**Twitter Sentiment Analysis**

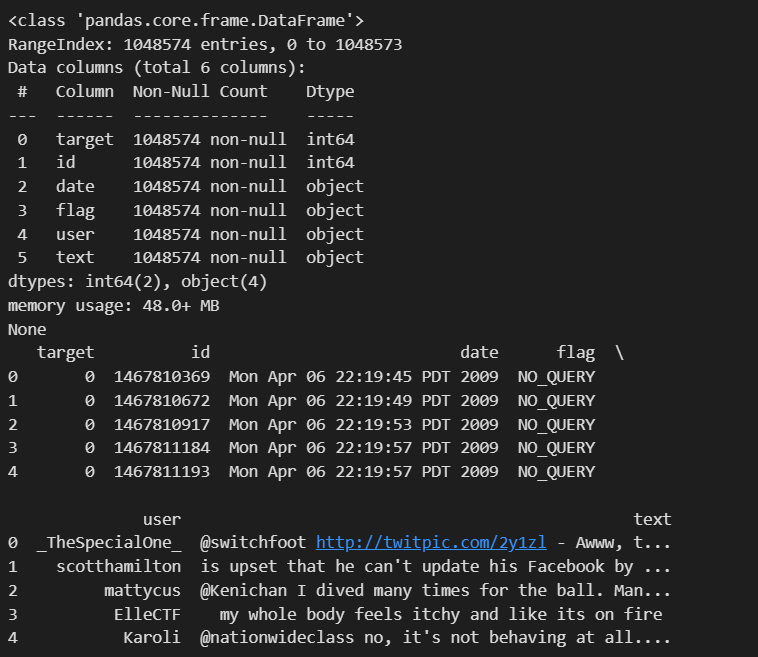
* **Vinod Ghanchi**

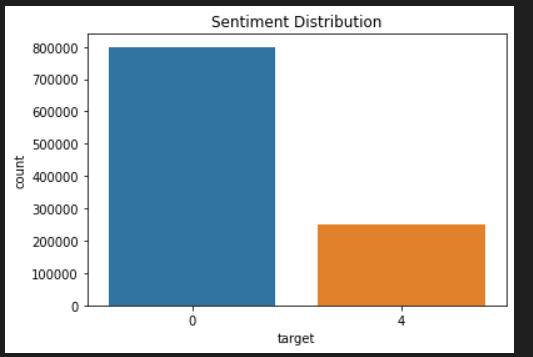
**Objective:**

The project aims to perform sentiment analysis on a dataset of tweets obtained from Twitter. Sentiment analysis involves determining the sentiment expressed in each tweet, categorizing them as either positive, negative, or neutral. The goal is to gain insights into the sentiment distribution and identify significant features contributing to sentiment predictions.

**1. Data Exploration:**

* The dataset was loaded from a CSV file containing columns such as "target," "id," "date," "flag," "user," and "text."
* Basic information about the dataset was displayed, including the data types and the presence of any missing values.





**2. Data Cleaning:**

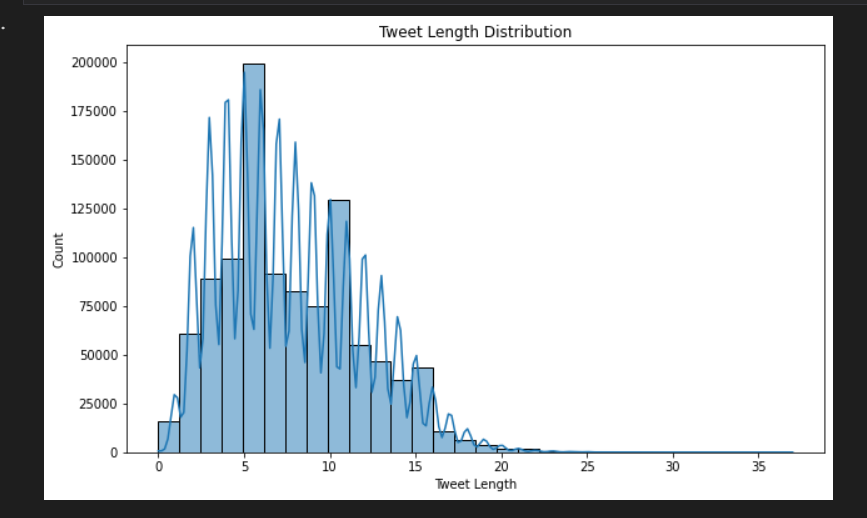
* Missing values were checked, and any NaN values were handled appropriately.
* Duplicate rows were removed from the dataset.
* Unnecessary columns ("id," "flag," "user") were dropped.

