JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

COVID 19 TWEETS GEO AND SENTIMENTAL ANALYSIS AND CLASSIFICATION

[1]Sarthak Patel, [2]Kalpan Shah, [3] Dr. J Amudhavel

[1]Student, Vellore Institute of Technology, Bhopal, India
[2]Student, Vellore Institute of Technology, Bhopal, India
[3]Professor, Vellore Institute of Technology, Bhopal, India

Email: amudhavel.j@vitbhopal.ac.in

Abstract: The promulgation of Covid-19 has caused widespread health concerns across the world. News and views on it are widely shared on social media. To use resources efficiently and appropriately, a realistic evaluation of the situation is required. In this research, I used Natural language processing (NLP) to do sentiment analysis on Covid-19 tweets in this study. The ability to identify Covid-19 emotions from tweets would enable more educated judgments to be made about how to handle the present pandemic scenario. Tweets are categorized as negative, positive, or neutral. This study attempts to evaluate and illustrate the global impact of the coronavirus (COVID-19) by using sentiment analysis techniques and methodologies on the Twitter dataset to comprehend both extremely favorable and very negative public opinion throughout the world.

Index Terms **– Covid-19, Twitter, Twitter API, Natural Language Processing (NLP), Twitter Sentiment Analysis**

I. INTRODUCTION AND BACKGROUND ANALYSIS

The Covid-19 outbreak has had a meteoric impact on the social as well economic strata of the whole world. On January 30, 2020, the World Health Organization designated it an epidemic [1]. Since then, it has expanded rapidly, causing significant health problems as well as agonizing deaths. The death toll had reached 636,633 as of May 31, 2020 [2]. In the following days, the infection will continue to spread.

People's lives are altering as a result of social media. With the rapid growth of Social Network Service (SNS), people may exchange news and ideas and communicate with each other on the Internet at any time and from anywhere [3]. People use major social media platforms like Twitter, Facebook, and Instagram to upload enormous amounts of geographical and temporal-based data, such as texts, photos, and videos, resulting in the growth of social big data. People on Twitter, for example, send out about 6000 tweets every second about a wide range of global events [4]. During the lockdown, social networking traffic rose dramatically. In terms of timely dissemination of Covid-19 news, Twitter has outperformed its competitors [5]. To this end, I have tried to answer 10 business questions:

1. Which countries are most people tweeting from?
2. What sources have people used more commonly to tweet
3. What are the common hashtags used in these tweets?
4. Is there any trend in the tweets daily?
5. Who are the most followed people/ accounts amongst these tweets?
6. What are the most common words in the tweets?
7. What is the sentiment in the tweets?
8. Which accounts mostly have a positive and negative tweet?
9. Does most favorited tweets have any sentiment pattern/ trend?
10. Which entities (people, locations, organizations) have been talked about the most?

Exploring and mining social big data to understand people's varied emotional states is crucial and relevant, especially as social media becomes a more vital part of people's everyday lives. Identifying the sentiment polarity is a basic technique of assessing human emotional states (positive, neutral and negative). The following are some similar works.

Stephen Wai Hang Kwok conducted sentiment analysis on covid-19 vaccination among Australian Twitter Users. This study focused to use extract topics and emotions on covid-19 vaccination on Twitter with the help of machine learning [6].

Liu *et al.* proposed a novel hierarchical neural network model based on dynamic word embeddings (HieNN-DWE) for document-level sentiment classification. The model consists of two layers: the first one uses bidirectional gated recurrent unit (BiGRU) and attention mechanism to encode sentences and in the second layer, both BiGRU and convolutional neural network (CNN) are employed to capture features in the sentences [7].

Wang *et al.* proposed the aspect-level sentiment capsules model (AS-Capsules), which could perform aspect detection and sentiment classification at the same time. Moreover, they added the attention mechanism to find out aspect related words and sentiment words without any linguistic knowledge [8].

Chatterjee *et al.* gathered large scale data from social media and proposed a novel Deep Learning based approach to detect emotions including happy, sad, and angry in texts [9].

In 2020, Lu *et al.* proposed an interactive rule attention network (IRAN) considering the influence of grammatical rules. IRAN simulates the grammatical functions at the sentence and also uses an attention network to learn attention information from context [10].

