**Technical Documentation**

**CAPSTONE CLASSIFICATION PROJECT**

**ON**

**CORONAVIRUS TWEETS SENTIMENTAL ANALYSIS**.

**Problem statement:**

This challenge asks you to build a classification model to predict the sentiment of COVID-19 tweets. The tweets have been pulled from Twitter and manual tagging has been done then.

**Algorithm followed:**

* importing the required libraries and packages
* Using Colab, mount the device and open the.csv [Comma Separated file].
* examining the data table.
* deleting duplicate, NAN, and null rows.
* removing the anomalies.
* Applying the concept of Data Wrangling and Data Visualization so we can analyse the dataset and retrieve necessary information.

***Step 1:***

* importing the required libraries:
* The most widely used Python package is NumPy [Numerical Python]. efficiently utilised for array calculation problems.
* Analysing tabular data is an excellent use of the Pandas library. It can be used for exploratory data analysis, including cleaning, wrangling, and manipulating data.
* With the use of the Matplotlib package, we can better visualise our tabular data in the form of [Pie Chart, Bar Graph, Line Graph, Histogram, etc.] which aids in doing the necessary analysis.

***Step 2:***

* Analysing the information sheet to identify any NULL/NAN/Missing values. As these numbers affect the correctness of the results, we wish to get rid of them. We frequently treat outliers as well; in essence, these are data points that are greatly dispersed from other data points and that would skew visual outcomes.

***Step 3:***

* Now that our dataset is clear of all ambiguity, we will go forward with the process of exploratory data analysis (EDA), which refers to the act of analysing data in order to produce visual results.

***Step 4:***

* Matplotlib effectively uses data frames and arrays to present the data for visualisation. It views axes and figures as objects. There are several stateful plotting APIs in it.

**Terminology that is frequently used:**

***Data Wrangling***

* Data wrangling is the process of organising and cleaning up large and disorganised data sets so they are simpler to access and analyse. the process of transforming and mapping data from one "raw" data type into another with the goal of improving its suitability and value for a range of downstream uses, including analytics.
* Functionalities performed under it:

Data exploration, handling NAN and missing values, removing duplicates, and data filtering.

**Data Visualization:**

It involves converting tabular data into graphic images so that users may quickly and easily understand it.

* It is employed because visualising graphical data yields considerably better outcomes than doing so for tabular data. The verbal and statistical descriptions of the data set are closely integrated during visualisation. This identifies problem areas and directs attention to them.

**Exploratory Data Analysis**

* Exploratory data analysis is a crucial procedure that entails performing early investigations on data in order to find patterns, identify anomalies, test hypotheses, and validate assumptions with the aid of summary statistics and graphical representations.

Some important functions used:

1. df.head( ):The Dataframe's top five rows are what are returned by default. The number of rows returned can be changed by inserting a value between the parentheses to override the default.
2. df.tail( ): The latest five rows of the Dataframe are returned by default. Position determines how you utilise this function.
3. df.describe( ):For each of the dataset's numerical columns, provide a statistical summary. This method computes some statistical measurements for the numerical values of the Series or DataFrame, such as percentile, mean, and standard deviation.
4. df.isnull( ).sum( ): Provide the total number of values missing from each column.
5. df.shape:It displays both the total number of dimensions and their respective sizes. The shape that is returned is the number of rows and columns because data frames are two-dimensional.
6. df.info( ): It helps in quickly gaining an overview of the dataset. To obtain a concise summary of the dataframe, use this function. This method prints details about a DataFrame, such as the non-null values, column and index dtypes, and memory utilisation.

**Tools using which we create graphical representation/Pictorial data**

***Lineplot:***

* on a different axis to depict the relationship between two data points, X and Y.

***Barplot :***

* A bar chart, often known as a bar graph, is a diagram that displays categorical data as rectangular bars with heights or lengths proportional to the values they stand for. You can plot the bars either vertically or horizontally. A vertical bar chart may also be referred to as a column chart.

***Boxplot:***

* A box plot is a visual representation of the location, dispersion, and skewness groups of quartiles of numerical data. The image of an "outlier" is the most significant one it offers.
* underlying distribution are provided, it is a useful addition to a boxplot or violinplot. On the basis of the category, a scatter plot is created.

***Pieplot:***

* A circular statistical graphic called a pie chart can only show one series of data at a time. The overall percentage of the provided data is represented by the chart's area. The percentage of the data pieces is represented by the area of the pie slices. Pie charts are frequently used in corporate presentations since they offer a quick summary of information like sales, operations, survey results, resources, etc.

***Countplot:***

* Utilizing bars, a method is utilised to display the numbers of observations in each category bin.

**Crossvalidation used:**

1- Count vectorizer

2- TF/IDF

**Model used for implementation:**

1- Logistic regression gridsearch cv

2- Decision tree classifier

3- Support vector machine

***Conclusion***

With the help of the above mentioned approach, we analysis the tweets sentiments during covid-19 pandemic all over the world

***Project by***

Amit s kashyap