MENTAL HEALTH OF INDIA DURING COVID

-by Shivam Kumar Giri

AIM

Study the given twitter data to understand the emotions of people against Covid-19 and classify the emotion based on tweets.

ABSTRACT

COVID-19 is a **humanitarian crisis on a global scale**. The virus continues to spread throughout the globe, placing health systems under unprecedented stress in the battle to save lives.
This Covid-19 pandemic has severely affected countries around the world. The intensity of the pandemic is increasing very fast in India. The number of new cases is increasing every day, every week. In a span of six months, the total number of **cases crossed 10 Million (1 Crore)** and total number of deaths is almost I lakh. It has been observed that the sudden outbreaks of such pandemics affect public mental states and emotions.

This pandemic also results in either constructive or destructive behavioral changes among people. Anger, Sadness, fear are the most common emotions witnessed among the people during several pandemics. Social media platform like Twitter and others have rich sources of information from people. Here we are going to build a **Machine Learning application which can understand emotions of people and classify it based on the sentiments of tweets.**

DATASET

The **dataset for training*** consists of Four Columns: author, tweet_id, **sentiment, content**. Only columns relevant to us is 'sentiment' and 'content'.

The Dataset for implementation is scattered into 28 files and each file contains dump of several tweets in json format. Each tweet has four attributes that are: **text, location, date and time.**

LIBRARIES USED

Here, the application is developed using python. I used the following libraries for implementation of the mental health detector:

Numpy: for Numerical computing

Pandas: for data manipulation and analysis.

Tensorflow: for dataflow and differentiable programming

Keras: for neural-network

json: for reading json file

os: for interacting with the operating system.

worlcloud: for creating word cloud.

Matplotlib: for data visualization

stopword: for the stop words

Sklearn: for classification algorithms and other Machine Learning metrics.

nltk: for word Tokenizer and stopwords.

re: for token search

- Dataset Downloaded from https://data.world/crowdflower/sentiment-analysis-in-text

BASIC DATA ANALYSIS

Shape of dataset: `(40000, 4)`
Columns of dataset: `'tweet_id', 'sentiment', 'author', 'content'`
Total Number of sentiments: 13

```
['empty', 'sadness', 'enthusiasm', 'neutral', 'worry', 'surprise',
'love', 'fun', 'hate', 'happiness', 'boredom', 'relief', 'anger']
```

Head Value of Data:

	tweet_id	sentiment	author	content
0	1956967341	empty	xoshayzers	@tiffanylue i know i was listenin to bad habi...
1	1956967666	sadness	wannamama	0Layin n bed with a headache ughhhh...waitio...
2	1956967696	sadness	coolfunky	Funeral ceremony...gloomy friday...
3	1956967789	enthusiasm	czareaquino	wants to hang out with friends SOON!
4	1956968416	neutral	xkilljoyx	@dannycastillo We want to trade with someone w...

We Removed the columns of tweet_id and author, as it was not much required as per our objective.

We also lower the case of the content section, removing all its punctuations

Now the data is ready for tokenization, we tokenize the data and add padding to it, using the TensorFlow using following code:

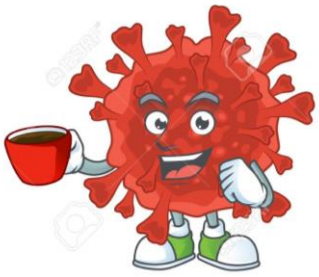
```
tokenizer = Tokenizer (num_words=vocab_size, oov_token=oov_tok)

tokenizer.fit_on_texts(training_sentences)


word_index = tokenizer.word_index


training_sequences = tokenizer.texts_to_sequences(training_sentences)

training_padded = pad_sequences(training_sequences,
maxlen=max_length, padding=padding_type, truncating=trunc_type)
```

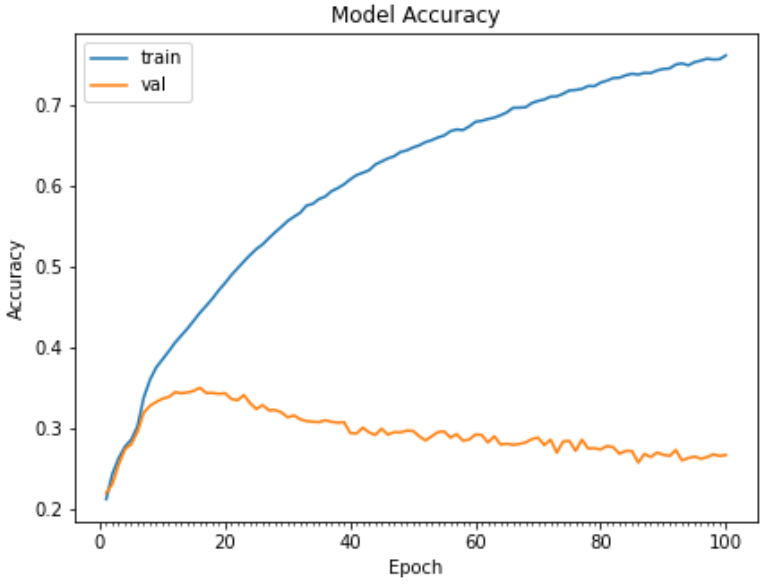


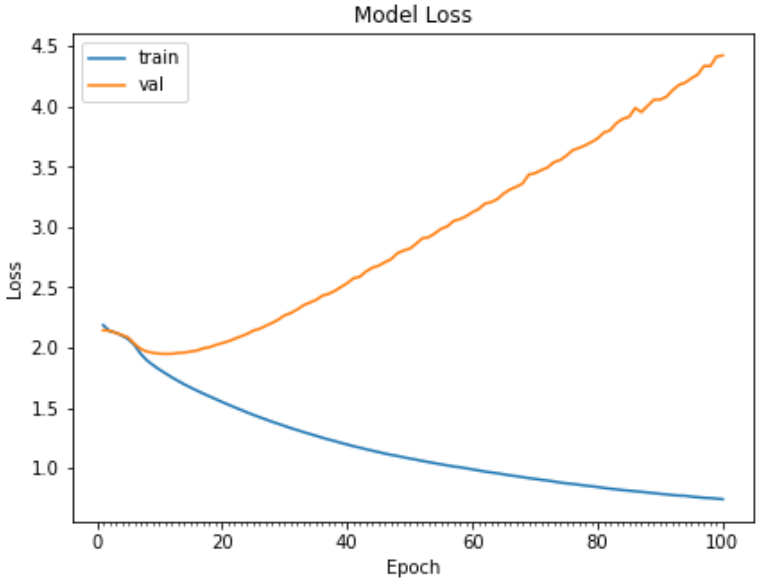
TRAINING AND VALIDATION

After Tokenization, I used the following model for training and validation:

```
Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_2 (Embedding)      (None, 100, 16)           160000
_________________________________________________________________
global_average_pooling1d_2 ( (None, 16)                0
_________________________________________________________________
dense_4 (Dense)              (None, 24)                408
_________________________________________________________________
dense_5 (Dense)              (None, 13)                325
=================================================================
Total params: 160,733
Trainable params: 160,733
Non-trainable params: 0
_________________________________________________________________
```

After compiling the model, we yield the following validation graph over accuracy and loses, we can see increase in training accuracy and reduced losses, but a subsequent downfall in validation accuracy, we do have 100 epochs:





IMPLEMENTATION OF APPLICATION IN GIVEN JSON DATA

We have read all the 28 Json files and created a DataFrame out of all the available file.

The shape of Data frame is **(496448,4);**

We have analyzed the following aspect of interpretation from the provided Data:

1. **Word Cloud from the data**
2. **Predicting the data and Emotion Analysis**
3. **Get information regarding location of Tweets**
4. **Get information regarding date of Tweets**
5. **Get information regarding time of Tweets.**
6. **Emotion of People w.r.t. their locations and dates.**
7. **Getting Trending Hashtags and their Trending Cycle**

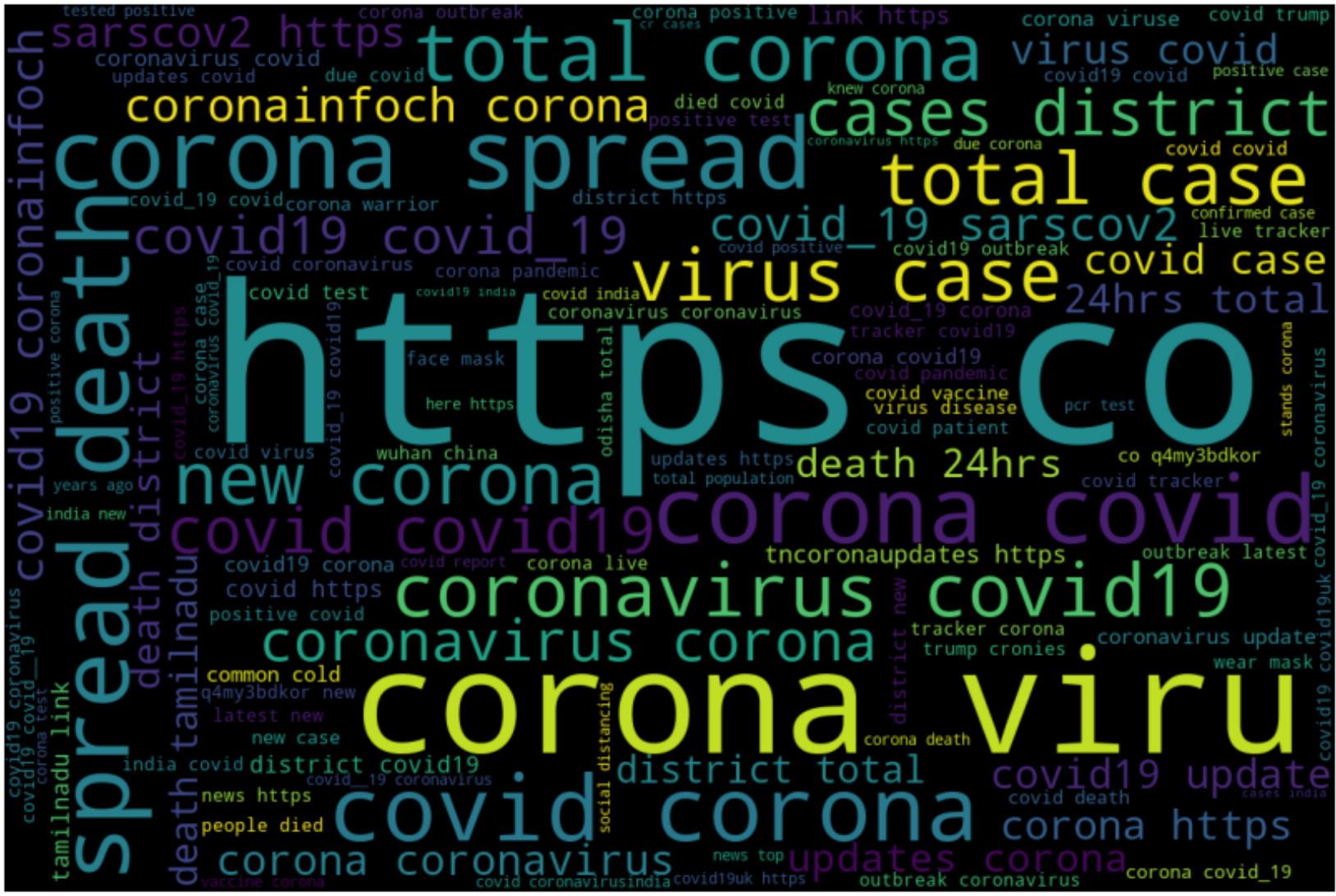
There are 4 columns of the data:

- Time of tweets
- Date of Tweet
- Location
- Text Content

We are processing the **data of 496448 tweets** to understands their emotions and classify them over 13: Different classes:

```
['empty', 'sadness', 'enthusiasm', 'neutral', 'worry', 'surprise',
'love', 'fun', 'hate', 'happiness', 'boredom', 'relief', 'anger']
```

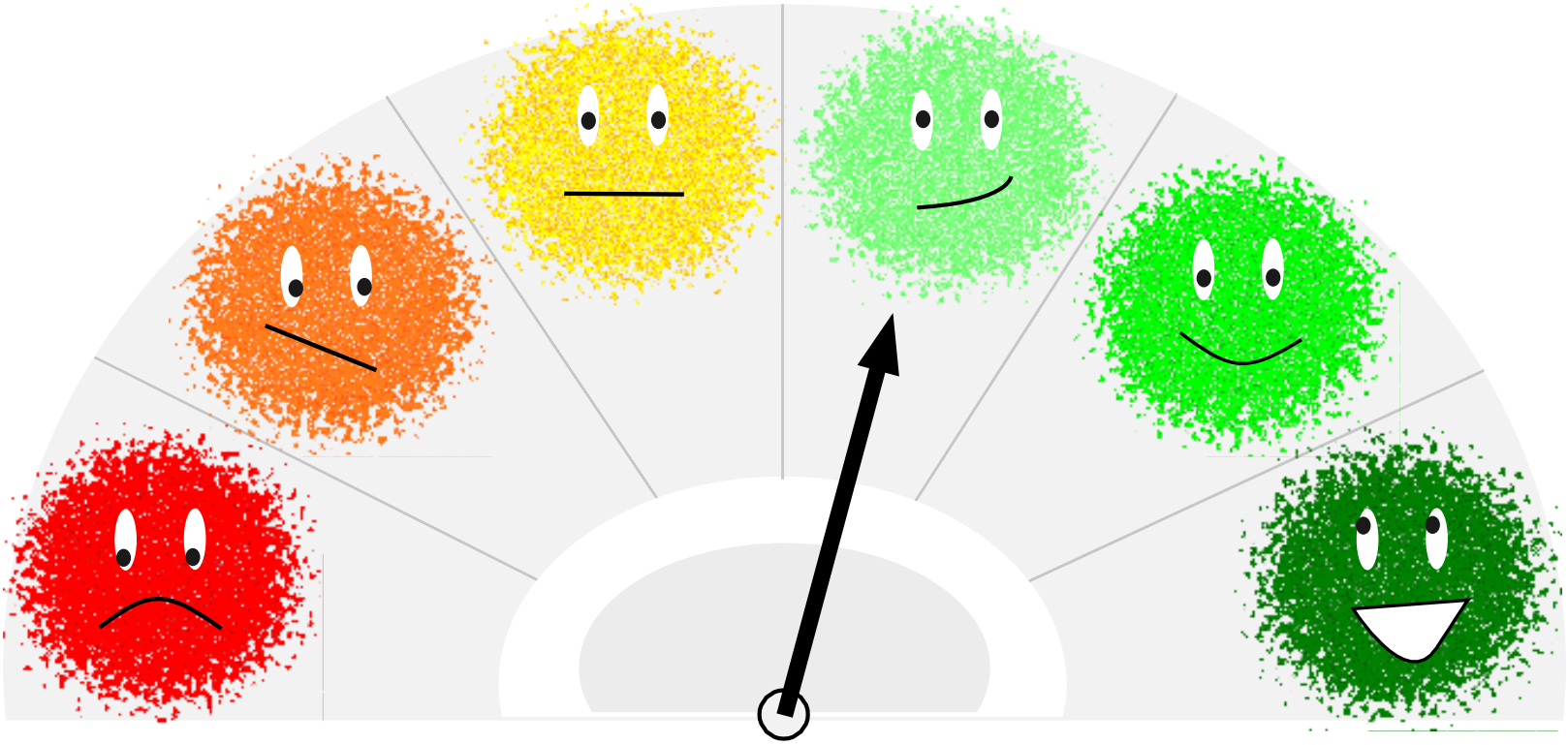
1. WORD CLOUD FROM THE DATA



We can interpret that the most common and used words in the tweets are corona, virus cases spreads and terms related to COVID -9. We can understand most of the sentiments are illustrated in the tweets are regd. this new Pandemic. And will see further the Emotions of people over regions and w.r.t. date and time of the tweets.

We can see that htts.com was the most used word indicating that Many People do share many links in the twitter making it the most used word

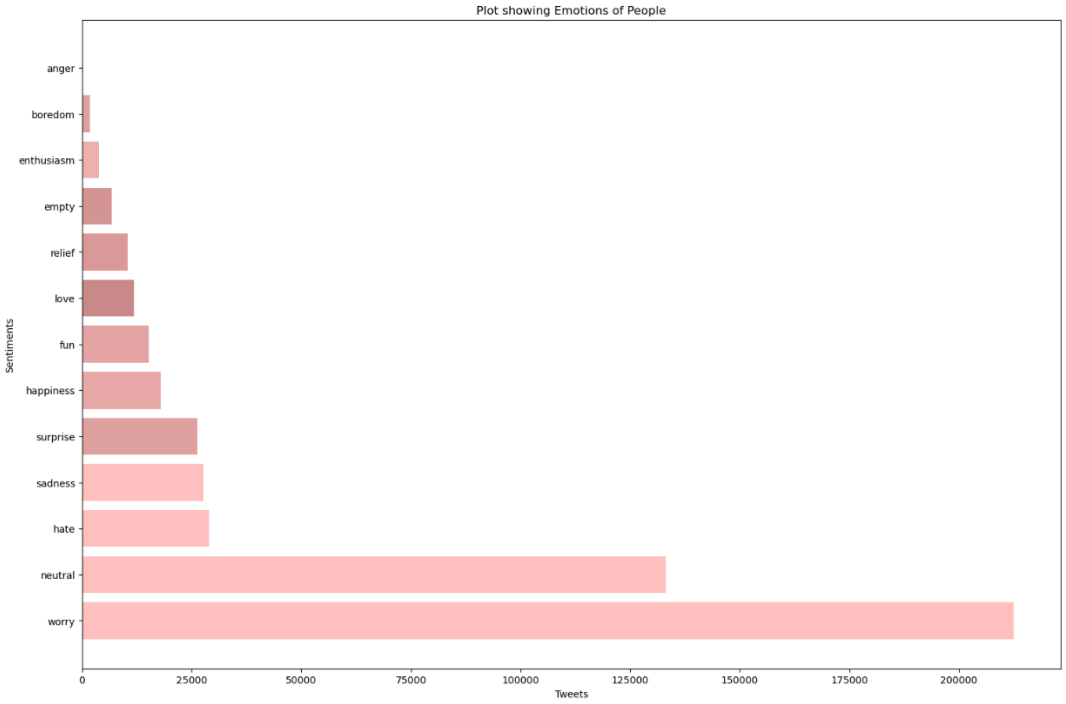
2. PREDICTING THE DATA AND EMOTION ANALYSIS