II. COVID-19 TWEETS DATASET SCENARIOS

Twitter data set has been collected using Twitter API which has a high-frequency hashtag (#covid19). The Twitter API allows you to read and write data from Twitter. As a result, one may use it to create tweets, go through profiles, and have access to one's followers' data as well as a large number of tweets on certain topics in specific regions [11]. A sample set of ~180k tweets for the period of July 24 to August 30, 20202 has been used. Dataset consists of 13 important attributes.

III. EXPLORATION DATA ANALYSIS AND DATA PRE-PROCESSING

To reduce the ambiguity of the result as there are chances that some non-verified user may be biased intentionally biased towards sentiment of the tweets. So, it's better to remove the non-verified user from the dataset. Out of over 180k tweets over 2,320 users are non-verified which is around 12.89% of the total users.

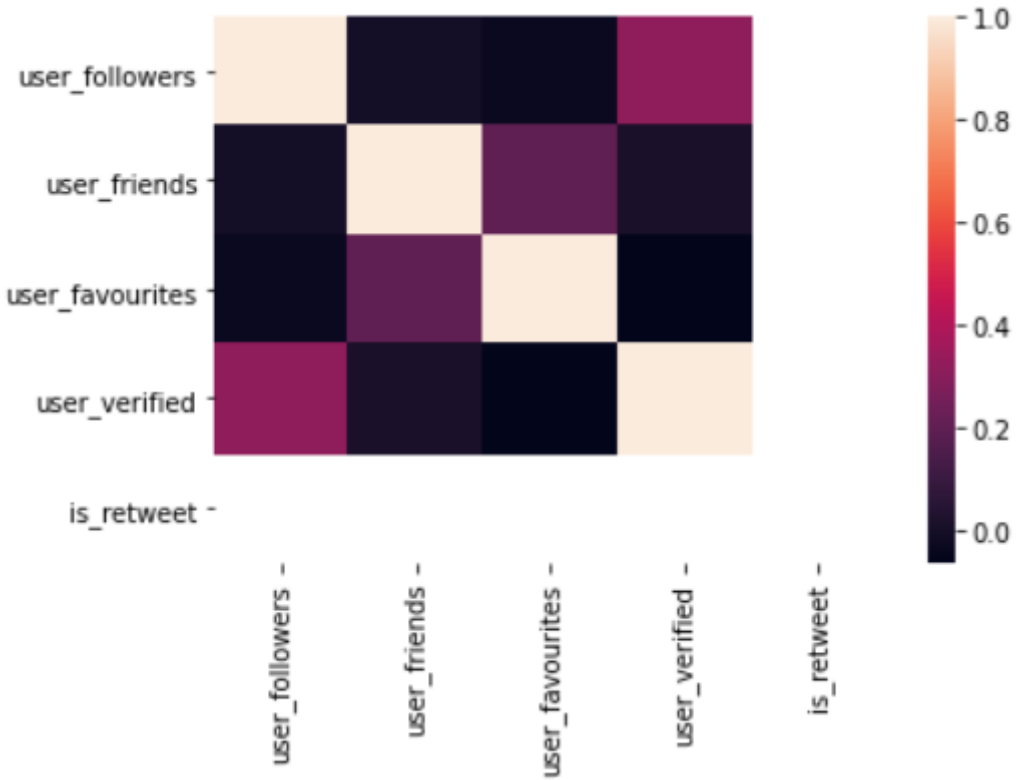
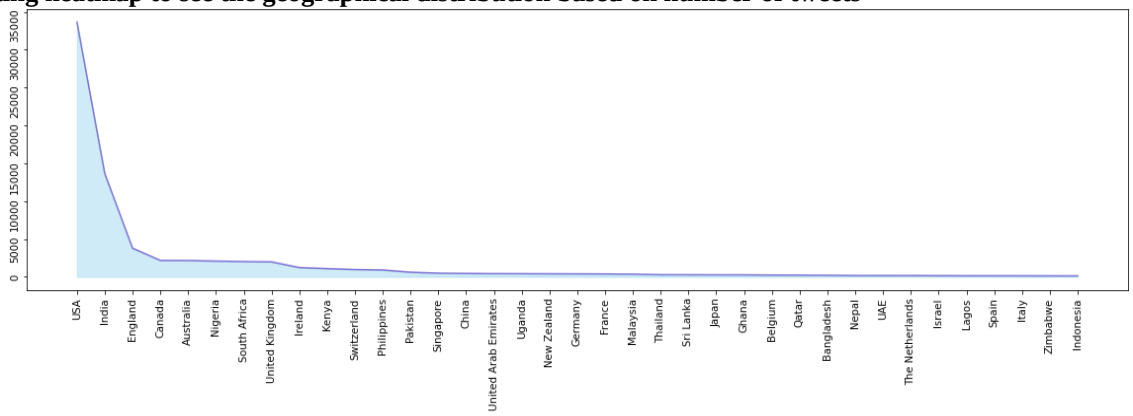


