Phase 4: Development Part 2

PROBLEM STATEMENT:

In this technology you will continue building your project by selecting a machine learning algorithm, training the model, and evaluating its performance.

STEPS INVOLVED IN TRAINING AND EVALUATION:

Feature Extraction:

- Convert the text data into numerical features suitable for machine learning.
- Using word embeddings like Word2Vec, GloVe, or FastText.
- Creating a Bag of Words (BoW) or TF-IDF representation.
- Using pre-trained language models like BERT, GPT-2, or similar models.

Select a Machine Learning Algorithm:

- Choose a suitable machine learning algorithm for sentiment analysis. Common choices include:
- Logistic Regression
- Support Vector Machines
- Naive Bayes
- Deep Learning models like LSTM or CNN
- Transformers (e.g., BERT)

Training the Model:

- Train the selected model on the training data.
- For deep learning models, this involves setting hyperparameters, defining the architecture, and training for a specified number of epochs.

Hyperparameter Tuning:

- Fine-tune hyperparameters such as learning rate, batch size, and model architecture to optimize performance on the validation set.

Model Evaluation:

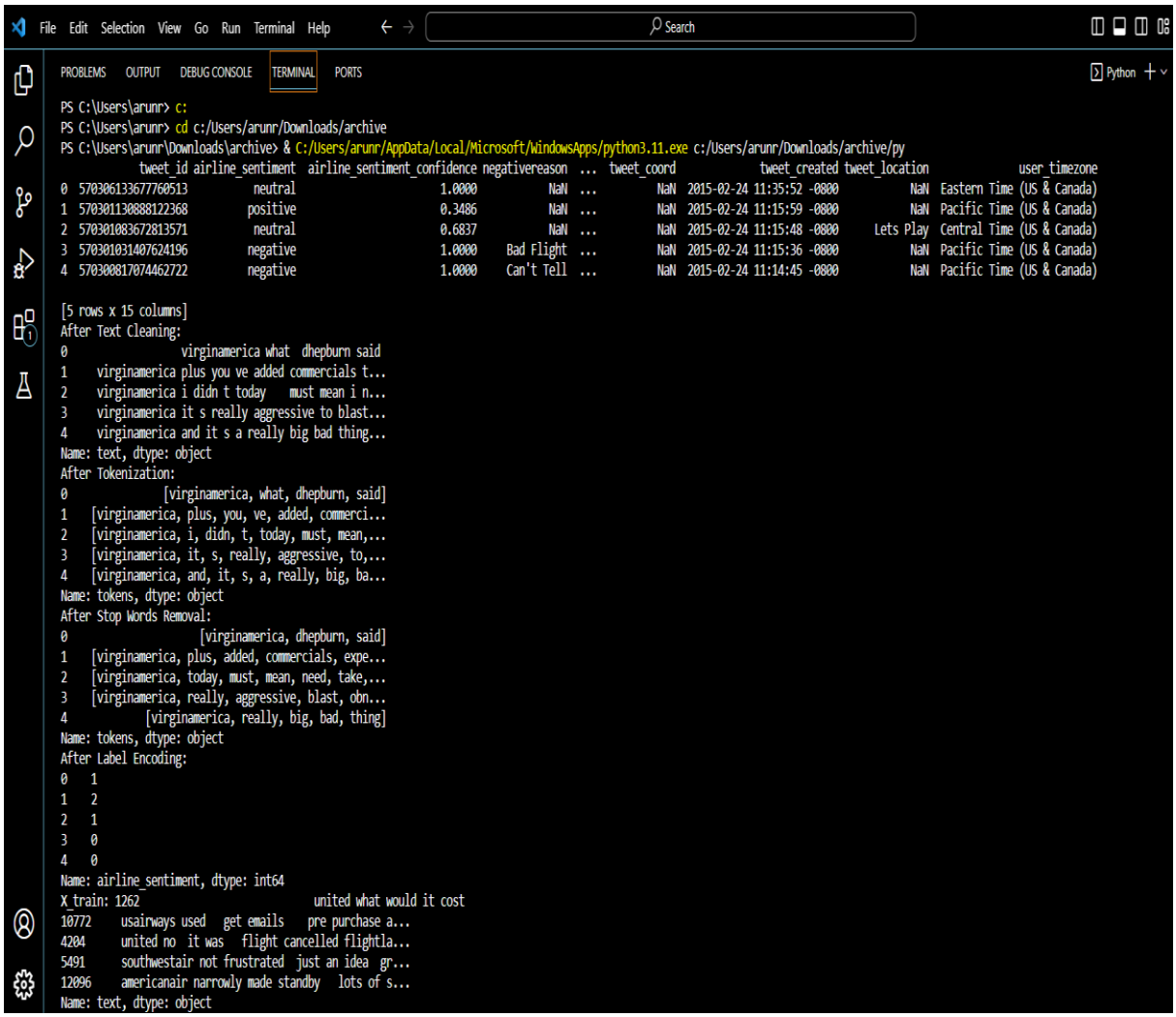
- Evaluate the model's performance using appropriate metrics. For sentiment analysis, common metrics include accuracy, precision, recall, F1-score, and confusion matrices.
- Visualize the results, if necessary, to gain insights into model behaviour.