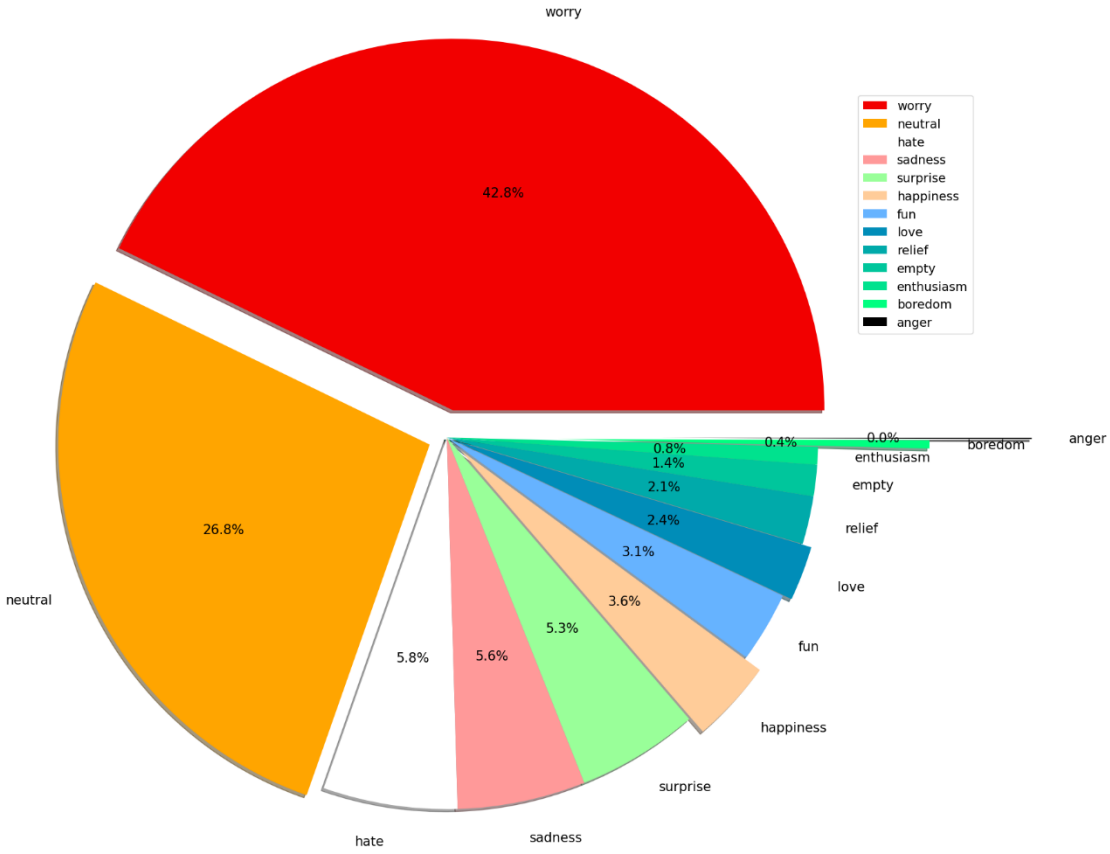


Here we predicted the emotions of people based on their tweets with model we trained earlier. We have added the result into a new column of database 'emotion'. The data is Tokenized first and then Padded and then the Emotion value of Tweet is evaluated. Head of new DataFrame is as below:

text	location	date	time	emotions
Curve flattening? Kenya records 48 new virus c...	IN	Sep 22	05:08:45	worry
Victoria and Melbourne Covid trend map: where ...	Erbil, Iraq	Sep 22	05:08:34	worry
NSW and Sydney Covid trend map: where coronavi...	Melbourne, Australia	Sep 22	05:08:33	worry
IT'S BAKE OFF DAY! 🙌\n\nWho else will be tuni...	Melbourne, Australia	Sep 22	05:06:02	surprise
@DanielAndrewsMP The Liberal party bots are ou...	Fareham	Sep 22	05:05:34	hate

We get the following conclusion from the prediction:

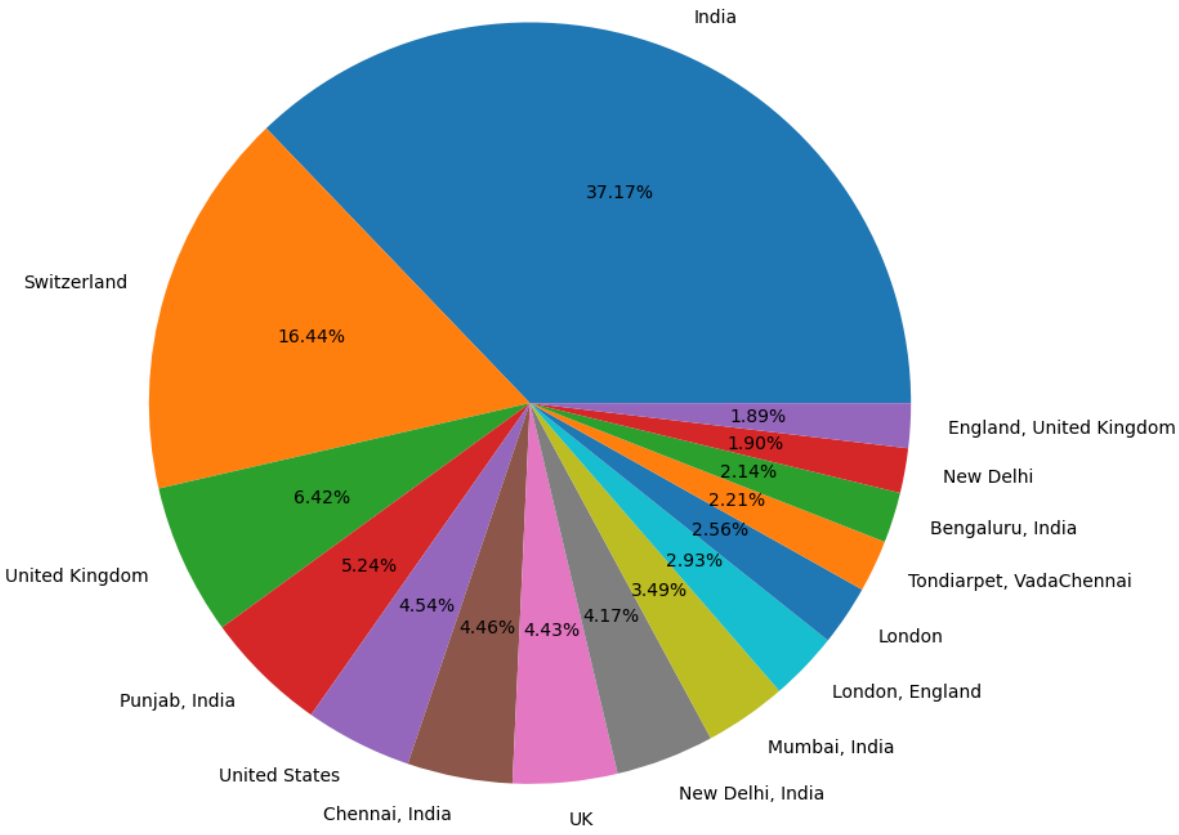




Hence, we can conclude that most of the people are over **Fear (Worry)** followed by **Neutral** over Covid19 Pandemic.

3. GET INFORMATION REGARDING LOCATION OF TWEETS

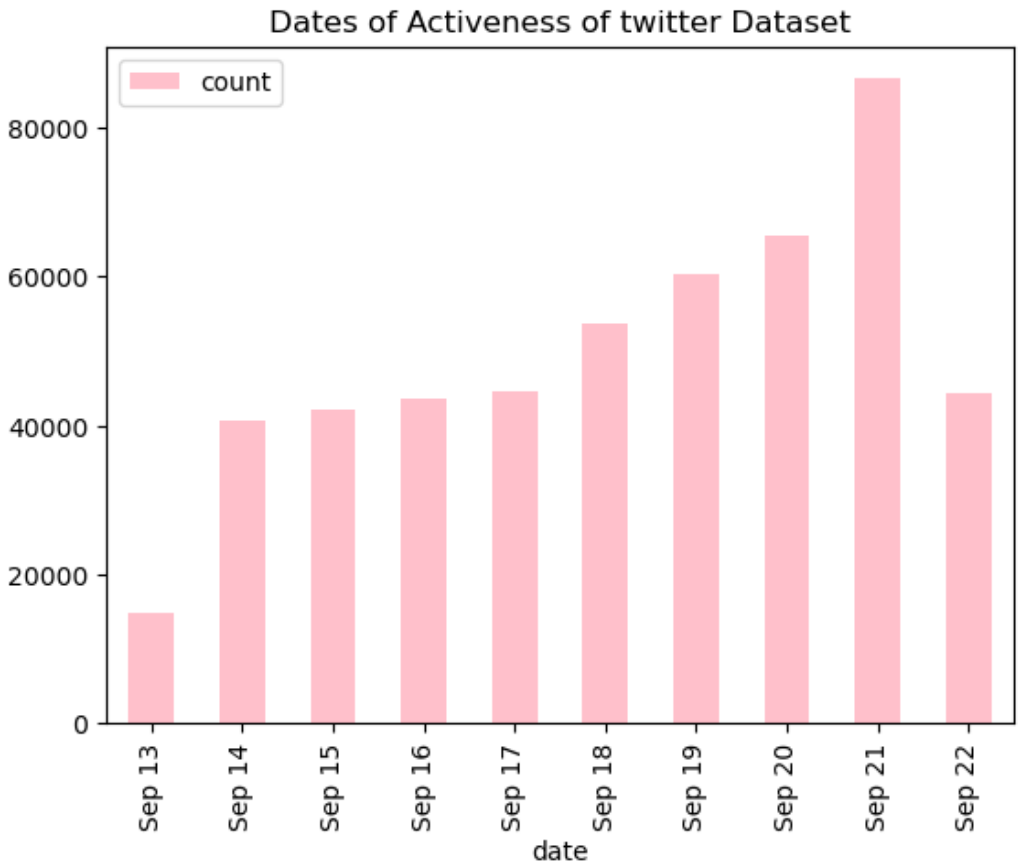
We have tweets from **3439 distinct location.** Location shares the maximum tweet are as follows (constitute **20% of the total tweets**):



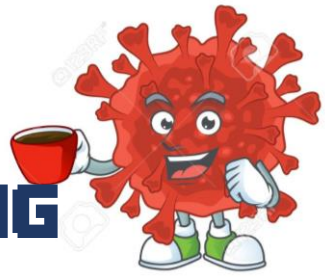
4. GET INFORMATION REGARDING DATE OF TWEETS

All the tweets range from **13 September to 22 September** in **the interval of 10 days**
The number of tweets is as follows:

date	count
Sep 13	14848
Sep 14	40580
Sep 15	42086
Sep 16	43735
Sep 17	44676
Sep 18	53779
Sep 19	60274
Sep 20	65575
Sep 21	86558
Sep 22	44337



