Phase 3

(Artificial intelligence)

Sentiment Analysis for Marketing

Loading and preprocessing a dataset is a Very first step in data analysis, machine learning and artificial intelligence. This process involves several stages, including data cleaning, data integration, data transformation, and data reduction. Let's discuss each of these stages in more detail:

1. Data Loading:

- Data loading is the initial step where you acquire the dataset you intend to work with. This can involve importing data from various sources such as CSV files, Excel spreadsheets, databases, web APIs, or other data storage formats.

- You may use libraries and tools like Pandas in Python or read functions in R to load and read the data into a data structure that can be manipulated and analyzed.

Program:

```
In [39]: import pandas as pd
         import numpy as np
         from sklearn.model_selection import train_test_split
         from sklearn.feature_extraction.text import TfidfVectorizer
         from sklearn.preprocessing import LabelEncoder
         from nltk.corpus import stopwords
         from nltk.tokenize import word_tokenize
         from sklearn.naive_bayes import MultinomialNB
         from sklearn.metrics import accuracy_score, classification_report
         from sklearn.metrics import classification_report, confusion_matrix
```

Importing Required Packages

```
In [47]: df=pd.read_csv(r"C:\Users\MUHILAN\OneDrive\Desktop\Tweets.csv")

In [48]: df
```

Reading the CSV Datasets and printing it

output:

Out[48]:

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline_sentiment_gold	
0	570306133677760513	neutral	1.0000	NaN	NaN	Virgin America	NaN	
1	570301130888122368	positive	0.3486	NaN	0.0000	Virgin America	NaN	
2	570301083672813571	neutral	0.6837	NaN	NaN	Virgin America	NaN	yv
3	570301031407624196	negative	1.0000	Bad Flight	0.7033	Virgin America	NaN	
4	570300817074462722	negative	1.0000	Can't Tell	1.0000	Virgin America	NaN	
...	...	...	...	...	...	...	...	
14635	569587686496825344	positive	0.3487	NaN	0.0000	American	NaN	Kristenl
14636	569587371693355008	negative	1.0000	Customer Service Issue	1.0000	American	NaN	
14637	569587242672398336	neutral	1.0000	NaN	NaN	American	NaN	:
14638	569587188687634433	negative	1.0000	Customer Service Issue	0.6659	American	NaN	Sr

jupyter Phase3 Last Checkpoint: 27 minutes ago (autosaved) Logout

ile Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

1	570301130888122368	positive	0.3486	NaN	0.0000	Virgin America	NaN	
2	570301083672813571	neutral	0.6837	NaN	NaN	Virgin America	NaN	yv
3	570301031407624196	negative	1.0000	Bad Flight	0.7033	Virgin America	NaN	
4	570300817074462722	negative	1.0000	Can't Tell	1.0000	Virgin America	NaN	
...	...	...	...	...	...	...	...	
14635	569587686496825344	positive	0.3487	NaN	0.0000	American	NaN	Kristenl
14636	569587371693355008	negative	1.0000	Customer Service Issue	1.0000	American	NaN	
14637	569587242672398336	neutral	1.0000	NaN	NaN	American	NaN	:
14638	569587188687634433	negative	1.0000	Customer Service Issue	0.6659	American	NaN	Sr
14639	569587140490866689	neutral	0.6771	NaN	0.0000	American	NaN	c

14640 rows × 15 columns

Printing the output of the Dataset

In [8]:
```
df.head()
```

Out[8]:

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline_sentiment_gold	name
0	570306133677760513	neutral	1.0000	NaN	NaN	Virgin America	NaN	cairdin
1	570301130888122368	positive	0.3486	NaN	0.0000	Virgin America	NaN	jnardino
2	570301083672813571	neutral	0.6837	NaN	NaN	Virgin America	NaN	yvonnalynn
3	570301031407624196	negative	1.0000	Bad Flight	0.7033	Virgin America	NaN	jnardino
4	570300817074462722	negative	1.0000	Can't Tell	1.0000	Virgin America	NaN	jnardino

Just printing head of the dataset

In [15]:
```
print(df.isnull().sum())
```

```
tweet_id                            0
airline_sentiment                   0
airline_sentiment_confidence        0
negativereason                   5462
negativereason_confidence        4118
airline                             0
airline_sentiment_gold          14600
name                                0
negativereason_gold             14608
retweet_count                       0
text                                0
tweet_coord                     13621
tweet_created                       0
tweet_location                   4733
user_timezone                    4820
dtype: int64
```

