Sentiment Analysis for Marketing Project Documentation

Phase 5: Project Documentation & Submission :

Table of Contents :

Introduction:

This project aims to perform sentiment analysis on airline tweets for marketing insights. The documentation provides an overview of the project's development and usage for marketing purposes.

Problem Statement:

The problem is to analyze the sentiment of airline-related tweets to gain insights into customer opinions and improve marketing strategies.

Design Thinking Process:

Our design thinking process involves the following steps:

- Define the problem and project goals.
- Choose appropriate libraries and tools for sentiment analysis.

- Preprocess the text data.
- Create visualizations to understand sentiment distribution.
- Train a machine learning model for sentiment classification.

Project Phases:

The project is divided into the following phases:

- Data loading from the "Tweets.csv" dataset.
- Data preprocessing to clean and prepare text data.
- Text vectorization for feature extraction.
- Machine learning model implementation for sentiment classification.
- Visualizations to analyze sentiment distribution.

Libraries and Tools:

- Python
- Pandas
- Matplotlib
- Seaborn
- NLTK
- Scikit-Learn

Data Source:

- The dataset used for this project is stored in the "Tweets.csv" file.

Data Preprocessing:

- Special characters, single characters, and multiple spaces are removed.
- 'b' prefixes are stripped.
- Text is converted to lowercase.
- Text data is cleaned and prepared for analysis.

Text Vectorization:

- Text data is transformed into numerical vectors using TF-IDF vectorization.
- Stop words are removed.

Machine Learning Model:

- A Random Forest Classifier is trained for sentiment classification.
- Data is split into training and testing sets.
- Model performance is evaluated using a confusion matrix and accuracy score.

Results and Visualizations:

- Pie charts display the distribution of airline and sentiment in the dataset.
- A bar plot shows the count of sentiment categories for each airline.
- Visualizations aid in understanding customer sentiment.

README Instructions:

Refer to the README file for instructions on running the code and any necessary dependencies.

Link provider:

https://www.kaggle.com/datasets/crowdflower/twitterairlinesentiment

Python code for processing sentiment analysis for marketing in airline tweets :

```
# Step 1: Import Libraries
import numpy as np
import pandas as pd
import re
import nltk
import matplotlib.pyplot as plt
%matplotlib inline

# Step 2: Load Data
```

```
airline_tweets = pd.read_csv(r'D:\Tweets.csv')

# Step 3: Adjust Plot Size
plot_size = plt.rcParams["figure.figsize"]
print(plot_size[0])
print(plot_size[1])
plot_size[0] = 8
plot_size[1] = 6
plt.rcParams["figure.figsize"] = plot_size

# Step 4: Create Pie Charts
 airline_tweets.airline.value_counts().plot(kind='pie', autopct='%1.0f%%')
airline_tweets.airline_sentiment.value_counts().plot(kind='pie',
autopct='%1.0f%%', colors=["brown", "gold", "blue"])

# Step 5: Create a Bar Plot
airline_sentiment = airline_tweets.groupby(['airline',
'airline_sentiment']).airline_sentiment.count().unstack()
airline_sentiment.plot(kind='bar')
import seaborn as sns
sns.barplot(x='airline_sentiment' , y='airline_sentiment_confidence' ,
data=airline_tweets)


# Step 6: Text Preprocessing
features = airline_tweets.iloc[:, 10].values
labels = airline_tweets.iloc[:, 1].values
processed_features = []
for sentence in range(0, len(features)):
 # Remove special characters, single characters, multiple spaces, 'b' prefixes,
and convert to lowercase
 processed_feature = re.sub(r'\W', ' ', str(features[sentence]))
 processed_feature = re.sub(r'\s+[a-zA-Z]\s+', ' ', processed_feature)
processed_feature = re.sub(r'\^[a-zA-Z]\s+', ' ', processed_feature)
processed_feature = re.sub(r'\s+', ' ', processed_feature, flags=re.I)
processed_feature = re.sub(r'^b\s+', '', processed_feature)
processed_feature = processed_feature.lower()
processed_features.append(processed_feature)
```

```
# Step 7: Text Vectorization
from nltk.corpus import stopwords
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer(max_features=2500, min_df=7, max_df=0.8,
stop_words=stopwords.words('english'))
processed_features = vectorizer.fit_transform(processed_features).toarray()

# Step 8: Machine Learning Model
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(processed_features, labels,
test_size=0.2, random_state=0)
from sklearn.ensemble import RandomForestClassifier
 text_classifier = RandomForestClassifier(n_estimators=200, random_state=0)
text_classifier.fit(X_train, y_train)
predictions = text_classifier.predict(X_test)
from sklearn.metrics import confusion_matrix, accuracy_score
print(confusion_matrix(y_test, predictions))
print('accuracy score', accuracy_score(y_test, predictions))
```