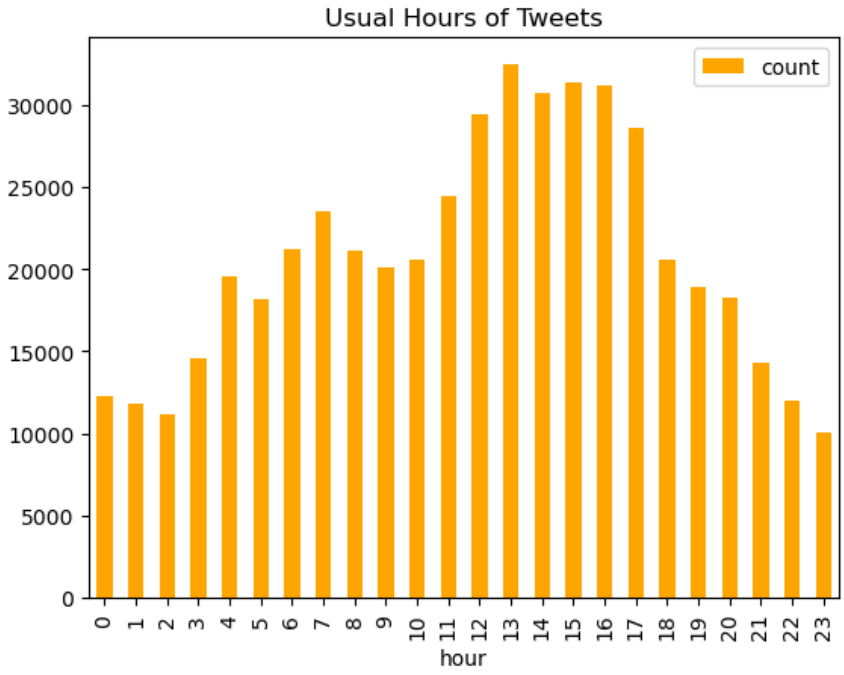
We can find that the number of **tweets gradually increased in number from 1st day and reached its peak at 21 September at 86558** which almost constitute about 18% of all the tweets.

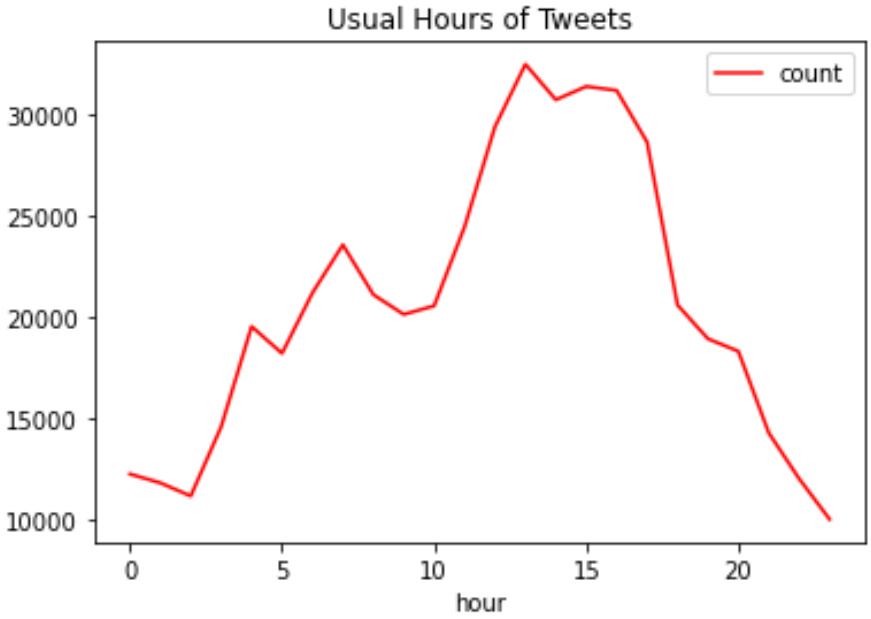


5. GET INFORMATION REGARDING TIME OF TWEETS.

We find Tweets **almost all time a day.** The number of tweets is as follow:

	count	hour
0	12249	0
1	11812	1
2	11163	2
3	14585	3
4	19531	4
5	18206	5
6	21190	6
7	23564	7
8	21106	8
9	20128	9
10	20548	10
11	24451	11
12	29392	12
13	32476	13
14	30722	14
15	31379	15
16	31183	16
17	28634	17
18	20593	18
19	18924	19
20	18308	20
21	14276	21
22	12014	22
23	10014	23

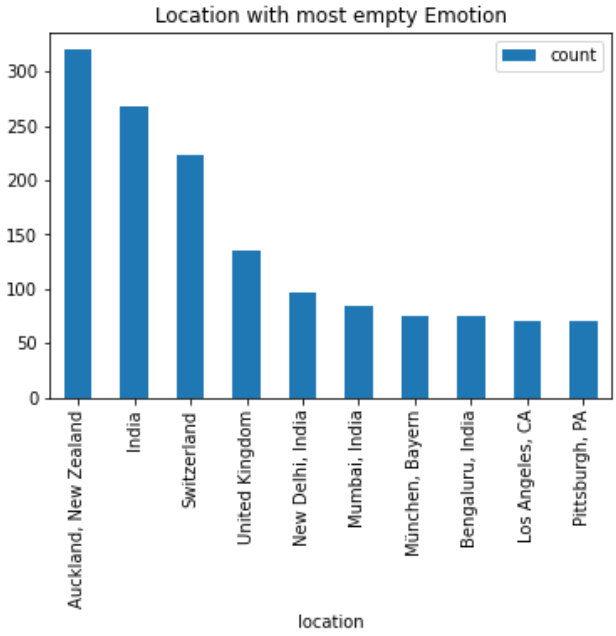


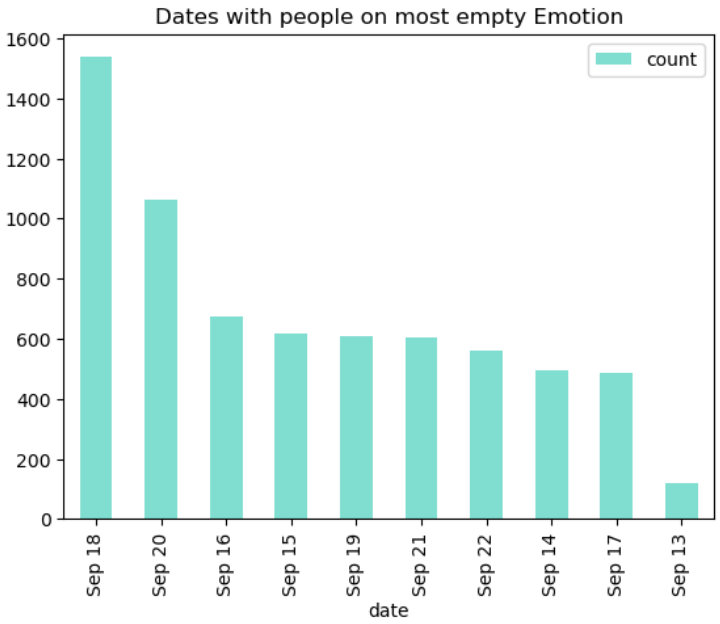


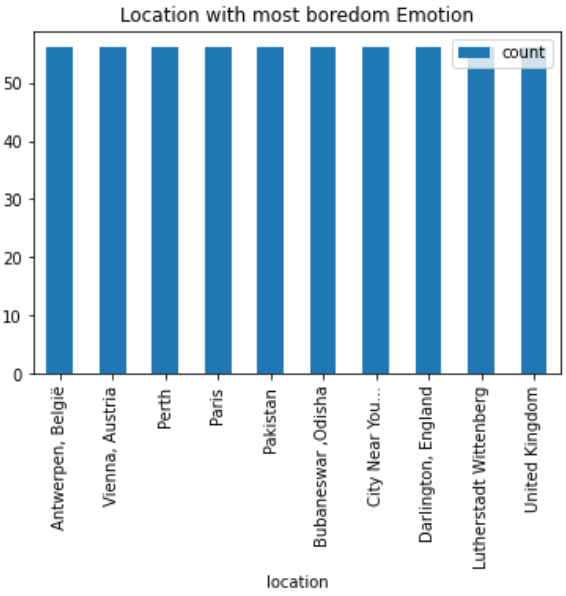
We can find out that people are mostly active at day time, with its **peak of 32476 at 12:00-16:00** or 12 PM- 4 PM in the afternoon and tweets gradually decrease in nights.

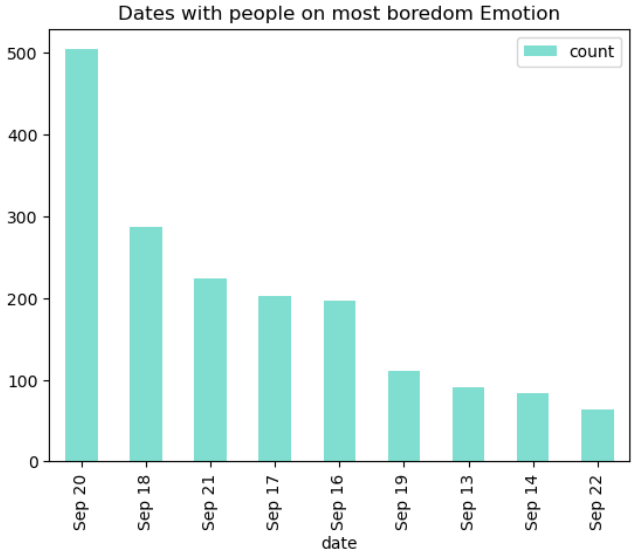
6. EMOTIONS OF PEOPLE W.R.T. THEIR LOCATIONS AND DATES.