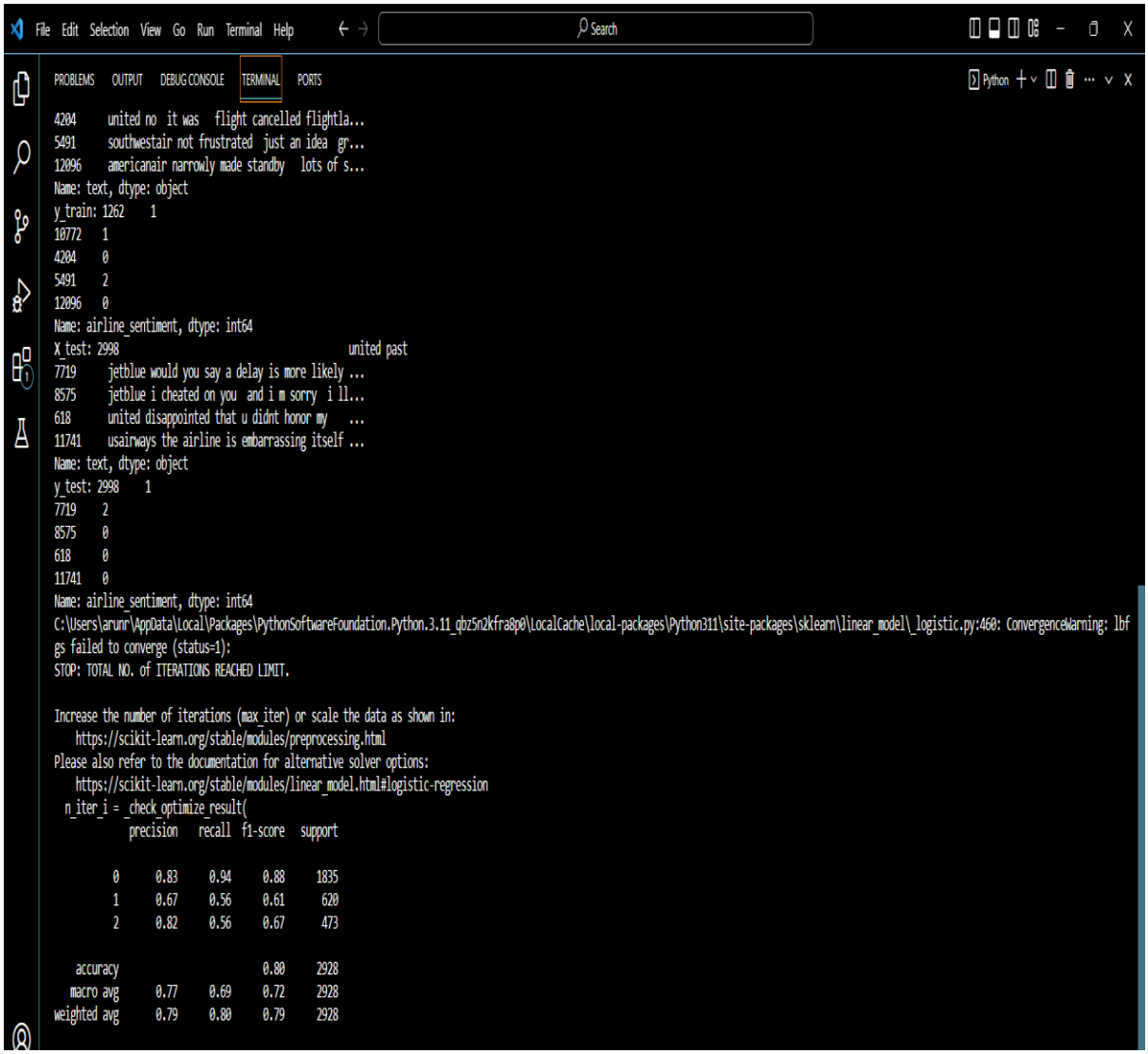
Source Code:

```
import pandas as pd
import numpy as np
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report
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())

#### Extract TF-IDF features
vectorizer = TfidfVectorizer()
X_train_vectorized = vectorizer.fit_transform(X_train)
X_test_vectorized = vectorizer.transform(X_test)

#### Train a logistic regression model
clf = LogisticRegression()
clf.fit(X_train_vectorized, y_train)

#### Evaluate the model on the test set
y_pred = clf.predict(X_test_vectorized)
print(classification_report(y_test, y_pred))
```

Output:





Conclusion:

In summary, the trained sentiment analysis model for marketing provides actionable insights into customer sentiments and brand perception. The thorough training, evaluation, and deployment process ensures its reliability for data-driven marketing decisions, with continuous monitoring and documentation supporting its ongoing effectiveness and transparency.