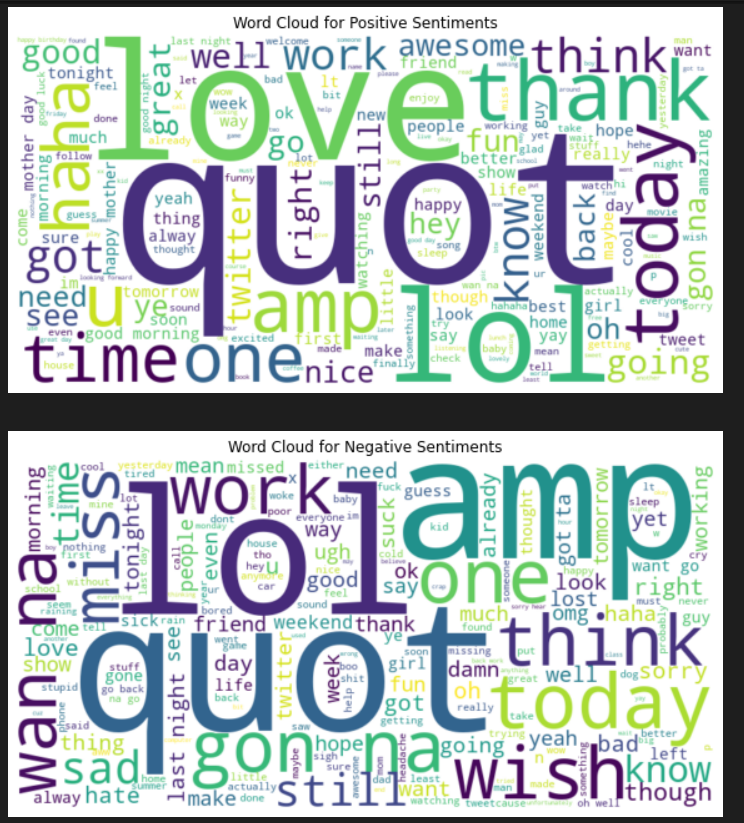
**3. Text Preprocessing:**

* Text data underwent preprocessing steps, including the removal of special characters, links, and conversion to lowercase.
* Tokenization and removal of stopwords were performed using NLTK.
* Lemmatization was applied to reduce words to their base or root form.

**4. Exploratory Data Analysis (EDA):**

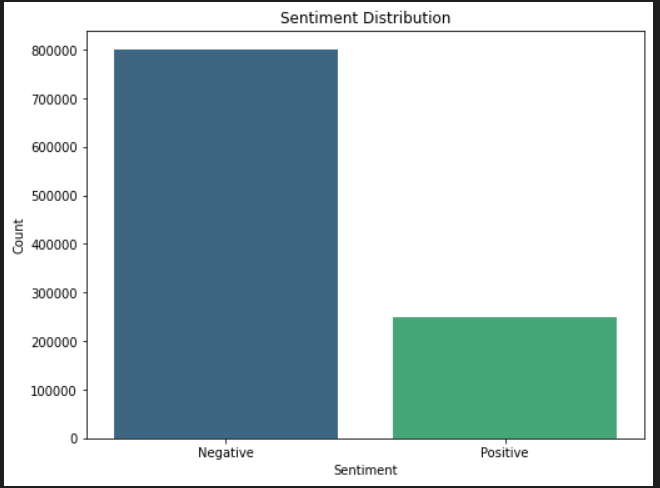
* Visualizations were created to understand the sentiment distribution, tweet length distribution, and word clouds for positive and negative sentiments.