We have calculated the emotions of people relative to the Locations they tweeted:



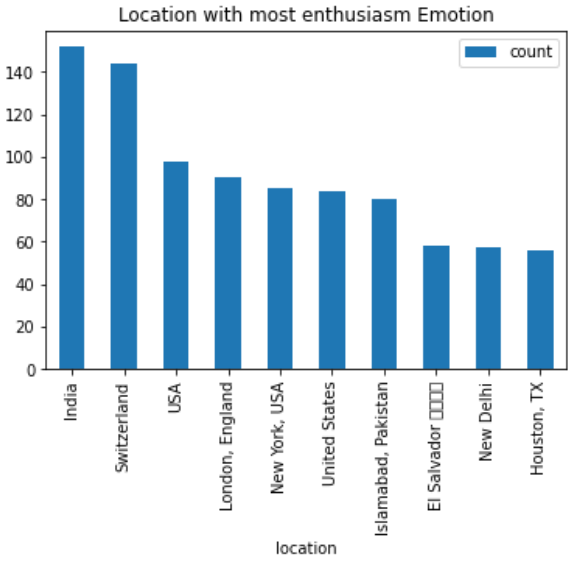


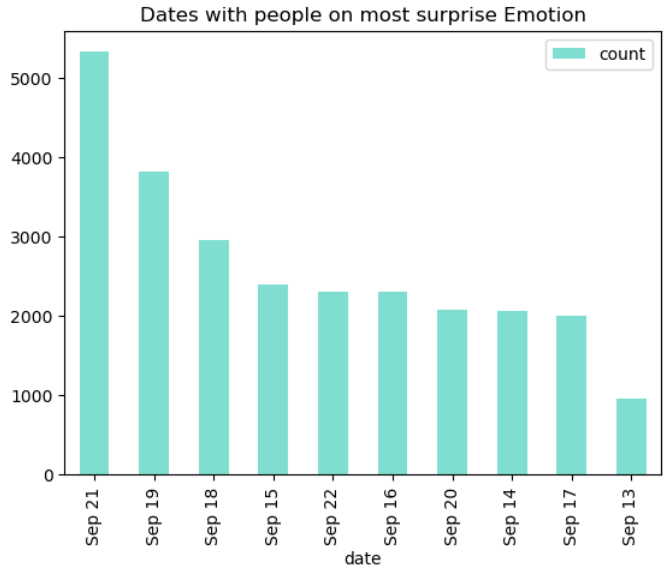


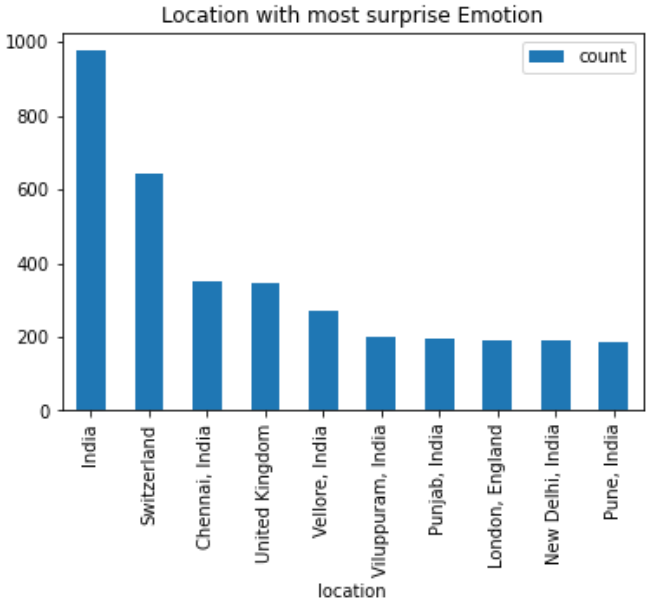


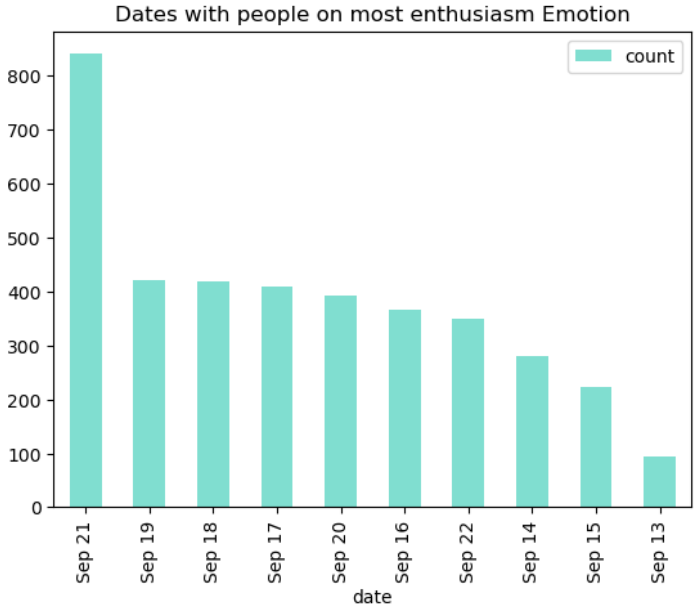
We can see that people **with boredom and empty, mostly in Auckland with 300+ Empty Emotional Tweets, followed by India, Switzerland** and Many other Parts of world with average 50+ tweet in each occasion, **but the dates of being boredoms and empty emotion in people is very irregular** as seen in graph

Mental health of India during COVID





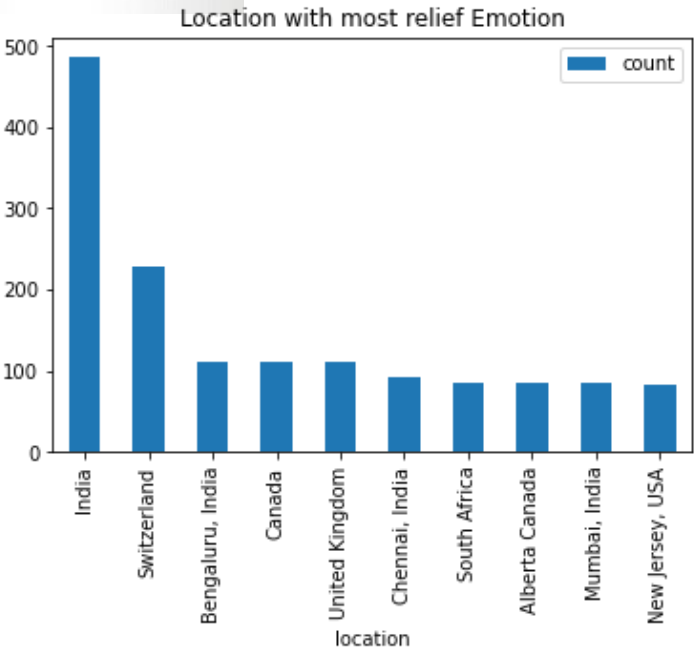


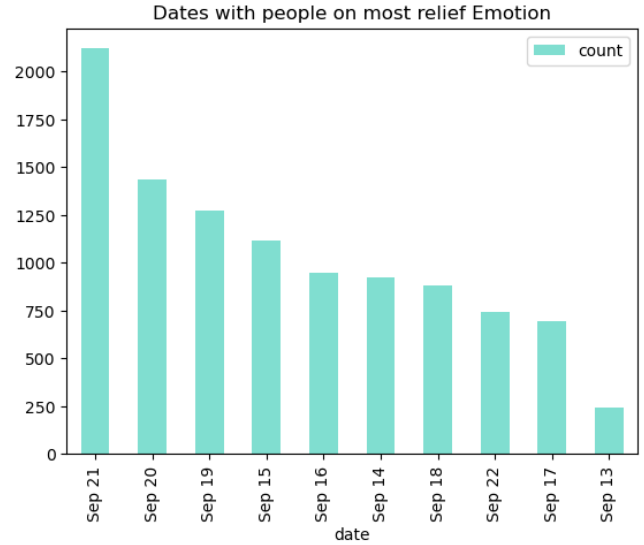


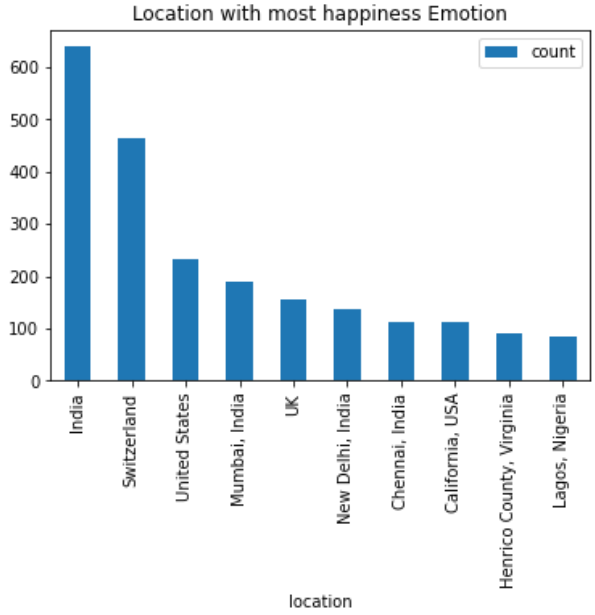
We can see **that India and Switzerland are leading in enthusiasm emotion each with 140+,** and the **dates with irregularity in trend, reaches its peak with 5000+ on September 21.**

