**Twitter Sentiment Analysis Documentation**

**Project Overview:**

The Twitter Sentiment Analysis project aims to analyze a dataset of tweets to understand the sentiments expressed by users on the Twitter platform. The primary objective is to classify tweets into positive, negative, or neutral categories and gain insights into public opinions and trends.

**Data Exploration:**

**Dataset Information:**

The dataset contains tweets with columns such as 'target,' 'ids,' 'date,' 'flag,' 'user,' and 'text.'

'target' is the sentiment label (0 for negative, 4 for positive).

**Key Observations:**

The dataset has a substantial number of tweets, providing a rich source of data for sentiment analysis.

'target' distribution shows an imbalance, with both positive and negative sentiments present.

**Data Cleaning:**

Checked for missing values, duplicates, and irrelevant information.

Ensured data quality by addressing anomalies and inconsistencies.

**Data Preprocessing:**

Text preprocessing involved removing stop words, special characters, and URLs.

Tokenization and lemmatization were applied to prepare the text for sentiment analysis.

**Exploratory Data Analysis (EDA):**

**Sentiment Distribution:**

Visualized the distribution of sentiment labels using a count plot.

Observed an imbalance between positive and negative sentiments.

**Word Frequency Analysis:**

Analyzed the frequency of words in tweets to identify common terms and themes.

Created word clouds for positive and negative sentiments to visualize most frequent words.

**Model Selection:**

Utilized a Multinomial Naive Bayes classifier with TF-IDF features and Count Vectorizer features.

Trained the model on a subset of the dataset.

**Model Evaluation:**

Evaluated model performance on both training and test sets.

Calculated accuracy and generated a classification report.