Coronavirus Tweet Sentiment Analysis

V Rupesh Kumar Patro
Shashank Maindola

Abstract:

In year 2020 the covid-19 caused by SARS-CoV-2 caused pandemic suddenly came out of no where and surged throughout the globe and affected everyday life. Starting with Wuhan and being a highly contagious new disease, the governments across the globe followed the lockdown procedures by imposing travel related curbs and shutting all the economic activities which required travelling and gathering. According to the International Telecommunication Union (ITU), during the COVID-19 pandemic, Internet user ranks grew to 4.9 billion in 2021, from 4.1 billion in 2019. Remote education, remote work and remote health services came within the reach of many. Social media platforms largely became one of the methods for expression and connecting. They saw a huge surge in usage and as per a report, Twitter saw 24% surge(from 152 million to 166 million) in usage during the lockdown (**source:- Bloomberg nNws**). Twitter took the internet by storm by becoming a key platform for some of the world's top experts to contribute to real-time knowledge-sharing and provide input on policymaking. Twitter became a method of expression for people by expressing their feelings regarding the vaccinations' safety and effectiveness. The proposed approach analyzes collected tweets' sentiments for sentiment classification using various feature sets and classifiers. The early detection of COVID-19 sentiments from collected tweets allow for a better understanding and handling of the pandemic.

Introduction:

During pandemic as the lockdowns were imposed across the globe, everyone took to the internet to socialise connect with the people. People expressed their concerns and feelings about the covid outbreak and its repercussions. There were people who were positive about the developments that were taking place during that time and so there were people who were neutral and so there were negative as well. The data set we are provided has a data of twitter collected over a period of 30 days i.e. from 16th March 2020 to 14th March 2020. The tweets are classified based on the sentiment of the user in 5 categories namely extremely positive, positive, neutral, negative and extremely negative. The tweets are from users across the globe. The screen name and user name of the users are coded for their privacy. There are altogether 6 features and 41157 rows. We evaluate the performance of machine learning (ML) classifiers using evaluation metrics (i.e., accuracy, precision, recall, and F1-score).

Data Set

Below is the info that is available in given dataset-

- **Username** – Coded Username
- **ScreenName** – Coded Screen name
- **Location** – Region of origin
- **TweetAt**- Tweet Timing.
- **OriginalTweet** – First tweet in the thread

- **Sentiment-Target variable** – Sentiment of tweet

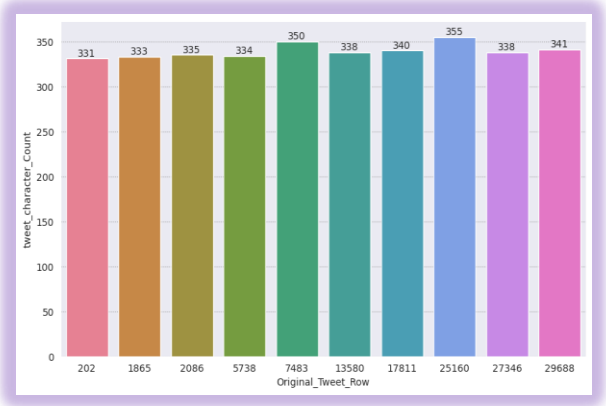
Steps for appliance energy prediction using machine learning

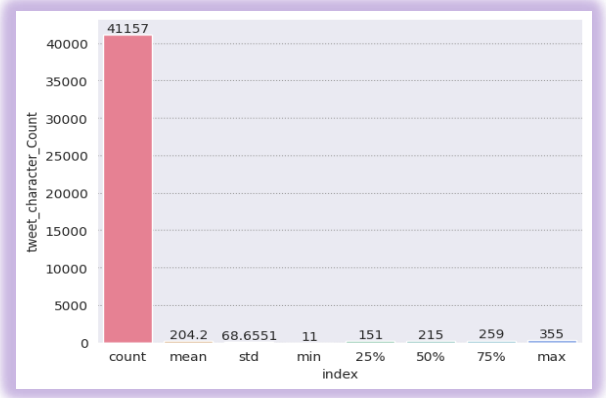
1. **EDA**
2. **Feature Engineering**
3. **Pre Processing**
4. **Model Implementation**