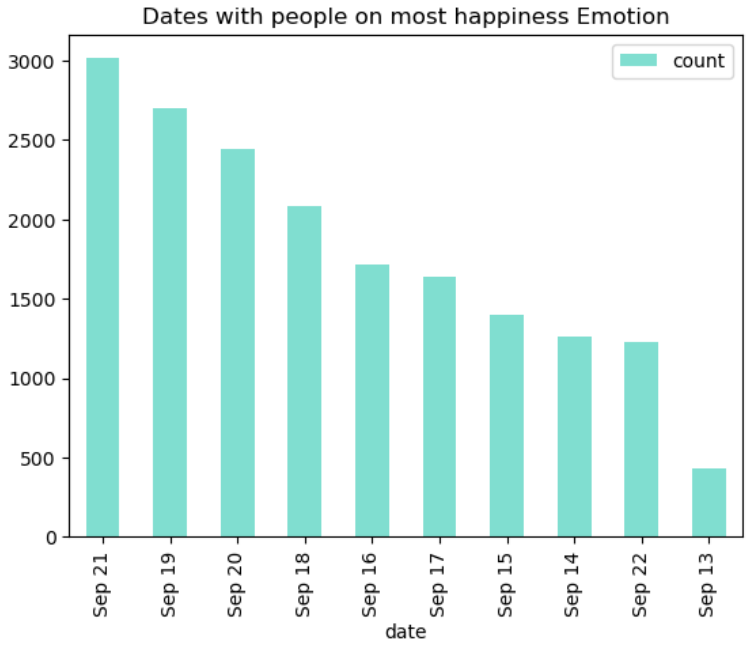
Again, we can see **India and Switzerland are leading in surprise emotion** each with 900+ and 600+ respectively, and tweets are on **its exceptional peak with 800+ (which is double the average of 300+) on 21st September.**

Mental health of India during COVID



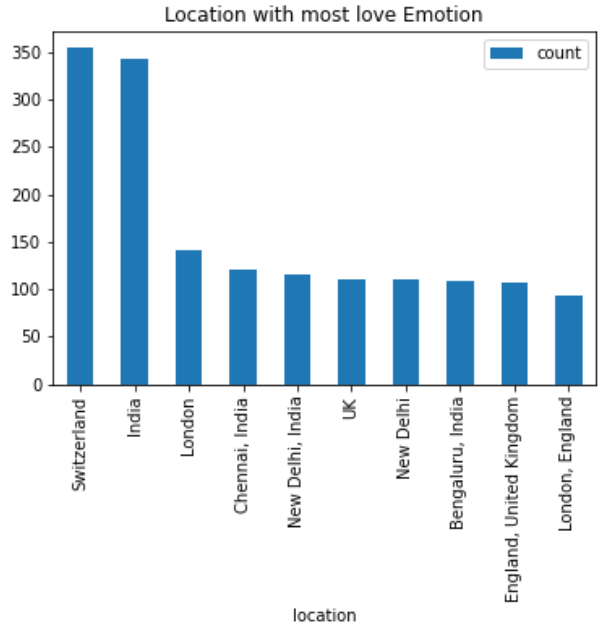


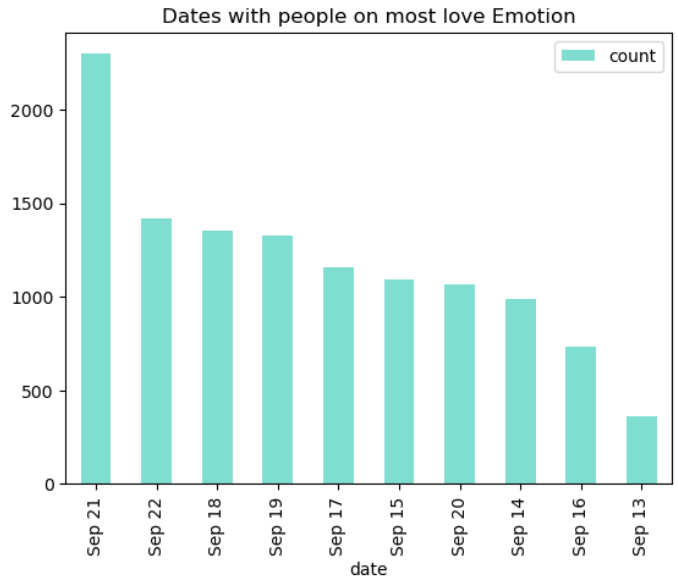


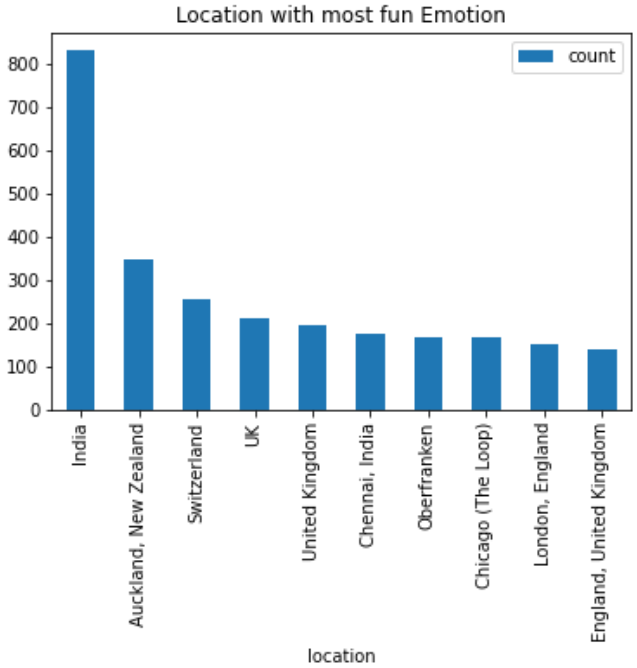


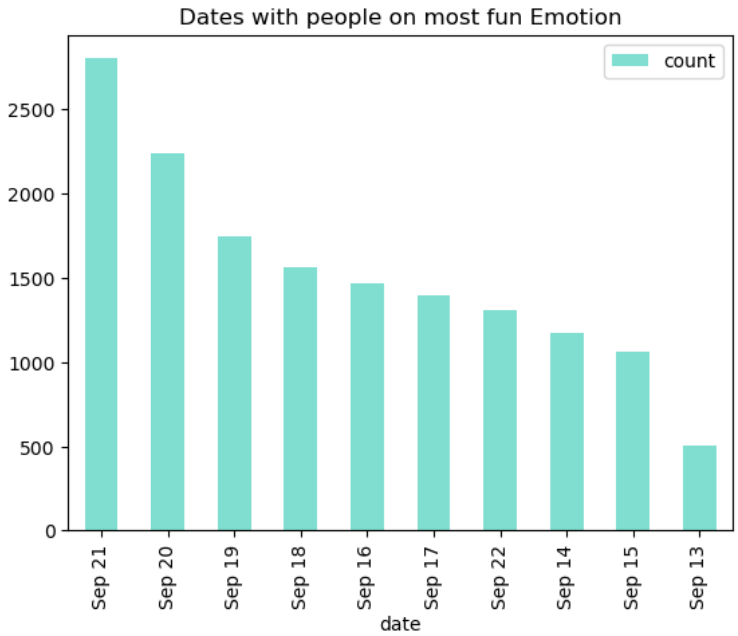
We can **see India and Switzerland clearly tops the relief and happiness Tweets with a good margin**, but the **dates of Tweets are with clear Irregular increase for Relief tweets,** and **peeks on 21 September and lowermost on 13 September.**

Mental health of India during COVID





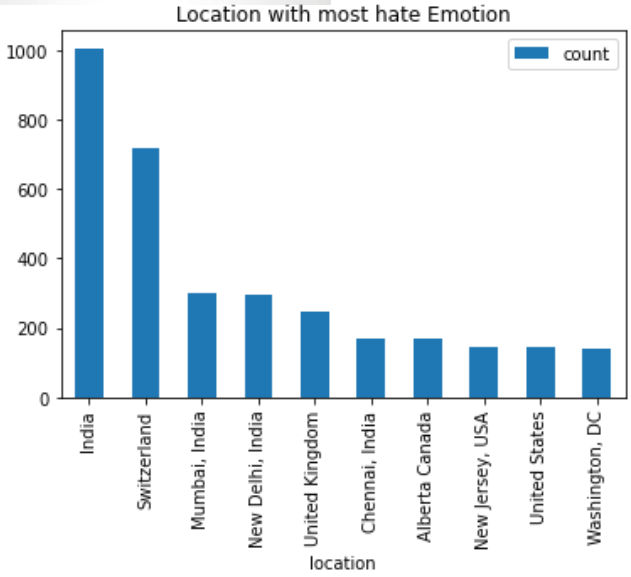


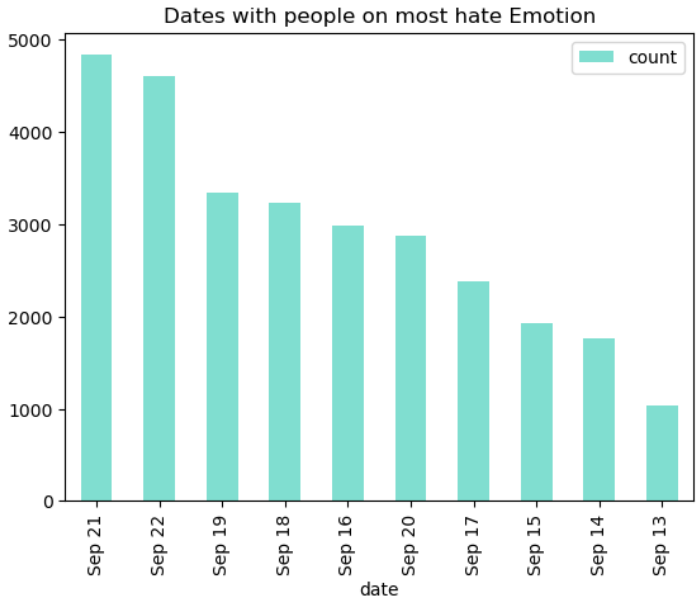


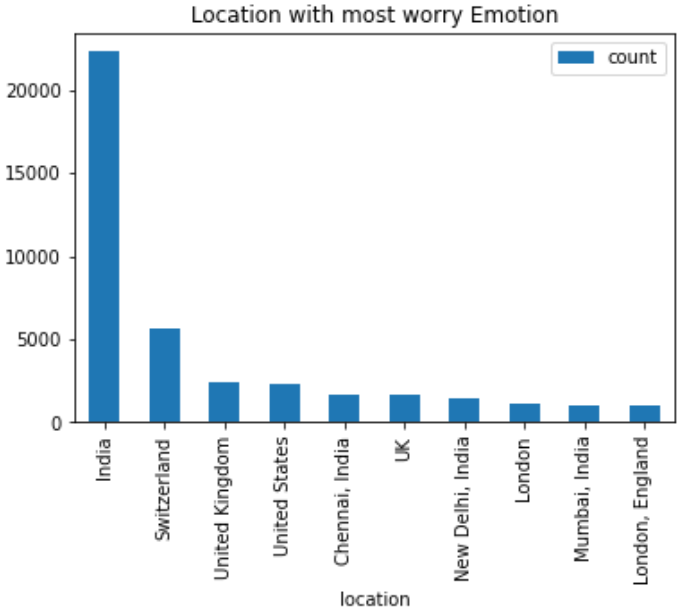
Switzerland and India Tops the Love Emotions with 350+ and 300+ respective with clear distinction from rest who are on an average of 100+. **India with clear distinction tops the list of fun Tweets with 800+,** followed by Auckland 300+

