**Twitter Sentiment Analysis Documentation**

**1. Project Overview**

This project involves analyzing tweets to determine their sentiment (positive, negative, or neutral) using the Sentiment140 dataset. The goal is to explore and preprocess the data, conduct exploratory data analysis (EDA), and build sentiment prediction models.

 **Description**: Loaded the dataset and assigned meaningful column names.

 **Key Variables**: sentiment, id, date, flag, user, text

 **Data Shape**: Displayed the shape of the dataset and the first few rows.

 **Sentiment Distribution**: Checked the distribution of sentiment labels.

**2.Data Exploration**

print(df.head())

print(df.shape)

print(df['sentiment'].value\_counts())

print(df.duplicated().sum())

* **Dataset Shape:** Provides the number of rows and columns.
* **Sentiment Distribution:** Shows the count of positive and negative tweets.
* **Duplicates:** Identifies and addresses duplicate entries to ensure data quality.

**3.Data Cleaning**

df.drop\_duplicates(inplace=True)

print(df.duplicated().sum())

**Handling Duplicates:** Removing duplicate tweets to prevent redundancy in analysis.

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

**Visualization:** Bar chart displaying the distribution of sentiment labels in the dataset.

**Temporal Trends:** Line graph showing how the number of tweets varies over time.

**5. Text Preprocessing**

**Text Cleaning:** Removed URLs, mentions, websites, stopwords, and punctuation. Applied stemming to normalize words.

**6. Word Frequency Analysis**

* **Word Cloud & Frequency:** Visualized frequent words in the dataset, with insights into prominent terms.

**7. Sentiment Trends Over Time**

* **Trends Analysis:** Line graph showing trends in positive and negative sentiments over time.

**8. Model Building and Evaluation**

**Random Forest Model:** Trained and evaluated. Achieved high accuracy with detailed classification report and confusion matrix.

**9.Conclusion and Insights**

* **Sentiment Distribution:** The dataset is imbalanced, with more positive tweets than negative ones. This could affect model performance.
* **Word Frequency Analysis:** Common words in the dataset reflect prevalent themes and sentiment-related terms.
* **Temporal Trends:** Sentiment trends vary over time, which could be useful for understanding shifts in public opinion.
* **Model Performance:** Both Random Forest and Naive Bayes classifiers performed well, with Random Forest slightly outperforming Naive Bayes.

**10. Future Work**

* **Handling Imbalanced Data:** Consider techniques like SMOTE or class weighting to address class