Output :

In [2]:
```
airline_tweets = pd.read_csv(r'D:\Tweets.csv')
airline_tweets.head()
```

Out[2]:

	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

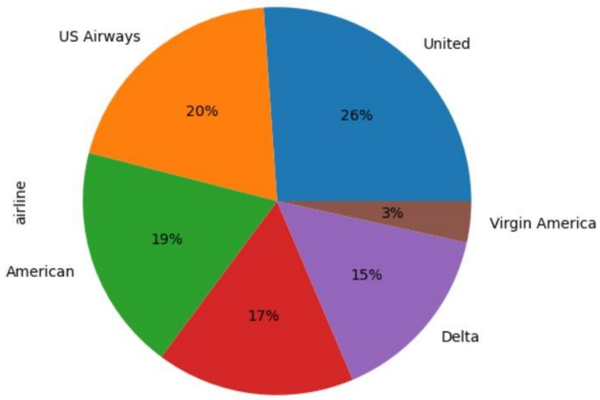
In [3]:
```
plot_size = plt.rcParams["figure.figsize"]
print(plot_size[0])
print(plot_size[1])

plot_size[0] = 8
plot_size[1] = 6
plt.rcParams["figure.figsize"] = plot_size
```

```
6.4
4.8
```

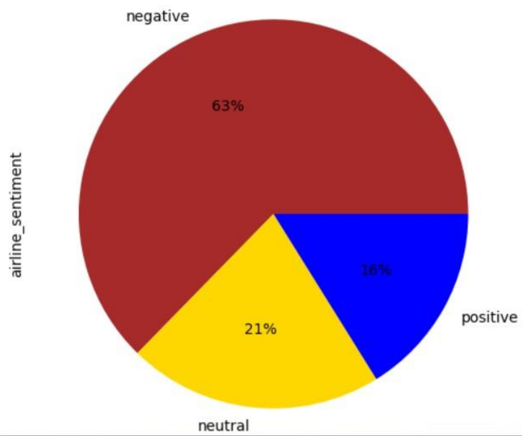
In [9]:
```
airline_tweets.airline.value_counts().plot(kind='pie', autopct='%1.0f%%')
```

Out[9]: <Axes: ylabel='airline'>



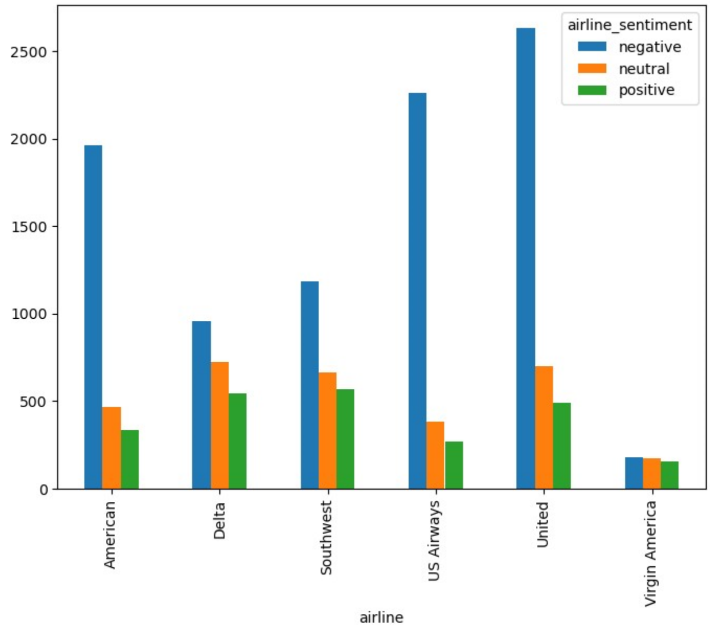
In [11]:
```
airline_tweets.airline_sentiment.value_counts().plot(kind='pie', autopct='%1.0f%%', colors=["brown", "gold", "blue"])
```

