**DATA ANALYST INTERN PROJECT PHASE – 2**

**Name: Sonam Singh**

**Project: 3**

**Topic: Twitter Sentiment Analysis**

**Company: Nexus Info**

**Date: 08/09/2024**

**INTRODUCTION**

This project involves the analysis of a Twitter Sentiment Dataset, named ‘Twitter Sentiment Analysis’ obtained from the Kaggle website. Kaggle is a popular online platform that hosts data science and machine learning competitions, provide datasets, and offers a collaborative environment for data scientists and analysts to share code, solutions, and insights.

**OBJECTIVE OF THE PROJECT**

The Twitter Sentiment Analysis project aims to gain insights into public opinions, trends, and sentiments shared on Twitter (now X), utilizing data analytics techniques.

**DATA METHODOLOGY**

The following steps outline the methodology adopted to analyze the Weather dataset –

1. **Data Loading:** The Twitter dataset was downloaded from Kaggle then the CSV file was loaded in Python. The dataset was converted into a Data Frame for easier manipulation and analysis.
2. **Data Exploration:** Basic descriptive statistics were calculated to understand the central tendencies, dispersion, and shape of the dataset’s distribution; and the dataset was checked for any missing values or inconsistencies that could affect the analysis.
3. **Data Analysis:** Various graph were generated to visualize the distributions and identify any relationships between the numerical features in the dataset.
4. **Documentation:** Each step of the analysis was documented, including the Python code and insights derived from the data. This documentation serves as a comprehensive guide to the project’s workflow and findings.

**DATA OVERVIEW**

The Twitter Sentiment analysis is a data analytics project that involves analyzing a dataset of tweets to determine the sentiment expressed in each tweet – whether it is positive, negative or neutral.

This is the sentiment dataset. It contains 1,600,000 tweets extracted using the Twitter api. The tweets have been annotated (0=negative, 4=positive) and they can be used to detect sentiment.

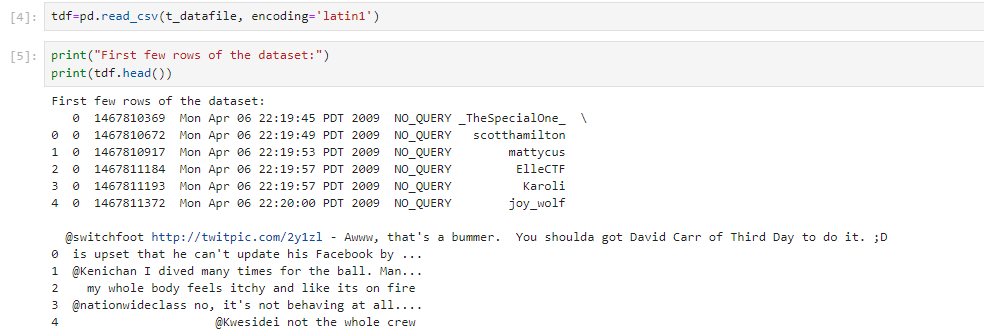
It contains the following six fields:-

* **Target:** The polarity of the tweet (*0 for negative; 2 for neutral; and 4 for positive*)
* **IDs:** The id of the tweet (*2087*)
* **Date:** The date and time of the tweet (*Sat May 16 23:58:44 UTC 2009*)
* **Flag:** The query (*lyx*). If there is no query, then this value is NO\_QUERY.
* **User:** The user that tweeted (*robotickilldozr*)
* **Text:** The text of the tweet (*Lyx is cool*).