Fig. 1 Correlation Matrix of attributes of dataset

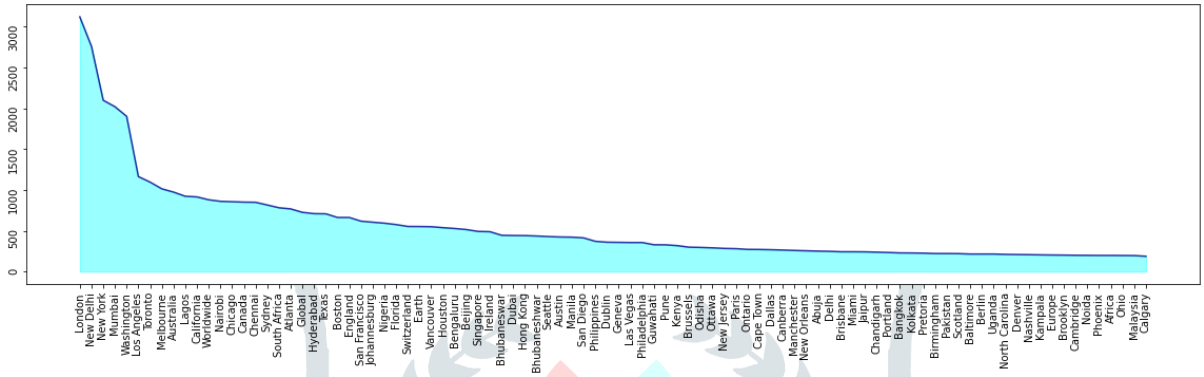
Fig. 1demonstrates the correlation among different attributes or columns of the dataset. The scale is between 0 to 1. The dependency among the attributes increases as we proceed among x and y axis of the graph.