1. Exploratory Data Analysis

We have broadly categorized our observations into following categories:-

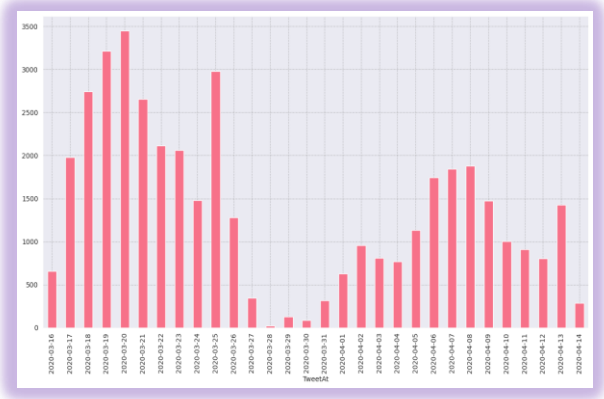
Tweets by Length



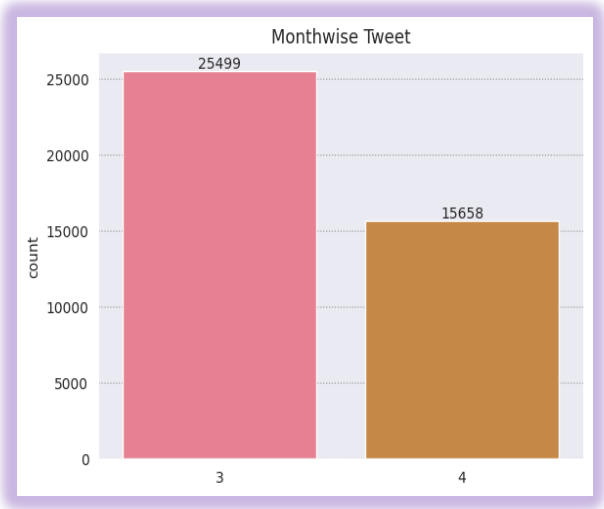


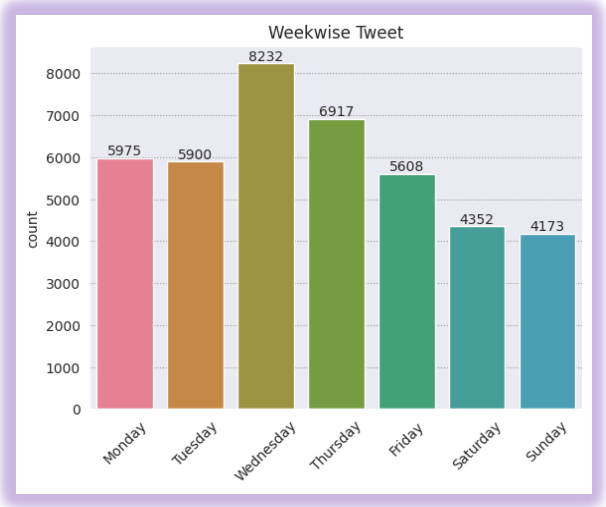
While the maximum tweet length is 355 and minimum is 11 the average remains around 205

Tweets by Time



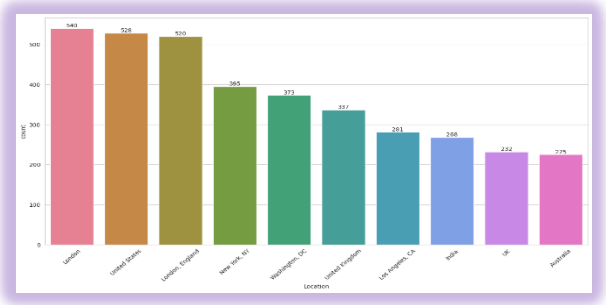
If we see day wise data them around third week the tweets were most followed by the second week of month as the lockdown followed up around third week of march but again dropped at the end then again rises. This trend might be due to the closing of financial year where people are very busy



