SENTIMENT ANALYSIS FOR MARKETING - PHASE 3

DATA LOADING:

Data loading is the initial step in the data analysis process. In this step, read and import the dataset into data analysis environment. Here's a breakdown of the data loading process

- **Import Necessary Libraries:**

  Before loading the dataset, import the appropriate libraries in Python script. Import the pandas library. This library is commonly used for data manipulation and analysis in Python.

- **Load the Dataset:**

  After importing the necessary libraries, use the pd.read_csv() function provided by pandas to load the dataset. This function is specifically designed to read data from a CSV file.

  Load the dataset from a file named "Tweets.csv" by passing the file's path as an argument to the pd.read_csv() function. This function reads the data from the CSV file and creates a pandas DataFrame, which is a two-dimensional, tabular data structure that's well-suited for data analysis.

  The loaded dataset is stored in the variable dataset, and it contains all the information from the CSV file, allows to perform various data manipulation and analysis tasks on it.

DATA PREPROCESSING:

Data preprocessing is a critical phase in any data analysis or machine learning project. It involves cleaning, transforming, and organizing the raw data into a format that is suitable for analysis, modeling, or visualization. Here is a detailed description of the data preprocessing steps:

- **Remove Unnecessary Columns**

  In this step, identify and remove columns from the dataset that are not needed for analysis. These columns might contain information that is irrelevant or redundant for specific task. The following columns are dropped,

  'airline', 'tweet_id', 'name', 'tweet_coord', 'tweet_created', 'user_timezone', 'tweet_location', 'negativereason_gold', 'retweet_count', and 'airline_sentiment_gold'. By doing this, the dataset becomes more concise and focused.

- **Handle Missing Values**

  Missing values, also known as NaN (Not a Number) or NULL values, can be problematic for data analysis.

  In this step, check for missing values in specific columns. If there are any rows with missing values in the 'airline_sentiment' or 'text' columns, remove those rows from the dataset.

- **Convert Sentiment Labels to Lowercase:**

  To ensure consistency and to avoid issues related to letter casing, convert the text in the 'airline_sentiment' column to lowercase. This ensures uniformity in sentiment labels.

- **Convert Text to Lowercase:**

  Similarly, convert the text in the 'text' column to lowercase. Converting text to lowercase is a common practice in text analysis and NLP (Natural Language Processing) to ensure that words are treated uniformly, regardless of their original casing.

- **Verify Data Changes:**

  After applying these preprocessing steps, print the first 10 rows of the processed dataset. This step allows to verify that the changes have been correctly applied to the dataset and that it is now ready for further analysis.