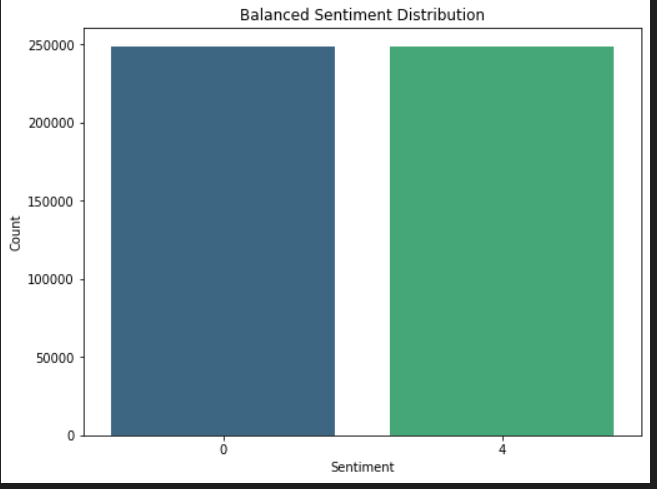




**5. Sentiment Distribution:**

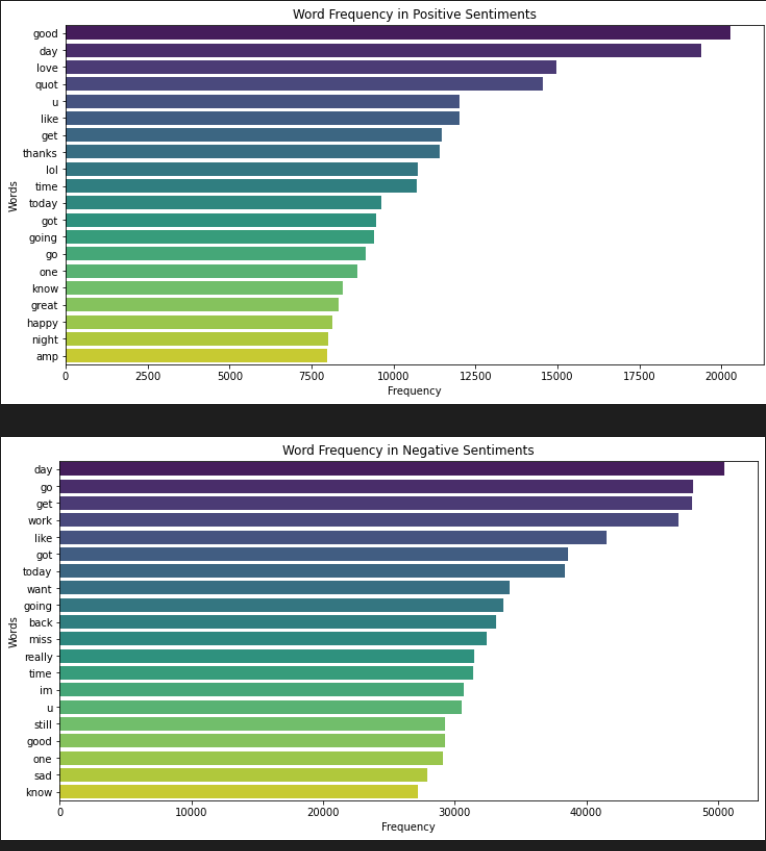
* The sentiment distribution was visualized after mapping numerical target values to sentiment labels (Negative, Neutral, Positive).