**DATA EXPLORATION**

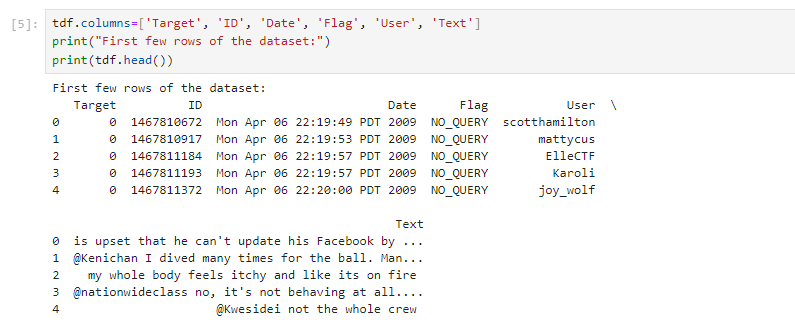
**LOADING THE DATASET**

The twitter sentiment dataset was accessed from Kaggle website. The CSV file was loaded in Python. The method ‘.head()’ was used to view first few rows of the dataset to get a clear picture of the dataset. The output for the same is –



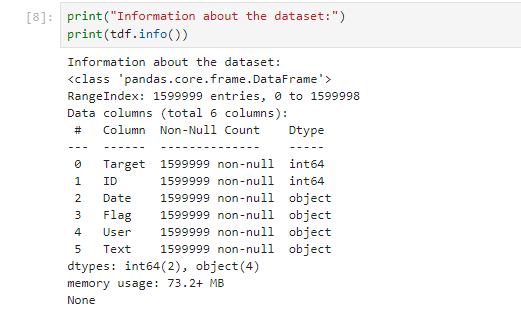
Figure

Now, we assign the column names to the respective columns to get the clear understanding of the dataset. Refer fig.2,



Figure

The dataframe was also created in Python for easier manipulation. The ‘.info()’ method in Python is used with pandas DataFrames to display a concise summary of the DataFrame. It provides essential information like data types of each column, about the DataFrame. The output for the same is –



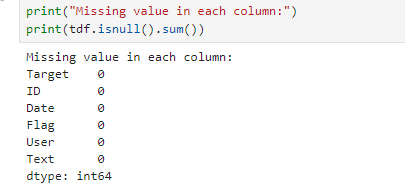
Figure

Thus, we can see from fig.,3 that the Target and ID columns have integer data type and rest of the columns have object data type.

**DATA CLEANING**

**CHECKING FOR MISSING VALUES**

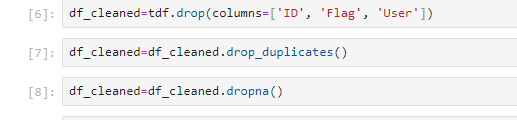
To ensure the integrity of the analysis, we first checked for missing values using the ‘isnull().sum()’ method in Python. This method provides a count of missing values in each feature. Fortunately, the Twitter sentiment dataset contained no missing values, allowing us to proceed with the analysis. The output for the same obtained from Python is –



Figure

**CLEANING THE DATASET**

Eliminating the unwanted columns from the dataset, i.e., removing ID, Flag, and User columns from the dataset by using ‘.drop()’ , ‘.drop\_duplicates()’ and ‘.dropna()’ options provided in Python. The code snippets are –



Figure

Also, cleaning the Text column by removing unwanted links (URLs), special characters and digits by using the following codes. Then checking for the modified output of the same-

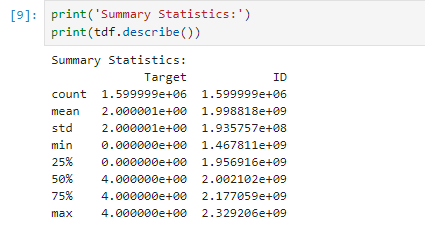


Figure

**EXPLORATORY DATA ANALYSIS (EDA)**

**SUMMARY STATISTICS**

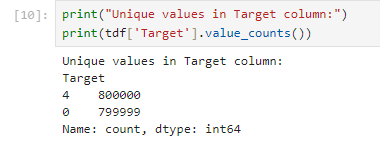
After confirming that the dataset contains no missing values, and further removing unwanted characteristics, the next step was to generate summary statistics for each numerical feature in the dataset. This was done using the ‘.describe()’ method in Python, which provides key statistics such as the mean, standard deviation, minimum, and maximum values, as well as, the 25th, 50th(median), and 75th percentiles for the numerical data. The summary statistics obtained from Python is given below –



Figure

**CHECKING FOR UNIQUE VALUES**

Now, we have to check for the unique values especially in the Target column. We will do it by using ‘.value\_counts()’ option provided in Python and the output along with the code snippet is –



Figure

We can see from fig. 8, that the Target column contains only two values 0 and 4, that too approximately equal in number. This implies that there is not a single observation or data with value 2 or ‘neutral’ sentiment.

