = df['date'].dt.day
df['time'] = df['date'].dt.strftime('%H:%M:%S')
df.drop(columns=['date'], inplace=True)
```

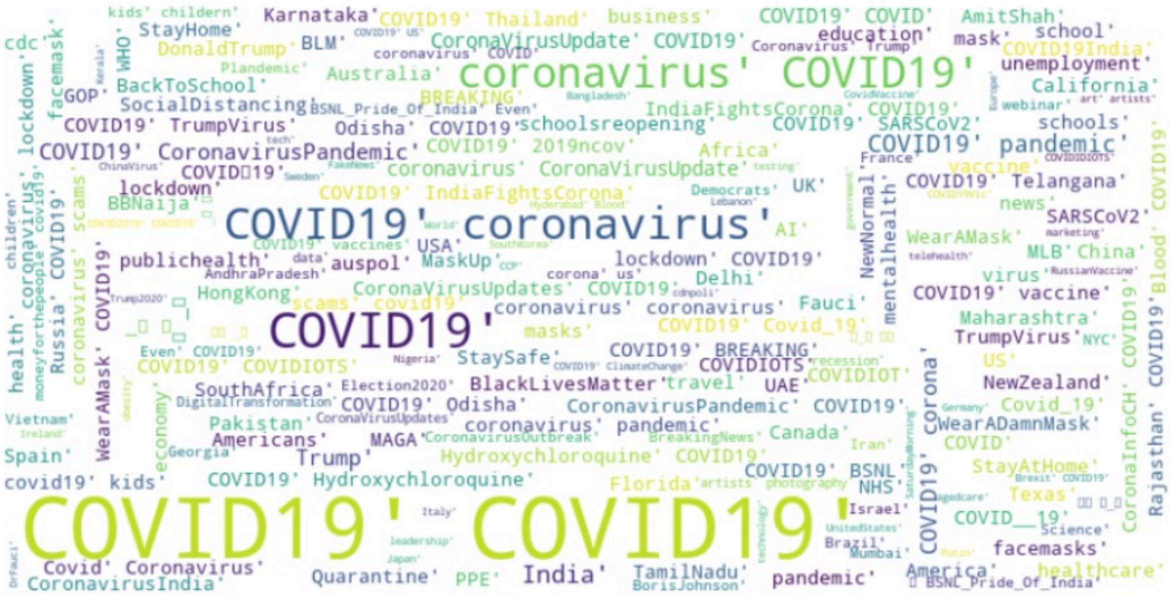
DATA EXPLORATION AND VISUALISATION



The line graph presented in the image illustrates the volume of COVID-19 related tweets over a specific time frame. Here are some observations and details based on the graph:

- **Time Frame**: The x-axis indicates the time period for the data, which spans from July 25, 2020, to August 29, 2020. This period likely captures a significant phase of the COVID-19 pandemic where public interest and discourse were high.
- **Tweet Volume:** The y-axis measures the number of tweets, with the scale reaching up to 17,500 tweets. The graph shows the daily volume of tweets related to COVID-19.
- **Trends and Patterns:** The graph displays a fluctuating pattern of tweet volumes with several noticeable spikes. These spikes may correspond to specific events or announcements related to the pandemic, such as policy changes, significant outbreaks, or news about vaccine development.
- **Peaks:** The highest peak occurs at the beginning of the time frame, with over 17,500 tweets in a single day. This suggests a particular event or high interest in COVID-19 related information on that day.
- Valleys: There are periods where the tweet volume significantly drops, indicating days with less activity. The lowest point appears to be around 2,300 tweets.

- **Overall Trend:** While there are fluctuations, there is no clear upward or downward trend over the entire period. Instead, the graph shows variability in public engagement with the topic of COVID-19 on Twitter.
- **Data Representation:** The graph uses a blue line with circular markers at each data point, making it easy to follow the changes in tweet volume over time.