Checking Null values in dataset

```
In [40]: print(df.describe())

             tweet_id  airline_sentiment_confidence  negativereason_confidence  \
count  2.000000e+00                      2.000000                   2.000000
mean   5.688328e+17                      0.928150                   0.796900
std    1.491659e+15                      0.101611                   0.287227
min    5.677780e+17                      0.856300                   0.593800
25%    5.683054e+17                      0.892225                   0.695350
50%    5.688328e+17                      0.928150                   0.796900
75%    5.693602e+17                      0.964075                   0.898450
max    5.698875e+17                      1.000000                   1.000000

       retweet_count
count            2.0
mean             0.0
std              0.0
min              0.0
25%              0.0
50%              0.0
75%              0.0
max              0.0
```

Describing dataset

Data Preprocessing

```
In [41]: # Remove any rows with missing data
df = df.dropna()

# Convert text to lowercase
df['airline_sentiment'] = df['airline_sentiment'].str.lower()

# Tokenization and vectorization using TF-IDF
vectorizer = TfidfVectorizer(max_features=5000)
X = vectorizer.fit_transform(df['airline_sentiment'])

# Label encoding for sentiment labels (if not already encoded)
sentiment_mapping = {'positive': 2, 'neutral': 1, 'negative': 0}
y = df['airline_sentiment'].map(sentiment_mapping)
```

Removing missing data, converting text to lowercase, Tokenization using TF-IDF, Label encoding for sentiment labels

```
In [42]: print (X)
print(y)

  (0, 0)	1.0
  (1, 0)	1.0
4206    0
9536    0
Name: airline_sentiment, dtype: int64
```

Printing the X and y

1. Data Cleaning:

- Data cleaning involves identifying and handling issues with the dataset, such as missing values, duplicates, outliers, and inconsistencies.
- Common data cleaning tasks include:
  - Handling missing data by imputation (replacing missing values with estimates) or removal.
  - Identifying and dealing with duplicate records.
  - Outlier detection and handling (e.g., removing or transforming outliers).
  - Standardizing or normalizing data to ensure consistency (e.g., converting categorical data to numerical format).
  - Correcting inconsistent or erroneous data entries.

2. Data Integration:

- Data integration is the process of combining data from multiple sources into a unified dataset for analysis.
- You might need to merge, join, or concatenate datasets, especially when working with data from diverse sources.
- Data integration may also involve resolving schema conflicts and data format discrepancies.

3. Data Transformation:

- Data transformation is about altering the format or structure of the data to make it more suitable for analysis or modeling.
- Common data transformation tasks include:
  - Feature engineering: Creating new features from existing ones to capture important information.
  - Encoding categorical variables: Converting categorical data (e.g., text labels) into numerical representations using techniques like one-hot encoding or label encoding.
  - Scaling or standardizing features: Bringing different features to a common scale to prevent certain features from dominating the analysis.
  - Reducing dimensionality: Reducing the number of features through techniques like Principal Component Analysis (PCA) or feature selection.

4.Data Reduction:

- Data reduction involves reducing the volume of data while retaining as much relevant information as possible.

- Techniques for data reduction include:

  - Aggregation: Summarizing data by grouping and aggregating values.

  - Sampling: Using a subset of the data for analysis, especially when dealing with large datasets.

  - Dimensionality reduction: Reducing the number of features while preserving as much variance as possible (e.g., using PCA).

  - Feature selection: Selecting a subset of the most relevant features for analysis or modeling.

The order in which you perform these steps may vary depending on the specific dataset and the goals of your analysis or machine learning project. The ultimate aim is to prepare the data in a clean, consistent, and informative form that is suitable for further analysis or modeling. Data preprocessing significantly impacts the quality and effectiveness of your data-driven work.

```
Name: airline_sentiment, dtype: int64

In [16]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

In [17]: model = MultinomialNB()
         model.fit(X_train, y_train)

Out[17]: ▾ MultinomialNB
         MultinomialNB()
```

Training and testing

```
In [18]: y_pred = model.predict(X_test)
         print(confusion_matrix(y_test, y_pred))
         print(classification_report(y_test, y_pred))

[[1]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         1

    accuracy                           1.00         1
   macro avg       1.00      1.00      1.00         1
weighted avg       1.00      1.00      1.00         1
```

Model Prediction