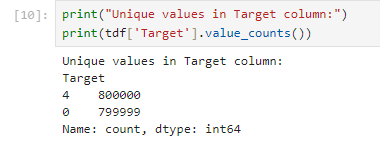
**SENTIMENT DISTRIBUTION**

The Target column contains the sentiments which are basically of three types –

1. 0=negative
2. 2=neutral
3. 4=positive

It is important to take a look at the distribution of these sentiments in our dataset which is explicitly large containing 1,600,000 data.

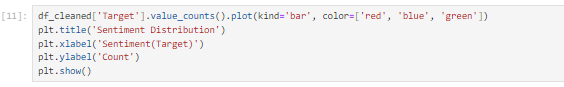
However, our dataset does not contain the type 2 sentiment i.e. Neutral. It is significantly equally distributed among Negative and Positive sentiments.



Figure

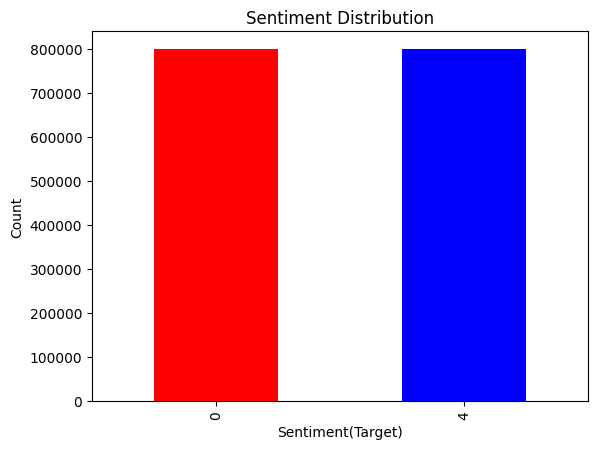
We can see that there are 800000 data which have positive sentiment i.e. having target value 4; and 799999 data have negative sentiment i.e. having target value 0.

The sentiment distribution in Bar Chart was calculated in Python by using –



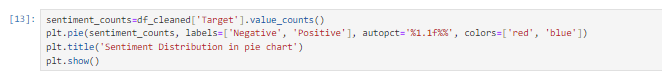
Figure

And the output of the plot is –



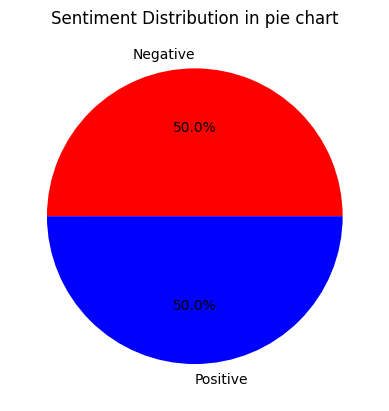
Figure

To obtain this output in the form of Pie graph, the codes used are –



Figure

And the pie chart is given by –



Figure

**CONCLUSION**

The Twitter Sentiment Analysis insights into the overall sentiment distribution within the dataset. The dataset comprised 1,600,000 tweets, which were categorized into two sentiments – positive and negative – indicating a lack of neutral sentiment. The sentiment distribution was nearly balanced , with 800,00 tweets classified as positive and 799,999 tweets classified as negative.

The EDA confirmed that the dataset contained no missing values, and after cleaning non-essential columns such as ID, Flag, and User were removed. Text preprocessing steps such as removing URLs, special characters and digits were performed to ensure high data quality for sentiment prediction.

This analysis paves the way for further exploration, such as building sentiment prediction models that could leverage the cleaned dataset to identify tweet polarity efficiently. By doing so, businesses and individuals can better understand public sentiment toward various topics, making informed decisions based on data-driven insights.