IV. DETAILED ANALYSIS OF TWEETS

4.1 Plotting heatmap to see the geographical distribution based on number of tweets

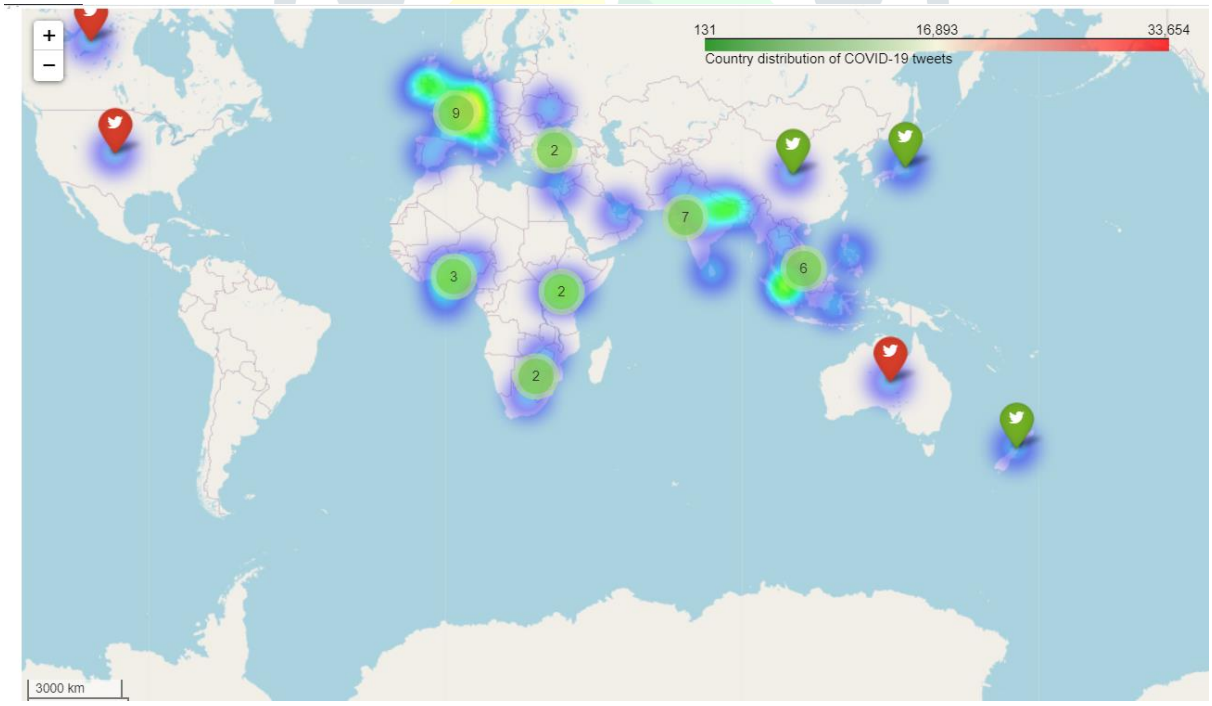


Tweets by country

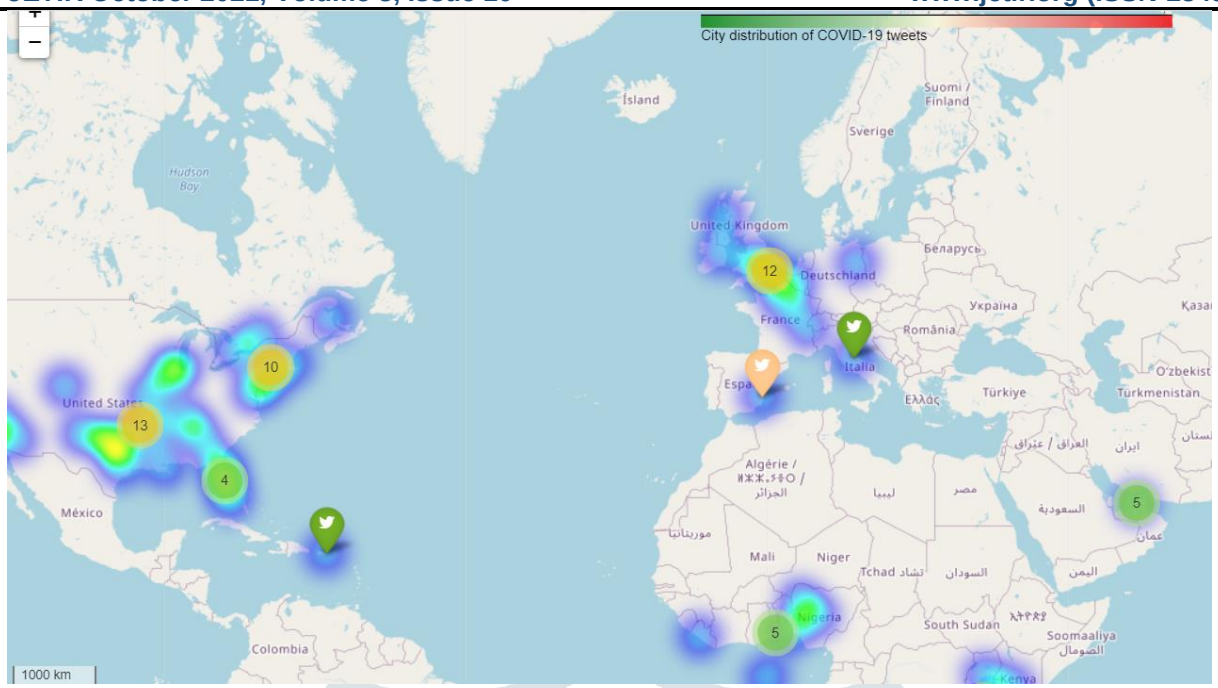


Tweets by city

We see that the US, India, and England are the top 3 countries with the highest tweets with **the** USA having significantly higher tweets (~20%). The curve is almost flattened beyond the top 3 countries with ~2k (or less) tweets per country. The top 5 cities with the highest tweets are London, New Delhi, New York, Mumbai, Washington with each having 2k-3k tweets.