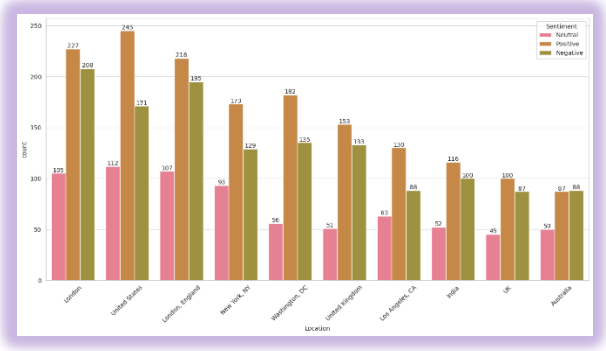


Out of the two months data march has almost 60% of the total tweets may due to sudden imposition of lockdown. It can be observed that people tweeted more on weekdays than the weekends as people are more online for work on weekdays and they check the social media frequently then

Tweets by Location

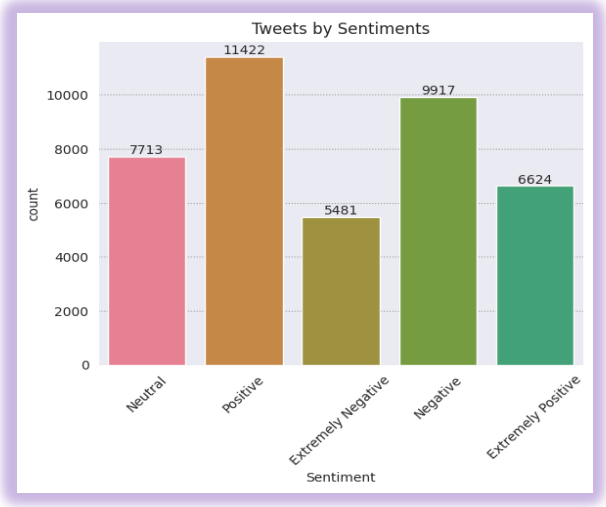


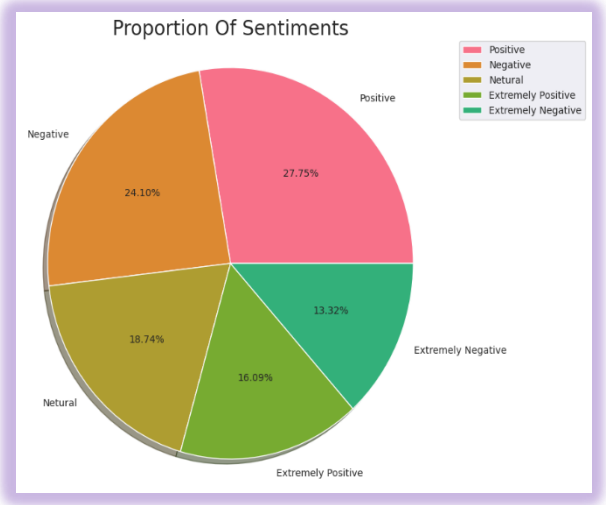
The top 10 tweets show that London has highest but if we see the third highest then we will notice it again mentions but as London as hence there may be many places which may be nicknamed but have the same reference. It has also been observed that western countries tweeted more than other parts of world.



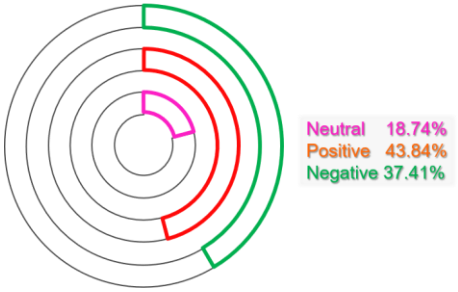
Though the no. of neutral tweets are also significant, almost everywhere people wrote positive tweets more than the negative tweets, which indicates people were optimistic even during the pandemic. Australia showed a reverse tread wherein no, of positive and negative tweets were equal.

Tweets by Sentiments



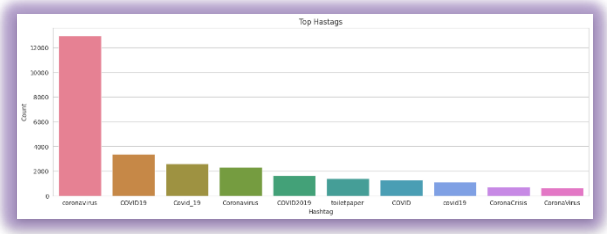


There were more people with extremely positive sentiments than the people who were extremely negative. Again, people with positive sentiments are more than the people with negative sentiments but a large number of neutral tweets signifies that there was a state of confusion.