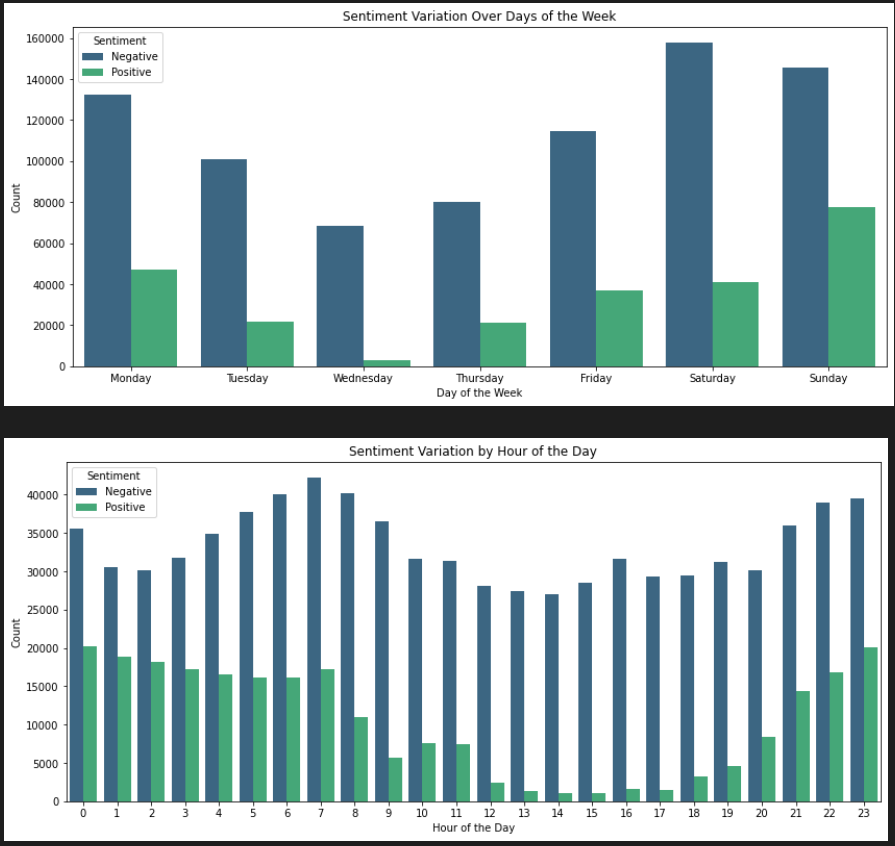
**6. Word Frequency Analysis:**

* Word frequency bar charts were created for both positive and negative sentiments, displaying the most common words.



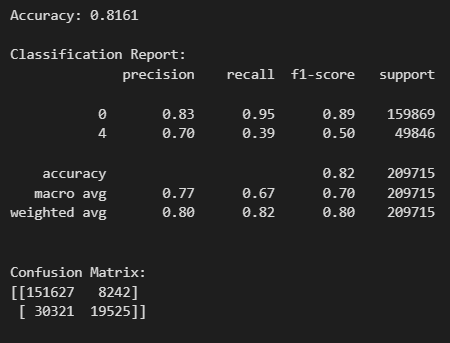
**7. Temporal Analysis:**

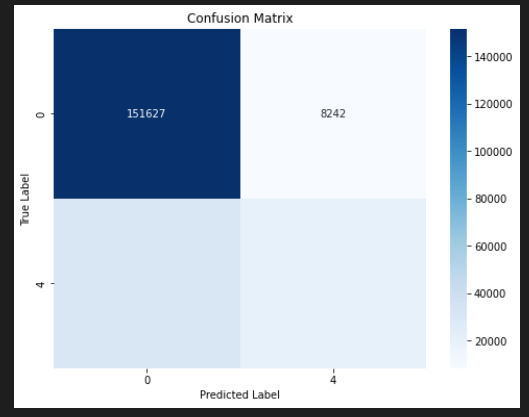
* Additional time-related features, such as day of the week and hour of the day, were extracted.
* Sentiment variation over days of the week and by hour of the day was visualized.



**8. Sentiment Prediction Model:**

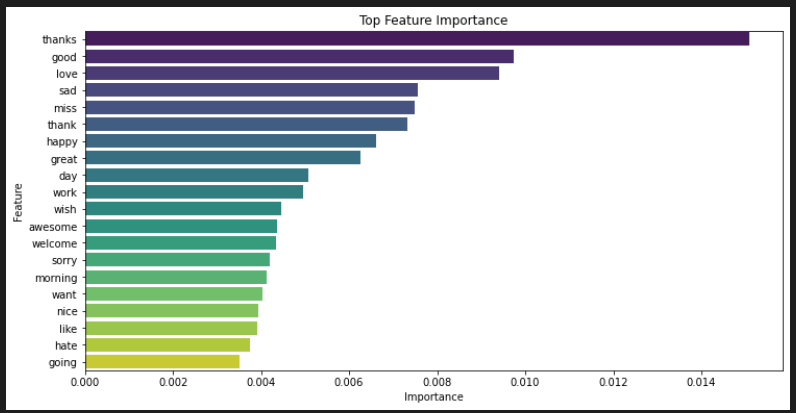
* A sentiment prediction model was built using a Random Forest Classifier with TF-IDF vectorization.
* The model's performance was evaluated using metrics such as accuracy, classification report, and confusion matrix.





**9. Feature Importance:**

* Feature importance analysis was conducted to identify the most significant words or phrases contributing to sentiment predictions.
* The top N important features were visualized using bar charts.



**10. User Interface:**- A simple user interface was developed for users to input custom text for sentiment analysis. - Sentiment prediction results were showcased in a user-friendly manner.