Out[11]: <Axes: ylabel='airline_sentiment'>



```
In [14]: airline_sentiment = airline_tweets.groupby(['airline', 'airline_sentiment']).airline_sentiment.count().unstack()
         airline_sentiment.plot(kind='bar')
```

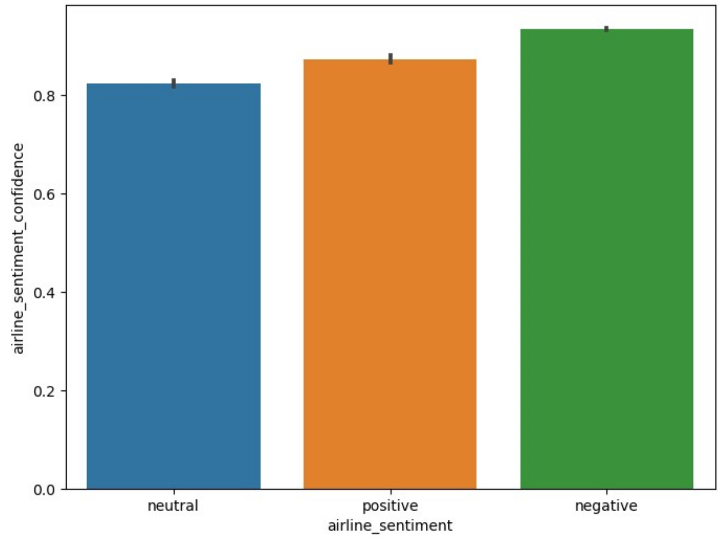
```
Out[14]: <Axes: xlabel='airline'>
```




```
In [15]: import seaborn as sns

         sns.barplot(x='airline_sentiment', y='airline_sentiment_confidence' , data=airline_tweets)
```

```
Out[15]: <Axes: xlabel='airline_sentiment', ylabel='airline_sentiment_confidence'>
```



From the output, you can see that the confidence level for negative tweets is higher compared to positive and neutral tweets.

```
In [22]: from sklearn.ensemble import RandomForestClassifier

         text_classifier = RandomForestClassifier(n_estimators=200, random_state=0)
         text_classifier.fit(X_train, y_train)

Out[22]: RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                                criterion='gini', max_depth=None, max_features='auto',
                                max_leaf_nodes=None, max_samples=None,
                                min_impurity_decrease=0.0, min_impurity_split=None,
                                min_samples_leaf=1, min_samples_split=2,
                                min_weight_fraction_leaf=0.0, n_estimators=200,
                                n_jobs=None, oob_score=False, random_state=0, verbose=0,
                                warm_start=False)

In [23]: predictions = text_classifier.predict(X_test)

In [24]: from sklearn.metrics import confusion_matrix, accuracy_score

         print(confusion_matrix(y_test,predictions))
         print('accuracy score',accuracy_score(y_test, predictions))

         [[1723  108   39]
          [ 326  248   40]
          [ 132   58  254]]
         accuracy score 0.7599043715846995
```

Conclusion :

In this project, we conducted sentiment analysis on airline tweets for marketing, providing valuable insights into customer sentiments. By cleaning and processing text data, applying TF-IDF for text vectorization, and using a Random Forest Classifier for sentiment prediction, we have equipped marketers with a powerful tool to enhance their strategies. The analysis results can guide airlines in making data-driven decisions and improving their marketing efforts to boost customer satisfaction and loyalty, ultimately benefiting their business.