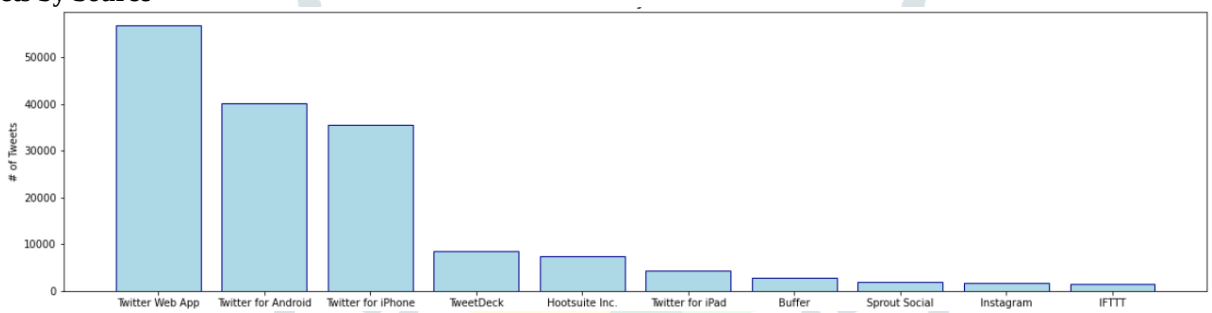
Heat base map of tweets by country



Heat base map of tweets by city

In the heatmaps above, we can see a high concentration in regions with higher tweets. The color (red, yellow, green) of **the** icon tells the intensity of tweets from high to for the different countries and cities and its number of tweets.

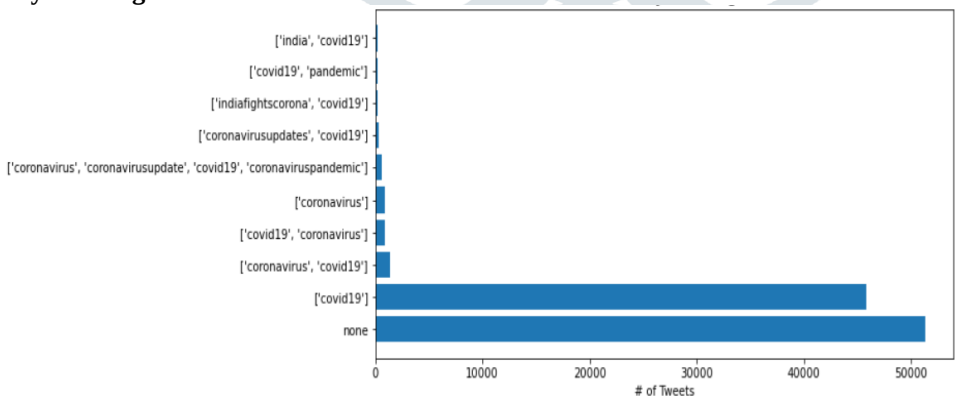
4.2 Tweets by Source



Tweets done from different devices of Twitter

32% of people have tweeted using the Web App, closely followed by Android users with 22%, and iPhone users with 20%.

4.3 Tweets by Hashtags



Tweets by Hashtags

As expected, ~70% of the tweets mention either covid19 or coronavirus as the hashtag.

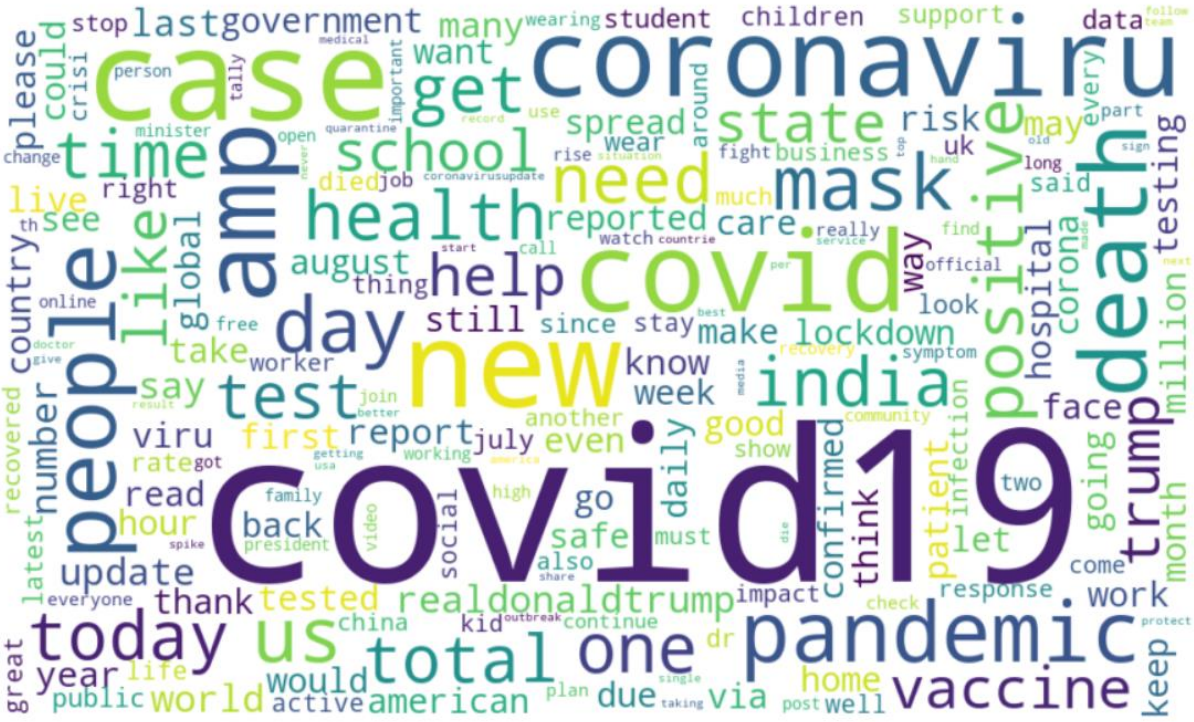
V. Natural Language Processing (NLP)