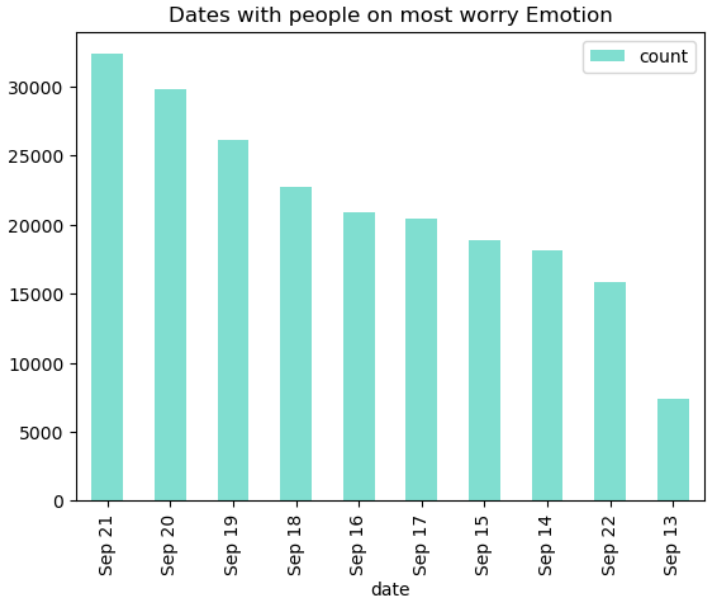
Both the Emotion of Fun and Love **have irregular increase with respect to Dates, and both peaks on 21 September and least on 13 September.**

Mental health of India during COVID





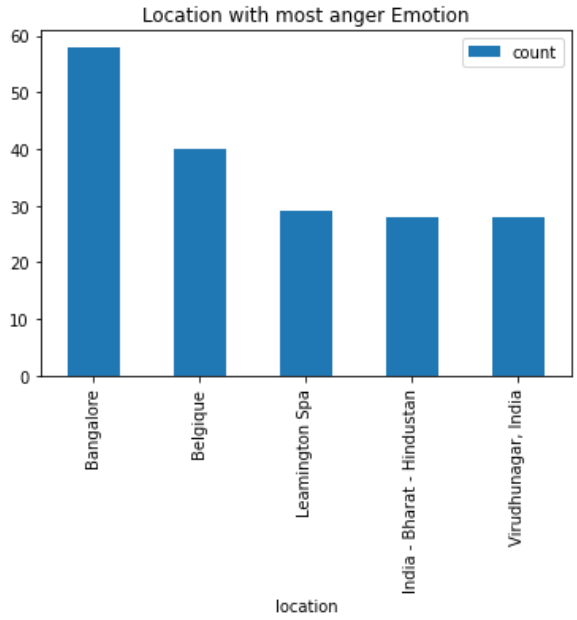


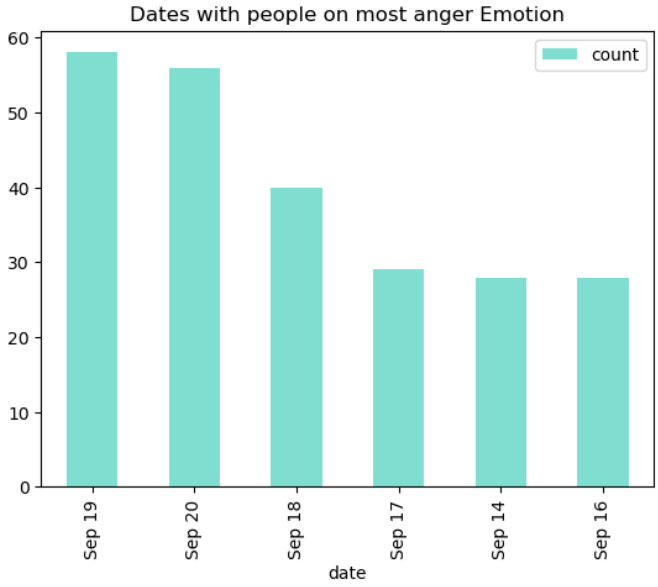


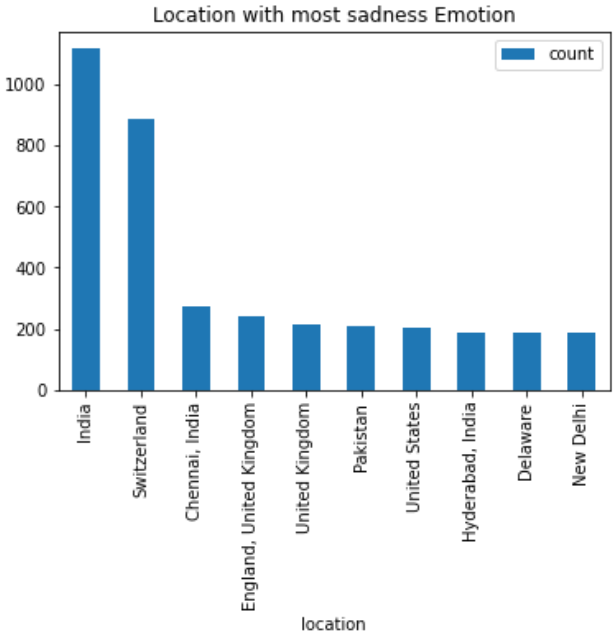
Taking the account of **hates tweets, India leads the row with 1000 tweets**, followed by Switzerland, and **hatred increases each day, with an irregular cycle.**

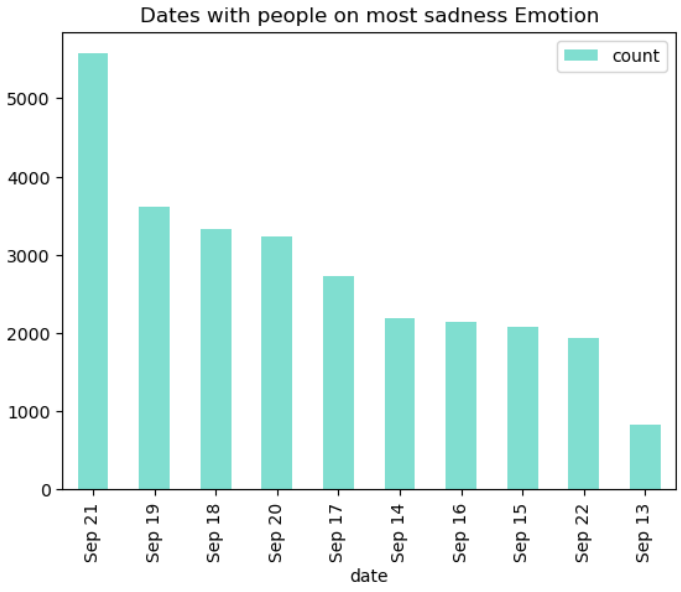
India clearly Emotion **tops the list** of Worry with **22000+ tweets with 5% of entire tweets we Gathered**, and **the Worry/Fear increases every day.**

Mental health of India during COVID



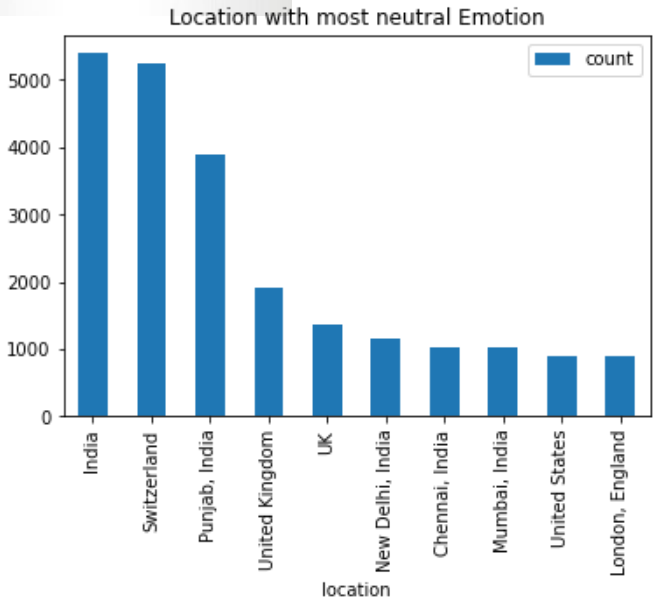


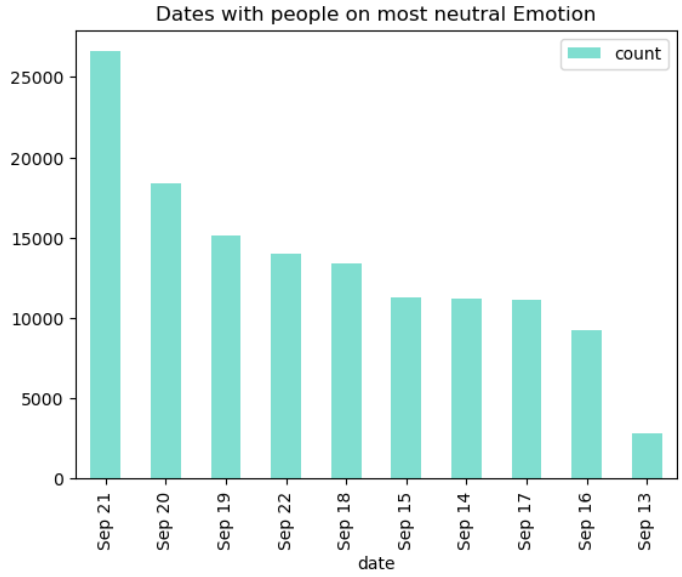




When it comes **to anger, Bangalore leading with 50+ tweets,** followed by Belgique, and most **anger Tweets on 19-20 September**.