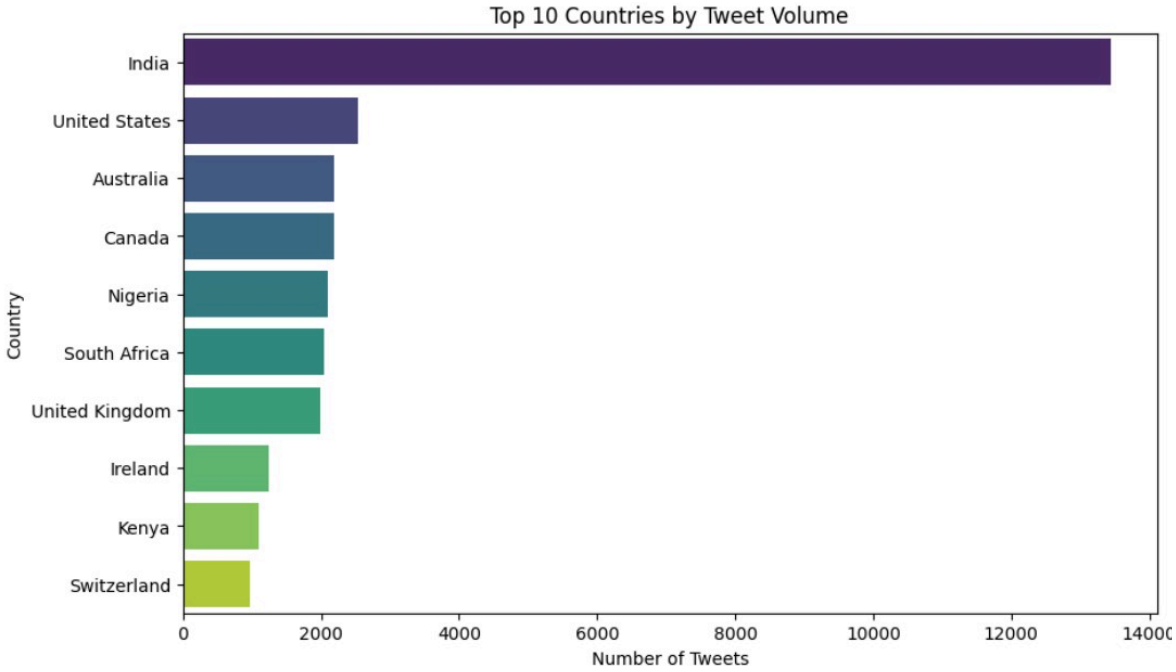


Word Cloud of COVID-19 Related Terms

The word cloud is a visual representation of the most frequently used terms in COVID-19 related tweets. Here are some observations:

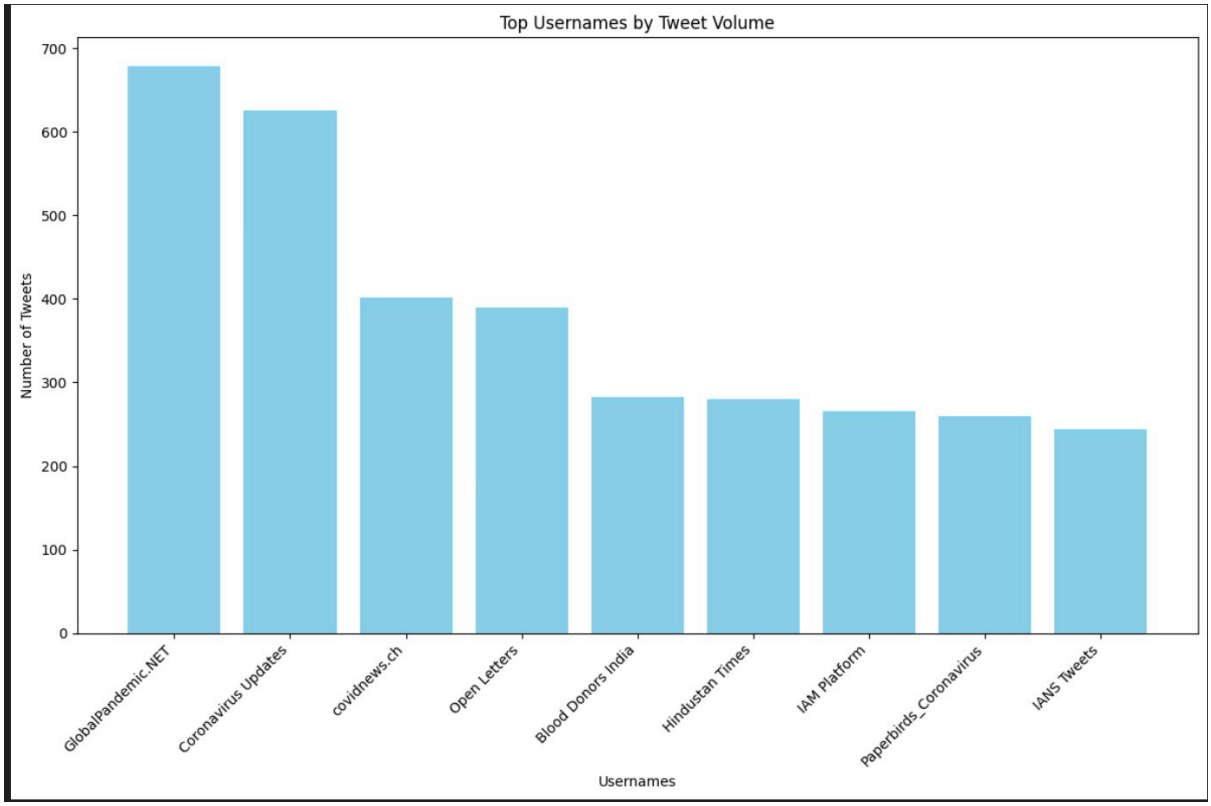
- **Prominent Terms:** The terms "COVID19" and "coronavirus" are the most prominent, indicating their high frequency in the dataset.
- **Diverse Topics:** The presence of words like "lockdown," "mask," "vaccine," and "pandemic" suggests a wide range of topics within the COVID-19 discourse.
- **Global Reach:** Geographic names such as "India," "Australia," and "USA" highlight the global impact of the pandemic and the diverse origins of the tweets.
- **Visual Appeal:** The word cloud's design, resembling a globe, emphasises the worldwide nature of the pandemic.

- **Colour Variation:** The use of different colours adds visual interest and may group related terms together, although the specific categorization is not indicated.



Top 10 Countries by Tweet Volume
The above chart visualises the top 10 countries by tweet volume. Following are the observations from the graph:

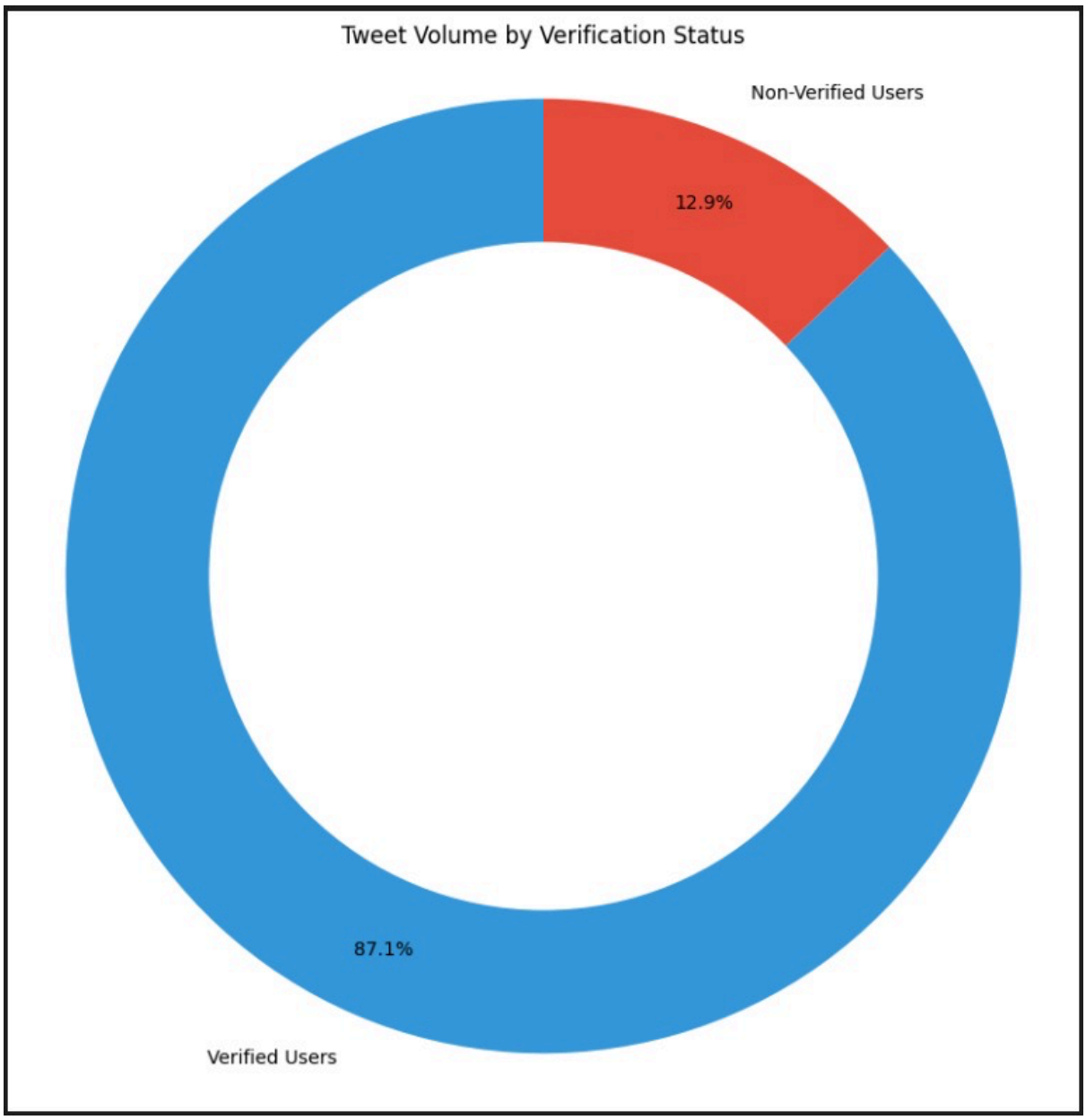
- **India's Predominance:** India has the highest number of tweets, significantly more than any other country, indicating a high level of engagement with COVID-19 related topics.
- **Other Active Countries:** The United States and Australia follow India, with substantial tweet volumes, but less than half of India's.
- **Variation in Engagement:** There is a noticeable drop in tweet volume after the top two to three countries, with others showing progressively fewer tweets.
- **Diverse Global Participation:** The chart includes countries from different continents, reflecting that the dataset contains the tweets regarding COVID 19 from all over the world.



Top Usernames by Tweet Volume

The above graph shows the top user names according to the number of tweets posted from the account. Following are the details about the graph:

- **X-Axis:** Represents different Twitter usernames of different accounts.
- **Y-Axis:** Shows the number of tweets associated with each username.
- **Bars:** Each bar corresponds to a username and its height tells the total number of tweets from that account in July and August 2020 regarding Covid 19
- **Data Insight:** The chart suggests that "Globalpandemic.NET" is the most active account in terms of tweets, followed by "Coronavirus Updates" and others. This may reflect the level of engagement or focus each account has on the topic at hand, which could be the COVID-19 pandemic based on the names of the top two accounts.



Tweet Volume by Verification Status

The pie chart shows the distribution of tweets from verified and non-verified Twitter users.

- **Segments**: There are two segments in the chart—one representing verified users(blue) and the other representing non-verified users(red).
- **Percentage Labels:** Each segment is labelled with a percentage that indicates the proportion of total tweets made by each group. The chart indicates that a significant majority of tweets (87.1%) are made by verified users, while a smaller portion (12.9%) are by non-verified users.

- **Implication:** This distribution implies that verified users are responsible for the bulk of the tweeting activity, at least within the data represented in the chart. This could suggest that verified users are more active.