Natural language processing (NLP) is an artificial intelligence field that aids computers in comprehending, interpreting, and manipulating human language [12]. NLP has come to be a handy tool in reducing the gap of understanding between human linguistics and understanding of the machine language. The primary disadvantages of NLP now are related to the fact that language is extremely difficult [13]. Because the process of comprehending and manipulating language is so complicated, it's typical to utilize a variety of approaches to address various problems before tying everything together.

Developers can use natural language processing (NLP) to organize and arrange knowledge for tasks including automatic summarization, translation, identification, connection extraction, sentiment analyzing, speech enhancer as well as recognition [14].

VI. Sentiment of Tweets

The tweets in a dataset are pre-processed by replacing other than alphabets and numerals with space and by removing links from tweets. The word cloud of most common words in tweets is created to calculate sentiments.

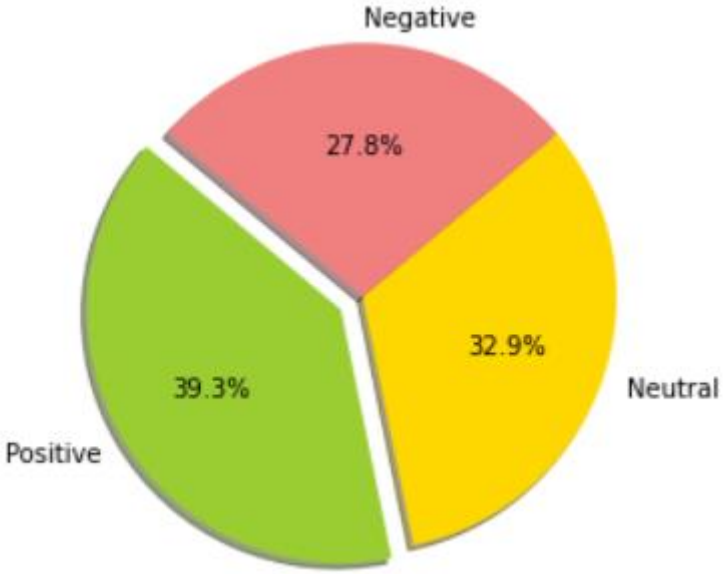


Word cloud of most common words in tweets

As expected, the most common words that pop up in these tweets are all corona-related viz. covid19, coronavirus, vaccine, death, mask, lockdown, school, health, etc.

6.1 Tweet's Sentiment Classification

Tweets are classified into three major groups: - Positive, Negative, and Neutral.

