Phase 3: Development Part 1

PROBLEM STATEMENT:

In these two technology projects you will begin building your project by loading and preprocessing the dataset. Perform different analysis as needed.

STEPS INVOLVED IN LOADING AND PREPROCESSING:

➢ Loading The Dataset:

- Accessing raw text data from various sources.
- Reading and ingesting text data into your analysis environment.
- Understanding the data's format and structure.

➢ Preprocessing The Dataset (Text Data):

1.Text Cleaning:

- This is the process of removing or correcting any noise, irrelevant characters, or artifacts in the text data. Common text cleaning operations include:
- Removing HTML tags, if the text contains web content.
- Handling special characters, punctuation, and non-alphanumeric characters.
- Converting text to lowercase to ensure uniformity.
- Handling whitespace and line breaks.

2.Tokenization:

- Tokenization is the process of breaking down the text into smaller units called tokens.
- These tokens can be words, phrases, or even individual characters, depending on the specific task. Tokenization is a crucial step for text analysis and natural language processing (NLP).
- It splits the text into meaningful units for further analysis.

3.Removal of Stop words:

- Stop words are common words (e.g., "the," "and," "is") that are often removed from text data because they don't typically carry meaningful information for many NLP tasks.

4.Label encoding:

- Label encoding is a common preprocessing technique used when dealing with categorical data in machine learning and data analysis.
- It involves converting categorical values into numerical values to make them compatible with machine learning algorithms that expect numeric inputs.

5.Stratified Sampling:

- Stratified sampling is a sampling technique commonly used in statistics and research to ensure that a representative sample is taken from a population.

- It is particularly useful when you have a diverse population with different subgroups, and you want to ensure that each subgroup is adequately represented in the sample

Source Code:

```
import pandas as pd
import numpy as np
from sklearn.model_selection import StratifiedShuffleSplit
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords

#### Load your dataset, assuming it's in a CSV file
data = pd.read_csv('Tweets.csv')

#### Inspect the first few rows of the dataset
print(data.head())

#### Text cleaning
data['text'] = data['text'].str.lower()
data['text'] = data['text'].str.replace('[^a-zA-Z]', ' ', regex=True)
print("After Text Cleaning:")
print(data['text'].head())

#### Tokenization (using NLTK as an example)
data['tokens'] = data['text'].apply(word_tokenize)
print("After Tokenization:")
print(data['tokens'].head())

#### Remove stop words (using NLTK as an example)
stop_words = set(stopwords.words('english'))
data['tokens'] = data['tokens'].apply(lambda tokens: [word for word in tokens
if word not in stop_words])
print("After Stop Words Removal:")
print(data['tokens'].head())

#### Label encoding (assuming you have 'sentiment' as the label column)
sentiment_mapping = {'negative': 0, 'neutral': 1, 'positive': 2}
data['airline_sentiment'] = data['airline_sentiment'].map(sentiment_mapping)
print("After Label Encoding:")
print(data['airline_sentiment'].head())

X = data['text']  # Input features
y = data['airline_sentiment']  # Target variable
```

```
#### Perform Stratified Sampling with a 80-20 split
stratified_split = StratifiedShuffleSplit(n_splits=1, test_size=0.2,
random_state=42)
for train_index, test_index in stratified_split.split(X, y):
    X_train, X_test = X[train_index], X[test_index]
    y_train, y_test = y[train_index], y[test_index]

print("X_train:", X_train.head())
print("y_train:", y_train.head())
print("X_test:", X_test.head())
print("y_test:", y_test.head())
```

Output:

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS
PS C:\Users\arunr\Downloads\archive> cd c:/Users/arunr/Downloads/archive
PS C:\Users\arunr\Downloads\archive> & C:/Users/arunr/AppData/Local/Microsoft/WindowsApps/python3.11.exe c:/Users/arunr/Downloads/archive/py
            tweet_id airline_sentiment  ...  tweet_location               user_timezone
0  570306133677760513          neutral  ...             NaN  Eastern Time (US & Canada)
1  570301130888122368         positive  ...             NaN  Pacific Time (US & Canada)
2  570301083672813571          neutral  ...       Lets Play  Central Time (US & Canada)
3  570301031407624196         negative  ...             NaN  Pacific Time (US & Canada)
4  570300817074462722         negative  ...             NaN  Pacific Time (US & Canada)

[5 rows x 15 columns]
After Text Cleaning:
0                virginamerica what  dhepburn said
1    virginamerica plus you ve added commercials t...
2    virginamerica i didn t today    must mean i n...
3    virginamerica it s really aggressive to blast...
4    virginamerica and it s a really big bad thing...
Name: text, dtype: object
After Tokenization:
0             [virginamerica, what, dhepburn, said]
1   [virginamerica, plus, you, ve, added, commerci...
2   [virginamerica, i, didn, t, today, must, mean,...
3   [virginamerica, it, s, really, aggressive, to,...
4   [virginamerica, and, it, s, a, really, big, ba...
Name: tokens, dtype: object
After Stop Words Removal:
0                   [virginamerica, dhepburn, said]
1   [virginamerica, plus, added, commercials, expe...
2   [virginamerica, today, must, mean, need, take,...
3   [virginamerica, really, aggressive, blast, obn...
4           [virginamerica, really, big, bad, thing]
Name: tokens, dtype: object
After Label Encoding:
0   1
1   2
2   1
3   0
4   0
Name: airline_sentiment, dtype: int64
X_train: 1262                          united what would it cost
10772     usairways used   get emails    pre purchase a...
4204      united no  it was   flight cancelled flightla...
5491      southwestair not frustrated  just an idea  gr...
12096     americanair narrowly made standby   lots of s...
```

```
Name: airline_sentiment, dtype: int64
X_train: 1262                          united what would it cost
10772     usairways used   get emails    pre purchase a...
4204      united no  it was   flight cancelled flightla...
5491      southwestair not frustrated  just an idea  gr...
12096     americanair narrowly made standby   lots of s...
Name: text, dtype: object
y_train: 1262     1
10772    1
4204     0
5491     2
12096    0
Name: airline_sentiment, dtype: int64
X_test: 2998                                        united past
7719      jetblue would you say a delay is more likely ...
8575      jetblue i cheated on you  and i m sorry  i ll...
618       united disappointed that u didnt honor my     ...
11741     usairways the airline is embarrassing itself ...
Name: text, dtype: object
y_test: 2998     1
7719     2
8575     0
618      0
11741    0
Name: airline_sentiment, dtype: int64
PS C:\Users\arunr\Downloads\archive>
```

Conclusion:

These steps collectively prepare the dataset for sentiment analysis, where the cleaned and pre processed data can be used to train and evaluate machine learning models or perform other sentiment-related tasks.