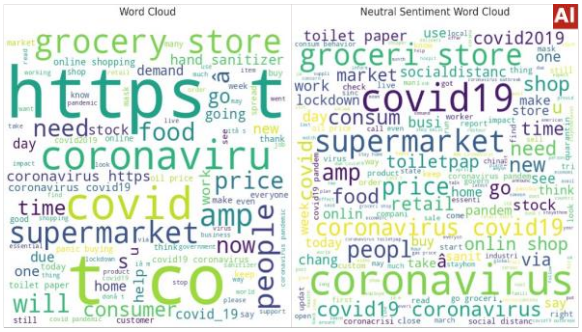


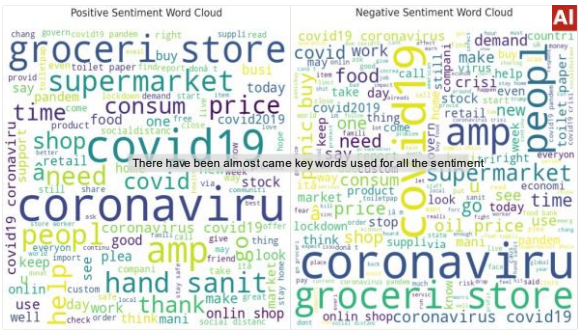
We collectively summed up the extreme sentiments with type of sentiments that they belong i.e. extremely positive with positive and extremely negative with negative and we can clearly see that while the positive sentiments are more than the negative sentiments, the neutral also contribute to significant percentage of 19%

Most common words and Hashtags in Tweets



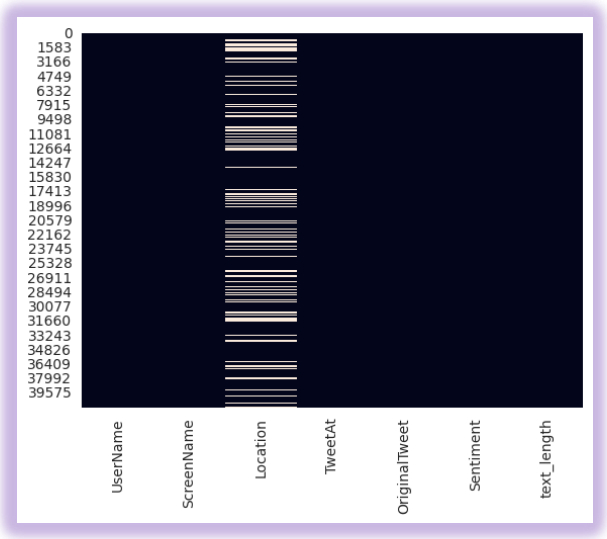
There are some words like ‘coronavirus’, ‘toilet paper’, having the maximum frequency in our dataset. There are various #hashtags in the tweets column. But they are almost the same in all sentiments, people might be discussing things but they are not giving us meaningful full information.





There has been use of almost same key words used for all the sentiment

2. Feature Engineering



While there were no duplicates, there were almost 20%missing values in the location field hence we imputed them with unknown