Coming **to Sadness, tweets from India leads the row with 1000+ tweets** followed by Switzerland and the **emotion of sadness increases each following day.**





We can almost **5000+ from India and almost same from Switzerland are being Neutral about COVID** and **the emotion of Neutrality increases almost every day.**

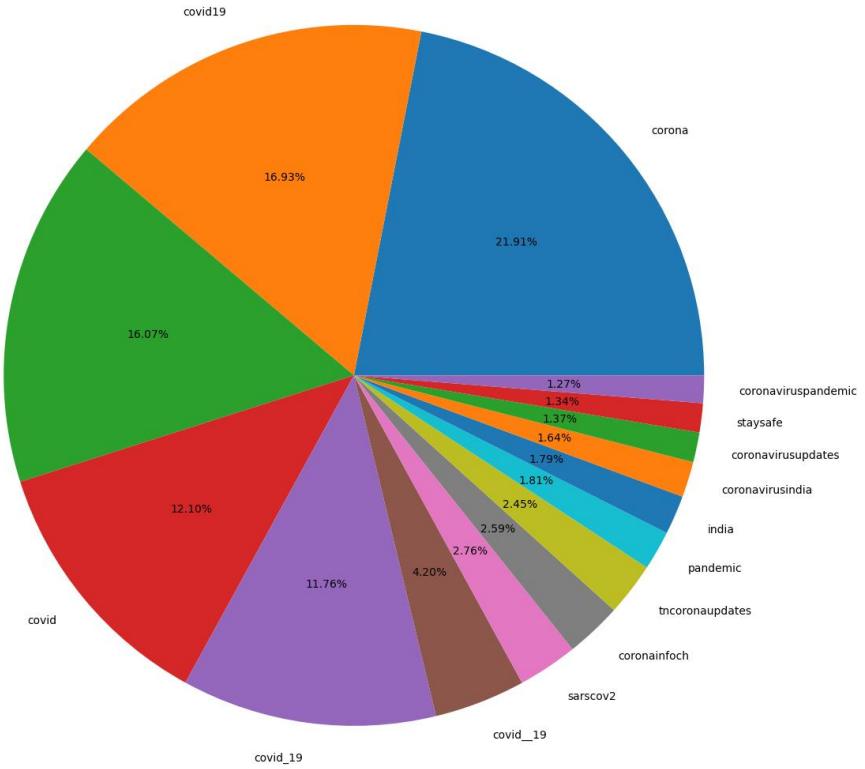
7. GETTING TRENDING HASHTAGS AND THEIR TRENDING CYCLE

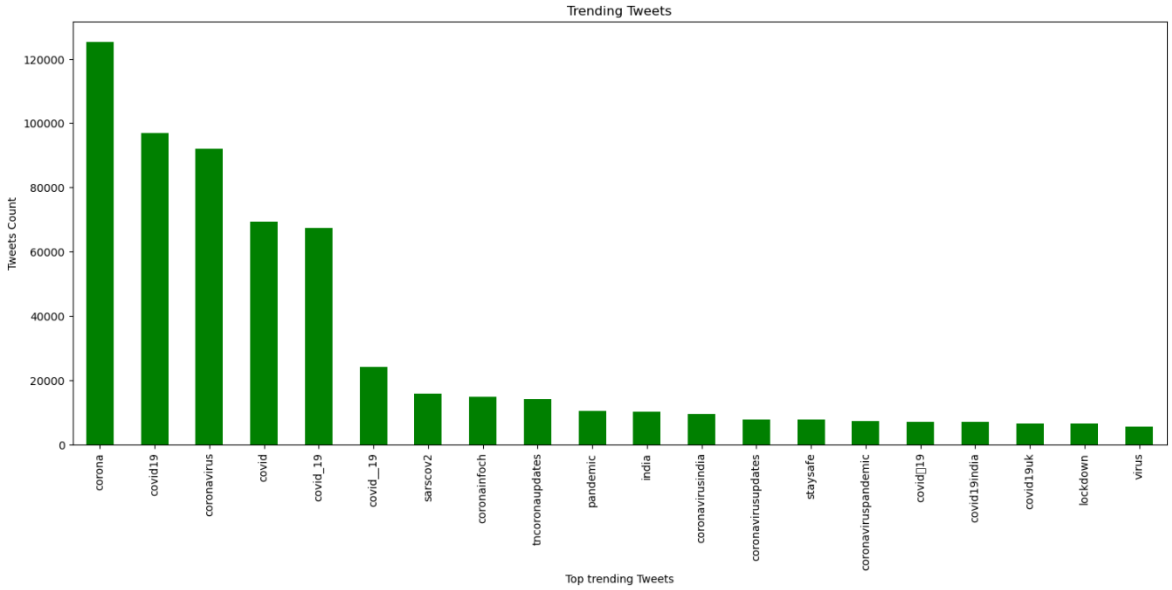
Hashtag is a type of metadata tag used on social networks such as Twitter and Instagram and other microblogging services. It lets users apply dynamic, user-generated tagging that helps other users easily find messages with a specific theme or content.

As per the analysis there are **1360194 hashtags** in the data of the tweets, out of this, **7675 hashtags are the unique hashtags**. **#corona** is most used Hashtag among all with **125391 hashtags**, which count almost to **¼ of the tweets**.

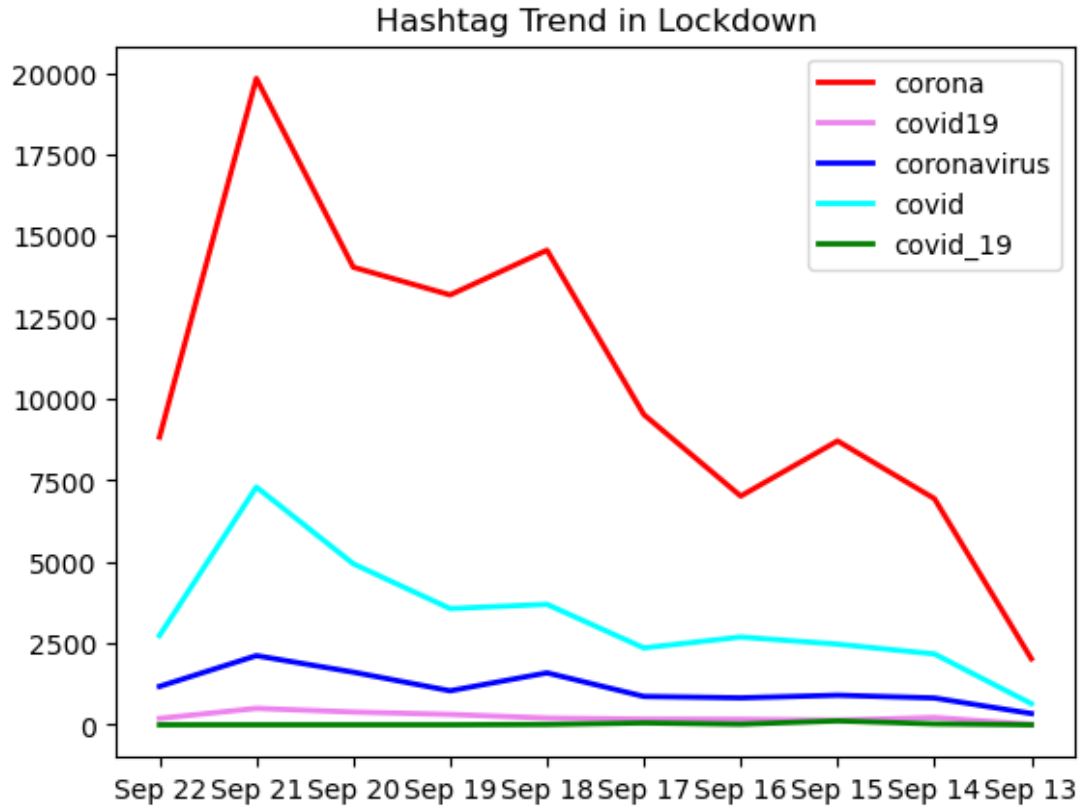
Out of these the most **trending hashtags with their number of occurrences** is as follows:

corona	125391
covid19	96883
coronavirus	91962
covid	69214
covid_19	67313
covid__19	24030
sarscov2	15782
coronainfoch	14828
tncoronaupdates	14017
pandemic	10374
india	10268
coronavirusindia	9397
coronavirusupdates	7864
staysafe	7661
coronaviruspandemic	7260
covid—19	7129
covid19india	6990
covid19uk	6644
lockdown	6513
virus	5630





The trend of top 5 hashtags are as follows:



Hence, we can see that **Hashtags increased on each the followings day** and become high trending on the later days, say 21 September, and was least on 13 September.

Also, we can see **the increase the use of hashtags** specially #corona and #covid with its peak on 21 September.

CONCLUSION

Hence, we analyzed the given tweets and try to carry out most of the Analysis with different aspects of location, time, tweets, date, hashtags and Emotions. Data is well Visualized and interpolated in various graphs, data is well studied and concluded **that Mental Health of Indian During COVID is Well prone to fear, sadness and heartedness, with a good amount to be neutral to COVID Epidemic.**

Thank You