**Name: Akomolafe Ayobami**

**Email: ayobamiakomolafe@gmail.com.**

**Report/Result Presentation of Covid-19 NLP Data-Set Analysis and Modeling**

**Overview**

The Dataset used for the analysis and modeling is the Coronavirus tweets NLP gotten from<https://www.kaggle.com/datatattle/covid-19-nlp-text-classification>. The dataset contains 41157 rows and 6 columns/attributes which include:

1. UserName
2. ScreenName
3. Location
4. Tweetat
5. OriginalTweet
6. Sentiment

The tweets have been pulled from Twitter and manual tagging has been done. The names and usernames have been given codes to avoid any privacy concerns.

The tasks performed on the dataset include:

1. Cleaning and preparing the data for statistical analysis,
2. Performing visualizations and statistical analysis of the dataset,
3. Showing the intrinsic statistical properties of the key features of the dataset,
4. Insights generations and,
5. Development of a prediction model (Text Sentiment Classification) using Classifiers Logistic Regression and LSTM.

**Data Cleaning Methodology**

Data cleaning is one of the most important steps in data analysis and modeling pipeline, such that your model and analysis can only be as good as your cleaning. The first step in the data cleaning process was to check for missing values. The missing values were discovered in the Location column and replaced/filled as **‘’UNKNOWN’’**. Next, the datatype of the Tweetat column which is a date column was changed from “object datatype” to “datetime datatype”. Also duplicate rows were checked for and dropped but however, no duplicate rows were discovered in the dataset. Lastly, the data formatting of the dataset was checked particularly for the “object datatype” columns which are the Sentiment and Location columns. The Sentiment column was well formatted but the Location column contained a lot of bad formats such as the same location represented with different names e.g. UK and United Kingdom.

**Statistical analysis**

The statistical analysis of the dataset by using the pandas describe method gives the following results: The “TweetAt”column contains 30 unique values with the most frequent one being 2020-03-20 which occurred 3448 times. The Location column contains 12220 unique values with the most frequent value being ‘’UNKNOWN’’ which occurred 8593 times. The Sentiment column has 5 unique values with the most frequent value being ‘’Positive’’ with a frequency value of 11422. The correlation between the “ScreenName” and “UserName” Column is 1 which means that they are highly correlated. This is not surprising since they both represent the same user; one of the columns (ScreenName) was therefore dropped.

# NLP

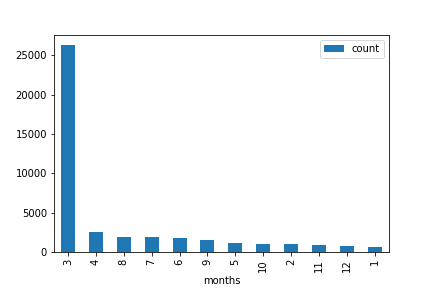
**Insight Generations and Visualizations**

# abc.png

Figure

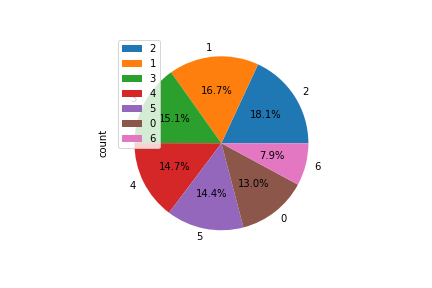
The Bar chart in Fig. 1 above shows the distribution of each of the sentiment class in the dataset. Out of the 41157 rows (Sentiments), the highest Sentiment is the Positive class sentiment with a total occurrence of 11422 (27.7522661%). The second highest used sentiment is the Negative Class occurring 9917 times (24.0955372%). The distribution of the other sentiments are Neutral: 7713(18.7404333%), Extremely Positive: 6624(16.0944684%) and Extremely Negative: 5481(13.317297%).

The Locations with the highest number of tweets and respective tweet counts, which also corresponds to the degree of activity per locations are: London: 5402, United States: 5283, London, England: 5204, New-York, NY: 3955, Washington DC: 3736, United-Kingdom: 3377, Los Angeles, CA: 2818, India: 2689 and UK: 232.



Figure

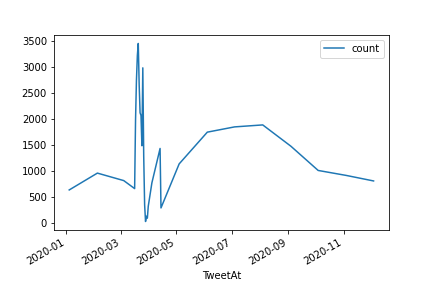
The Bar chart above in Fig 2. shows the distribution of the number of tweets in the respective months. From the bar chart above, the month with the highest number of tweets is March (3).The bar chart also shows that there is a huge gap between the number of tweets made in the highest month (March) and the second highest month (April). The month with lowest number of tweets is January (1).

***Monday: 0, Tuesday: 1, Wednesday: 2, Thursday: 3, Friday: 4, Saturday: 5, Sunday: 6***

Figure

The Pie chart in Fig 3. above shows the percentage tweets of the over-all tweets made during each day of the week. It reveals that the day with the highest number of tweets is Wednesday which is followed by Tuesday. The day with the lowest number of tweets is Sunday. This at first was surprising considering that more tweets are expected to occur during weekends. However considering the pandemic and everyone being on lockdown and at home, there really isn’t much difference between a weekday and a weekend.

The dataset contains tweets of over a 30 day period. The date/day with the highest number of tweets is 2020-03-20 with a total tweet count of 3448 tweets. This is followed by 2020-03-19, with a total tweet count of 3215. The date with the lowest tweets made is 2020-03-28 with just 23 tweets. The Line graph below shows the distribution of dates and respective tweets.



Figure

**Predictive Model Results**

Two modeling algorithms were built: **LSTM and Logistic Regression**.

* LSTM is a form of Recurrent Neural Network algorithm which was built with one embedding layer, one LSTM layer with 64 units, two dense layers each of 64 neuron and relu activation function with a final dense layer output with 5 neurons and softmax activation function. After training with the train dataset with the described network using 10 epochs, a train accuracy of 92% and test accuracy of 72% was obtained.
* Logistic Regression gave a train accuracy of 28%. The low accuracy is most likely due to the use of a multiclass classification, since logistic regression is primarily meant for binary classifications. Another reason for the low accuracy is inability of the Model to reach convergence due to the high number of dimensionality (69 characters).

|  |  |
| --- | --- |
|  |  |