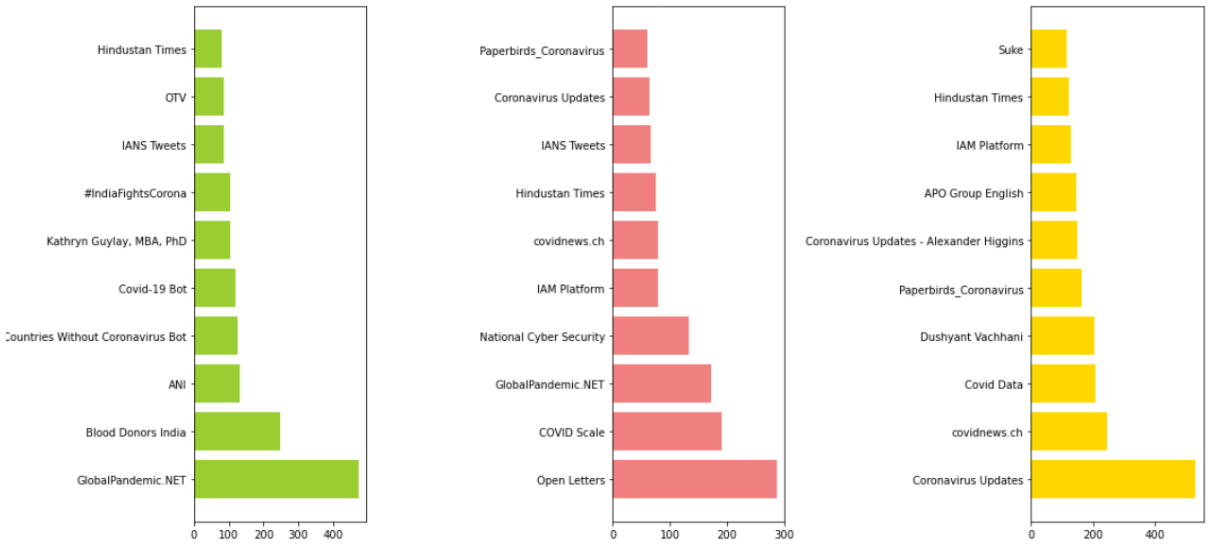


Tweet's sentiment distribution

A total of ~70k tweets have a positive sentiment making the largest pie in the chart with ~40% tweets.

6.2 Classification of top 10 Twitter accounts

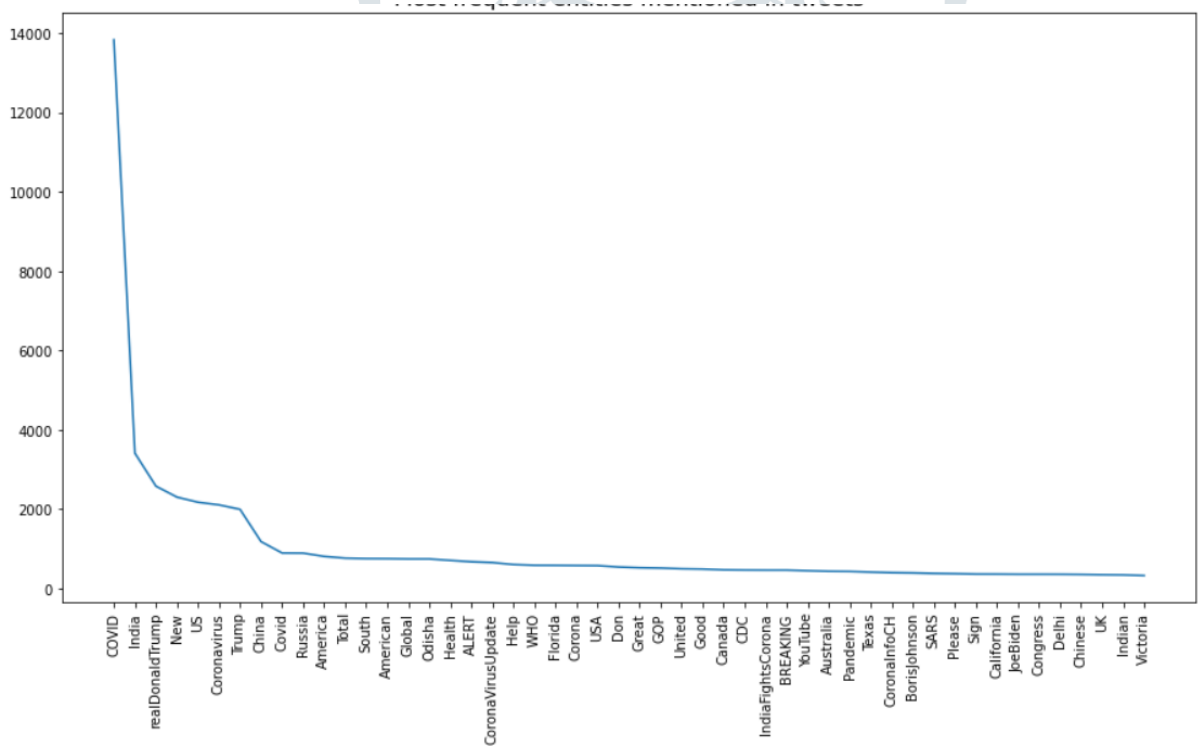
Twitter accounts are classified based on their sentiments.