Firstly, we removed certain features like temperature in individual rooms,

3. Pre Processing

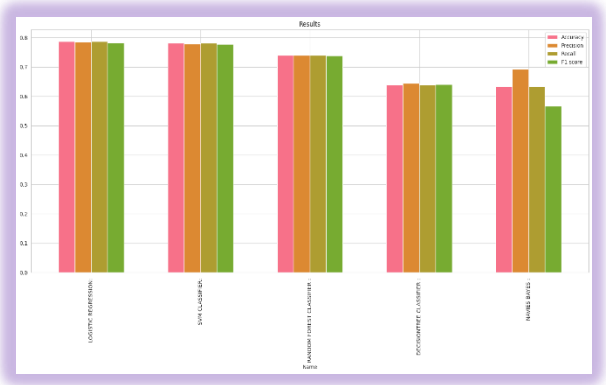
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41157 entries, 0 to 41156
Data columns (total 7 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   UserName       41157 non-null  int64
 1   ScreenName     41157 non-null  int64
 2   Location       32567 non-null  object
 3   TweetAt        41157 non-null  object
 4   OriginalTweet  41157 non-null  object
 5   Sentiment      41157 non-null  object
 6   text_length    41157 non-null  int64
dtypes: int64(3), object(4)
memory usage: 2.2+ MB
```

Here we have no need of any scaling of Dataset because here only check sentiment of people, hence we analyze only User sentiment tweet according to the scenario of Covid-19.
We used TF-IDF method as classic approach of converting input data from its raw format (i.e. text) into vectors of real numbers which is the format that ML models support TF-IDF is calculated as the product of two values: term frequency (TF) and Inverse document frequency (IDF). The term frequency (TF) measures how frequently a term appears in a document. The inverse document frequency (IDF) measures how important a term is in the entire corpus, by penalizing the terms that appear in many documents.

```
X_train.shape :  (32925, 63453)
X_test.shape :  (8232, 63453)
y_train.shape :  (32925,)
y_test.shape :  (8232,)
```

4. Model Implementation

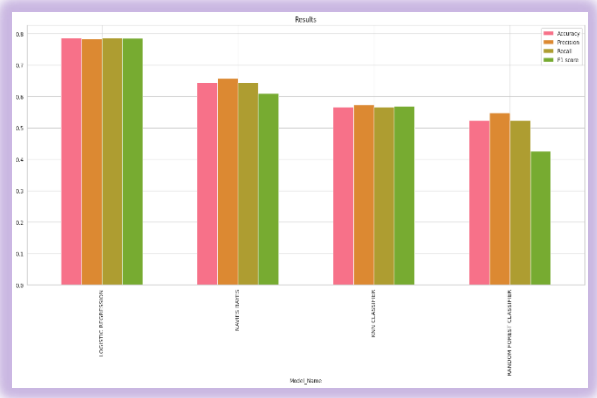
We used Logistic Regression, SVM Classifier, Random Forest Classifier, Decision Tree Classifier , Navies Bias algorithms for model training.



Name	Train_Time	Accuracy	Precision	Recall	F1 score
LOGISTIC REGRESSION:	7.152557e-07	0.787293	0.785991	0.787293	0.783083
SVM CLASSIFIER:	4.768372e-07	0.780977	0.778950	0.780977	0.776826
RANDOM FOREST CLASSIFIER :	1.192093e-06	0.739796	0.739662	0.739796	0.738312
DECISIONTREE CLASSIFIER :	9.536743e-07	0.638605	0.644412	0.638605	0.640423
NAVIES BAYES :	9.536743e-07	0.633139	0.692338	0.633139	0.566547

Out of these logistic regression, SVM classifier and Random Forest classifier gave us the most promising results. Now we will use Hyper parameter optimization and check which one algorithm gives the best result.

We used GridSearchCV for hyperparameter tuning. GridSearchCV uses each and every combination to build and evaluate the model performance. Since the dependent variable has only 3 unique values hence we used GridSeachCV.



Model_Name	Accuracy	Precision	Recall	F1 score
LOGISTIC REGRESSION	0.7871	0.7848	0.7871	0.7851
NAVIES BAYES	0.6449	0.6580	0.6449	0.6095
KNN CLASSIFIER	0.5668	0.5740	0.5668	0.5689
RANDOM FOREST CLASSIFIER	0.5232	0.5481	0.5232	0.4263

We can conclude that Logistic regression is the best model for our dataset, followed closely by Navies Bayes, KNN Classifier and Random Forest classifier did not give a good result compared to others.

Conclusion

- The majority of the tweets were around 250 characters long, indicating that there was a lot of interest in COVID-19 among the general public.
- More positive tweets than neutral or negative ones were tweeted globally.
- People tweeted more in March than in April since many nations imposed lockdown during this time.
- The United States and England were the two

countries with the most tweets.

- We saw inconsistent responses from Australia during the pandemic, with nearly equal numbers of positive and negative tweets.
- Logistic regression algorithm gave the most accurate model fit

References-

1. Analyticsvidya
2. GeeksforGeeks
3. Stackoverflow
4. Kaggle