Positive, negative, and Neutral tweets

'GlobalPandemic.NET' has the highest number of positive tweets while 'Open Letters' leads the list for having maximum negative tweets and 'Coronavirus Updates' for neutral tweets.

6.3 Checking most frequent entities on tweets



Most frequent entities

In terms of entities that have been most talked about, as expected, COVID tops the list again. Interestingly, other entities to notice are Donald Trump (3rd and 7th on the list), Joe Biden, Boris Johnson, Congress, WHO, CDC.
We also see a few countries like India, the US, China, Russia, and cities/ states like Odisha (a city in India), Florida, Texas, etc. that have been quite talked about.

6.4 Looking into tweets talking about 'Trump'

```
Negative    0.475422
Positive    0.264246
Neutral     0.260333
Name: sentiment, dtype: float64
```

Out of all the tweets that mention 'Trump', almost 50% have a negative tone. When inspected most of these tweets had people talking about the increasing corona cases. Some of the tweets talking about the country's inefficiency in promoting awareness about coronavirus, having faulty ventilators, lack of much needed preventive measures etc.

VII. Conclusion and Closing Thoughts

- US (New York, Washington), India (New Delhi, Mumbai), and England (London) have the highest tweets. These are also the most populated and metropolitan cities indicating that people in these cities are more active on Twitter.
- People using Twitter on mobile are 1.5x of the ones using the web when combined the Android and iPhone users collectively making ~45% of tweets.
- Looking at the tweets' trend, people have tweeted more following a spike in covid19 cases especially when cases on June 24 were the highest to date, we see a huge spike in the number of tweets on June 25.
- As expected, most of the news channels like CNN, National Geographic, CGTN, NDTV, and Times of India are the top 5 Twitter accounts to be followed with CNN having over 50m followers, and National Geographic having ~25m followers.
- While doing text analysis, we captured the words like covid19, coronavirus, pandemic, vaccine, death, mask, etc. which were most talked about.
- Overall, there were more positive tweets (~40%) than negative or neutral, indicating people still have great hopes for the world to become normal soon again. When analyzing positive tweets further, we found 'GlobalPandemic.NET' leads the list for having the highest number of positive tweets while 'Open Letters' leads the negative tweet list and 'Coronavirus Updates' for neutral tweets.
- Apart from the most obvious word 'COVID', the other entities to be most talked about were Donald Trump, Joe Biden, Boris Johnson, Congress, WHO, CDC. A few countries like India, the US, China, Russia, and cities/ states like Odisha (a city in India), Florida, Texas, etc. were very frequently mentioned as well.

REFERENCES

[1] https://wchh.onlinelibrary.wiley.com/doi/10.1002/psb.1843

[2] https://www.worldometers.info/coronavirus/

[3] Asur, S., & Huberman, B. A. (2010, August). Predicting the future with social media. In *2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology* (Vol. 1, pp. 492-499). IEEE.

[4] Miller, D., Sinanan, J., Wang, X., McDonald, T., Haynes, N., Costa, E., ... & Nicolescu, R. (2016). *How the world changed social media* (p. 286). UCL press.

[5] Wu, S., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011, March). Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web* (pp. 705-714).

[6] Kwok, S. W. H., Vadde, S. K., & Wang, G. (2021). Tweet topics and sentiments relating to COVID-19 vaccination among Australian Twitter users: Machine learning analysis. *Journal of medical Internet research*, *23*(5), e26953.

[7] Liu, F., Zheng, L., & Zheng, J. (2020). HieNN-DWE: A hierarchical neural network with dynamic word embeddings for document level sentiment classification. *Neurocomputing*, *403*, 21-32.

[8] Wang, Y., Sun, A., Huang, M., & Zhu, X. (2019, May). Aspect-level sentiment analysis using as-capsules. In *The World Wide Web Conference* (pp. 2033-2044).

[9] Chatterjee, A., Gupta, U., Chinnakotla, M. K., Srikanth, R., Galley, M., & Agrawal, P. (2019). Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*, *93*, 309-317.

[10] Lu, Q., Zhu, Z., Zhang, D., Wu, W., & Guo, Q. (2020). Interactive rule attention network for aspect-level sentiment analysis. *IEEE Access*, *8*, 52505-52516.

[11] Bucher, T. (2013). Objects of intense feeling: The case of the Twitter API. *Computational Culture*, (3).

[12] Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. I. (2012, April). BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 102-107).

[13] Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, *37*(1), 51-89.

[14] Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544-551.